BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 8 2010, pages 1050–1056
doi:10.1093/bioinformatics/btq085

Genetics and population analysis

Advance Access publication February 25, 2010

Post hoc power estimation in large-scale multiple testing problems

Sonja Zehetmayer and Martin Posch*

Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria

Associate Editor: Jeffrey Barrett

ABSTRACT

Background: The statistical power or multiple Type II error rate in large-scale multiple testing problems as, for example, in gene expression microarray experiments, depends on typically unknown parameters and is therefore difficult to assess a priori. However, it has been suggested to estimate the multiple Type II error rate *post hoc*, based on the observed data.

Methods: We consider a class of *post hoc* estimators that are functions of the estimated proportion of true null hypotheses among all hypotheses. Numerous estimators for this proportion have been proposed and we investigate the statistical properties of the derived multiple Type II error rate estimators in an extensive simulation study.

Results: The performance of the estimators in terms of the mean squared error depends sensitively on the distributional scenario. Estimators based on empirical distributions of the null hypotheses are superior in the presence of strongly correlated test statistics.

Availability: R-code to compute all considered estimators based on *P*-values and supplementary material is available on the authors web page http://statistics.msi.meduniwien.ac.at/index.php?page=pageszfnr

Contact: martin.posch@meduniwien.ac.at

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

Received on October 29, 2009; revised on February 3, 2010; accepted on February 23, 2010

1 INTRODUCTION

In genomic and proteomic research thousands of hypotheses are tested simultaneously. Numerous procedures have been proposed to control multiple Type I error rates, e.g. the false discovery rate (FDR) or the family-wise error rate. The multiple Type II error rate or the statistical power, in contrast, have received less attention so far. The power of a multiple testing procedure can be defined in several ways (Senn and Bretz, 2007). For example, one can consider the probability to reject at least one alternative hypothesis or the probability to reject all alternative hypotheses. We consider an intermediate approach, the average power, defined as the arithmetic mean of the elementary power values of all alternative hypotheses. The average power can be interpreted as the expectation of the proportion of rejected alternative hypotheses among all alternative hypotheses and corresponds to the so-called false negative rate (FNR) (e.g. Delongchamp *et al.*, 2004; Pawitan *et al.*, 2005). The FNR is defined as the expectation of the false negative proportion (FNP), the proportion of retained true alternative hypotheses among all true alternative hypotheses and is related to the average power by Average Power = 1 − FNR. The FNR depends on a number of parameters as the sample size, the effect sizes and the proportion of true null hypotheses among all hypotheses. Therefore, the FNR of an experiment is unknown a priori and can only be guessed based on preliminary assumptions. However, if data have already been observed, the FNR can be estimated based on the observed test statistics. This concept of '*post hoc* power analysis' has been criticized for the case of a single hypothesis test (e.g. Hoenig and Heisey, 2001), where the *post hoc* power is just a 1:1 function of a single *P*-value. However, in experiments with a large number of hypotheses one can utilize the empirical distribution of the test statistics to estimate the FNR. Under suitable assumptions (e.g. if the elementary test statistics are sufficiently independent), the FNR is asymptotically equivalent to the FNP such that an estimator for the FNR can also be used to estimate the FNP. An essential element in the estimation of the FNR is the estimation of the proportion of true null hypotheses among all null hypotheses, denoted by π_0. Numerous estimators for π_0 have been proposed in the literature and we assess their performance in the estimation of the FNR.

The FNR is a measure for the fraction of undetected alternative hypotheses and is therefore a crucial parameter to interpret negative findings in experiments where a large number of hypotheses is investigated. In addition, in large-scale multiple testing problems, FNR estimates can be used to define stopping rules for sequential testing. For example, one could continue sampling until the estimated FNR falls below a prespecified threshold. As shown in Posch *et al.* (2009) under suitable assumptions such sequential testing asymptotically does not inflate the FDR if the sample size is increased for all hypotheses simultaneously and only the test at the final interim analysis determines which hypotheses are rejected.

The nomenclature for the FNR is not consistent in literature: we label the expected proportion of retained true alternatives under all alternatives 'FNR' according to Pawitan *et al.* (2005) and Norris and Kahn (2006). Delongchamp *et al.* (2004) label this quantity 'fraction of genes not selected', Craiu and Sun (2008) non-discovery rate. Additionally, the term FNR has also been used to denote the proportion of false negatives among all retained hypotheses (Sarkar, 2004), a quantity that has also been labeled, e.g. 'false non-discovery rate (FNDR)' (Genovese and Wasserman, 2002).

In Section 2.1, we introduce a family of estimators of the FNR that depend on the data only through the number of

*To whom correspondence should be addressed.

rejected null hypotheses, the critical value applied in the multiple testing procedure and an estimate of the proportion of true null hypotheses π_0. In addition, we consider an estimator that is based on estimation of local FDRs. In Section 2.2, we review published and implemented estimators for π_0. These π_0 estimators and the resulting estimators of the FNR are evaluated in a simulation study in Section 3 for a variety of scenarios including different dependency structures of test statistics across hypotheses as independence, weak dependence and equi-correlation. Finally, we give several real data examples in Section 4.

2 METHODS

2.1 Estimating the FNR and FNP

Consider a multiple testing procedure to test m hypotheses of which m_0 (m_1) are true null (alternative) hypotheses. Then the proportion of true null hypotheses among all hypotheses is $\pi_0 = m_0/m$. Assume that for each of the elementary hypotheses a test with P-value $p_i, i = 1, \ldots, m$, is defined.

Let γ denote the critical value applied to the unadjusted elementary P-values. For example, for the Bonferroni test $\gamma = \alpha/m$ while for the Benjamini–Hochberg (BH) test $\gamma = d\alpha/m$, where $d = \text{argmax}_i\{p_{(i)} \leq i\alpha/m\}$ and $p_{(1)}, \ldots, p_{(m)}$ denote the ordered P-values. We set $\gamma = 0$ if $p_{(i)} > i\alpha/m$ for all $i = 1, \ldots, m$.

Let $R = R(\gamma) = \#\{p_i \leq \gamma\}$ denote the total number of rejected hypotheses and V the number of rejected true null hypotheses. Then the FDR, defined as the expected proportion of erroneous rejections among all rejections, is given by $\text{FDR} = E(V/\max\{R, 1\})$ (Benjamini and Hochberg, 1995). Let $S = R - V$ denote the number of rejected true alternative hypotheses. The FNP is defined as

$$\text{FNP} = 1 - \frac{S}{m_1} \tag{1}$$

and the FNR is given by

$$\text{FNR} = E(\text{FNP}) = 1 - \frac{E(S)}{m_1}. \tag{2}$$

2.1.1 A class of FNR estimators based on estimates of π_0 To estimate the FNR, we consider estimators of the quantities $E(S)$ and m_1. Let $\hat{\pi}_0$ denote an estimate of π_0 (in Section 2.2, we discuss several such estimators). Let $\hat{m}_1 = (1 - \hat{\pi}_0)m$ and estimate $E(S) = E(R) - E(V)$ in two steps: $E(R)$ is estimated by the observed number of rejections R, and V by $mF_0(\gamma)\hat{\pi}_0$, where F_0 denotes the cumulative distribution function (c.d.f) of the P-values corresponding to the true null hypothesis. Thus, an estimator of the FNR is given by

$$\widehat{\text{FNR}} = 1 - \frac{R(\gamma) - mF_0(\gamma)\hat{\pi}_0}{m(1 - \hat{\pi}_0)}. \tag{3}$$

If the null distribution of the P-values is estimated from the data then F_0 in (3) is replaced by its estimator $\hat{F}_0$. Note that $\pi_0 = 1$ implies that for all hypotheses the null hypothesis holds. Then $m_1 = 0$ and we define $\text{FNR} = 0$, since in this case no false negative decision can be made. Similarly, for $\hat{\pi}_0 = 1$ we set $\widehat{\text{FNR}} = 0$.

If we additionally assume that the P-values are uniformly distributed under H_0 then the estimator simplifies to

$$\widehat{\text{FNR}} = 1 - \frac{R(\gamma) - m\gamma\hat{\pi}_0}{m(1 - \hat{\pi}_0)}. \tag{4}$$

This estimator has been considered by Posch *et al.* (2009), who use the Storey (2002) method to estimate π_0, and Delongchamp *et al.* (2004), who apply a method by Hsueh *et al.* (2003) to estimate π_0. If γ is chosen to control the FDR at level α, an alternative estimator is given by

$$\widehat{\text{FNR}} = 1 - \frac{R(\gamma)(1 - \alpha)}{m(1 - \hat{\pi}_0)}, \tag{5}$$

which has been applied by Norris and Kahn (2006) and Craiu and Sun (2008) using the Storey estimator or the Smoother estimator (Storey and Tibshirani, 2003) of π_0, respectively.

A large variety of estimators for π_0 have been proposed, mainly as a tool to estimate the FDR. By (3) each such estimator defines also an estimator for the FNR. Given the underlying π_0 estimators are consistent as $m \to \infty$ and the observations are sufficiently independent across hypotheses (and some additional technical assumptions), the FNR estimators defined by (3) are consistent as well; if the π_0 estimator is asymptotically biased, the asymptotic bias of the resulting FNR estimator has the opposite sign (see Section 1 in the Supplementary Material for details). In Section 2.2, we give a review of estimators for π_0 and investigate the properties of the resulting FNR estimators for finite m in Section 3.

2.1.2 FNR estimators based on estimates of local FDRs An alternative estimator of the FNR can be defined using Efron's (2007b) empirical Bayes estimator of the density of the test statistics and estimates of the so-called local FDR. Consider the test of two-sided hypotheses $H_i\colon \mu_i = 0$ versus $H_i'\colon \mu_i \neq 0$, $i = 1, \ldots, m$. Then z-values z_i are calculated by taking the standard normal quantile of the one-sided P-values. According to a Bayesian mixture model, it is assumed that the z-statistics belong to one of two classes, either corresponding to the true null hypothesis or the true alternative. The prior probability that a z-statistic corresponds to the true null hypothesis is π_0. Denoting the density of the z-statistics under the null (alternative) hypothesis by $f_0(z)$ ($f_1(z)$), the mixture density $f(z)$ of the z-statistics is given by

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z). \tag{6}$$

The local FDR for hypothesis i is defined as the posterior probability in the Bayesian mixture model that for hypothesis i the null hypothesis holds. The local FDR is given by $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$, $i = 1, \ldots, m$, such that $f_1(z) = (1 - \text{fdr}(z))f(z)/\int_{-\infty}^{\infty}(1 - \text{fdr}(z))f(z)\text{d}z$ (Efron, 2007b). Let $c_{1-\gamma}$ denote the $(1-\gamma)$-quantile of the standard normal distribution. Based on this representation of $f_1(z)$, we define an FNR estimator by

$$\text{FNR} = \int_{c_{\gamma/2}}^{c_{1-\gamma/2}} f_1(z)\text{d}z = \frac{\int_{c_{\gamma/2}}^{c_{1-\gamma/2}}(1 - \text{fdr}(z))f(z)\text{d}z}{\int_{-\infty}^{\infty}(1 - \text{fdr}(z))f(z)\text{d}z} \tag{7}$$

replacing $\text{fdr}(z)$ and $f(z)$ by appropriate estimates $\widehat{\text{fdr}}(z)$ and $\hat{f}(z)$. Since $\text{fdr}(z) = \pi_0 f_0(z)/f(z)$, the estimate of fdr depends on estimates of $\hat{\pi}_0$, $\hat{f}_0$ and $\hat{f}$. As in Efron (2007b), we estimate $\hat{f}(z)$ by a natural spline and consider two estimators for $f_0(z)$ and π_0. First, it can be assumed that the null density is a normal distribution with mean δ and variance σ^2 and that the z-values corresponding to alternative hypotheses have support outside a known interval $[-x_0, x_0]$ only. Then maximum likelihood estimates (MLEs) of δ, σ^2, and π_0 can be derived based on all observations falling into $[-x_0, x_0]$. The threshold x_0 is chosen based on a heuristic algorithm. For the second option, it is assumed that $f_0(z)$ is a standard normal distribution [the 'theoretical null distribution' which corresponds to the assumption of uniformly distributed P-values in (4)] and π_0 can be estimated as above.

For the actual computation we approximate $\widehat{\text{fdr}}$ and $\hat{f}$ by piecewise constant functions using the R-package *locfdr* such that the integral simplifies to a sum. We consider two FNR estimators: LocThe based on the theoretical null distribution and LocMLE based on the estimated null distribution. Note that this FNR estimator is in [0,1] by construction. If $\widehat{\text{fdr}}(z) = 1$ for all z, this suggests that the global null hypothesis holds and we set $\widehat{\text{FNR}} = 0$.

Additionally, we consider two FNR estimators based on (7) where $\text{fdr}(z)$ and $f(z)$ are replaced by corresponding estimates from the fdrtool package. See the paragraph on the fdrtool package in Section 2.2 for details.

2.1.3 Relationship between the two types of FNR estimators The FNR estimators based on π_0 estimators are closely related to those based on local FDRs. Each term in (3) corresponds to a term in (7): $R(\gamma)/m$ is an estimator of the c.d.f. of the (two-sided) P-values at γ given by $\int_{c_{\gamma/2}}^{c_{1-\gamma/2}} f(z)\text{d}z$, the term $F_0(\gamma)\hat{\pi}_0$ is an estimator for $F_0(\gamma)\pi_0 = \int_{c_{\gamma/2}}^{c_{1-\gamma/2}} \text{fdr}(z)f(z)\text{d}z$. Finally, $\hat{\pi}_0$ in the denominator of (3) is an estimator of $\pi_0 = \int_{-\infty}^{\infty} \text{fdr}(z)f(z)\text{d}z$. Thus, the estimators based on (3) and (7) have essentially the same structure. An advantage of the estimator based on (7) is that this FNR estimator is

in [0,1] by construction. Additionally, the estimated local FDR's are clipped to the range [0,1], which can reduce the variability of the resulting estimates especially if the actual local FDR is close to 1. Finally, while in (3), $E(R(\gamma))$ is estimated by the empirical distribution function of the P-values, in (7) more sophisticated estimators are used. However, these estimators rely on specific assumptions and may be inefficient if these assumptions are violated.

2.2 Estimators for π_0

In this section, we give a short description of the considered π_0 estimators. They were chosen based on their availability in R (R Development Core Team, 2009). As in Section 2.1.2, we assume that the distribution of the z-statistics and P-values can be modeled via a mixture model. According to Equation (6), the mixture density $f(p)$ of the P-values is given by

$$f(p)=\pi_0 f_0(p)+(1-\pi_0)f_1(p). \tag{8}$$

We use the same symbols to refer to densities of z and P-values and distinguish them by the argument only. If we additionally assume that the P-values under the null hypothesis are uniformly distributed on $[0,1]$, then

$$f(p)=\pi_0+(1-\pi_0)f_1(p). \tag{9}$$

2.2.1 Estimators based on P-value density estimation In this section, we consider estimators of π_0 that are based on estimates of the density $f(p)$. Under the mixture model (9), $\hat{\pi}_0=\min_p f(p)$ gives an estimate of π_0 which is positively biased if $\min_p f_1(p)>0$. If f_1 takes its minimum at $p=1$, the estimate for π_0 simplifies to $f(1)$. The estimators for π_0 considered in the following are based on different estimators $\hat{f}$ of the mixture density f.

Beta Uniform model (Bum): Pounds and Morris (2003) fit the mixture of a uniform and a beta distribution $f(x)=\lambda+(1-\lambda)\alpha x^{\alpha-1}$ to the observed P-values. Since the MLEs $\hat{\lambda}$ and $\hat{a}$ (obtained by numerical optimization) appear to have a high variability, $\hat{\lambda}$ is not a useful estimate for π_0. However, the density estimate $\hat{f}=\hat{\lambda}+(1-\hat{\lambda})\hat{a}x^{\hat{a}-1}$ appears to be less variable such that Pounds and Morris propose to estimate π_0 by $\hat{\pi}_0=\hat{f}(1)$. *R-package oompa at http://bioinformatics.mdanderson.org/Software/OOMPA*.

Convest: Langaas *et al.* (2005) use a non-parametric maximum likelihood estimator $\hat{f}$ of the P-value density. It is based on the mixture model (9) and the assumption that f_1 is decreasing and convex. π_0 is then estimated by $\hat{\pi}_0=\hat{f}(1)$. *R-function convest in bioconductor package limma*.

Poisson regression approach (Pre): Broberg (2005) constructs a histogram of the P-values and models the number of hits in each subinterval by an inhomogeneous poisson process. The expected number of hits in each subinterval is then fitted by a polynomial of low degree in the midpoints of the subintervals based on Poisson regression. This leads to an estimate for f and π_0 is estimated by $\hat{\pi}_0=\min_p \hat{f}(p)$. *R-function p0.mom in bioconductor package SAGx*.

Spacing LOESS histogram (Splosh): Pounds and Cheng (2004) estimate the density f by smoothing the slope of the empirical distribution function of the P-values with local regression and set $\hat{\pi}_0=\min_p \hat{f}(p)$. *Splosh.R provided at http://www.stjuderesearch.org/depts/biostats/splosh.html*.

Empirical Bayes (EfronThe, EfronMLE): as described in Section 2.1.2, Efron (2007b) proposes to estimate π_0 based on a mixture model of the z-transformed P-values. We consider two such estimators for π_0: EfronThe, derived under the assumption of the theoretical null distribution, and EfronMLE, where also the null distribution is estimated from the data. For a reliable estimation of the null distribution Efron suggests to use the method if $\pi_0>0.9$, only. *R-package locfdr*.

Fdrtool (Pfndr, Zfndr, Tfndr, Ppct0, Zpct0, Tpct0): Strimmer (2008) applies a truncated maximum likelihood approach to estimate π_0 from the empirical distribution of either P-values, z-scores or t-statistics. If P-values are provided, the theoretical null distribution (i.e. uniform distribution on [0,1]) is assumed. For z-statistics f_0 is assumed to follow a $N(0,\sigma_0^2)$ distribution, and σ_0 and π_0 are estimated with the truncated maximum likelihood approach. Similarly, for t-statistics f_0 is assumed to follow a t-distribution and the corresponding degrees of freedom and π_0 are estimated. As in Efron's approach, an interval $[-x_0,x_0]$ is chosen which is assumed to contain only test statistics from true null hypotheses. Two different methods to determine x_0 are proposed: the FNDR method is based on minimizing the estimated FNDR which is defined as the expected proportion of retained true alternative hypotheses under all retained hypotheses. For the second method, x_0 is chosen such that a predefined fraction (default: 0.75) of the observed test statistics lies in the interval $[-x_0,x_0]$. Then $\hat{\pi}_0$ is obtained via MLE of the truncated data. We denote the π_0 estimators applying the minimization of the FNDR by Pfndr, Zfndr and Tfndr (based on either P-values, z-statistics or t-tests, respectively). Likewise, Ppct0, Zpct0 and Tpct0 denote the estimators based on the prefixed truncation fraction of 0.75. As outlined in Section 2.1.2, we also considered two FNR estimators, LocZfndr and LocZpct0, which are based on local FDR estimators from the fdrtool package: both local FDR estimators are based on z-statistics. For LocZfndr, the interval $[-x_0,x_0]$ is chosen based on FNDR estimates as described above. For LocZpct0, the central 75% of z-values are used to estimate the null distribution. *R-package fdrtool*.

EbayesThresh (ETLapl, ETCau): Johnstone and Silverman (2004) propose an empirical Bayes approach to estimate π_0 for sparse data. They assume that the observations for each hypothesis i are drawn independently from a normal distribution with mean μ_i and variance 1. The prior distributions of the μ_i are mixtures of a probability mass of π_0 at zero with a given symmetric heavy-tailed distribution. The mixing weight π_0 is estimated based on the marginal maximum likelihood. We considered the π_0 estimates resulting from a Laplace (ETLapl) and quasi-Cauchy prior (ETCau). *R-package EbayesThresh*.

2.2.2 Estimation based on the c.d.f. of the P-values The proportion of true null hypotheses can be estimated via

$$\hat{\pi}_0(\lambda)=\frac{\#\{p>\lambda\}}{m(1-\lambda)}, \tag{10}$$

where λ is a tuning parameter. This approach goes back to Schweder and Spjotvoll (1982), who gave a graphical motivation. The nominator gives the observed number of P-values larger than λ, the denominator is the expected number of P-values larger than λ, given that for all hypotheses the null hypothesis holds. Assuming that no P-value larger than λ corresponds to an alternative hypothesis, (10) is an unbiased estimator. There is a trade-off between bias (which decreases for $\lambda\to 1$) and variance (which increases for $\lambda\to 1$). Several choices for λ have been considered.

Storey: Storey (2002) proposes $\lambda=0.5$.

Bootstrap: Storey (2002) suggests to choose λ such that the bootstrap estimate of the mean squared error of $\hat{\pi}_0(\lambda)$ is minimized. *R-function pval.estimate.eta0 in package fdrtool*.

Smoother: Storey and Tibshirani (2003) fit a natural cubic spline y with 3 degrees of freedom to $\hat{\pi}_0(\lambda)$ and set $\hat{\pi}_0=y(1)$. *R-function pval.estimate.eta0 in package fdrtool*.

Lowest Slope (LSL): the lowest slope estimator (Benjamini and Hochberg, 2000) is based on the slope of the c.d.f. of the P-values $\hat{F}(p)$. Let $S_i=(1-p_{(i)})/(m+1-i)$ denote the slope of the line from $\left(p_{(i)},\hat{F}(p_{(i)})\right)$ to the point $(1,1)$. Here, $p_{(i)}$ denote the ordered P-values. Let i denote the smallest i such that $S_i<S_{i-1}$ and define $\hat{\pi}_0=\min[1,1/m+1/(S_i m)]$. *R-function pval.estimate.eta0 in package fdrtool*.

Howmany: Meinshausen and Rice (2006) suggest an upper confidence bound $\hat{\pi}_\alpha$ for π_0 such that $P(\hat{\pi}_\alpha\geq\pi_0)\geq 1-\alpha$. The bound can be written as

$$\hat{\pi}_\alpha=\inf_{t\in(0,1)}\hat{\pi}_0(t)+\beta_\alpha\sqrt{t(1-t)}, \tag{11}$$

where $\hat{\pi}_0(\lambda)$ is defined in (10), $\beta_\alpha=a^{-1}[E^{-1}(1-\alpha)+b]$, E is the c.d.f. of the Gumbel distribution, $a=\sqrt{2m\log\log m}$, and $b=2\log\log m+0.5\log\log\log m-0.5\log 4\pi$ (where π denotes the circle constant). It is shown that under suitable conditions this bound is a consistent estimator of π_0 as $m\to\infty$. For finite m, the estimator can be improved if the infimum in (11) is taken over $(\epsilon,1-\epsilon)$ for some small ϵ in the order of $1/m$. For the simulation study in Section 3, we set $\alpha=0.5$ such that the estimate (11) is median

unbiased if the confidence bound has exact coverage probability. *R-package howmany*.

2.2.3 Other Estimators *Jin*: as for the EfronMLE estimator, the null distribution is estimated from the data. Under the true null hypothesis, the z-scores are assumed to be $N(\mu_0, \sigma_0^2)$ distributed. The parameters μ_0, σ_0 as well as π_0 are estimated based on the empirical characteristic function of the z-transformed P-values (Jin and Cai, 2007). Under suitable conditions, if $1-\pi_0$ is asymptotically larger than $1/\sqrt{m}$, it is shown that the resulting estimate is consistent for $m \to \infty$. *R-function provided by the authors: http://www.stat.cmu.edu/~jiashun/Research/software/NullandProp*.

Location-based estimator (Lbe): based on the mixture model (9), Dalmasso *et al.* (2005) construct a family of estimators

$$\hat{\pi}_0(\varphi) = \frac{\frac{1}{m}\sum_{i=1}^{m}\varphi(p_i)}{E_0(\varphi(P))}, \tag{12}$$

where φ is a real-valued function and $E_0(\varphi(P))$ denotes the expectation over the P-value distribution under the null hypothesis. Dalmasso *et al.* (2005) give conditions for φ that lead to a non-negatively biased estimate of π_0 and propose the choice $\varphi(p) = -\log(1-p)$. Note that the estimator (10) can be written in the form (12) with $\varphi(p) = I(p > \lambda)$, where $I(\cdot)$ denotes the indicator function. *Bioconductor package LBE*.

Moment generating function (Mgf): based on the mixture model (9), Broberg (2005) constructs an estimator for π_0 using the estimated moment generating function of the P-value distribution. The moment generating function is represented as a weighted sum of the moment generating function of the uniform distribution and the unknown P-value distribution under the alternative. The latter is estimated by a recursive algorithm. *R-function p0.mom in the Bioconductor package SAGx*.

2.3 Comparison of FNR estimators

By (3) each estimate of π_0 leads to an estimate of the FNR. Either the theoretical null distribution F_0 can be used in the estimator or an estimated null distribution $\hat{F}_0$. We investigate the FNR estimators defined by (3) resulting from different estimators of π_0 in a simulation study and report the square root of the mean squared error (RMSE) and the bias for each estimator under a wide range of scenarios. Additionally, we include the estimators LocThe, LocMLE, LocZfndr and LocZpct0 based on local FDRs defined in Section 2.1.2. Note that the π_0 estimators corresponding to the LocThe, LocMLE, LocZfndr and LocZpct0 estimators are the same as for the EfronThe, EfronMLE, Zfndr and Zpct0 estimators, respectively, and are not reported separately in the tables and figures.

For the simulation, we consider the test of m null hypotheses $H_i : \mu_i = 0$ for the mean of normally distributed observations with mean μ_i and variance σ_i^2 against the alternatives $H_i' : \mu \neq 0, i = 1, \ldots, m$, with two-sided one-sample t-tests. The critical value is determined by the BH step-up procedure at level $\alpha = 0.05$ applied to elementary P-values based on the central t-distribution. Note that the BH procedure has been shown to control the FDR also under certain dependency structures (Benjamini and Yekutieli, 2001). For $\hat{\pi}_0$ estimators involving tuning parameters, the default values recommended by the respective authors are used except for the howmany estimator where we set $\alpha = 0.5$ and for the estimators based on Efron's locfdr procedure (EfronThe, EfronMLE, LocThe and LocMLE) where the value of the degrees of freedom for fitting the estimated density $f(z)$ is set to 14. Additionally, we restrict the π_0 and FNR estimates to the interval [0,1].

The simulations are performed for six reference scenarios. Here, we assume that $m = 10000$, $n = 20$ and consider three different proportions of true null hypotheses: $\pi_0 \in \{0.9, 0.95, 0.99\}$. For the alternative hypotheses, we assume that the data are $N(\delta_j, 1)$ distributed where the δ_j are alternating $-\Delta, -3\Delta/4, -\Delta/2, -\Delta/4$ and $\Delta, 3\Delta/4, \Delta/2, \Delta/4$ for the $(1-\pi_0)m$ alternatives. For the reference scenarios, we consider $\Delta \in \{1, 2\}$. The actual FNRs in these scenarios are given in Table 1. For the scenarios with correlated test statistics, the FNRs are practically identical.

To investigate the impact of correlated test statistics on the properties of the FNR estimates, we performed the simulations for independent test statistics, equi-correlated test statistics (as in Benjamini *et al.*, 2006) and a block correlation structure (Storey *et al.*, 2004). For the latter, we assume that the test statistics are correlated in blocks of 10 hypotheses. Within one block, the correlation between the test statistics of hypotheses H_j and H_i is ρ, if $i, j \leq 5$ or $i, j > 5$, and $-\rho$ if $i \leq 5$ and $j > 5$ or vice versa. For each block ρ is drawn from a uniform distribution on $[0, 1]$.

Table 1. The actual FNRs under different scenarios applying the BH test at level $\alpha = 0.05$ and assuming independent test statistics

	π_0		
Δ	0.9	0.95	0.99
1	0.67	0.74	0.88
2	0.24	0.27	0.34

3 RESULTS

3.1 Comparison of the RMSE of the FNR estimators

Table 2 shows the maximum RMSE and maximum bias of $\widehat{\text{FNR}}$ for $\pi_0 = 0.9, 0.99$ for independent and equi-correlated test statistics. The maximum is taken over the alternatives Δ in $\{1, 2\}$. Tables for all reference scenarios (RMSE and bias for $\hat{\pi}_0$ and $\widehat{\text{FNR}}$) can be found in the Supplementary Material. In all considered scenarios, $\hat{\pi}_0$ has much lower RMSE than $\widehat{\text{FNR}}$. None of the FNR estimators has a uniformly lowest RMSE. Under independence, the Convest, LocThe and Pfndr estimator have the lowest *maximal* RMSE across the six considered scenarios (Section 2.3). For the scenario with block correlation, we observe somewhat larger RMSE than in the independent case (Supplementary Material). For the equi-correlated case (with $\rho = 0.5$), all estimators with the exception of the LocMLE and EfronMLE estimator have distinctively larger RMSE and bias compared with the independent case. Also the Jin estimator shows low RMSE in scenarios with equi-correlation but, as discussed below, its RMSE appears to depend very sensitively on the underlying parameters. Figure 1 shows the RMSE of $\widehat{\text{FNR}}$ of the Convest, Pfndr, LocThe, LocMLE and EfronMLE estimators when varying one of the parameters m, π_0, Δ or ρ in the reference scenario $m = 10000$, $\pi_0 = 0.9$, $\rho = 0$, $\Delta = 2$, $n = 20$. The figures for the remaining FNR estimators as well as for the π_0 estimators are given in the Supplementary Material.

Due to the factor $1/(1-\hat{\pi}_0)$ in the FNR estimators defined by (3), the RMSE is large for π_0 close to one even though the RMSE of $\hat{\pi}_0$ decreases with π_0. If $\pi_0 = 1$, with a large probability $\hat{\pi}_0 = 1$ and the FNR is correctly estimated as 0. However, if the estimate $\hat{\pi}_0$ is less than 1, the FNR estimate is typically 1 (since under the global null hypothesis the BH test guarantees that with probability $1-\alpha$ no hypothesis is rejected such that $R(\gamma) = 0$). Thus, in this setting, the distribution of $\widehat{\text{FNR}}$ is concentrated on 0 and 1 which is reflected in a large RMSE. For $\pi_0 << 1$ the estimators LocMLE and EfronMLE based on estimated null distributions have a larger RMSE than the estimators based on theoretical null distributions.

Increasing the effect sizes of the alternative hypotheses does not lead to a consistent decrease of the RMSE of the FNR estimators. In contrast, for all FNR estimators in Figure 1, the RMSE reaches a maximum for $\Delta \sim 1$. For increasing m, the RMSE of most estimators slightly decreases. For increasing ρ, the (absolute) bias and RMSE

Table 2. The maximum RMSE (maximum bias) of the considered $\widehat{\text{FNR}}$ estimators for independent or equi-correlated ($\rho = 0.5$) test statistics for $\pi_0 = 0.9, 0.99$ and $m = 10\,000$

	Independence		Equi-correlation	
	$\pi_0 = 0.9$	$\pi_0 = 0.99$	$\pi_0 = 0.9$	$\pi_0 = 0.99$
Storey	0.10 (−0.08)	0.42 (−0.21)	0.52 (−0.30)	0.71 (−0.52)
LSL	0.55 (−0.54)	0.44 (−0.40)	0.53 (−0.50)	0.45 (−0.35)
Bootstrap	0.11 (0.05)	0.32 (0.21)	0.48 (−0.25)	0.70 (−0.51)
Smoother	0.16 (−0.09)	0.56 (−0.33)	0.51 (−0.27)	0.70 (−0.51)
Pfndr	0.09 (−0.08)	0.20 (−0.05)	0.51 (−0.29)	0.71 (−0.53)
Mgf	0.11 (−0.10)	0.33 (−0.15)	0.52 (−0.32)	0.71 (−0.53)
Pre	0.28 (0.27)	0.29 (0.22)	0.37 (0.18)	0.51 (0.36)
Lbe	0.31 (−0.16)	0.59 (−0.36)	0.51 (−0.28)	0.71 (−0.50)
Ppct0	0.12 (−0.11)	0.30 (−0.14)	0.52 (−0.39)	0.72 (−0.55)
Tpct0	0.22 (−0.12)	0.59 (−0.38)	0.53 (−0.42)	0.72 (−0.56)
Zfndr	0.61 (−0.60)	0.22 (−0.15)	0.39 (−0.18)	0.46 (0.18)
Convest	0.07 (−0.05)	0.20 (0.11)	0.37 (−0.11)	0.70 (−0.52)
Splosh	0.22 (−0.12)	0.47 (0.45)	0.55 (0.51)	0.65 (0.65)
Bum	0.14 (0.14)	0.22 (0.22)	0.25 (0.13)	0.66 (−0.44)
EfronMLE	0.24 (−0.22)	0.51 (−0.31)	0.15 (−0.08)	0.31 (0.12)
Jin	0.17 (−0.16)	0.24 (−0.12)	0.19 (−0.05)	0.33 (0.28)
Howmany	0.14 (−0.13)	0.25 (−0.18)	0.37 (−0.15)	0.48 (−0.17)
Zpct0	0.28 (−0.24)	0.53 (−0.30)	0.49 (−0.38)	0.74 (−0.59)
Tfndr	0.67 (−0.67)	0.33 (−0.22)	0.50 (−0.27)	0.66 (−0.47)
EfronThe	0.09 (−0.08)	0.43 (−0.22)	0.67 (−0.67)	0.88 (−0.88)
LocThe	0.08 (−0.07)	0.21 (0.06)	0.33 (−0.20)	0.41 (−0.16)
LocMLE	0.22 (−0.21)	0.24 (0.03)	0.14 (−0.08)	0.28 (0.22)
ETLapl	0.37 (0.37)	0.43 (0.43)	0.38 (0.27)	0.46 (0.24)
ETCau	0.39 (0.39)	0.41 (0.41)	0.39 (0.29)	0.42 (0.24)
LocZfndr	0.56 (−0.56)	0.22 (−0.16)	0.39 (−0.18)	0.47 (0.24)
LocZpct0	0.27 (−0.23)	0.57 (−0.35)	0.49 (−0.39)	0.72 (−0.56)

The maximum is taken over the alternatives Δ in $\{1, 2\}$. The bias of the scenario with the largest absolute bias is reported. 5000 simulation runs have been performed per scenario.

for both $\hat{\pi}_0$ and $\widehat{\text{FNR}}$ increase, with the exception of the RMSEs of EfronMLE and LocMLE which are nearly constant in ρ.

In several of the plots the Jin estimate shows an erratic behavior (Supplementary Material). The estimator involves a tuning parameter γ and appears to depend quite sensitively on its value. In the simulation study, we used $\gamma = 0.1$ as recommended in Jin and Cai (2007). However, for some scenarios, choosing a slightly different value has a large impact on the RMSE (data not shown).

In the simulation study, we applied the BH procedure which does not rely on a π_0 estimate but is strictly conservative if $\pi_0 < 1$. This choice guarantees that the actual FNR is the same regardless of the π_0 estimator considered for the FNR estimation and allows for a better comparability of the investigated methods. Additional simulations for the Storey, Pfndr and Convest methods, where the FDR was controlled based on the same π_0 estimator as used in the FNR estimation, gave very similar RMSEs (data not shown).

Normalization: It is well known that normalization can reduce the correlation in microarray datasets. We made further simulations for the five estimators considered above where the observations were standardized per chip such that they have mean 0 and variance 1. This standardization had hardly any impact in the independent and block correlated case. In the equi-correlated case, though, the correlation is practically removed by standardization and the resulting FNR estimators have similar RMSEs as in the independent case. Note, however, that standardization may introduce a bias if a larger fraction of genes is differentially expressed and the effect of over- and underexpressed genes does not cancel out. To demonstrate that standardization cannot remove the impact of correlation in general, we performed additional simulations for the Convest, Pfndr, LocThe, LocMLE and EfronMLE estimators under a block correlation structure with larger block size. When increasing the block size to 100, standardization has practically no impact on the RMSE, but the RMSE (across $\Delta = 1, 2$) increases compared with the scenarios with block size 10 even after normalization: for $\pi_0 = 0.9$ to 0.18, 0.19, 0.16 (for Convest, Pfndr and LocThe) and to 0.3, 0.36 (for LocMLE and EfronMLE); for $\pi_0 = 0.99$ to 0.38, 0.38, 0.34 (for Convest, Pfndr and LocThe) and to 0.39, 0.57 (for LocMLE and EfronMLE). Note that in contrast to the equi-correlated case, for block correlations with large blocks, the estimators based on the empirical null hypothesis show no advantage.

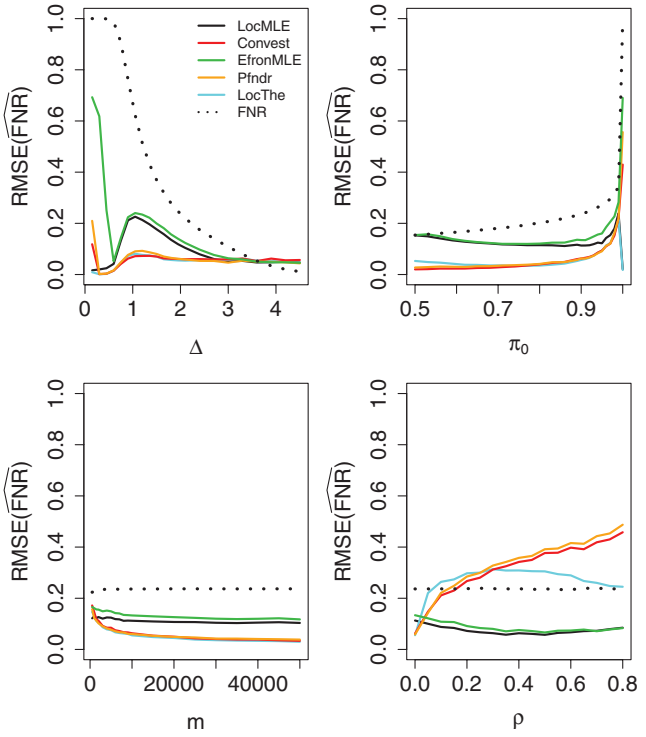

Fig. 1. RMSE of $\widehat{\text{FNR}}$ estimators. In the scenario with $m = 10000$, $\pi_0 = 0.9$, $\Delta = 2$, $\rho = 0$ in each of the graphs one of the parameters is varied. In the graph with varying Δ only values $\Delta > 0$ are plotted ($\Delta = 0$ corresponds to the case $\pi_0 = 1$). The graphs with varying ρ refers to the case of equi-correlated test statistics with correlation coefficient ρ. The dotted line shows the true FNR. 500 simulation runs have been performed for each value on the x-axis.

3.2 Assessing correlations

The simulation study shows that the RMSE of the estimators depends on the strength of dependence of the observations across genes. While the results under block correlation and independence match closely, strong correlation (as in the equi-correlation scenario) leads for all estimators based on theoretical null distributions to much larger RMSEs. To assess the correlation between the observed expression levels of different genes, one can investigate

the distribution of pairwise correlation coefficients for all pairs of genes (see, e.g. Efron, 2007a; Owen, 2005). As in Efron (2007a), we apply Fisher's z-transformation to the correlation coefficients and standardize the approximate theoretical standard deviation (SD) $\sqrt{1/(n-3)}$ such that the theoretical distribution of the transformed correlation coefficients is $N(0,1)$ given that the true correlation is zero for all genes. The observed mean $\hat{\mu}$ and SD $\hat{\sigma}$ of the transformed correlation coefficients can then be compared with the theoretical values $\mu=0$ and $\sigma=1$ under the assumption of independence.

We performed a simulation study (1000 runs) to investigate how reliable the mean and variance of the pairwise correlation coefficients can be estimated. Assuming a per-group sample size of $n=20$, we estimated the mean and SD of pairwise correlation coefficients under different correlation structures. For computational feasibility, we consider the pairwise correlations between the first 5000 genes in each dataset, only. The mean correlation coefficients showed very low variability under independence and block correlation (with block sizes up to 250) where 95% of the mean z-transformed correlation coefficients were in $(-0.0008, 0.0006)$. Under equi-correlation (with $\rho=0.2, 0.5$ corresponding to z-transformed values of 0.84 and 2.26), the variation was somewhat larger and 95% of the z-transformed average correlation coefficients were in (0.45, 1.34) or (1.46, 3.25), respectively. This indicates that the order of magnitude of the mean correlation can be well estimated in the considered scenarios. Similarly, under independence the observed SD of the z-transformed correlations was close to the nominal value 1 under independence. However, under equi-correlation ($\rho\in\{0.2, 0.5\}$) the estimated SD was somewhat lower [95% of the estimates in (0.94, 0.99) or (0.83, 0.92)]. Under block correlation for block sizes of 10, 20, 100 and 250, 95% of the estimated SDs lie in (1.01, 1.014), (1.02, 1.03), (1.08, 1.21) and (1.14, 1.59), respectively.

Standardizing the observations per chip as described in the previous section centers the distribution of z-transformed correlation coefficients around zero (Efron, 2010). Consequently, the estimated mean correlation coefficients for the standardized data are close to zero under independence and under equi-correlation. For the estimated SDs, the impact of standardization is more intricate. While for equi-correlated data, the SD becomes practically 1 with very low variation, for block correlation standardization has only a marginal impact on the SD.

4 REAL DATA EXAMPLES

We estimate the FNR of two gene expression microarray experiments that were reanalyzed in Pavlidis *et al.* (2003). Based on a dataset by Gruvberger *et al.* (2001), we compare gene expression measurements of breast cancer in patients with positive and negative estrogen receptor status by two-sided two-sample t-tests ($m=3389$, $n=28$ per group). In the second example based on Huang *et al.* (2001), we compare gene expression measurements of patients with papillary versus normal thyroid carcinoma ($m=12558$, $n=8$ per group). Due to the outliers and skewed distributions of expression values, the latter comparison was performed with Wilcoxon tests. As in the simulation study, the BH method was used to control the FDR at 5%. The considered estimators give rather divergent π_0 and FNR estimates for the Gruvberger data (Table 3). However, they can be classified into two groups: the estimators based on the assumption of a uniform null distribution of the P-values (Convest, LocThe and Pfndr) and the estimators based on estimated null distributions (EfronMLE and LocMLE). Within each group the results are similar.

Table 3. Estimates of π_0 and the FNR for the microarray datasets from Gruvberger *et al.* (2001) with $m=3389$, $n=28/28$, and Huang *et al.* (2001) with $m=12558$, $n=8/8$ for standardized and non-standardized data, respectively

	Gruvberger				Huang			
	Non-standardized ($R=163$)		Standardized ($R=296$)		Non-standardized ($R=50$)		Standardized ($R=110$)	
	$\hat{\pi}_0$	$\widehat{\text{FNR}}$	$\hat{\pi}_0$	$\widehat{\text{FNR}}$	$\hat{\pi}_0$	$\widehat{\text{FNR}}$	$\hat{\pi}_0$	$\widehat{\text{FNR}}$
Convest	0.77	0.80	0.63	0.77	0.93	0.95	0.94	0.85
LocThe		0.78		0.73		0.98		0.92
Pfndr	0.77	0.80	0.65	0.76	0.93	0.95	0.93	0.88
LocMLE		0.35		0.14		0.91		0.87
EfronMLE	0.96	0.27	0.96	0.12	0.95	0.93	0.94	0.86

The differences between the estimates from the two classes of estimators may be due to either a large fraction of alternative hypotheses or correlation between the test statistics of different hypotheses. As described in Section 3.2, we investigated the distribution of pairwise correlation coefficients between pairs of genes and computed pairwise Pearson's (Spearman's) correlations for the Gruvberger (Huang) dataset. For the Gruvberger dataset, the distribution of the z-transformed correlation coefficients indicates much stronger dependence than for the Huang dataset [Gruvberger: group 1 (2) $\hat{\mu}=1.32$ (1.44), $\hat{\sigma}=1.5$ (1.73); Huang: group 1 (2), $\hat{\mu}=0.014$ (0.02), $\hat{\sigma}=1.08$ (1.06)]. After chip-wise standardization (Section 3), the average of the z-transformed pairwise correlations in the Gruvberger dataset is close to zero but the SD is still larger than the nominal value 1 [group 1 (2) $\hat{\mu}=0.002$ (0.007), $\hat{\sigma}=1.38$ (1.52)]. This indicates a strong dependence of observations, in the order of the magnitude of block correlation with block size 250 (Section 3.2). Thus, in this dataset standardization cannot remove the correlations. For the Huang dataset, the standardization is performed by subtracting the chip-wise median and dividing by the interquartile range. Since there is little correlation observed in the raw data, standardizing hardly changes the distribution of correlation coefficients [group 1 (2), $\hat{\mu}=0.05$ (0.04), $\hat{\sigma}=1.08$ (1.05)]. Note that in both examples after standardization a larger number of hypotheses can be rejected and the FNR estimates become smaller (Table 3).

The example shows that the choice of the null distribution may be crucial for the estimation of the FNR. However, this also holds for hypothesis testing: for example, if the rejection threshold γ in the Gruvberger dataset is chosen such that the FDR estimate from the locfdr package is 0.05, one can reject 111 hypotheses assuming the theoretical null hypothesis but only seven when choosing the empirical null distribution.

5 CONCLUSIONS

We investigated a family of FNR estimators which is based on the estimated proportion of true null hypotheses π_0 as well as estimators based on local FDR estimates. For the former, one can show that given the observations are sufficiently independent across hypotheses, the asymptotic FNR estimates are consistent as $m\to\infty$ if the underlying π_0 estimate is consistent. This holds, e.g. for the

Storey (assuming that the tuning parameter λ approaches 1 at an appropriate rate), Lbe, Howmany and Jin estimators. However, for finite m the simulation studies show that the estimation error in estimating the FNR is considerable larger than for the estimators of π_0. A reliable estimation of the FNR is difficult, especially if the number of alternative hypotheses is small. In these settings, the FNP is still highly variable and the FNR estimators are unreliable.

Since the proposed FNR estimators are based on univariate *P*-values, they can be applied to a wide range of statistical tests including multi-stage procedures for which group-sequential *P*-values can be defined (Victor and Hommel, 2007; Zehetmayer *et al.*, 2005, 2008).

The Convest, LocThe and Pfndr estimators that are based on theoretical null distributions showed the most favorable characteristics for independent test statistics in the considered scenarios, but have a large RMSE under strong dependence. In contrast, the EfronMLE and LocMLE estimators based on estimated null distributions are more robust in the equi-correlated scenarios, but show no advantage in the block correlated scenarios or if the proportion of alternative hypotheses is large. The latter comes from the fact that the estimation of the null distribution is based on the assumption that for most hypotheses the null hypothesis holds.

ACKNOWLEDGEMENTS

We thank Peter Bauer and the two referees for many helpful suggestions.

Funding: Austrian Science Fund FWF (grant numbers P18698-N15 and T 401-B12).

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

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

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

Benjamini,Y. *et al.* (2006) Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**, 491–507.

Broberg,P. (2005) A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, **6**, 199–219.

Craiu,R. and Sun,L. (2008) Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. *Stat. Sin.*, **18**, 861–879.

Dalmasso,C. *et al.* (2005) A simple procedure for estimating the false discovery rate. *Bioinformatics*, **21**, 660–668.

Delongchamp,R.R. *et al.* (2004) Multiple-testing strategy for analyzing cDNA array data on gene expression. *Biometrics*, **60**, 774–782.

Efron,B. (2007a) Correlation and large-scale simultaneous significance testing. *JASA*, **102**, 93–103.

Efron,B. (2007b) Size, power and false discovery rates. *Ann. Stat.*, **35**, 1351–1377.

Efron,B. (2010) Correlated z-values and the accuracy of large-scale statistical estimates. *JASA*, in press.

Genovese,C. and Wasserman,L. (2002) Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. B*, **64**, 499–517.

Gruvberger,S. *et al.* (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res.*, **61**, 5979–5984.

Hoenig,J. and Heisey,D. (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am. Stat.*, **55**, 19–24.

Hsueh,H. *et al.* (2003) Comparison of methods for estimating the number of true hypotheses in multiplicity testing. *J. Biopharm. Stat.*, **13**, 675–689.

Huang,Y. *et al.* (2001) Gene expression in papillary thyroid carcinoma reveals highly consistent profiles. *Proc. Natl Acad. Sci. USA*, **98**, 15044–15049.

Jin,J. and Cai,T. (2007) Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *JASA*, **102**, 495–506.

Johnstone,I. and Silverman,B. (2004) Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.*, **32**, 1594–1649.

Langaas,M. *et al.* (2005) Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Stat. Soc. B*, **67**, 555–572.

Meinshausen,N. and Rice,J. (2006) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Stat.*, **34**, 373–393.

Norris,A.W. and Kahn,C.R. (2006) Analysis of gene expression in pathophysiological states: balancing false discovery and false negative rates. *Proc. Natl Acad. Sci. USA*, **103**, 649–653.

Owen,A.B. (2005) Variance of the number of false discoveries. *J. R. Stat. Soc. B*, **67**, 411–426.

Pavlidis,P. *et al.* (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Pawitan,Y. *et al.* (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*, **21**, 3017–3024.

Posch, M. *et al.* (2009) Hunting for significance with the false discovery rate. *JASA*, **104**, 836–840.

Pounds,S. and Cheng,C. (2004) Improving false discovery rate estimation. *Bioinformatics*, **20**, 1737–1745.

Pounds,S. and Morris,S. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

R Development Core Team (2009) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org.

Sarkar,S.K. (2004) FDR-controlling stepwise procedures and their false negatives rates. *J. Stat. Plan. Infer.*, **125**, 119–137.

Schweder,T. and Spjotvoll,E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, **69**, 493–502.

Senn,S. and Bretz,F. (2007) Power and sample size when multiple endpoints are considered. *Pharm. Stat.*, **6**, 161–170.

Storey,J.D. *et al.* (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. B*, **66**, 187–205.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479–498.

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

Strimmer,K. (2008) A unified approach to false discovery rate estimation. *BMC Bioinformatics*, **9**, 303–317.

Victor,A. and Hommel,G. (2007) Combining adaptive designs with control of the false discovery rate - a generalized definition for a global p-value. *Biometrical J.*, **49**, 94–106.

Zehetmayer,S. *et al.* (2005) Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics*, **21**, 3771–3777.

Zehetmayer,S. *et al.* (2008) Optimized multi-stage designs controlling the false discovery or the family wise error rate. *Stat. Med.*, **27**, 4145–4160.