A resampling-based empirical Bayes method for precise false discovery rate estimation under dependence

Stijn Hawinkel (stijn.hawinkel@ugent.be)
Ghent University

Luc Bijnens
Janssen Pharmaceutical companies of Johnson and Johnson

Olivier Thas
Ghent University

Research Article

Keywords: Correlation, Flow cytometry, Microbiome, Null distribution, Resampling, Simultaneous inference

Posted Date: June 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1760657/v1

RESEARCH

A resampling-based empirical Bayes method for precise false discovery rate estimation under dependence

Stijn Hawinkel[1*], Luc Bijnens[2,3] and Olivier Thas[1,3,4]

*Correspondence:
stijn.hawinkel@ugent.be
[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
Full list of author information is available at the end of the article

Abstract

Many methods exist for controlling or estimating the false discovery rate (FDR), which is the mean of the false discovery proportion (FDP). However, in the presence of correlation between the test statistics, the FDP may show considerable variation between repeated experiments. Here we first provide an intuitive explanation for this phenomenon by demonstrating an increased sampling variability of the test statistics when tests are dependent. Next, we propose an empirical Bayes solution based on resampling techniques that is applicable to many types of test statistics, and allows to estimate the null distribution more precisely. This reduces the variability of the FDP and even yields an increase in sensitivity. Furthermore the implicit assumptions made by empirical Bayes approaches to false discovery rate estimation are explored. We demonstrate our approach on the differential expression problem in RNA-sequencing, and on a data integration problem of microbiome sequence count and flow cytometry data, where the test statistics are correlated by design. Our algorithm is available as the R-package `reconsi`.

Keywords: Correlation; Flow cytometry; Microbiome; Null distribution; Resampling; Simultaneous inference

1 Introduction

In contemporary biological research, often hundreds to thousands of features are measured on a smaller number of samples. Statistical hypotheses are then tested on each of these features, posing a huge multiple testing problem. Limiting the probability of making at least one false rejection (the family-wise error rate, FWER) would require very conservative tests. For that reason, the quantity of choice to control is the number of false rejections among all rejections: the false discovery proportion (FDP) [1]. Different successful strategies have been developed to control or estimate the expected FDP over many experiments, which is called the *false discovery rate* [1, 2]. For biological or technical reasons, the test statistics of the different features may be correlated. In this article we consider the motivating example of differential absolute abundance testing in microbiomics. There, test statistics are correlated because they make use of the same estimate of total cell concentration for all features. Also correlations between the features in omics datasets, e.g. gene co-expression networks in transcriptomics, can introduce correlation between test statistics. This correlation has the undesirable consequence of inflating the variability of the FDP [3, 4]. This means that even though *on average* the false discovery rate is controlled

at a preset level, the outcome of the analysis of a single experiment can be overly conservative, or worse, overly liberal. Past research efforts for inference under dependence often focused on the correlation structure of the test statistics [3, 5–8]. A common approach is then to find a low dimensional approximation of this correlation structure, and correct for it to eliminate the effect of the dependence. Still, many results only hold for linear models [6, 7], limiting their applicability. Some methods even require the correlation structure of test statistics to be known [5]. In other cases, this correlation structure needs to be estimated from the data [3, 8, 9], which can be challenging for high-dimensional, non-normal data. Another stream of publications provides expressions to estimate the FDP of a single experiment and its variance directly [5, 10]. Still, no algorithm or software implementation to tackle the variability of the FDP in a general setting is available.

Another important proposal is to use a custom estimator of the null distribution of all test statistics combined, when the test statistics are correlated [3, 11]. These publications champion the estimation of an *empirical*, univariate null based on the central part of the histogram of the observed test statistics. Performing inference based on this empirical null is then another way of conditioning on the given experiment and reducing instability of statistical inference. Yet the definition of "central" test statistics is somewhat arbitrary, and the procedure may break down when too many null hypotheses are false. In addition, the conditions for modelling all test statistics of a single experiment through a univariate distribution have not been thoroughly explored. Another suggestion is to estimate the null distribution after Fourier transformations, using characteristic functions [12].

Permutations have often been used for multiplicity correction, as they can provide additional information on the joint null distribution. A well known example in the context of -omics data is the SAM method, which uses permutations for estimating the false discovery rate for a given significance threshold [13, 14]. [3] mentioned the use of permutations for the estimation of a conditional false discovery rate, but only for binned test statistics and without explicit estimation of the null distribution. [15] used the posterior probabilities of false findings, estimated through the empirical Bayes framework of [2], as weights in several multiplicity correction methods. One of their applications is to downweigh permutation test statistics from non-null features in the estimation of the overall null distribution, because these non-null features would have a more dispersed permutation distribution. This idea is somewhat similar to the one presented here, although we do not pool the test statistics from all permutations. Rather, we treat every permutation instance separately, which will allow us to account for correlation between test statistics.

The organisation of this article is as follows: Section 2 discusses the problem of the inflated variability of the FDP in the presence of correlation. Section 3 explores the conditions for univariate normality of the null distribution, and then outlines our empirical Bayes strategy for estimating this null distribution through resampling. Next, a small simulation study is presented to demonstrate the efficacy of our method. In Section 4, case studies on microbiomics and transcriptomics data are presented, including realistic simulation studies and real data analyses. Section 5 contemplates on some of the key points and provides an outlook for the future.

2 Multiple testing with correlated test statistics

2.1 Multiple testing

Suppose a large number (p) of features is measured for a smaller number (n) of samples. These features may be gene expressions, species abundances, substrate concentrations or

others. These measurements result in an $n \times p$ matrix $\mathbf{Y}$, with element y_{ij} the observed value of feature j in sample i. In addition, an $n \times d$ matrix $\mathbf{X}$ may be available with baseline sample variables, e.g. treatment group or a continuous covariate. Frequently, the scientific question is to test a statistical hypothesis H_j on every feature j, i.e. on every vector $\mathbf{y}_{.j}$. This statistical hypothesis may either only involve $\mathbf{y}_{.j}$, or relate to the association of $\mathbf{y}_{.j}$ with one or more variables from $\mathbf{X}$. Whichever statistical test is used, the resulting test statistic can often be converted to a z-statistic through

$$Z_j = \Phi^{-1}\big(D_j(T_j)\big), \tag{1}$$

where D_j is the distribution function of test statistic T_j under null hypothesis H_j and Φ^{-1} is the standard normal quantile function. Statistical testing then proceeds by comparing the observed z-statistic Z_j with the quantiles of the standard normal distribution, and rejecting the j-th null hypothesis at the significance level α when $|Z_j| > \Phi^{-1}(1-\alpha/2)$. Simply performing each individual test at the α significance level leads to many false positives. This is the multiple testing problem. Several procedures have been developed to control some risk of false rejections. Call R the total number of hypotheses that are rejected (the "discoveries"), of which V are in fact true null hypotheses. Call p_0 the number of true null hypotheses and $\pi_0 = \frac{p_0}{p}$ the fraction of true null hypotheses. One popular risk measure is the family-wise error rate (FWER), which is defined as the chance of having at least one false positive: $\mathrm{P}(\mathrm{V} \geq 1)$. Yet methods controlling the FWER typically have small sensitivity when p is large. As an alternative, [1] defined the false discovery proportion (FDP):

$$\mathrm{FDP} = \frac{V}{R}. \tag{2}$$

It is the fraction of the number of rejected null hypotheses R for which the null hypothesis is in reality true, and it is set to 0 when R=0 by convention. The FDP is a random variable, whose realization is unique to every experiment. In experiments where many hypotheses are being tested (p large), controlling the expected FDP is much less conservative than controlling the FWER. Rather than limiting the probability of making at least one false discovery, the researcher accepts a small but controlled proportion of false findings among the discoveries. The expected FDP is commonly called the *false discovery rate* (FDR). Throughout the literature, slightly different definitions of the FDR exist (see [16] for an extended discussion). [1] defined the FDR as

$$FDR = E(FDP|R > 0)P(R > 0). \tag{3}$$

The factor P(R>0) implies that the FDP is considered to be zero when no discoveries are made (R = 0). [16] defined the marginal FDR (mFDR) as

$$mFDR = E(V)/E(R). \tag{4}$$

Definition (4) was rejected by [1] because it would be impossible to control the false discovery rate when $p_0 = p$ [4]. Still, the case $p_0 = p$ is very unlikely for large p [17]. The relationship between both quantities is given by [18]

$$\frac{E(V)}{E(R)} = E\left(FDP\right) + \frac{\mathrm{Cov}\left(FDP, R\right)}{E(R)}, \tag{5}$$

and hence the FDR and the mFDR are only equivalent when the FDP and the total number of rejections R are independent. The method that we will present greatly reduces this dependence Cov (FDP, R), almost equalizing FDR and mFDR and thus allowing both to be controlled at a prespecified level. Moreover, our method reduces the variance of the FDP as compared to many competitor methods.

2.2 A mixture model for estimating the false discovery rate

Different procedures have been developed to control or estimate the false discovery rate. [1] proposed a step-down method based on p-values that guarantees control of the FDR. [19] extended this procedure to make it work for any form of dependence between the p-values. [2] proposed a procedure to estimate the probability of a false discovery given the observed test statistic, $P(H_0 \text{ is true}|Z = z)$. It is known as the *local* false discovery rate or fdr. Their method is based on the premise that the observed test statistics in a single experiment are drawn from a mixture distribution of null and non-null densities:

$$f(z) = \pi_0 h(z) + (1 - \pi_0)a(z), \tag{6}$$

where $h(z)$ is the density of the null distribution, $a(z)$ is the density of the test statistics under false null hypotheses and π_0 is the proportion of null hypotheses that are true. The fdr can then be estimated directly by invoking Bayes' theorem:

$$\begin{aligned} fdr(z) &= P(H_0 \text{ is true}|Z = z) \\ &= \frac{g(z|H_0)P(H_0 \text{ is true})}{g(z)} = \frac{h(z)\pi_0}{f(z)}. \end{aligned} \tag{7}$$

The corresponding tail-area false discovery rate (Fdr) reflects the false discovery rate when declaring tests with z-values falling below a certain threshold significant, and is defined (for z-values lying in the lower tail) as:

$$Fdr(z) = P(H_0 \text{ is true}|Z \leq z) = \frac{H(z)\pi_0}{F(z)} \tag{8}$$

We can see that when rejecting all hypotheses for which $Z \leq z$, $H(z)\pi_0$ = E(V) and F(z) = E(R) (note that R is random since we fix the desired FDP, such that the number of rejections R varies across experiments). Hence the tail-area false discovery rate agrees with the mFDR [3].

2.3 Dependence inflates the variability of the false discovery proportion

In genomics applications, the test statistics are often correlated. This correlation may result from associations between the columns of $\mathbf{Y}$, e.g. due to co-expression of certain genes. In microbiome studies correlations may be caused by microorganisms living in the same niche. Alternatively, the common dependence of the test statistics on a matrix of regressors $\mathbf{X}$ or another common factor may engender some correlation between test statistics. Whatever its source, the dependence between the test statistics inflates the variability of the FDP [3, 4], as shown in Figure 1. The dependence also causes the FDR and mFDR to diverge, as the FDP tends to be higher when more discoveries are made (see Figure S1).

An explanation for the large variance of the FDP lies in the greater within-experiment variability of the test statistics under the null hypothesis in the presence of correlation. This is illustrated in Figure 2. When the null z-values within an experiment are dispersed and/or shifted with respect to the standard normal density, many of them are declared significant when the standard normal distribution is assumed for h(z), resulting in a FDP larger than the nominal level. Conversely, when the null z-values are less variable than expected under the standard normal, there are less extreme test statistics and the FDP is lower than the nominal level. This phenomenon also explains the divergence between the FDR and the mFDR. When the z-values are dispersed and/or off center, many discoveries R are made, of which many are false ($FDP = V/R$ large). Hence the dependence between test statistics leads to $\text{Cov}(FDP, R) > 0$, such that mFDR $>$ FDR according to (5). This implies that discoveries in an experiment with many discoveries are less reliable than those from an experiment with few discoveries, which is not desirable.

Ideally, one would want a multiplicity correction method for which 1) the FDP varies little around the nominal false discovery rate and 2) if the FDP varies, it does so independently from the total number of discoveries such that $FDR = mFDR$.

3 An empirical Bayes resampling method for reducing the variance of the false discovery proportion

3.1 The univariate normality null model

We start by formulating the statistical model which forms the basis of empirical Bayes methods for false discovery rate estimation. First, we explore the conditions under which the vector of test statistics for which the null hypothesis holds true, which is denoted by $\mathbf{Z}_0$, can be approximated by a univariate normal distribution h in the mixture distribution (6). Next we demonstrate that the mean and the variance of this normal distribution vary between repeated experiments, motivating the estimation of these two parameters.

We assume that the vector $\mathbf{Z}_0$ has a joint multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}_0$ with all diagonal elements equal to one. Thus all individual test statistics in $\mathbf{Z}_0$ have a marginal standard normal distribution, but they may be correlated. This statement holds over repeated experiments. Suppose now that there exists an underlying latent variable that varies between experiments. This latent variable may be e.g. related to particular conditions in the lab, to sample preparation, or to a variable common in the calculation of the test statistics, such as the estimated total cell count in flow cytometry (see Section 4.1 below). We now look at the h component in the mixture distribution (6) as the distribution of the test statistics in $\mathbf{Z}_0$ for a given experiment, or, more precisely, for a given latent variable. This component distribution h is no longer equal to the marginal null distribution (over repeated experiments) of the individual test statistics. We will refer to this distribution as the collapsed distribution.

We further assume that the latent variable stochastically depends on the sample mean of all p test statistics, say $\bar{Z}$. We first investigate the conditional distribution of $\mathbf{Z}_0$, given $\bar{Z}$. Since both $\mathbf{Z}_0$ and $\bar{Z}$ have normal distributions, so has $\mathbf{Z}_0 \mid \bar{Z}$. We thus only have to look for its mean and variance. We will use the notation $\bar{Z} = p^{-1}\mathbf{1}^t\mathbf{Z}$, $\sigma^2_{\bar{Z}} = \text{Var}(\bar{Z}) = p^{-2}\mathbf{1}^t\mathbf{\Sigma}\mathbf{1}$, with $\mathbf{1}$ a column vector with all elements equal to 1 and $\mathbf{\Sigma} = \text{Var}(\mathbf{Z})$, $\mathbf{\Sigma}_0 = \text{Var}(\mathbf{Z}_0)$ and $\mathbf{\Sigma}_{0\bar{Z}} = \text{Cov}(\mathbf{Z}_0, \bar{Z})$. We then find

$$\begin{aligned} E(\mathbf{Z}_0 \mid \bar{Z}) &= E(\mathbf{Z}_0) + \text{Cov}(\mathbf{Z}_0, \bar{Z})\text{Var}(\bar{Z})^{-1}(\bar{Z} - E(\bar{Z})) \\ &= \mathbf{0} + (\sigma^2_{\bar{Z}})^{-1}(\bar{Z} - E(\bar{Z}))\mathbf{\Sigma}_{0\bar{Z}} \end{aligned} \tag{9}$$

and

$$\begin{aligned}\text{Var}(\mathbf{Z}_0 \mid \bar{Z}) &= \text{Var}(\mathbf{Z}_0) - \text{Cov}(\mathbf{Z}_0, \bar{Z})\text{Var}(\bar{Z})^{-1}\text{Cov}(\bar{Z}, \mathbf{Z}_0) \\ &= \boldsymbol{\Sigma}_0 - (\sigma^2_{\bar{Z}})^{-1}\boldsymbol{\Sigma}_{0\bar{Z}}\boldsymbol{\Sigma}^t_{0\bar{Z}}.\end{aligned} \tag{10}$$

The condition for the collapsed distribution of $\mathbf{Z}_0$ to be a univariate normal distribution, is that all components of $\mathbf{Z}_0$ have the same normal distribution (i.e. the same mean and variance) and are independent when conditioned on $\bar{Z}$ (i.e. with covariance 0). The mean, variance and covariance can be rewritten in scalar form as

$$\begin{aligned}E(Z_{0j} \mid \bar{Z}) &= \frac{\bar{Z} - E(\bar{Z})}{\sigma^2_{\bar{Z}} p} \sum_{i=1}^{p} \sigma_{ij} \\ \text{Var}(Z_{0j} \mid \bar{Z}) &= 1 - \frac{\left(\sum_{i=1}^{p} \sigma_{ij}\right)^2}{\sum_{i=1}^{p}\sum_{k=1}^{p} \sigma_{ik}} = 1 - \frac{\left(\sum_{i=1}^{p} \sigma_{ij}\right)^2}{p^2\sigma^2_{\bar{Z}}} \\ \text{Cov}(Z_{0j}, Z_{0i} \mid \bar{Z}) &= \sigma_{ij} - \frac{\left(\sum_{k=1}^{p} \sigma_{jk}\right)\left(\sum_{k=1}^{p} \sigma_{ik}\right)}{\sum_{i=1}^{p}\sum_{k=1}^{p} \sigma_{ik}} = \sigma_{ij} - \frac{\left(\sum_{k=1}^{p} \sigma_{jk}\right)\left(\sum_{k=1}^{p} \sigma_{ik}\right)}{p^2\sigma^2_{\bar{Z}}},\end{aligned} \tag{11}$$

with σ_{ij} the covariance between the ith and jth component of $\mathbf{Z}$. We see that after conditioning on $\bar{Z}$, the expected values of the test statistics are no longer 0, but depend on $\bar{Z}$, and the variances have shrunk below 1. Now the question is under which conditions the mean and variance are constant over the features, and the covariance zero. Under compound symmetry, meaning that all covariances are equal (all $\sigma_{ij} = \sigma$), it is easy to see that the mean and variance are constant, and the covariance zero. For slightly less stringent assumptions, univariate normality cannot be shown exactly (see Supplementary Section 2.11), but it is nevertheless often used as an approximation; e.g. [3]. From (11) it is evident that this approximation works best when all features have similar patterns of covariance with other features. This type of correlation structure is common, as it is for instance engendered by a common dependence of all test statistics on a single variable, such as a regressor or the total cell count for flow cytometry data. Yet even when the conditions for an i.i.d. normal distribution for all features are not fulfilled, often the univariate normal assumption holds well enough to iable statistical inference (see Supplementary Section 3.1.2), and a custom normal distribution for h improves upon the assumption of standard normality. It thus makes sense to estimate the two parameters of the normal distribution h, which we discuss next.

3.2 General procedure

For a given value z of the test statistic, (7) and (8) give expressions for fdr and Fdr, in which π_0, $h(z)$ and $f(z)$ need to be estimated from the data. However, the dependence between tests increases the sampling variability of the test statistics, which in turn increases the variance of such density estimators. Reducing the variance of the FDP can thus be achieved by reducing the variance of these estimators. We will focus on the estimation of the null density function $h(z)$; π_0 and $f(z)$ will be estimated in a conventional way (see below). Since $h(z)$ can be assumed to be normal, it can be described by its mean and variance $\boldsymbol{\beta} = (\mu, \sigma^2)$, which vary between repeated experiments. Our approach aims at minimising the expected estimation loss of this normal distribution. This estimation loss is

quantified through the Kullback-Leibler divergence, and is minimised by the following two Bayes estimators [20]:

$$\begin{aligned}\hat{\mu} &= E(\mu \mid \mathbf{Z}_0) \\ \hat{\sigma}^2 &= E(\sigma^2 \mid \mathbf{Z}_0) + \text{Var}(\mu \mid \mathbf{Z}_0).\end{aligned} \tag{12}$$

We condition on the vector $\mathbf{Z}_0$ of test statistics belonging to true null hypotheses, assuming for a moment that we have this information. This choice is justified as only these test statistics contain relevant information on the null parameters $\boldsymbol{\beta}$. Let $\mathcal{J}_0$ denote the set of indexes $j \in \{1, 2, \ldots, p_0\}$ referring to the corresponding features. To achieve (approximate) normality of $\mathbf{Z}_0$, we will also condition on the mean test statistic $\bar{Z}$ of the experiment. Invoking Bayes' theorem, the estimator of the mean μ can be written as (suppressing dependence on σ^2):

$$\begin{aligned}E\left(\mu \mid \mathbf{Z}_0, \bar{Z}\right) &= \int \mu g_\mu(\mu \mid \mathbf{Z}_0, \bar{Z}) d\mu \\ &= \int \mu \frac{f_z(\mathbf{Z}_0 \mid \mu, \bar{Z}) g_\mu(\mu \mid \bar{Z})}{f_z(\mathbf{Z}_0 \mid \bar{Z})} d\mu \\ &= \int \mu \frac{f_z(\mathbf{Z}_0 \mid \mu, \bar{Z})}{f_z(\mathbf{Z}_0 \mid \bar{Z})} dG_\mu(\mu \mid \bar{Z})\end{aligned} \tag{13}$$

in which $g_\mu(\mu \mid \bar{Z})$ is the prior density of μ, $G_\mu(\mu \mid \bar{Z})$ is the corresponding distribution function, $f_z(\mathbf{Z}_0|\mu, \bar{Z})$ is the likelihood evaluated in $\mathbf{Z}_0$, and $f_z(\mathbf{Z}_0 \mid \bar{Z})$ is the marginal likelihood. Suppose we have a sample of B i.i.d. observations from $G_{\mu|\bar{Z}}$, say $\mu_1, \ldots, \mu_B$, then the unknown distribution function $G_{\mu|\bar{Z}}$ in (13) can be replaced by the empirical distribution function $\hat{G}_{\mu|\bar{Z}}$, which is a step function with step height $\frac{1}{B}$. This gives an estimate of $E\left(\mu \mid \mathbf{Z}_0, \bar{Z}\right)$:

$$\hat{E}\left(\mu \mid \mathbf{Z}_0, \bar{Z}\right) = \sum_{b=1}^{B} \mu_b \frac{f_z(\mathbf{Z}_0 \mid \boldsymbol{\beta}_b, \bar{Z})}{f_z(\mathbf{Z}_0 \mid \bar{Z})} d\hat{G}_\mu(\mu_b \mid \bar{Z}) = \sum_{b=1}^{B} \mu_b \frac{f_z(\mathbf{Z}_0 \mid \boldsymbol{\beta}_b, \bar{Z})}{f_z(\mathbf{Z}_0 \mid \bar{Z})} \frac{1}{B}. \tag{14}$$

An analogous reasoning holds for σ^2, thus completing our estimator $\hat{\boldsymbol{\beta}} = (\hat{\mu}, \hat{\sigma}^2)$ (details on the calculation of the (marginal) likelihood will be provided further down). We propose to construct the sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_B$ by resampling the observed data. Since G represents the distribution of μ under the null hypothesis, these resampling procedures (permutations or bootstrap) must mimic the joint null distribution of the test statistics given $\bar{Z}$. We achieve this by resampling the rows from $\mathbf{Y}$ and/or $\mathbf{X}$ to generate datasets for which the null hypothesis of interest holds, while retaining the correlation between their columns (see Supplementary Section 2.1 for details). Resampling procedures are conditional on the data, such that we are automatically conditioning on the latent factor represented by $\bar{Z}$, which is a prerequisite for (approximate) normality of the test statistics. For every resampling instance $b = 1, \ldots, B$, a vector of test statistics $\mathbf{Z}_b$ is calculated in the same way as for the original observed test statistics and a normal distribution is fitted to its elements. The resulting parameter estimate is denoted by $\boldsymbol{\beta}_b$ and is considered as a randomly sampled

element from the prior $G_{\boldsymbol{\beta}|\bar{Z}}$. The prior is thus estimated from the observed data, which gives our procedure an empirical Bayes flavour. An exploration of samples $\boldsymbol{\beta}_b$ from the prior distribution is presented in Supplementary Section 2.2. Note that $\boldsymbol{\beta} = (0, 1)$, corresponding to the standard normal distribution, describes the ensemble of all $\mathbf{Z}_b$'s well [3]; this is illustrated in Figure S4.

The likelihood $f_z(\mathbf{Z}_0 \mid \boldsymbol{\beta}_b, \bar{Z})$ can be written as

$$f_z(\mathbf{Z}_0|\boldsymbol{\beta}_b, \bar{Z}) = \prod_{j=1}^{p} h(z_j \mid \boldsymbol{\beta}_b, \bar{Z})^{I(j\in\mathcal{J}_0)}, \tag{15}$$

which makes use of conditional independence of the test statistics given $\boldsymbol{\beta}$ and $\bar{Z}$, i.e. relying on the assumption from Section 3.1 that $\boldsymbol{\beta}$ captures all dependence between test statistics. The marginal likelihood is $f_z(\mathbf{Z}_0 \mid \bar{Z}) = \sum_{b=1}^{B} f_z(\mathbf{Z}_0|\boldsymbol{\beta}_b, \bar{Z}) d\hat{G}_{\boldsymbol{\beta}} = \frac{1}{B}\sum_{b=1}^{B} f_z(\mathbf{Z}_0|\boldsymbol{\beta}_b, \bar{Z})$. With the notation

$$w(\mathbf{Z}_0, \boldsymbol{\beta}_b) = \frac{f_z(\mathbf{Z}_0|\boldsymbol{\beta}_b, \bar{Z})}{\sum_{b=1}^{B} f_z(\mathbf{Z}_0|\boldsymbol{\beta}_b, \bar{Z})} \tag{16}$$

we can now write

$$\hat{\mu} = \hat{E}(\mu|\mathbf{Z}_0, \bar{Z}) = \sum_{b=1}^{B} \mu_b w(\mathbf{Z}_0, \boldsymbol{\beta}_b). \tag{17}$$

Analogously, we find that

$$\hat{\sigma}^2 = \hat{E}(\sigma^2|\mathbf{Z}_0, \bar{Z}) + \widehat{\mathrm{Var}}(\mu|\mathbf{Z}_0, \bar{Z}) = \sum_{b=1}^{B} \sigma_b^2 w(\mathbf{Z}_0, \boldsymbol{\beta}_b) + \sum_{b=1}^{B} (\mu_b^2 - \hat{\mu}^2) w(\mathbf{Z}_0, \boldsymbol{\beta}_b). \tag{18}$$

Given the estimate $\hat{\boldsymbol{\beta}}$, the local and tail-area false discovery rates can be estimated as in (7) and (8). The density $f(z)$ of (6) is estimated using kernel density estimation, and $F(z)$ is obtained as its integral. Estimation of π_0 follows [17]; see also Supplementary Section 2.5.

The strategy of weighing different models to obtain a final consensus model is also known as *Bayesian model averaging* [21]. Intuitively, more weight is given to null distributions with parameter $\boldsymbol{\beta}_b$ that are more likely to have generated $\mathbf{Z}_0$. We call estimators (17) and (18) relying on (15) the "oracle posterior mean null", because it makes use of the unknown information about which null hypothesis is true or not.

3.3 Expectation-Maximization algorithm

Obviously, in reality the set of true null hypotheses $\mathcal{J}_0$ is not known. However, some collection of null z-values is needed to evaluate the likelihoods $f_z(\mathbf{Z}_0|\boldsymbol{\beta})$. We propose to replace $I(j \in \mathcal{J}_0)$ by some estimator of the probability that the null hypothesis corresponding to a given test statistic z_j is true: $\hat{P}(H_j \text{ is true}|Z = z_j)$. One way to achieve this would be to use some central part of the observed z-value distribution, e.g.

$\hat{P}(H_j \text{ is true}|Z = z_j) = I(z_j \in [q_{0.25}(\mathbf{Z}), q_{0.75}(\mathbf{Z})])$ with $q_{0.25}(\mathbf{Z})$ and $q_{0.75}(\mathbf{Z})$ indicating the first and third quartile of $\mathbf{Z}$ respectively [11]. However, this puts a hard boundary between z-values within the central region, that are all considered to belong to true null hypotheses, and z-values outside this region, that are all considered to belong to false null hypotheses. In reality, $h(z)$ and $a(z)$ may partly overlap, and hence we propose estimating $\hat{P}(H_j \text{ is true}|Z = z_j)$ as in (7): $\hat{P}\big(H_j \text{ is true}|z_j, \hat{f}(z), \hat{\boldsymbol{\beta}}, \hat{\pi}_0\big) = \widehat{\text{fdr}}(z)$. This smoothly weighs the contributions of the z-values to the likelihood. This concept is known as *weighted likelihood* [22]. The weighted likelihoods are thus calculated as

$$f_z(\mathbf{Z}|\boldsymbol{\beta}, \widehat{\text{fdr}}(\mathbf{Z}), \bar{Z}) = \prod_{j=1}^{p} h(z_j \mid \boldsymbol{\beta}, \bar{Z})^{\widehat{\text{fdr}}(z_j)}. \tag{19}$$

Since $I(j \in \mathcal{J}_0)$ is in fact a latent variable, our estimation procedure is an Expectation-Maximization (EM) algorithm [23]. It will iterate between the estimation of $\text{fdr}(z_j)$ as the expected class membership of $\mathcal{J}_0$ (the E step), and the estimation of $\boldsymbol{\beta}$ by minimising the estimation loss (the M-step). In addition, the estimation of the parameter π_0 needs to be included in the iterative procedure as it depends on $\hat{\boldsymbol{\beta}}$ and is required for the estimation of $\text{fdr}(z_j)$.

3.4 Iterative algorithm

In summary, the entire estimation procedure is listed below:

1 Calculate a test statistic for every column of $\mathbf{Y}$: T_j, $j = 1, \ldots, p$.
2 Randomly resample (permute or bootstrap) rows from $\mathbf{Y}$ and/or $\mathbf{X}$ (B times) and calculate the corresponding test statistics: T_j^b, $b = 1 \ldots, B$, $j = 1, \ldots, p$.
3 Convert all test statistics to z-statistics $Z_j = \Phi^{-1}\big(D_j(T_j)\big)$ and $Z_j^b = \Phi^{-1}\big(D_j^b(T_j^b)\big)$, with D_j the marginal distribution function of T_j under the null hypothesis.
4 Estimate $f(z)$ from $\mathbf{Z}$ through kernel density estimation. Estimate $F(z)$ as $\hat{F}(z) = \int_{-\infty}^{z} \hat{f}(e)de$. Compute the sample mean and sample variance for every $\mathbf{Z}_b$ and denote the vector with these estimates as $\boldsymbol{\beta}_b$.
5 Set $\hat{p}_0 = 1$ and $\hat{\boldsymbol{\beta}} = (0, 1)$ as starting values.
6 Estimate $\widehat{\text{fdr}}(z_j) = \min\left(\frac{h(z_j|\hat{\boldsymbol{\beta}})\hat{\pi}_0}{\hat{f}(z_j)}, 1\right)$.
7 Calculate the weights $w(\widehat{\text{fdr}}(z), \boldsymbol{\beta}_b, \mathbf{Z})$ using equations (16) and (19).
8 Estimate $\boldsymbol{\beta}$ from $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_B$ through (17) and (18).
9 Estimate $\hat{\pi}_0(\hat{\boldsymbol{\beta}}, \mathbf{Z})$ using the procedure of [17] (see Supplementary Section 2.5).
10 Repeat steps 6-9 until convergence. Convergence is assumed when the squared change in $\hat{\pi}_0$ and the square root of the mean squared change in $\widehat{\text{fdr}}(\mathbf{Z})$ are both smaller than 10^{-4}.
11 Calculate $\widehat{\text{Fdr}}(\mathbf{Z})$, and perform inference based on $\widehat{\text{Fdr}}(\mathbf{Z})$ or $\widehat{\text{fdr}}(\mathbf{Z})$.

3.5 Exploratory simulation study

The efficacy of our method for multiple testing correction is demonstrated in a small simulation study.

3.5.1 Data generation

Multivariate standard normal data with $n = 50$ samples and $p = 1000$ features are generated. Three different simulation settings for the correlations between the features are

considered: 0, 0.25 and 0.5. The samples were evenly split into two groups and for 25% of the features, the mean was increased by 0.6 in one group. 100 Monte-Carlo instances were generated.

3.5.2 Competitor methods

The Wilcoxon rank sum test and the two-sample t-test were applied to test all features for association with the grouping factor, and we performed multiplicity correction in the following ways. Our resampling method for estimation of the null distribution h was used with 1000 permutations of the grouping factor. For illustrative purposes, we also included two oracle estimators of the null distribution. The first is the "oracle MLE null", which estimates the mean and variance of the null density as the sample mean and variance of $\mathbf{Z}_0$. The other oracle estimator is the "oracle posterior mean null" from (17) and (18). In addition, h was estimated through the standard normal null, as an empirical null (including an asymmetric version) [2], and using the Fourier transform [12]. In all cases, the tail-area false discovery rate $\text{Fdr}(z)$ was used for inference. The `FAMT` method was used as a representative of the factor analysis based approaches [7]. Finally also Benjamini-Hochberg was applied to the raw p-values, as well as its adaptive version which includes estimation of the fraction of true null hypotheses [19].

3.5.3 Performance metrics

Features were considered significant if they had an estimated Fdr or adjusted p-value <0.1. For all methods the proportion of false null hypotheses rejected (the true positive proportion, TPP) and the FDP were calculated for each Monte-Carlo run. The sensitivity, FDR and mFDR were then approximated as a (weighted) average over these runs for every scenario. Finally, the mean squared error (MSE) of the FDP with respect to the nominal level was calculated for every scenario.

3.5.4 Exploratory simulation results

The results for the Wilcoxon rank sum test are shown in Figure 3. The variances of the FDP and the TPP increase with increasing correlation between the features when the Benjamini-Hochberg correction or the standard normal null are used (see also Figure S10). For the methods based on the resampling null and oracle null, there is very little variability in the FDP or TPP, and the FDR and mFDR are almost identical and equal to the nominal level. The sensitivity of our resampling method is equal to that of the standard normal null under independence between features, but increases with the correlation between the features. In addition, our resampling method is insensitive to an imbalance in the direction of the effects, unlike the empirical null and factor analysis (see Figure S9 and Supplementary Section 3.1 for details). Similar results were obtained for the two-sample t-test (see Figure S11).

Figure S24 shows another simulation study for a scenario without a grouping factor (one-sample t-test) and where the nonparametric bootstrap is used in step 2 of our algorithm. The results are very similar to those with permutations described above. Additional simulations explored the effect of different correlation structures, more extreme values for π_0 and varying magnitudes of n and p. Figure S14 and S15 show that our resampling method has a smaller variance of the FDP than its competitor methods under different correlation structures in the data, although our method becomes slightly liberal when the correlation

structure is not compound symmetry. In addition, our method's sensitivity increases with the correlation strength. Figures S19 and S20 reveal that our method, as well as the empirical null, the Fourier null and factor analysis become slightly liberal as π_0 approaches 1. This is not surprising as this violates the assumptions of the mixture mode (6). Figures S21 and S23 show that our resampling method achieves accurate false discovery rate estimation when either n or p are sufficiently large, as do other methods.

4 Case studies

4.1 Case study 1: integrating sequence count and flow cytometry data

In microbiome studies, the bacterial composition of sample is determined by sequencing marker genes. The marker genes uniquely identify bacterial species, henceforth referred to as taxa. Because of all the technical manipulations, the eventual total number of sequence counts is unrelated to the initial number of cells or biomass. Hence, inference can only be related to the *relative abundances* of the taxa. This poses problems of compositionality: an increase in the relative abundances of one taxon inevitably leads to the decrease of others.

Flow cytometry provides an easy way to count cell numbers in microbial samples. The product of the relative abundance from the sequencing assay with this total cell count can serve as an estimate of the absolute cell concentrations of each taxon, eliminating the compositional effect [24–26]. Taxon-wise tests for differences in cell concentration between two or more sample groups are then said to look for "differential *absolute* abundance" (DAA), in analogy to the classical "differential (relative) abundance" (DRA).

The DAA test statistics of the taxa are correlated, because a change in a flow cytometry count affects all test statistics in the same way. Additionally, biological correlations between taxa (e.g. due to competition or cross-feeding) can also cause the relative abundances to be correlated. Even though the source, type and strength of the correlation between the test statistics are unknown, it is crucial to address it in the multiplicity correction of the DAA tests. In this section we apply our new resampling-based empirical Bayes method to several simulated and real datasets. Four microbiome datasets are considered, from engineered ("Props2016") [25], freshwater ("Props2018") [27], human ("Vandeputte2017") [24] and forest ("Rivett2018") [26] origin. The grouping factors to test for absolute differential abundance for the first 3 datasets are: 'reactor cycle', 'lake' and 'health status', respectively. For the Rivett2018 dataset no obvious grouping factor was available, but this dataset will be used as a template for parametric simulation and real data shuffling (see next section).

4.1.1 Simulation study setup

Data were generated through parametric as well as non-parametric simulation and real data reshuffling, as in [28]. For the parametric simulation, it is assumed that the sequence counts follow the negative binomial distribution. The parameters of this distribution are estimated from the real datasets through maximum likelihood. Next, pairs of mean and dispersion parameter estimates are sampled from this pool of estimates, and synthetic sequence counts are drawn from the corresponding negative binomial distributions. This is done by treating all taxa as independent in one setting, and by using an estimated taxon correlation network in another [28]. Fold changes to the relative abundances of DAA taxa were set to 1 (null scenario), 5 and 10; differential abundances are introduced "with compensation" as in [28], such that the abundances of null taxa are left unchanged. The sequence count data in the

non-parametric simulation are generated using `SimSeq` as described in [29]. In both cases, the flow cytometry counts were drawn from the pool of observed counts from each dataset, so differential abundance is only introduced in the sequence count part of the data. For the parametric simulation, tests for relative as well as absolute differential abundance are performed, resulting in the following four scenarios of correlation between test statistics: "None" (no correlation between taxa, test for DRA), "FC" (no correlation between taxa, test for DAA), "Cor" (correlation between taxa, test for DRA), "FC + Cor" (correlation between taxa, test for DAA). In the non-parametric simulation, correlation between taxa can be automatically included by resampling complete samples (correlation inherited from the original data) or feature counts can be sampled independently, and tests for differential relative as well as absolute abundance were performed, again leading to "None", "Cor", "FC" and "Cor + FC" scenarios. Sample sizes were set to 20, 50 and 100 for the parametric simulations, and to 40 and 56 for the non-parametric simulations. In both cases, true proportions of null taxa were set to 0.5, 0.75 and 0.9, and the number of taxa was p=1000. In a third type of simulations, mock groups were created in the real datasets by repeatedly, randomly assigning the samples to two groups. DAA was then tested with respect to this mock grouping factor. This mimics a complete null scenario, while retaining correlation between features. Sample sizes of 20, 50 and 100 were used for this scenario. In all settings, 100 Monte-Carlo instances were generated, and both the two-sample t-test and the Wilcoxon rank sum test were used for testing for differential abundance. The same multiplicity corrections and diagnostics as in Section 3.5 were used, with the addition of the SAM method [14]. $B = 1000$ permutations were used for SAM and for our resampling procedure. For the Wilcoxon rank sum test, the ties were randomly broken for the calculations of the test statistics in the resampling procedure (step 2 of the algorithm), but mid-ranks were used in the calculation of the test statistics in the original dataset (see Supplementary Section 2.8 for a motivation).

4.1.2 Simulation study results

An example of the simulation results of parametric simulation on flow cytometry data is shown in Figure 4; an exhaustive presentation can be found in Supplementary Section 4.4. Our multiplicity correction method based on the resampling null distribution, Benjamini-Hochberg correction, and the standard normal null are the only methods that consistently control the FDR and mFDR with little variability of the FDP, and among these our method achieves the highest sensitivity. Our method also shows little discrepancy between the FDR and mFDR.

Surprisingly, the oracle MLE method performs worse than our resampling null method when using the Wilcoxon rank sum tests, as does the empirical null method (see also Figures S25-S40). This is likely caused by the discreteness of the test statistic distribution, resulting from the high zero frequency in the data. This reveals another advantage of using resampling: the ties among zero observations are broken at random in the permuted datasets, which renders the estimated null distribution more smooth. On the other hand, when mid-ranks are used for calculating the test statistics in the permuted datasets, the performance of our method deteriorates (see Supplementary Section 2.8).

In conclusion, for DAA testing of microbiome data, our resampling method provides a slight increase in power with respect to classical methods. On the other hand, our method accurately controls the false discovery rate, unlike the methods based on the empirical or Fourier null.

4.1.3 Real data analysis: testing for differential absolute abundance

Benjamini-Hochberg correction, the standard normal null, empirical null and resampling null multiplicity correction methods as discussed in Section 4.1.1 were applied to three datasets to test for DAA (Props2016, Props2018 and Vandeputte2017). The Venn diagrams of significant taxa resulting from the Wilcoxon rank sum tests are shown in Figure S42. The fitting of the empirical null failed for the Props2018 dataset. Our new method results in more discoveries than with the Benjamini-Hochberg correction or the standard normal null distribution, which is in line with the simulation results.

4.2 Case study 2: differential expression testing in presence of gene co-expression networks

In addition, a set of simulations was based on three RNA-sequencing (RNA-Seq) datasets generated on human cell lines (”neuroblastoma”) [30], human brain tissue (”GTEx”) [31] and human neuroblastoma tumors (”Zhang”) [32]. The grouping factors to test for differential expression were ”ethanol or nutlin treatment”, ”brain region” and ”MYCN amplification” respectively. In this dataset, gene co-expression causes dependence between features (genes).

4.2.1 Simulation study setup

Synthetic datasets were generated and analysed with the same settings as for microbiome data, but only the ”None” and ”Cor” scenarios. The only differences were that the parametric simulations were done with fold changes of 1, 1.25 and 1.5, and the non-parametric simulations were done with samples sizes 10, 30 and 92. The estimated correlations between features and test statistics were in general stronger than for the microbiome data (see Figures S43 and S60).

4.2.2 Simulation study results

Figures 5 and S44-S58 show that our resampling method has a higher power than the competitor methods, and good control of the FDR with little variability of the FDP in some settings. Yet, like competitor methods, our method may become slightly liberal in some other settings, especially with strong correlation. Still in these cases, the MSE of the FDP is still comparatively low (see Figure 6), indicating that the FDP is never too far above the nominal level. The FDR control becomes better at larger sample sizes (see Figures S52 and S56), suggesting that violation of the exchangeability assumption of permutation and slow convergence of the studentized test statistics to asymptotic behaviour may be to blame (see also Supplementary Section 3.1.1 for a demonstration with Gaussian data). Another cause may be the inequality of the correlations between the features, violating the condition laid out in Section 3.1.

Hence, for differential expression testing of RNA-Seq data, our resampling method provides a considerable increase in power with respect to classical methods. It is slightly liberal in some settings, but the low variability of its FDP ensures that no experiments have excessively large numbers of false discoveries.

5 Discussion

Test statistics from large scale hypothesis testing are frequently correlated. Their correlations may originate from technical factors, reflect true biological interactions or they may

simply result from the construction of the test statistics. This correlation causes all statistical tests to have more similar results, i.e. either many or few hypotheses are rejected and the inference is either overly conservative or overly liberal. The fact that the inference over many repeated experiments is on average still correct, is cold comfort to a researcher who is concerned about the reproducibility of his or her own experiments.

We have presented an empirical Bayes method that reduces the variability of the false discovery proportion, while increasing the sensitivity. The method boils down to generating many null distributions using permutations or bootstraps, and then selecting those that best match the distribution of the observed test statistics. For this aim we use intuitive and well established concepts related to weighted likelihood and Bayesian model averaging. Given the power of modern computers, the additional computation times are negligible compared to the cost of gathering more data (see Supplementary Table S3). The only restriction is the sample size, which must be sufficiently large to allow for sufficient variability in the resampling step. We have demonstrated the superiority of our method over other approaches through simulation studies in realistic settings of differential absolute abundance testing for the microbiome, and differential expression testing for the transcriptome.

The assumption of univariate normality for the null component of the mixture model (the ”collapsed null distribution”), which underlies empirical Bayes methods for false discovery rate estimation, has so far only been proven approximately, and under the assumption of bivariate normality for any pair of test statistics [3]. We have investigated the normality of the collapsed null distribution under the assumption of multivariate normality of all test statistics and conditioning on the experiment. We have shown that in this case, univariate normality only holds for the restrictive case of compound symmetry of the covariance matrix of the test statistic. However, we have empirically demonstrated that univariate normality is also well approximated in other cases, and that empirical Bayes methods for false discovery rate estimation perform well even when the assumption of compound symmetry is not met. However, caution is indicated in case of strong departures from this assumption, e.g. a block correlation structure. We have also shown that when the univariate normality assumption holds, the collapsed null distribution is always narrower than the theoretical standard normal null distribution. The often observed widening of the collapsed null distribution [3] must thus be seen as a correction for its lack-of-fit to the univariate null distribution due to unmet assumptions. As opposed to the field of false discovery rate control, empirical Bayes false discovery rate estimation methods cannot be proven mathematically for general settings, but have proven their value in practice. In this work we provide a formal framework for the univariate treatment of test statistics of multiple tests, to facilitate further theoretical development.

In recent years, the false discovery proportion has become a popular measure in controlling false findings in large genomics studies. Several methods have been developed to control its expectation: the false discovery rate. Yet in the presence of correlation between test statistics, the number of discoveries and the false discovery proportion often covary. This leads to a discrepancy between the expected false discovery proportion of a single experiment (the classical false discovery rate), and the expected proportion of false findings reported in the whole literature (the *marginal* false discovery rate). Choosing which one to control is not obvious. Fortunately, we show that while the empirical Bayes approach to false discovery rate estimation by [2] aims to control the marginal false discovery rate, our resampling null approach has the appealing property of breaking the association between number of discoveries and the proportion of false discoveries. This means that the

marginal and classical false discovery rates converge, and that findings in experiments with many feature discoveries become as reliable as findings in experiments with few discoveries. Added to that, our resampling method reduces the variability of the false discovery proportion across experiments, contributing to reproducibility of experiments.

The objectives of our research were twofold: on the one hand we wanted to repeat existing warnings about the instability of simultaneous inference under dependence. We explain this phenomenon by demonstrating an increased sampling variability of null test statistics when the tests are dependent. On the other hand we provide a solution based on empirical Bayes concepts, which comes down to estimating the null distributions through resampling, and basing the inference on this estimate. We hope this work and the `reconsi` R-package provide a practical, powerful and widely applicable method for multiplicity correction under dependence.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
All code and data used for producing the results in the paper are available on the journal's website

Competing interests
Stijn Hawinkel was funded by Janssen Pharmaceutical companies of Johnson and Johnson. Luc Bijnens is currently employed by Janssen Pharmaceutical companies of Johnson and Johnson. Olivier Thas has no competing interests to declare.

Funding
Stijn Hawinkel was funded by Janssen Pharmaceutical companies of Johnson and Johnson.

Authors' contributions
SH conceived the idea for the algorithm; SH and OT developed the method; SH wrote the manuscript; OT and LB provided commens for the manuscript.

Acknowledgments
We are grateful to Ruben Props, Doris Vandeputte and Damian Rivett for sharing their data. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center). We thank Jelle Goeman for fruitful discussions.

Author details
[1]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [2]Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson, Beerse, Belgium. [3]Data Science Institute, I-Biostat, Hasselt University, Hasselt, Belgium. [4]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) **57**(1), 289–300 (1995)
2. Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.: Empirical Bayes Analysis of a Microarray Experiment. Journal of the American Statistical Association **96**(456), 1151–1160 (2001). doi:10.1198/016214501753382129. https://doi.org/10.1198/016214501753382129
3. Efron, B.: Correlation and Large-Scale Simultaneous Significance Testing. Journal of the American Statistical Association **102**(477), 93–103 (2007). doi:10.1198/016214506000001211. https://doi.org/10.1198/016214506000001211
4. Schwartzman, A.: Comment: FDP vs FDR and the Effect of Conditioning. Journal of the American Statistical Association **107**(499), 1039–1041 (2012). doi:10.1080/01621459.2012.712876. Pmid: 24976660. https://doi.org/10.1080/01621459.2012.712876
5. Fan, J., Han, X., Gu, W.: Estimating False Discovery Proportion Under Arbitrary Covariance Dependence. J Am Stat Assoc **107**(499), 1019–1035 (2012). doi:10.1080/01621459.2012.720478. 24729644
6. Leek, J.T., Storey, J.D.: A general framework for multiple testing dependence. Proc. Natl. Acad. Sci. USA **105**(48), 18718–18723 (2008). doi:10.1073/pnas.0808709105. http://www.pnas.org/content/105/48/18718.full.pdf
7. Friguet, C., Kloareg, M., Causeur, D.: A Factor Model Approach to Multiple Testing Under Dependence. Journal of the American Statistical Association **104**(488), 1406–1415 (2009). doi:10.1198/jasa.2009.tm08332. https://doi.org/10.1198/jasa.2009.tm08332

8. Fan, J., Han, X.: Estimation of the false discovery proportion with unknown dependence. Journal of the Royal Statistical Society. Series B, Statistical methodology **79**(4), 1143–1164 (2017). doi:10.1111/rssb.12204
9. Sun, W., Cai, T.T.: Large-Scale Multiple Testing under Dependence. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **71**(2), 393–424 (2009)
10. Efron, B.: Correlated z-values and the accuracy of large-scale statistical estimates. J Am Stat Assoc **105**(491), 1042–1055 (2010). doi:10.1198/jasa.2010.tm09129. 21052523
11. Efron, B.: Large-Scale Simultaneous Hypothesis Testing. Journal of the American Statistical Association **99**(465), 96–104 (2004). doi:10.1198/016214504000000089. https://doi.org/10.1198/016214504000000089
12. Jin, J., Cai, T.T.: Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons. Journal of the American Statistical Association **102**(478), 495–506 (2007). doi:10.1198/016214507000000167. https://doi.org/10.1198/016214507000000167
13. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, ??? (1993)
14. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA **98**(9), 5116–5121 (2001). doi:10.1073/pnas.091062498. 091062498
15. Guo, X., Pan, W.: Using weighted permutation scores to detect differential gene expression with microarray data. Journal of Bioinformatics and Computational Biology **03**(04), 989–1006 (2005). doi:10.1142/S021972000500134X. https://doi.org/10.1142/S021972000500134X
16. Tsai, C.A., Hsueh, H.m., Chen, J.J.: Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data. Biometrics **59**(4), 1071–1081 (2003). doi:10.1111/j.0006-341X.2003.00123.x. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2003.00123.x
17. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences **100**(16), 9440–9445 (2003). doi:10.1073/pnas.1530509100. http://www.pnas.org/content/100/16/9440.full.pdf
18. Heijmans, R.: When does the expectation of a ratio equal the ratio of expectations? Statistical Papers **40**, 107–115 (1999). doi:10.1007/BF02927114
19. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Statist. **29**(4), 1165–1188 (2001). doi:10.1214/aos/1013699998
20. Champion, C.J.: Empirical Bayesian estimation of normal variances and covariances. Journal of Multivariate Analysis **87**(1), 60–79 (2003)
21. Buckland, S.T., Burnham, K.P., Augustin, N.H.: Model selection: An integral part of inference. Biometrics **53**, 603–618 (1997)
22. Hu, F., Zidek, J.V.: The weighted likelihood. The canadian journal of statistics **30**(3), 347–371 (2002)
23. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39**(1), 1–38 (1977)
24. Vandeputte, D., Kathagen, G., Hoe, K.D., Vieira-Silva, S., Valles-Colomer, M., Sabino, J., Wang, J., Tito, R.Y., Commer, L.D., Darzi, Y., Vermeire, S.Š., Falony, G., Raes, J.: Quantitative microbiome profiling links gut community variation to microbial load. Nature **551**, 507 (2017)
25. Props, R., Kerckhof, F.-M., Rubbens, P., De Vrieze, J., Hernandez Sanabria, E., Waegeman, W., Monsieurs, P., Hammes, F., Boon, N.: Absolute quantification of microbial taxon abundances. The ISME Journal **11**, 584–587 (2016). Short Communication
26. Rivett, D.W., Bell, T.: Abundance determines the functional role of bacterial phylotypes in complex communities. Nature Microbiology **3**(7), 767–772 (2018). doi:10.1038/s41564-018-0180-0
27. Props, R., Schmidt, M.L., Heyse, J., Vanderploeg, H.A., Boon, N., Denef, V.J.: Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. Environ. Microbiol. **20**(2), 521–534 (2018). doi:10.1111/1462-2920.13953
28. Hawinkel, S., Mattiello, F., Bijnens, L., Thas, O.: A broken promise: Microbiome differential abundance methods do not control the false discovery rate. Briefings in Bioinformatics, 104 (2017). doi:10.1093/bib/bbx104. /oup/backfile/content$_{p}ublic/journal/bib/pap/10.1093_{b}ib_{b}bx104/1/bbx104.pdf$
29. Benidt, S., Nettleton, D.: Simseq: a nonparametric approach to simulation of rna-sequence datasets. Bioinformatics **31**(13), 2131–2140 (2015). doi:10.1093/bioinformatics/btv124. http://bioinformatics.oxfordjournals.org/content/31/13/2131.full.pdf+html
30. Assefa, A.T., Paepe, K.D., Everaert, C., Mestdagh, P., Thas, O., Vandesompele, J.: Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. Genome Biol **19**, 96 (2018). doi:10.1186/s13059-018-1466-5. 30041657
31. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F.: The Genotype-Tissue Expression (GTEx) project. Nature Genetics **45**, 580 (2013)
32. Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Zhang, W., Thierry-Mieg, D., Wang, J., Furlanello, C., Devanarayan, V., Cheng, J., Deng, Y., Hero, B., Hong, H., Jia, M., Li, L., Lin, S.M., Nikolsky, Y., t. Oberthuer, A., Qing, T., Su, Z., Lababidi, S., Lancashire, L.J., Li, Y., Lu, X.X., Luo, H., Ma, X., Ning, B., Noguera, R., Peifer, M., Phan, J.H., Roels, F., Rosswog, C., Shao, S., Shen, J., Theissen, J., Tonini, G.P., Vandesompele, J., Wu, P.-Y., Xiao, W., Xu, J., Xu, W., Xuan, J., Yang, Y., Ye, Z., Dong, Z., Zhang, K.K., Yin, Y., Zhao, C., Zheng, Y., Wolfinger, R.D., Shi, T., Malkas, L.H., Berthold, F., Wang, J., Tong, W., Shi, L., Peng, Z., Fischer, M.: Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol **16**(1), 133 (2015). doi:10.1186/s13059-015-0694-1. 694

Supplementary material
The R-package `reconsi` implementing our empirical Bayes method is available from BioConductor. Additional supporting information, including source code and data to reproduce the results is available from the journal's website.

Additional Files
Additional file 1 — Supplementary Material
Background details and exhaustive simulation results

Additional file 2 — Code and data
All code and data used to produce the results from the paper

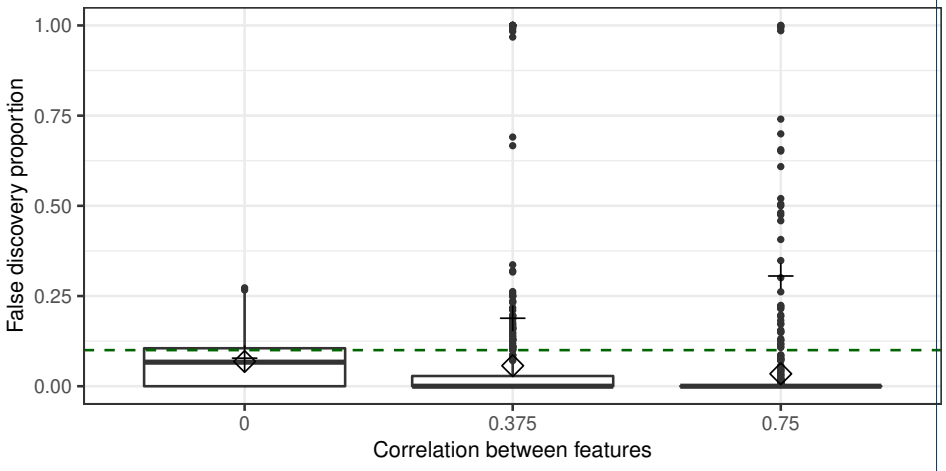

Figure 1 Boxplots of the false discovery proportion (y-axis) after Benjamini-Hochberg correction under increasing correlation between the features of the response matrix (x-axis). The number of simulated features per dataset was 1,000, 500 Monte-Carlo instances were generated. The dashed line indicates the nominal false discovery rate, diamonds indicate the approximated FDR, crosses the approximated mFDR.

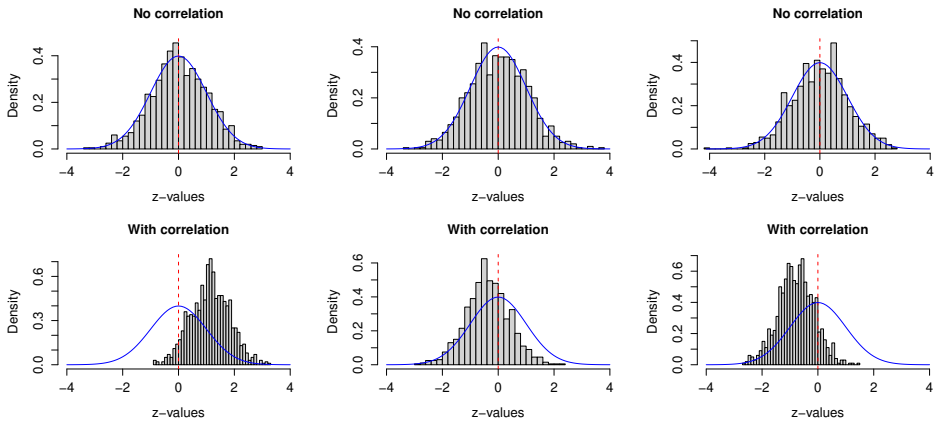

Figure 2 Histograms of z-statistics under the null hypothesis for three experiments without correlation (top) and three experiments with correlation between feature outcomes (bottom). Vertical dashed line indicates 0, smooth curve indicates standard normal density.

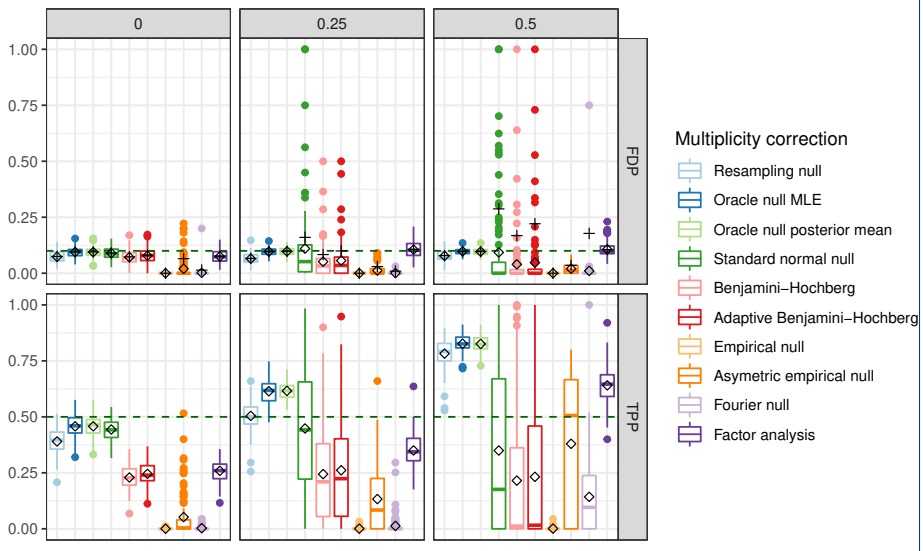

Figure 3 Boxplots of the false discovery proportion (FDP, top) and true positive proportion (TPP, bottom) of multiplicity correction methods on several multivariate normal data with increasing correlation strengths (top labels). In the top panels, diamonds indicate estimated FDR, crosses the estimated mFDR. This figure appears in color in the electronic version of this article.

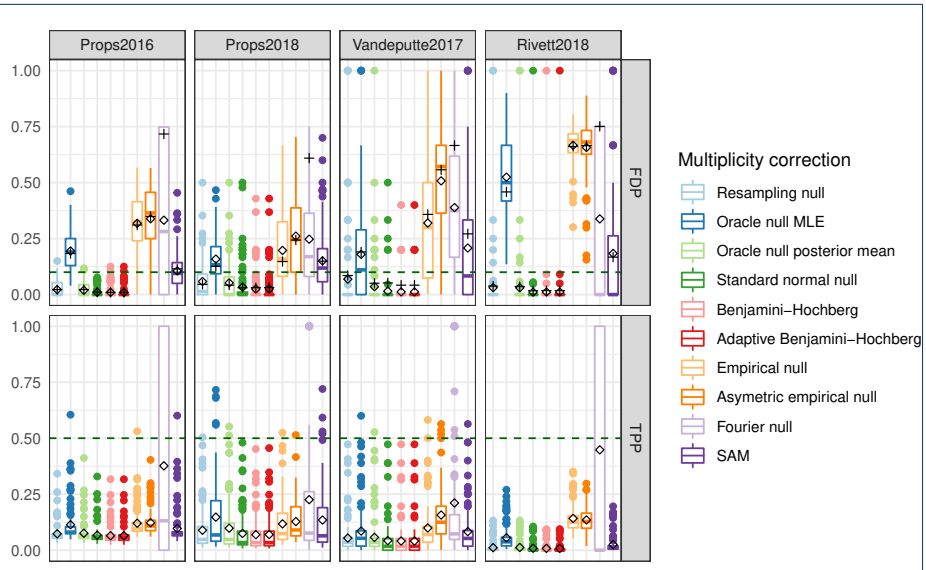

Figure 4 Boxplots of false discovery proportion (FDP, top) and true positive proportion (TPP, bottom) under parametric simulation as a function of template dataset (top panels). The fold change was 10, the sample size 100, the proportion of null taxa was 0.75. Wilcoxon rank sum test was used to test for DAA, sequence counts were generated with taxon correlation networks. Wilcoxon rank sum test was used to test for differential absolute abundance. In the top panel, diamonds represent estimated FDR, crosses the estimated mFDR. This figure appears in color in the electronic version of this article.

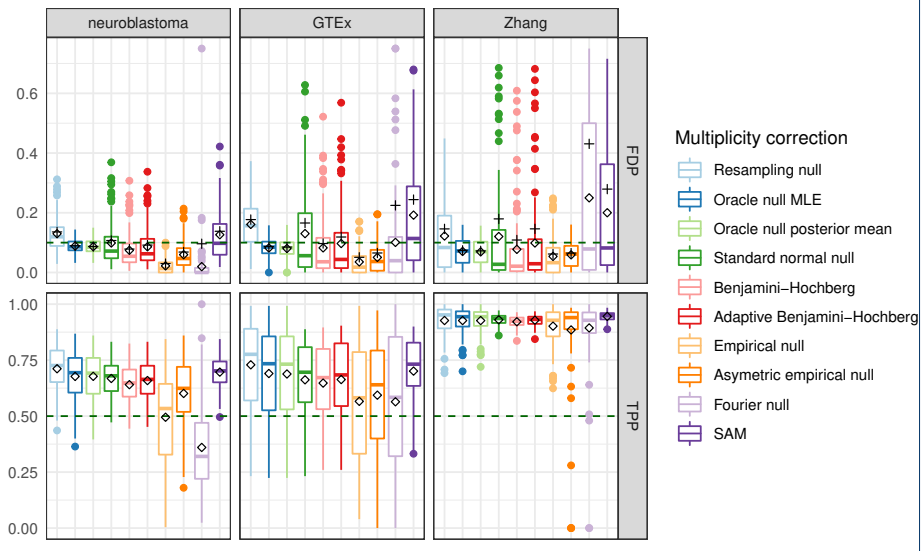

Figure 5 Boxplots of false discovery proportion (FDP, top) and true positive proportion (TPP, bottom) under parametric simulation of RNA-seq data, as a function of template datasets (top), for a sample size of 100, a fold change of 1.25 and proportion of null features of 0.75. Wilcoxon rank sum test was used to test for differential expression. In the top panel, diamonds represent estimated FDR, crosses the estimated mFDR. This figure appears in color in the electronic version of this article.

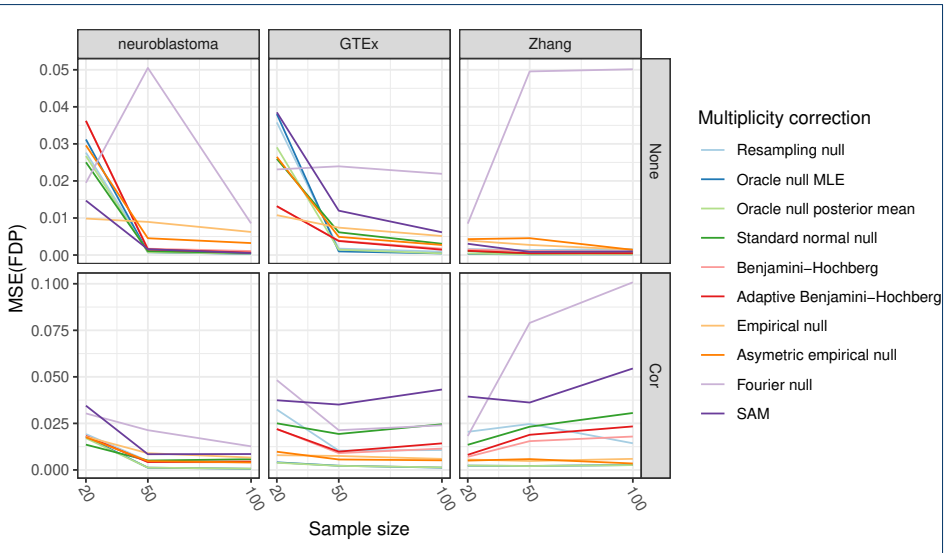

Figure 6 MSE of the FDP (y-axis) under parametric simulation of RNA-Seq data for several template datasets (top panels), correlation scenarios (side panels) and sample sizes (x-axis). The proportion of null features is 0.75, the fold change is 1.25 and the Wilcoxon rank sum test was used to test for differential expression. This figure appears in color in the electronic version of this article.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MultiplicityDependenceSupp.pdf
- CodeAndDataFlow.tar.gz