

Nonparametric relevance-shifted multiple testing procedures for the analysis of high-dimensional multivariate data with small sample sizes

Cornelia Frömke, Ludwig A. Hothorn, Siegfried Kropf

Suggested citation:

Frömke, Cornelia, Ludwig A. Hothorn, and Siegfried Kropf. 2008. "Nonparametric relevance-shifted multiple testing procedures for the analysis of high-dimensional multivariate data with small sample sizes." *BMC Bioinformatics* 2008 (9:54). <https://doi.org/10.25968/opus-1132>.

Abstract

Background: In many research areas it is necessary to find differences between treatment groups with several variables. For example, studies of microarray data seek to find a significant difference in location parameters from zero or one for ratios thereof for each variable. However, in some studies a significant deviation of the difference in locations from zero (or 1 in terms of the ratio) is biologically meaningless. A relevant difference or ratio is sought in such cases. Results: This article addresses the use of relevance-shifted tests on ratios for a multivariate parallel two-sample group design. Two empirical procedures are proposed which embed the relevanceshifted test on ratios. As both procedures test a hypothesis for each variable, the resulting multiple testing problem has to be considered. Hence, the procedures include a multiplicity correction. Both procedures are extensions of available procedures for point null hypotheses achieving exact control of the familywise error rate. Whereas the shift of the null hypothesis alone would give straight-forward solutions, the problems that are the reason for the empirical considerations discussed here arise by the fact that the shift is considered in both directions and the whole parameter space in between these two limits has to be accepted as null hypothesis. Conclusion: The first algorithm to be discussed uses a permutation algorithm, and is appropriate for designs with a moderately large number of observations. However, many experiments have limited sample sizes. Then the second procedure might be more appropriate, where multiplicity is corrected according to a concept of data-driven order of hypotheses.

Methodology article

Open Access

Nonparametric relevance-shifted multiple testing procedures for the analysis of high-dimensional multivariate data with small sample sizes

Cornelia Frömke*¹, Ludwig A Hothorn² and Siegfried Kropf*³

Address: ¹Department of Biometry, Hannover Medical School, Carl-Neuberg-Str. 1, D-30625 Hannover, Germany, ²Institute of Biostatistics, Leibniz University of Hannover, Herrenhäuserstr. 2 D-30419 Hannover Germany and ³Institute for Biometry and Medical Informatics, Otto von Guericke University Magdeburg, Leipziger Str. 44, D-39120 Magdeburg, Germany

Email: Cornelia Frömke* - froemke.cornelia@mh-hannover.de; Ludwig A Hothorn - hothorn@biostat.uni-hannover.de; Siegfried Kropf* - siegfried.kropf@medizin.uni-magdeburg.de

* Corresponding authors

Published: 27 January 2008

Received: 15 June 2007

BMC Bioinformatics 2008, 9:54 doi:10.1186/1471-2105-9-54

Accepted: 27 January 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/54>

© 2008 Frömke et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: In many research areas it is necessary to find differences between treatment groups with several variables. For example, studies of microarray data seek to find a significant difference in location parameters from zero or one for ratios thereof for each variable. However, in some studies a significant deviation of the difference in locations from zero (or 1 in terms of the ratio) is biologically meaningless. A relevant difference or ratio is sought in such cases.

Results: This article addresses the use of relevance-shifted tests on ratios for a multivariate parallel two-sample group design. Two empirical procedures are proposed which embed the relevance-shifted test on ratios. As both procedures test a hypothesis for each variable, the resulting multiple testing problem has to be considered. Hence, the procedures include a multiplicity correction. Both procedures are extensions of available procedures for point null hypotheses achieving exact control of the familywise error rate. Whereas the shift of the null hypothesis alone would give straight-forward solutions, the problems that are the reason for the empirical considerations discussed here arise by the fact that the shift is considered in both directions and the whole parameter space in between these two limits has to be accepted as null hypothesis.

Conclusion: The first algorithm to be discussed uses a permutation algorithm, and is appropriate for designs with a moderately large number of observations. However, many experiments have limited sample sizes. Then the second procedure might be more appropriate, where multiplicity is corrected according to a concept of data-driven order of hypotheses.

Background

Nowadays, high-dimensional multivariate data are used in agriculture, biology and medicine. A recent example are microarray data, where two groups, for example normal and diseased tissue, are compared using tens of thousands of genes. The aim is to identify those genes with relevant

over- or under-expression. Therefore, only two-sided tests are considered here. Nevertheless, directional one-sided relevance-shifted versions are also available [1]. Distinguishing between formal statistical significance and biological relevance is a frequently discussed issue [2]. One reason is that the commonly used point-zero null-hypothesis

esis $H_0: \mu_2 - \mu_1 = 0$ is often biologically inappropriate, because depending on sample size and variance, biologically irrelevant small differences can be marked as statistically different. Therefore, the relevance-shifted null-hypothesis $H_0: \mu_2 - \mu_1 - \delta = 0$ should be used. Hereby, the problem of the choice of δ appears [3]. Instead of absolute relevance margins, the use of relative margins may be more appropriate in some applications. For example, a compound will be declared potentially mutagenic in the Ames mutagenicity assay if the number of revertants is at least double of those in the control; this is the so-called two-fold rule [4]. Another example is the characterization of the anti-neoplastic activity of a new compound by its ratio of the mean tumor volume under treatment to that of the control [5]. For microarray experiments, a specific fold-change may be of interest as well. For example, Guo et al. [6] searched genes which were significant to the unadjusted α of 5% and have a fold change greater than 1.5. To analyze the relevance-shifted hypothesis in gene expression data, Li and Wong [7] propose using a confidence interval to estimate the fold change. However, this confidence interval requires the normality assumption and does not account for multiplicity. Both problems will be addressed in the subsequent proposals.

The problem occurs with the validation of normal assumptions for high dimensional data with small sample sizes. It seems to be hopeless to empirically characterize the distribution as multivariate normal. Hence nonparametric methods may be advantageous. In the current literature, examples for the analysis of multivariate studies using nonparametric methods can be found [8,9]. The present article focusses on nonparametric approaches as well. Relevance-shifted Wilcoxon rank statistics are used as basic test statistics in both approaches considered in this paper. Parametric test procedures for relevance-shifted hypotheses can be found in Frömke [1].

Since a local decision is provided for each of the thousands of genes, the resulting multiplicity problem has to be considered, too. Otherwise, the probability to falsely reject a null hypothesis increases dramatically. To overcome the problem of multiplicity, several approaches are discussed in the literature. Aside from the classical family-wise error rate (FWER) – the probability to reject at least one true null hypothesis – the false discovery rate (FDR) introduced by Benjamini and Hochberg [10] is often used [11], giving the expected proportion of falsely rejected hypotheses among the set of all rejected hypotheses. According to the definitions, the FDR criterion usually delivers more 'significant' genes because – in contrast to the FWER – a small rate of falsely positive results is tolerated. Therefore, the main arguments in favour of the FDR (or against the FWER) are the low detection rate of the FWER procedures for high-dimensional data combined

with a still sufficient type I error control for screening purposes. Nevertheless, many authors emphasize that the FWER criterion is necessary for confirmatory purposes [12-16]. Also the argument of the larger power of FDR procedures has to be qualified. As pointed out by Dudoit et al. [15] and Speed [16], the present FDR controlling procedures are usually based on independence assumptions between the single test statistics which are not acceptable in gene expression data (particularly, the Benjamini-Hochberg procedure) or they are corrected for that problem and are then computer intensive and/or so much reduced in their power that the advantage with respect to the FWER procedures is more or less lost. Here, we focus on two FWER controlling procedures (in the strong sense, i.e., keeping the FWER under any constellation of true and false local hypotheses) and we will demonstrate that they may be well applied also in high-dimensional data.

The simplest method to correct for multiplicity in both parametric and nonparametric settings is the α -adjustment of Bonferroni. Here the unadjusted p -values of the individual tests are compared with α/m , where m denotes the dimensionality of the data, that is, the number of observed variables. The modification by Holm [17] uses the threshold α/m only for the comparison with the smallest of the m individual p -values. The next smallest are compared to $\alpha/(m-1)$, $\alpha/(m-2)$, ..., $\alpha/1$. If one of the p -values does not fall below the corresponding threshold, then the statistical procedure will stop and this null hypothesis as well as all succeeding ones will be accepted. However, even for low-dimensional data, the Bonferroni-type methods are known to be conservative. Furthermore, the potential improvement using the Holm method is minimal in high-dimensional data with only a small portion of variables with effects. One reason for the conservativeness is that the Bonferroni correction does not utilize the correlation structure of the variables. In rank tests with their discrete distributions, we have the additional problem that the procedures usually cannot fully exploit the prespecified error level but have to switch to the next possible p -value less than or equal to the given threshold. Particularly for very small sample sizes and a high dimension m , it might thus even be impossible to reject the null hypothesis using Bonferroni type methods. Both procedures extended in this paper utilize the covariance structure as they are based on permutations of the whole multivariate observation vectors although the technical procedure does not show that explicitly.

In the following text the two original procedures, which will be extended in this article, are briefly presented. The first procedure is the well-known permutation algorithm of Westfall and Young [18], as proposed by Dudoit et al. [19] for the analysis of microarray data. Just as the Bonferroni-Holm correction, this method is a step-down procedure

cedure. However, it consists of a permutation algorithm to compute the null distribution of the p -values. By permuting the variables, the algorithm takes their dependence structure into account. Given certain data conditions, above all not too small samples, this algorithm is probably the most powerful testing procedure for high-dimensional data in the current literature.

The second procedure is discussed in Kropf et al. [20]. This procedure belongs to the class of procedures with a data-driven order of hypotheses. These procedures consist of sequential testing of the variables at the unadjusted error level until the first nonsignificant result occurs. The order of testing is derived from the data themselves by means of selector statistics calculated for each variable (variables sorted for decreasing values of the selector statistics). The original procedure, which will be extended in this paper, is as follows:

1. For each of the m variables separately, compute the interquartile range using the combined data of the two samples. Order the m variables for decreasing values of their empirical interquartile ranges. The interquartile ranges serve as selector statistics.
2. Calculate the two-sample Wilcoxon test at the unadjusted level α in this order as long as significance is attained. Stop at the first nonsignificant result.

Assuming that all variables are measured on an equal scale and have similar variability within group, a large variability over the pooled samples for some of the variables could be a hint for large group differences. Therefore, the interquartile ranges of the pooled samples for each variable are used as selector statistics. The proof for the exact control of the familywise error rate utilizes, roughly speaking, the independence of rank and order statistics. If – contrary to the assumption – the variables have heterogeneous variability, then the procedure loses power. For example, Frömke [1] presents simulation studies, where the standard deviations vary by factor of 1.5 and the procedure loses approximately 10% power. The loss in power increases with increasing variability of standard deviations. However, the procedure controls the type I error in any case. A parametric counterpart of this procedure based on the theory of spherical distributions can be found in Kropf and Läuter [21].

Both of these nonparametric procedures have been shown to achieve the exact control of the familywise error rate under a point null hypothesis in the strong sense, where the observation vectors

$$\mathbf{x}_{ik} = \left(x_{ik1}, \dots, x_{ikj}, \dots, x_{ikm} \right)' \sim F_m^{(i)}(\mathbf{x}) \quad (1)$$

$(i = 1, 2; k = 1, \dots, n_i)$

belong to identical multivariate continuous distributions

$$F_m^{(1)}(\mathbf{x}) = F_m^{(2)}(\mathbf{x}). \quad (2)$$

In this paper, we are interested in a slightly different situation. The model (1) is additionally restricted by the assumption that the independent and continuous vectors \mathbf{x}_{1k} and \mathbf{x}_{2k} only have positive components and that their distribution functions are equal except for a different scaling characterized by a vector $\theta = (\theta_1, \dots, \theta_m)'$, that is

$$F_m^{(1)}(\mathbf{x}) = F_m^{(2)}(\mathbf{x}/\theta), \quad (3)$$

where the operator '/' indicates a componentwise division of vectors. Thus, θ_j denotes the true ratio of the treatment medians of variable j .

For each of the m variables, it shall be tested, whether the two-sided null hypothesis

$$H_{0,j}: \theta_{lower} \leq \theta_j \leq \theta_{upper} \quad (4)$$

can be rejected in favor of the alternative

$$H_{1,j}: \theta_j < \theta_{lower} \text{ or } \theta_j > \theta_{upper} \quad (5)$$

where $\theta_{lower} \leq 1$ and $\theta_{upper} \geq 1$ denote the lower and the upper relevance threshold.

In both procedures considered here, this multiplicative model (3) is traced back to an additive one by a variable-wise logarithmic transformation $y = \ln(x)$. So it changes to

$$F_m^{*(1)}(\mathbf{y}) = F_m^{*(2)}(\mathbf{y} - \delta) \quad (6)$$

with

$$\delta = (\delta_1, \dots, \delta_m)' \text{ and } \delta_j = \ln(\theta_j) \quad (j = 1, \dots, m) \quad (7)$$

and the null hypotheses are correspondingly transformed into

$$H_{0,j}^*: \delta_{lower} \leq \delta_j \leq \delta_{upper} \quad (8)$$

and the alternative hypotheses are given as

$$H_{1,j}^*: \delta_j < \delta_{lower} \text{ or } \delta_j > \delta_{upper}, \quad (9)$$

where $\delta_{lower} = \ln(\theta_{lower}) \leq 0$ and $\delta_{upper} = \ln(\theta_{upper}) \geq 0$ denote the lower and the upper relevance threshold. In practice, the choice of δ_{lower} and δ_{upper} is dependent on the experimental question. For example, in microarray experiments the thresholds can be set to $-\delta_{lower} = \delta_{upper} = 0.4055, 0.6931$ or 0.7885 . This is equivalent to testing for a fold-change in gene expression of 1.5, 2 or 2.2 [6,22,23]. So an obvious approach would be to use the above mentioned two procedures after the logarithmic transformation and an additional shift by the relevance thresholds. However, this is associated with some problems. The shifted one-sided tests control the familywise error rate on the threshold which was used for shifting. But here we have two one-sided tests and two relevance thresholds, the lower and the upper one, and it is necessary to find some combination rule. It is likely that the second threshold (which is not used for the shift at that moment) is far enough from the first one so that a type I error caused by the one-sided test at the opposite border of the null space is unlikely. But the whole parameter space between both thresholds belongs to the null hypothesis as well and there is no proof that the two basic procedures control the type I error or are conservative under these conditions though one would expect a monotone behaviour of the rejection probability for increasing deviations from the exact null point. Finally, a correction is necessary for the selector statistic in the second procedure with data dependent sequential testing. Otherwise, variables with no shift or a only small one (but within the tolerance region) could have larger expected values for the selector than variables under the alternative hypothesis. The procedure would then loose its power by stopping prematurely.

The modifications of the exact procedures for point null hypotheses described in the following section have been adapted to these problems in an empirical manner. No exact proof exists for the control of the familywise error rate. Therefore, results of simulation experiments are presented after the detailed description of the modified procedures and their demonstration in examples. An R package for the methods is available [24].

Results and Discussion

Algorithm

Relevance-shifted permutation algorithm

We will first introduce a relevance-shifted modification of the permutation algorithm for step-down minP adjusted p -values of Westfall and Young [18] for point null hypotheses. More strictly speaking, we are starting from a proposal from Ge et al. [25] which delivers the same results as that of Westfall and Young but is less time consuming. Whereas the original algorithm requires two permutation runs, one for the calculation of raw p -values and a second one for multiplicity adjusting, Ge et al. [25] share the permutations of both runs.

In order to detect the deviation from the null hypothesis at both relevance thresholds, two passes are needed for each variable, one for relevant decrease and another one for relevant increase. Finally, out of the two one-sided p -values, a two-sided one is computed for each variable. The passes themselves consist of two parts. The relevance-shifted permuted unadjusted p -values from Wilcoxon's rank sum test are computed first. Then the unadjusted p -values are corrected for multiplicity. As mentioned above, we will use the log transformed observations and relevance thresholds.

The proposed algorithm is given here in detail for the *test on decrease*:

Part 1: Permutation algorithm for raw p -values

- Fix the thresholds $\delta_{lower} = \ln(\theta_{lower})$ and $\delta_{upper} = \ln(\theta_{upper})$.
- Create the pseudosample vectors $\mathbf{y}_{ik}^* = (y_{ik1}^*, \dots, y_{ikm}^*)$, with $y_{1kj}^* = y_{1kj} + \delta_{lower}$ and $y_{2kj}^* = y_{2kj}$ ($i = 1, 2; k = 1, \dots, n_i; j = 1, \dots, m$).
- In the b^{th} permutation step, $b = 0, \dots, B$ ($b = 0$ corresponds to unpermuted data) do:

- For each variable, compute the one-sided Wilcoxon rank sum statistic W_{1b}, \dots, W_{mb} for the pseudosamples:

$$W_{jb} = \sum_{k=1}^{n_2} r_{2jkb}$$

where ranks are computed over both groups and r_{2jkb} denotes the k th ranked observation of the second group and the j th variable with the pseudosamples to test for decreases.

- Permute the $n_1 + n_2 = N$ pseudosample vectors \mathbf{y}_{ik}^* ($i = 1, 2; k = 1, \dots, n_i$).

- Calculate the one-sided raw p -values for hypothesis $H_{0,j} : \delta_j \geq \delta_{lower}$ as

$$p_{j,lower}^* = \frac{\#\{b: b > 0 \text{ and } W_{jb} \leq W_{j0}\}}{B} \quad (10)$$

for $j = 1, \dots, m$.

Part 2: Permutation algorithm for step-down minP adjusted p -values

- Re-number the m variables such that $p_{1,lower}^* \leq \dots \leq p_{m,lower}^*$.

- Prepare three matrices for further computation:

The matrix W of size $m \times B$ includes the test statistics from the B permutation steps from Part 1 (renumbered and without the values for $b = 0$)

$$W = \begin{pmatrix} W_{11} & \dots & W_{1b} & \dots & W_{1B} \\ \vdots & & \vdots & & \vdots \\ W_{j1} & \dots & W_{jb} & \dots & W_{jB} \\ \vdots & & \vdots & & \vdots \\ W_{m1} & \dots & W_{mb} & \dots & W_{mB} \end{pmatrix}. \quad (11)$$

Two empty matrices $P = (p_{jb})$ of size $m \times B$ and $Q = (q_{jb})$ of size $(m + 1) \times B$ are filled successively from the bottom to the top in the course of the following algorithm.

- Set $q_{m+1,b} = 1$ for $b = 1, \dots, B$.
- For $j = m, m - 1, \dots, 1$ do:

Compute the B one-sided raw p -values p_{j1}, \dots, p_{jB} for hypothesis $H_{0,j}$ (row j in matrix P) as

$$p_{jb} = \frac{\#\{b': W_{jb'} \leq W_{jb}\}}{B}, \quad (12)$$

which is in row j of matrix W for each W_{jb} the proportion of test statistics $W_{jb'}$, equal to or smaller than W_{jb} .

- Determine the j th row of matrix Q as the successive minima

$$q_{jb} = \min(q_{j+1,b}, p_{jb}), \quad b = 1, \dots, B. \quad (13)$$

Compute the adjusted p -value for hypothesis $H_{0,j} : \theta_j \geq \theta_{lower}$:

$$\tilde{p}_{j,lower}^* = \frac{\#\{b: q_{jb} \leq p_{j,lower}^*\}}{B}. \quad (14)$$

- Enforce monotonicity of $\tilde{p}_{j,lower}^*$:

$$\begin{aligned} \tilde{p}_{1,lower}^* &:= \tilde{p}_{1,lower}^* \\ \tilde{p}_{j,lower}^* &:= \max(\tilde{p}_{j-1,lower}^*, \tilde{p}_{j,lower}^*) \\ &\text{for } j = 2, \dots, m. \end{aligned} \quad (15)$$

- Revoke the renumbering of the variables in the beginning of Part 2.

For the *test on increase*, repeat the entire procedure with the

pseudosample vectors $\mathbf{y}_{ik}^* = (\gamma_{ik1}^*, \dots, \gamma_{ikm}^*)$, where $\gamma_{1kj}^* = \gamma_{1kj} + \delta_{upper}$ and $\gamma_{2kj}^* = \gamma_{2kj}$ ($j = 1, \dots, m$) and the rank sum test on increase to achieve the one-sided multiplicity

adjusted p -values on increase $\tilde{p}_{j,upper}^*$. Finally, the two-sided adjusted p -values are given by $\tilde{p}_j^* = \min(2 \cdot \tilde{p}_{j,lower}^*, 2 \cdot \tilde{p}_{j,upper}^*)$.

Procedure with a data-driven order of relevance-shifted hypotheses
An empirical extension for the nonparametric procedure of Kropf et al. [20] for relevance-shifted hypotheses will now be proposed:

- Select the two relevance thresholds $\delta_{lower} = \ln(\theta_{lower})$ and $\delta_{upper} = \ln(\theta_{upper})$.

- Determine the pseudosample vectors $\mathbf{y}_{ik}^* = (\gamma_{ik1}^*, \dots, \gamma_{ikm}^*)$ with $\gamma_{1kj}^* = \gamma_{1kj} + \delta_{lower}$ and $\gamma_{2kj}^* = \gamma_{2kj}$ ($i = 1, 2; k = 1, \dots, n_i; j = 1, \dots, m$) and calculate for each variable the one-sided Wilcoxon rank sum statistic $W_{j,lower}$

$= \sum_{k=1}^{n_2} r_{2jk}$ ($j = 1, \dots, m$), again using the ranks determined over the two combined pseudosamples,

- Replace δ_{lower} by δ_{upper} and repeat exactly the former step to compute $W_{j,upper}$ ($j = 1, \dots, m$).

- Use the permutation algorithm described in the previous procedure or suitable tables to derive the corresponding unadjusted one-sided p -values $p_{j,lower}$ and $p_{j,upper}$ respectively, for the variablewise Wilcoxon statistics.

- Compute the unadjusted two-sided p -values p_j as $p_j = \min(2 \cdot p_{j,lower}, 2 \cdot p_{j,upper})$ ($j = 1, \dots, m$) for each variable.

- In order to prepare the determination of selector statistics, calculate the sample medians for the j th (logarithmic but not shifted) variable in sample 1 and 2, \tilde{Y}_{1j} and \tilde{Y}_{2j} , respectively, and, once again, derive pseudosample values by

$$Y_{1kj}^{**} = \begin{cases} Y_{1kj} + d_{lower} & \text{if } \tilde{y}_{2j} - \tilde{y}_{1j} < 0 \\ Y_{1kj} + d_{upper} & \text{if } \tilde{y}_{2j} - \tilde{y}_{1j} \geq 0 \end{cases}$$

$$Y_{2kj}^{**} = \begin{cases} Y_{2kj} - \tilde{y}_{2j} + \tilde{y}_{1j} + d_{upper} & \text{if } 0 \leq \tilde{y}_{2j} - \tilde{y}_{1j} < d_{upper} \\ Y_{2kj} - \tilde{y}_{2j} + \tilde{y}_{1j} + d_{lower} & \text{if } d_{lower} < \tilde{y}_{2j} - \tilde{y}_{1j} < 0 \\ Y_{2kj} & \text{else,} \end{cases}$$

($k = 1, \dots, n_1$ or $k = 1, \dots, n_2$, respectively; $j = 1, \dots, m$).

- Compute a selector statistic for each variable as the interquartile range IRQ_j (difference of percentiles 75 and 25) from the combined sample values Y_{ikj}^{**} , ($i = 1, 2; k = 1, \dots, m$).
- Sort the m p -values p_j for decreasing values of the corresponding selector statistics IRQ_j .
- In this order, compare the corresponding p -value with the unadjusted α for each variable j . The original variable has a significantly relevant ratio of medians if $p_j < \alpha$ and all previously tested null hypotheses have been rejected, too.
- Stop at the first non-significance and accept the null hypothesis for all further variables.

The different formulae for the selector statistic depending on the difference of the two group medians (positive or negative, within or without the tolerance region) ensure an appropriate sorting of the variables giving the procedure a high power.

In the following sections, the Bonferroni-Holm procedure will be used for comparison. The unadjusted p -values will also be taken from two-sample Wilcoxon tests with the pseudosample values as in the above two methods. The one-sided p -values $p_{j,lower}$ and $p_{j,upper}$ will be determined separately for each of the m variables, each with the corresponding shift in the pseudosamples. These unadjusted p -values can be either taken from the first pass of the minP algorithm or from the second procedure. Then – as above – two-sided p -values will be calculated using $p_j = \min(2 \cdot p_{j,lower}, 2 \cdot p_{j,upper})$ ($j = 1, \dots, m$) and will be used as the basis for the Bonferroni-Holm adjustment.

Testing

Performance on simulated data

To confirm the control of the FWER, extended simulation studies were applied to the new procedures. A small part of the results for two-sided testing is shown in the following two tables. All scenarios were tested with three levels of relevance thresholds. For comparison with Kropf et al. [20], in one type of setting the thresholds were set to θ_{lower}

$= \theta_{upper} = 1$ ($\delta_{lower} = \delta_{upper} = 0$). In this case, the procedure with a data-driven order of relevance-shifted hypotheses reduces to the exact nonparametric procedure with a data-driven order of point-zero hypotheses for two unpaired samples applied to the logarithmized data. In the remaining two types of settings, the thresholds were set to $\theta_{lower}^{-1} = \theta_{upper} = 1.5$

($\delta_{lower} = \delta_{upper} = 0.4055$) or to the extreme case of $\theta_{lower}^{-1} = \theta_{upper} = 5$ ($\delta_{lower} = \delta_{upper} = 1.6094$). Fifty observed variables were tested in all simulated situations.

To test if the procedures control the FWER in the weak sense, which is in the special case where all null hypotheses are true and the simulated FWER is less or equal to the selected α , 25 variables were generated with expected values of $\mu_{1j} = \mu_{2j} / \theta_{upper} = 100$ and true standard deviations of $\sigma_{1j} = \sigma_{2j} / \theta_{upper}$. The remaining 25 had $\mu_{1j} / \theta_{upper} = \mu_{2j} = 100$ and $\sigma_{1j} / \theta_{upper} = \sigma_{2j}$.

Furthermore, the control in the strong sense is important. In this case, the FWER is protected if some null hypotheses may be true and others false but the probability to reject any true null hypothesis is less or equal to α . For the assessment of the control, 45 variables were simulated under the null hypothesis and had the same true mean values as for the weak control; 22 were set to a non-relevant decrease and 23 to an increase. From the remaining five variables, two had a relevant ratio of treatment means with $\mu_{1j} = 100 \cdot \theta_{upper} + 50$ and $\mu_{2j} = 100$ with $\sigma_{1j} = 10 \cdot \theta_{upper} + 5$, $\sigma_{2j} = 10$ and the other three had $\mu_{1j} = 100$ and $\mu_{2j} = 100 \cdot \theta_{upper} + 50$ with $\sigma_{1j} = 10$, $\sigma_{2j} = 10 \cdot \theta_{upper} + 5$.

All variables had equal pairwise correlations ρ and equal variances 'on a logarithmic scale'. Together with the sample size per group, these parameters differed between the individual simulation settings and are noted in the table. If not stated otherwise, the random numbers were generated from a normal distribution, the nominal FWER was set $\alpha = 5\%$ and the empirical FWER was computed with 10,000 simulation runs each. The modified Westfall-Young permutation algorithm is shown as 'permutation' in the following tables and figures and the procedure with a data-driven order of hypotheses is shown as 'selector'.

Table 1 presents the results of several simulation series for balanced multivariate normal samples at a nominal α level of 5% with varying relevance thresholds, sample sizes, variances, and pairwise correlation coefficients.

In Table 2, similar settings to the above were simulated. However, the random numbers were generated from a

Table 1: Simulation results of the FWER for normal distributed data. This table shows simulation results of the relevance-shifted Westfall-Young permutation algorithm using 500,000 permutation runs ('permutation') and the procedure with a data-driven order of relevance-shifted hypotheses ('selector') for different levels of sample sizes, variances and correlation coefficients using normal distributed data.

$q_{lower}^{-1} = \theta_{upper}$	n_i	σ_{ij}	ρ	selector		permutation	
				weak	strong	weak	strong
1	5	10	0.1	3.42	3.05	0	0
1.5	5	10	0.1	2.53	1.98	0	0
5	5	10	0.1	2.53	1.87	0	0
1	10	10	0.1	4.39	4.22	4.65	4.12
1.5	10	10	0.1	2.70	2.81	3.30	3.14
5	10	10	0.1	2.70	2.48	3.30	3.01
1	10	10	0.5	4.26	4.22	4.74	4.61
1.5	10	10	0.5	3.08	3.07	3.36	3.71
5	10	10	0.5	3.08	2.89	3.36	3.59
1	5	10	0.9	3.12	3.13	1.42	0.91
1.5	5	10	0.9	2.83	2.61	1.37	1.45
5	5	10	0.9	2.83	2.80	1.37	1.44
1	5	15	0.1	3.40	2.52	0	0
1.5	5	15	0.1	2.48	1.73	0	0
5	5	15	0.1	2.48	1.98	0	0
1	10	15	0.1	4.27	4.21	4.65	4.10
1.5	10	15	0.1	2.67	2.78	3.30	3.11
5	10	15	0.1	2.67	2.51	3.30	2.97
1	20	10	0.9	5.01	4.96	4.90	4.58
1.5	20	10	0.9	4.04	3.97	3.29	3.64
5	20	10	0.9	3.90	3.97	3.29	3.64

multivariate skewed distribution. For the generation of the random numbers, first univariate non-normal distributed samples with a priori selected expected value, standard deviation, skewness and kurtosis were created by application of a polynomial data transformation proposed by Fleishman [26]: A random variate $X \sim N(0, 1)$ is transformed into the polynomial $Y = a + bX + cX^2 + dX^3$.

$$\gamma_1 = \frac{E((Y-\mu)^3)}{\sigma^3(Y)}$$

The dependence of the skewness

$$\gamma_2 = \frac{E((Y-\mu)^4)}{\sigma^4(Y)}$$

kurtosis on the constants a, b, c and d is described in Fleishman's paper. An underlying covariance matrix for the simulated vectors is created as follows: Let \mathbf{x} denote an m -dimensional vector, where all components are iid with skewness γ_1 and kurtosis γ_2 . Now determine a transformation matrix \mathbf{R} of size $m \times m$, such that $\Sigma = \mathbf{R}'\mathbf{R}$ (for example with Cholesky decomposition). Then the transformed vector $\mathbf{y} = \mathbf{R}'\mathbf{x}$ has the variance-covariance matrix Σ . Using this method, sample vectors with $\gamma_1 = 2$ and $\gamma_2 = 7$ were produced for the simulation series in Table 2.

The results in the tables show that the new procedures control the FWER empirically. Likewise, the FWER is protected for two-sided testing in further simulated situations, including other settings of the true mean values, skewed data, variances and correlations among the variables.

Extended results for one-sided testing using the procedure with a data-driven order of relevance-shifted hypotheses are also given [1]. Small increases of the FWER occurred in that case. The largest increase for the nominal α level of 5% was 6.3%. Error rates for the permutation algorithm corresponding to the one-sided case have not yet been analyzed.

The control of the FWER is a premise of a statistical test. However, the aim of the experiments discussed here is to find variables which discriminate two kinds of treatments with a high probability. Hence, graphical representations of the simulation results in terms of the power of the new procedures compared to a standard technique will now be shown.

The simulation setting is nearly the same as for the control of the FWER in the strong sense. However, the setting of the expected values of variables under H_0 was changed.

Table 2: Simulation results of the FWER for non-normal distributed data. This table shows the results for similar settings as in Table 2. Again, different levels of sample sizes, variances and correlation coefficients were tested, however non-normal distributed data was generated.

$Q_{lower}^{-1} = \theta_{upper}$	n_i	σ	ρ	selector		permutation	
				weak	strong	weak	strong
1	5	10	0.1	3.31	2.87	0	0
1.5	5	10	0.1	2.40	1.57	0	0
5	5	10	0.1	2.40	1.78	0	0
1	10	10	0.1	4.15	4.32	4.88	4.18
1.5	10	10	0.1	2.48	2.64	3.46	3.42
5	10	10	0.1	2.48	2.37	3.46	3.15
1	5	10	0.9	3.19	2.96	0.80	0.39
1.5	5	10	0.9	2.78	2.45	0.83	1.12
5	5	10	0.9	2.78	2.64	0.84	1.07
1	10	10	0.9	4.33	4.48	4.65	4.25
1.5	10	10	0.9		3.59	3.26	0.33
5	10	10	0.9	3.51	3.49	3.26	0.33
1	5	5	0.9	3.16	3.13	0.80	0.46
1.5	5	5	0.9	2.77	2.8	0.84	1.12
5	5	5	0.9	2.77	2.44	0.84	1.12
1	5	15	0.9	3.17	2.73	0.80	0.31
1.5	5	15	0.9	2.76	2.44	0.80	1.09
5	5	15	0.9	2.76	2.56	0.84	1.03

For the control of the FWER, these variables had a ratio of means set to one of the margins of the null hypothesis because this choice resulted in the largest empirical FWER compared to variables with ratios closer to 1. A more realistic setting was selected for the simulation of the power, where a variable under H_0 received a random ratio of means. This random value was a number $\theta_{lower} \leq \tau \leq \theta_{upper}$. Two sets of values were created to generate τ . One set included all values between 1 and θ_{upper} in steps of 0.05. To receive an equal amount of ratios in the second set, all values between 1 and Q_{lower}^{-1} in steps of 0.05 were computed and the second set took the inverse of these values. The sets were combined and τ was chosen separately for each variable. If $\tau < 1$ then expectation values were set to $\mu_{1j} = 100/\tau$ and $\mu_{2j} = 100$ and the standard deviations were set to $\sigma_{1j} = \sigma_{2j}/\tau$. Otherwise the true mean values were $\mu_{1j} = 100$ and $\mu_{2j} = \tau \cdot 100$ with $\sigma_{1j} = \sigma_{2j}/\tau$.

As for the simulations of the FWER, $\alpha = 5\%$ and each result consisted of 10,000 simulation runs. In the tested scenarios, the thresholds were set to $Q_{lower}^{-1} = \theta_{upper} = 2$ ($\delta_{lower} = \delta_{upper} = 0.6931$) and $\sigma_{ij} = 10$.

All further settings of the parameters are given in the captions of the figures. The figures show the ratio of detected

false null hypotheses as an estimation of the proportional power, which is defined as the average probability of rejecting the false null hypotheses [19]. The power of the exact relevance-shifted Wilcoxon rank sum test on ratio with the multiplicity correction of Bonferroni-Holm ('Bonferroni-Holm') is plotted together with simulation results of the two new procedures.

It can be seen from Figure 1 that both new procedures achieve a higher power than the Bonferroni-Holm correction, irrespective of the correlation among the variables. While the power of the Bonferroni-Holm correction is constant for increasing correlation coefficients, the power of the new procedures increases. In Figure 2, the dependency of the three procedures on the relevance thresholds is shown. It can be clearly seen that the ratio of expected values has to be increased for all procedures in order to acquire a comparable power with increasing distance of the thresholds from the neutral value 1. In this and further simulation studies (results not shown here), the required ratio of expected values is approximately a multiple of the upper relevance threshold. The power is only smaller in

the special case of $Q_{lower}^{-1} = \theta_{upper} = 1$, as here all ratios of variables under H_0 are set to the margins of the thresholds. To achieve a power of around 50%, for example the procedure with a data-driven order of relevance-shifted hypotheses requires a ratio of expected values of 1.25 for

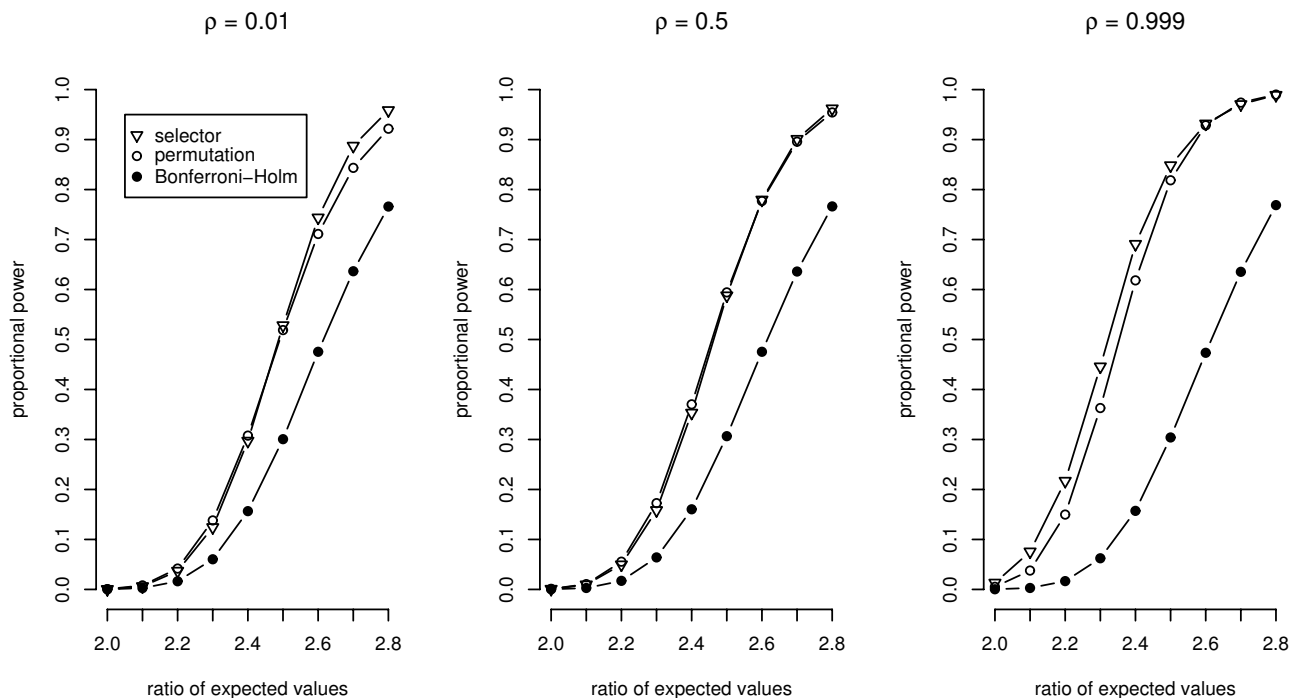


Figure 1
Power for varying correlation structure among variables. The three plots show the proportional power of the three procedures using a varying correlation structure among the variables. In each setting a sample size of $n_i = 7$ was used.

$Q_{lower}^{-1} = \theta_{upper} = 1$. By multiplying this ratio of expected values with the upper relevance thresholds 2 or 5 (giving the ratios 2.5 and 6.25, respectively), the power is around 55% in both cases.

In Figure 3 the dependency of the three procedures on the sample sizes is shown. For small sample sizes, say 4 to 6, the procedure with a data-driven order of hypotheses is better than the permutation algorithm and the Bonferroni-Holm correction. With a sample size of 7 or more, the permutation algorithm achieves a higher proportional power. The Bonferroni-Holm correction can only be applied in this simulation setting if the sample size is at least 7. If the sample sizes are reduced to 6 per group, the smallest possible two-sided Bonferroni-Holm adjusted p -value is 0.108, and thus no significant variables can be achieved with $\alpha = 5\%$. The power of the Bonferroni-Holm correction also increases with increasing sample sizes. In the observed simulation setting a sample size of 10 is required to be better than the procedure with a data-driven order of hypotheses.

In most microarray experiments several thousand variables are tested. Hence, simulations presented in Figure 4

were carried out including 5,000 variables as well. Basically, the simulation setting was the same as the setting presented in Figure 3. However, the number of variables was set to 5,000, where 50 variables were tested under H_1 . And as the power decreases with an increasing number of variables, the expected values were set to 1/2.5 and 2.5 for 25 variables under H_1 each. The simulations of the permutation algorithm including 5,000 variables were time consuming. Therefore, simulations were carried out up to a sample size of 10 per group.

As in Figure 3, the procedure with a data-driven order of hypotheses is more powerful than the permutation algorithm if the sample sizes are small. However, using a larger sample size the permutation algorithm is preferable. The Bonferroni-Holm correction achieves no power, because the procedure is too discrete. If an experiment consists of 5,000 variables, a sample size of 12 per group is required to achieve a two-sided p -value of 3.7%. For example, using a sample size of 11 per group, the smallest achievable two-sided p -value is 14.2%. Irrespective of the effect size, this p -value cannot be less than $\alpha = 5\%$.

The choice of the procedure with the best power does not only depend on the sample size. In particular with an increasing α , the permutation algorithm and the Bonfer-

roni-Holm correction are more powerful than the procedure with a data-driven order of hypotheses with sample sizes as low as 7 or 10. The choice is also dependent on the correlation among the variables as shown in Figure 1. Fur-

thermore, the power of the procedure with a data-driven order of hypotheses reduces drastically in the case of variance heterogeneity among the variables. To be powerful,

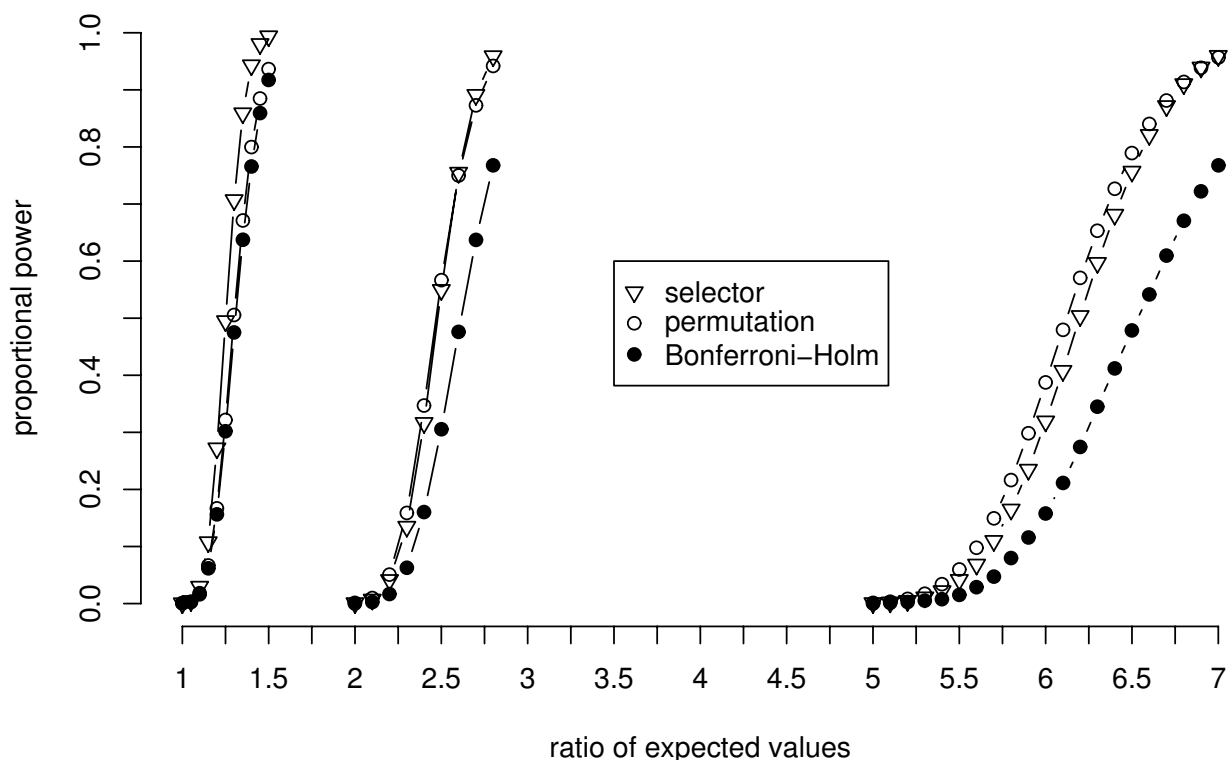


Figure 2

Power for varying levels of relevance thresholds. The figure shows the proportional power for varying levels of relevance thresholds. The power was computed using a sample size of $n_i = 7$ and a correlation among the variables of $\rho = 0.3$. To simulate the power for the left curves the thresholds were set to $\theta_{lower} = \theta_{upper} = 1$, for the curves in the middle $\theta_{lower}^{-1} = \theta_{upper}$ = 2 were chosen and the power curves on the right were computed using $\theta_{lower}^{-1} = \theta_{upper} = 5$.

the procedure requires approximately homogeneous variances after the logarithmic transformation. Corresponding simulation results to these influences are given in Frömke [1]. Although the impact of a varying number of variables was not examined, it can be assumed to have significant effects as well.

Implementation

Method comparison using a publicly available dataset

This section illustrates the application of the two procedures using a subset of the microarray study published by Khan et al. [27]. The entire data set consists of four subgroups of small, round blue cell tumors (SRBCTs) of childhood. Cell lines are available for all four subgroups, and biopsy material is available for two subgroups. The subset of the original study used here incorporates the biopsy material, which consists of 13 samples of the

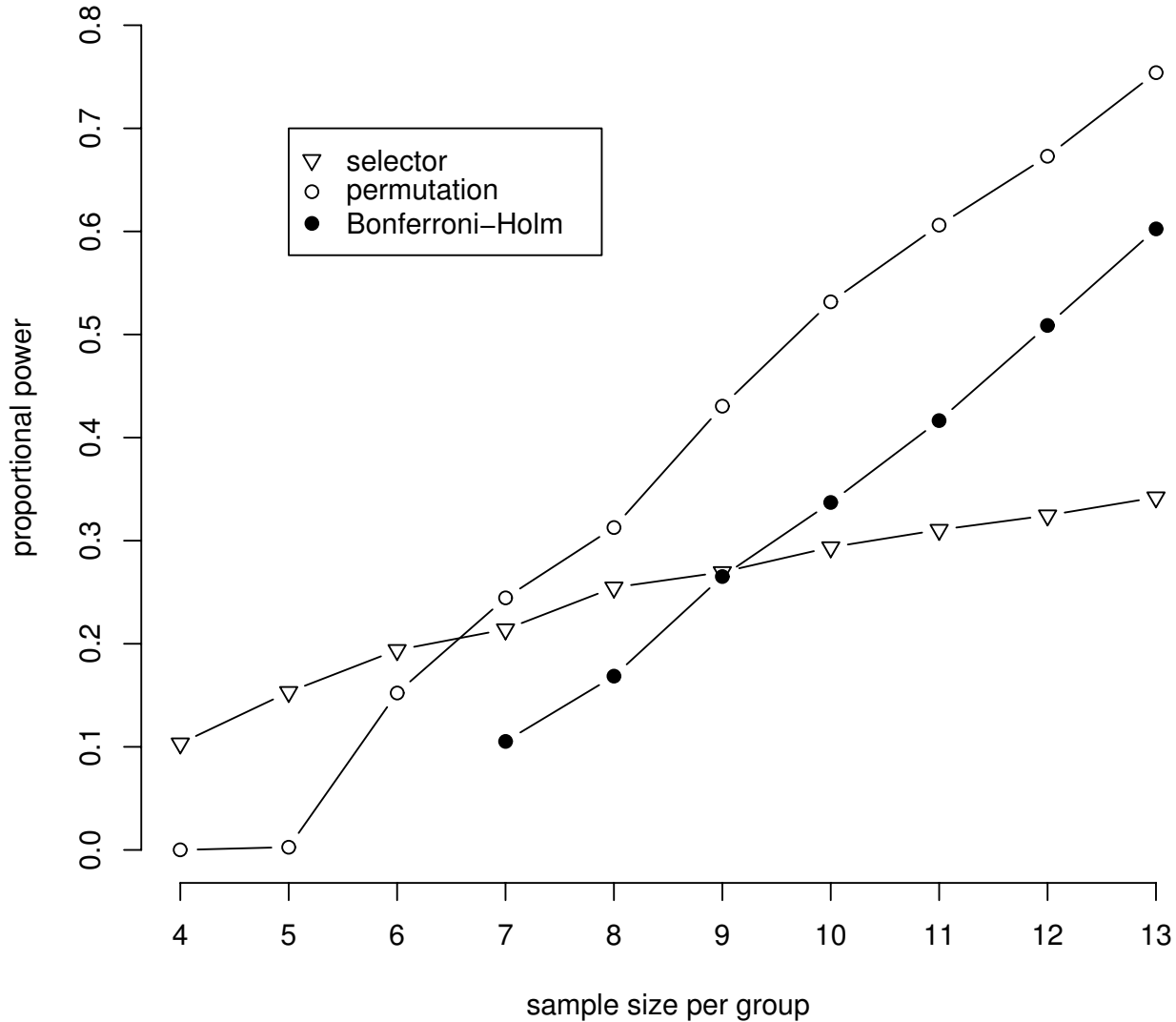


Figure 3
Power for varying levels of sample sizes per group using 50 variables. The figure shows the power for varying levels of samples sizes per group. All settings were simulated using a ratio of expected values for variables under H_1 of 2.35 and a correlation among the variables of $\rho = 0.3$.

Ewing family of tumors (EWS) and 10 samples of the rhabdomyosarcoma (RMS). Furthermore, all 2,308 genes of the original data set will be analyzed. For the following analysis, significant two-fold under- or over expressions to an $\alpha = 5\%$ are sought. Hence, the thresholds are set to $q_{lower}^{-1} = \theta_{upper} = 2$ corresponding to $-\delta_{lower} = \delta_{upper} = 0.6931$.

The results of the relevance-shifted Westfall-Young permutation algorithm, the procedure with a data-driven order of relevance-shifted hypotheses and the Bonferroni-Holm correction are listed in Table 3. The last column provides a ranking number. These ranks are taken from the analysis methods supplement [27], where the top 96 genes were ranked according to importance using artificial neural network techniques.

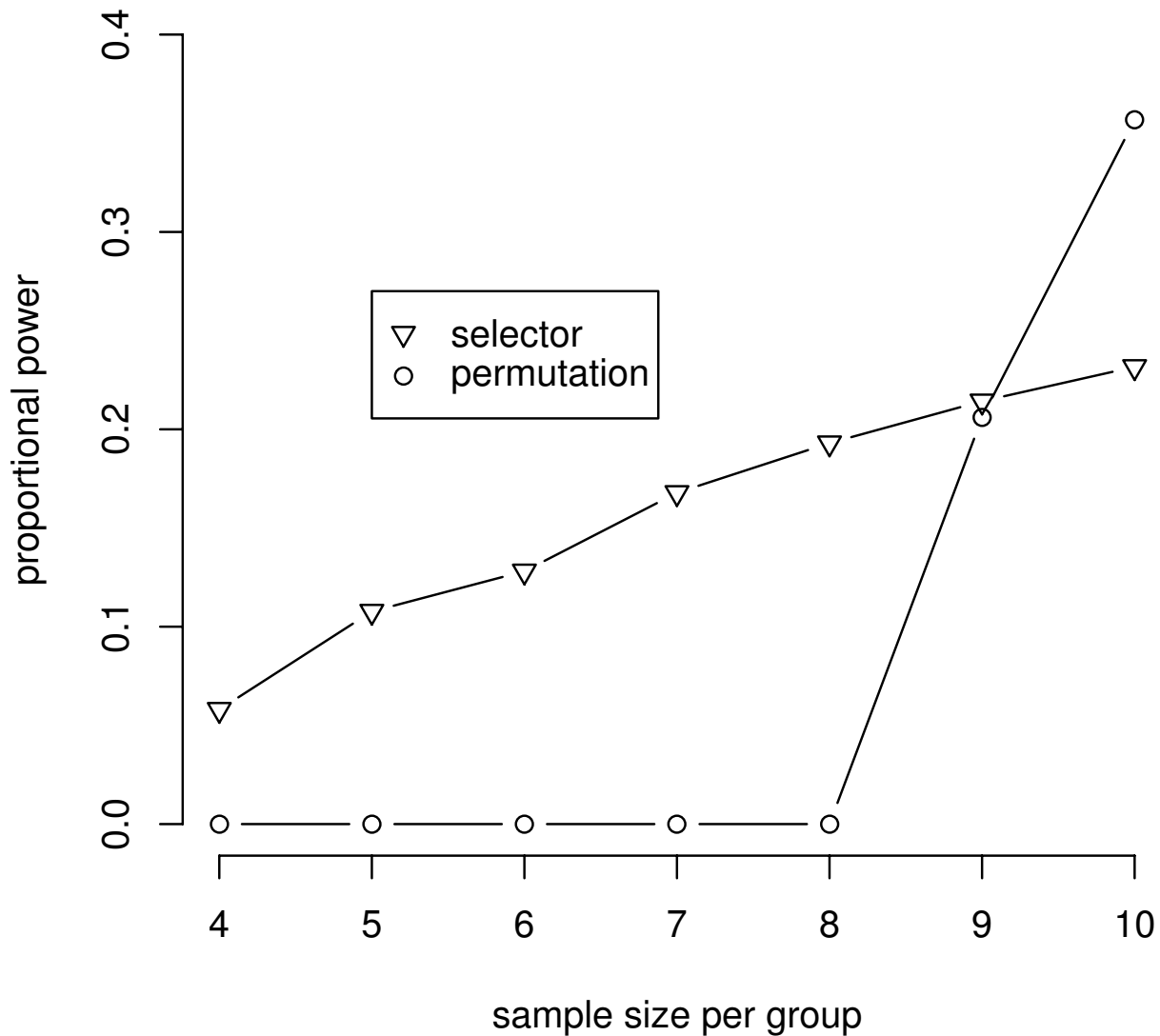


Figure 4
Power for varying levels of sample sizes per group using 5,000 variables. As in Figure 3 the power for varying levels of samples sizes per group are presented. The simulation settings were the same using in Figure 3. Only the expected values of the 50 variables under H_1 were set to 2.5 and 1/2.5 and results consisted of 1,000 simulation runs. The simulations carried out for the permutation algorithm were time consuming. Therefore, the power computed for sample size 10 per group was simulated using 100 runs and the number of permutations in each run was restricted to 100,000.

On top the table shows the results for the five significant genes found by the relevance-shifted Westfall-Young permutation algorithm after 500,000 permutation runs. In contrast, the procedure with a data-driven order of rele-

vance-shifted hypotheses detects three significant genes, where one of them was also found using the above method. Three genes are also found with the Bonferroni-Holm adjustment. They are completely different genes

Table 3: Results from the publicly available dataset. Results from the publicly available dataset: This table shows the results of the relevance-shifted Westfall-Young permutation algorithm using 500,000 permutation runs, the procedure with a data-driven order of relevance-shifted hypotheses and the Bonferroni-Holm correction. Note that the fourth and the sixth column are necessary for the procedure with a data-driven order of relevance-shifted hypotheses only.

procedure	image id.	ratio	selector statistic	p-value	test decision	ranking
permutation	770394	0.051	-	0.00278	(reject H_0)	6
	814260	0.032	-	0.00278	(reject H_0)	75
	244618	24.918	-	0.00348	(reject H_0)	7
	207274	4259.257	-	0.01983	(reject H_0)	2
	43733	0.040	-	0.04832	(reject H_0)	9
selector	207274	4259.257	6.881	0.00003	reject H_0	2
	122159	169.102	6.204	0.02136	reject H_0	40
	296448	1445.051	6.041	0.00145	reject H_0	1
	34849	0.728	5.239	1.00000	accept H_0 , stop procedure	-
Bonferroni-Holm	770394	0.051	-	0.00403	(reject H_0)	6
	244618	24.918	-	0.00403	(reject H_0)	7
	814260	0.032	-	0.00403	(reject H_0)	75

compared to the former procedure, but they were also found using the modified Westfall-Young method.

In this analysis, the permutation algorithm detects more significant variables than both the procedure with a data-driven order of relevance-shifted hypotheses and the α -adjustment of Bonferroni-Holm. As shown in the former section, this can be explained with the general performance of these three methods for the present case of moderately large sample sizes in both groups. However, the procedure with a data-driven order of sequential testing is the only one that found the gene 296448, which according to Khan et al. [27] is the most important one.

Conclusion

The comparison of two groups of individuals with many variables is a common problem in biological studies. In the current literature, procedures are proposed which perform local tests for each variable and correct for multiplicity. Most of these procedures test the point-zero or point-one null hypotheses of a difference or ratio in treatment effects of 0 or 1, respectively. A parametric procedure is available for relevance-shifted hypotheses [7]. In this article, two nonparametric procedures are proposed which perform a local relevance-shifted test on ratio on the two samples for each variable and include a multiplicity correction as well. They are extensions of the Westfall-Young permutation algorithm [18] and of a sequential procedure with data-driven order of hypotheses [20], which consider point-null hypotheses in their original form.

Both new procedures utilize the correlation structure. In the proofs of the original versions, this can be seen in the fact that they consider permutations of the whole observation vectors and not separate permutations for single variables. In the technical procedures, the influence of the correlation among the variables is not seen explicitly

because univariate test statistics and selector functions are calculated. But it is present in the ordering of variables, which is part of both procedures in some way. When the variables are highly correlated, then the relation of their Wilcoxon test statistics or interquartile ranges effectively reflects the relation in the degree of violation of the corresponding null hypothesis. The less these correlations are, the more this relation is overlaid with random influences.

As not all modifications, applied to the point-null versions, could be covered by the theoretical considerations, simulation experiments were carried out for the control of the FWER and for the assessment of the power. In these experiments, the FWER was always controlled for the two-sided test versions discussed in this paper.

The power of the two new proposals and of the Bonferroni-Holm method was similar to the original procedures for point-null hypotheses (cf. Kropf et al. [20]). The procedure with data-driven sequential hypothesis testing uses a nonparametric measure of variability in the pooled samples as an additional source of information. This provides an advantage in small samples if the variances of the different variables are more or less homogeneous in the data. This advantage is lost and even reversed with increasing sample sizes. As discussed in Kropf et al. [20], this is due to the fact, that the probability to detect a difference in the unadjusted tests (which is the major input in the other test procedures) increases faster than the probability of the correct ordering of variables with and without deviations from the null hypothesis. Therefore, this ordering becomes the critical part in the sequential procedure for at least moderately large samples. However, data from microarray and proteomics experiments are commonly characterized by a very large number of variables and small sample sizes. The analysis of such experiments using standard multivariate approaches is inappropriate. The

proposed procedures can be used instead, particularly if relevance shifted hypotheses are of interest.

Authors' contributions

CF basically developed the modifications of the procedures, carried out the simulation studies and prepared the draft of the paper. LAH initiated the investigations, collected the relevant literature and essentially contributed to the modification of the permutation algorithm. SK delivered basic parts for the discussion of the multiple testing problem, contributed to the modification of the procedure with data-driven order and took part in the preparation of the final version of the paper. All authors read and approved the final manuscript.

Acknowledgements

The authors are very grateful to Dr. Frank Bretz and to the anonymous reviewers for their helpful comments and constructive suggestions.

References

- Frömke C: **Relevance-shifted tests for high dimensional data with small sample sizes.** PhD thesis 2006 [<http://www.biostat.uni-hannover.de/research/thesis/>]. University of Hannover, Institute of Biostatistics
- Hauschke D, Schall R, Luus HG: **Statistical Significance.** In *Encyclopedia of Biopharmaceutical Statistics* 1st edition. Edited by: Chow SC. New York: Marcel Dekker; 2000:493-497.
- Lange S, Freitag G: **Choice of delta: requirements and reality – Results of a systematic review.** *Biom J* 2005, **47**:12-27.
- Cariello NF, Piegorsch WW: **The Ames test: The two-fold rule revisited.** *Mutat Res* 1996, **369**:23-31.
- Hothorn LA: **Statistical analysis of in vivo anti cancer experiments: Tumor growth inhibition.** *Drug Inf J* 2006, **40**:229-238.
- Guo L, Fang H, Collins J, Fan X, Dial S, Wong A, Mehta K, Blann E, Shi L, W T, Dragan YP: **Differential gene expression in mouse primary hepatocytes exposed to the peroxisome proliferator-activated receptor a agonists.** *BMC Bioinformatics* 2006, **7**(Suppl 2):S18.
- Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: model validation, design issue and standard error application.** *Genome Biol* 2001, **2**:research0032.1-0032.11.
- Schaid DJ, McDonnell SK, Hebring SJ, Cunningham JM, Thibodeau SN: **Nonparametric tests of association of multiple genes with human disease.** *Am J Hum Genet* 2005, **76**:780-793.
- Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18**:1454-1461.
- Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Statist Soc B* 1995, **57**:289-300.
- Polacek DC, Passerini AG, Shi C, Francesco NM, Manduchi E, Grant GR, Powell S, Bischof H, Winkler H, Stoeckert CJ, Davies PF: **Fidelity and enhanced sensitivity of differential transcription profiles following linear amplification of nanogram amounts of endothelial mRNA.** *Physiol Genomics* 2003, **13**:147-156.
- Westfall PH, Kropf S, Finos L: **Weighted FWE-controlling methods in high-dimensional situations.** In *Recent Developments in Multiple Comparison Procedures Volume 47.* Edited by: Benjamini Y, Bretz F, Sarkar S. Institute of Mathematical Statistics Lecture Notes-Monograph Series; 2004:143-154.
- Chich JF, David O, Villers F, Schaeffer B, Lutomski D, Huet S: **Statistics for proteomics: Experimental design and 2-DE differential analysis.** *Journal of Chromatography B* 2007, **849**:261-272.
- Witt E, McClure J: *Statistics for Microarrays. Design, Analysis and Inference* Chichester: John Wiley & Sons; 2004.
- Dudoit S, H YY, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** 2000. [Technical report #578]
- Speed T: *Statistical analysis of gene expression microarray data* Boca Raton: Chapman & Hall/CRC; 2003.
- Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Statist* 1979, **6**:65-70.
- Westfall PH, Young SS: *Resampling-based multiple testing: Examples and methods for p-value adjustment* New York: John Wiley & Sons; 1993.
- Dudoit S, Shaffer JP, Boldrick JC: **Multiple hypothesis testing in microarray experiments.** *Statist Sci* 2003, **18**:71-103.
- Kropf S, Läuter J, Eszlinger M, Krohn K, Paschke R: **Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses.** *J Statist Plann Inference* 2004, **125**:31-48.
- Kropf S, Läuter J: **Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data.** *Biom J* 2002, **44**:789-800.
- Zimmermann N, King NE, Laporte J, Yang M, Mishra A, Pope SM, Muntel EE, Witte DP, Pegg AA, Foster PS, Hamid Q, Rothenberg M: **Dissection of experimental asthma with DNA microarray analysis identifies arginase in asthma pathogenesis.** *The Journal of Clinical Investigations* 2003, **111**:1863-1874.
- Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Boguski MS, Lashkari D, Shalon D, Botstein D, Brown PO: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283**:83-87.
- Frömke C.: **See R program nprrest.** [http://www.biostat.uni-hannover.de/software/index_en.html].
- Ge Y, Dudoit S, Speed TP: **Resampling-based multiple testing for microarray data hypothesis.** *Test* 2003, **12**:1-44.
- Fleishman AI: **A method for simulating non-normal distributions.** *Psychometrika* 1978, **43**:521-532.
- Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7**:673-679.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

