Biometrika
2015, Vol. 102, No. 3, pp. 695-704.

Efficient estimation of the number of false positives in high-throughput screening

Holger Rootzén and Dmitrii Zholud*

Abstract

This paper develops tail estimation methods to handle false positives in multiple testing problems where testing is done at extreme significance levels and with low degrees of freedom, and where the true null distribution may differ from the theoretical one. We show that the number of false positives, conditional on the total number of positives, approximately has a binomial distribution, and find estimators of its parameter. We also develop methods for estimation of the true null distribution, and techniques to compare it with the theoretical one. Analysis is based on a simple polynomial model for very small p-values. Asymptotics which motivate the model, properties of the estimators, and model checking tools are provided. The methods are applied to two large genomic studies and an fMRI brain scan experiment.

Keywords: Correction of p-values, extreme value statistics, false discovery rate, SmartTail, high-throughput screening, multiple testing, positive false discovery rate.

1 Introduction

The purpose of high-throughput screening in bioscience is to identify interesting candidate cases for further study, and it differs from classical testing in several ways. First, it involves many thousands of hypotheses. Second, to get a manageable number of positives, testing is done at extreme significance levels. Third, individual tests are often based on very few observations. Fourth, the true null distributions in these very complex experiments often deviate from the theoretical ones. This paper develops methods to handle false positives in high-throughput screening.

We prove that the conditional distribution of the number of false positives given that there are r positives is asymptotically binomial, observe that the success probability parameter of this binomial distribution coincides with the Storey (2002) positive false discovery rate, and develop methods to estimate it. We also introduce new estimators of Efron's local false discovery rate and other error control parameters and methods to estimate the true null distribution and to compare it with the theoretical null distribution. We provide confidence intervals for both independent and dependent p-values.

Earlier approaches use either fully parametric models or empirical distribution functions. However, trusting that a model is accurate far out in the tails can lead to a high bias; for the genome screening data considered below, it gave estimates which were clearly wrong. Furthermore, in high-throughput

*Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mails: hrootzen@chalmers.se and dmitrii@chalmers.se

screening the empirical distribution function estimator is typically based on a small number of observations and has high variance. Our approach is semi-parametric: we use a parametric model, but only for the tails of the distributions. This makes it possible to obtain both low bias and low variance.

The high-throughput screening experiments considered here lead to the asymptotics n fixed, $m \to \infty$, $\alpha \to 0$, where n is the number of observations used in the individual tests, m is the number of tests, and α is the significance level. In particular a Bonferroni procedure takes $\alpha = \eta/m$, with η fixed. Then $\alpha \to 0$ as $m \to \infty$. In practice, a small α may instead be mandated by limited capacity for further study of the positives.

Let P denote a generic p-value and write H_0 and H_1 for the null and the alternative hypothesis. Our basic model, the extreme tail mixture model, is that there exist $c_i, \gamma_i, u_i > 0$ such that under H_i the p-values have cumulative distribution functions

$$F_i(x) = c_i x^{1/\gamma_i}, \quad 0 \le x \le u_i, \quad i = 0, 1. \tag{1}$$

This model is derived in Section 2. Further we assume that for $\pi_0, \pi_1 > 0$, $\pi_0 + \pi_1 = 1$,

$$F(x) = \mathrm{pr}(P \le x) = \pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}, \quad 0 \le x \le u = \min(u_0, u_1). \tag{2}$$

Mixture models of this type are standard in the multiple testing literature. Still, our methods apply also if the number of true and false null hypotheses are nonrandom.

Dudoit and van der Laan (2008) and Kerr (2009) are recent useful reviews of the area. Knijnenburg et al. (2009) suggests using generalized Pareto approximations to improve efficiency of permutation test in bioinformatics. Presumably the most common approach is the false discovery rate error control procedure of Benjamini and Hochberg (1995). However, in screening studies the aim is to select interesting cases for further study, and then error control may be less natural. The estimation approach to multiple testing has already attracted significant interest (Storey, 2002, 2003, Efron et al., 2001, Efron, 2004, 2008, Ruppert et al., 2007, Jin and Cai, 2007), and the true null distribution is often different from the theoretical one (Efron et al., 2001, Jin and Cai, 2007). Fan et al. (2007) consider uniform normal approximations of t-distributions when both $n \to \infty$ and $m \to \infty$, but such approximations are inaccurate for small n.

2 Basic theory

In this section we give conditions that ensure that the model (1) holds asymptotically as $u \to 0$, and that the limiting binomial distribution of the number of false positives holds. The proofs are given in the Supplementary Material.

Let G_t, G_0 and G_1 be the cumulative distribution functions of the test statistic under the theoretical null hypothesis, the true null hypothesis, and the alternative hypothesis, respectively, and write x_t^*, x_0^*, x_1^* for their right endpoints. Let $\bar{G} = 1 - G$ and write $\bar{G}^{\leftarrow}$ for the right continuous inverse of $\bar{G}$. Then $\bar{G}_0\{\bar{G}_t^{\leftarrow}(x)\}$ is the true null distribution of p-values, and $\bar{G}_1\{\bar{G}_t^{\leftarrow}(x)\}$ is the alternative distribution of p-values. If $G_0 = G_t$ and the distributions are continuous, then the true null distribution $\bar{G}_0\{\bar{G}_t^{\leftarrow}(x)\}$ is $U(0,1)$.

Theorem 2.1. *Let $i = 0$ or 1 and suppose that G_t and G_i belong to the max domains of attraction of extreme value distributions with shape parameters ξ_t and ξ_i, respectively. If (i) $\xi_t, \xi_i > 0$ or (ii) $\xi_t, \xi_i < 0$ and $x_t^* = x_i^* < \infty$ then, for some constants $\gamma_i > 0$,*

$$F_i(x) = G_i\{G_t^{\leftarrow}(x)\} = c_i(u) x^{1/\gamma_i}\{1 + o(1)\}, \tag{3}$$

with the $o(1)$ term uniform in $\epsilon \le x/u \le 1$ as $u \to 0$, for any $\epsilon > 0$.

It can be shown that (3) holds also for $\xi_t = \xi_i = 0$, under suitable further conditions.

If G_t and G_0 satisfy the conditions of Theorem 1, then, by (3), equation (1) applies for $i = 0$ and sufficiently small u_i. Now, it is typically known that G_t satisfies the conditions, and extremal domain of attraction theory motivates that they usually also hold for G_0. Thus equation (1) for $i = 0$ should be valid in most testing problems.

The motivation for (1) for $i = 1$ is more delicate since it requires that G_1 belongs to a max domain of attraction. From a Bayes or empirical Bayes perspective (Efron, 2008), the arguments are the same for G_1 as for G_0. From a frequentist point of view, however, one might believe in a few major effects plus a larger number of unimportant ones, expressed to a random amount and measured with error. If the major effects are large compared to the cutoff for tests, then testing is in the center of their distribution, and the extremal motivation is less compelling. However, then these effects often would be clearly visible, and careful analysis of false positives less important. Otherwise, testing is still in the tail of G_1, and the extreme value arguments motivate equation (1) also for G_1. Below we provide methods to use the data to distinguish between the cases.

The motivation for (1) given by Theorem 2.1 is mathematical. Empirical motivation is given by the examples below and from extensive experience from extreme value statistics.

For t- and F-tests assumption (i) of Theorem 2.1 holds with $\gamma_i = 1$ under quite general conditions, regardless of non-normality or dependence of the data, see Zholud (2014). The convergence in (3) is fast for low degrees of freedom, but slower for larger ones.

We next show that a conditional binomial distribution of the number of false positives is widely applicable. Let m_0 be the number of true null hypotheses, m_1 be the number of false null hypotheses, and $m = m_0 + m_1$ be the total number of tests. Let $\alpha = \alpha_m$ be the critical level, Q_0 and Q_1 be the distributions of the numbers of true and false positives, respectively, and P_r be the conditional distribution of the number of false positives given that there is a total of r positives. Write $\mathrm{Bin}(r, p)$ for a binomial distribution with r trials and success probability p, and $Po(\lambda)$ for a Poisson distribution with parameter λ. Let $||P-Q|| = \sup_A |P(A)-Q(A)|$ be the variation distance between the probability distributions P and Q, and let pFDR be positive false discovery rate of Storey (2003),

$$\mathrm{pFDR} = \frac{\pi_0 F_0(\alpha_m)}{\pi_0 F_0(\alpha_m) + \pi_1 F_1(\alpha_m)}. \tag{4}$$

Theorem 2.2. *Suppose there exist constants $0 < c \le C < \infty$ such that $c \le m_i F_i(\alpha_m) \le C$, that $||Q_i - Po\{\pi_i F_i(\alpha_m)\}|| \to 0$ as $m \to \infty$, for $i = 0, 1$, and that the number of false positives is independent of the number of true positives. Then, for r fixed,*

$$||P_r - Bin(r, pFDR)|| \to 0, \quad m \to \infty. \tag{5}$$

In particular (5) holds if the p-values are mutually independent and one of the following conditions is satisfied: (i) $m_0, m_1 \to \infty$ are non-random, and $c \le m_i F_i(\alpha_m) \le C$ and $\pi_i = m_i/m$, for $i = 0, 1$, or (ii) the model (2) holds, $m \to \infty$ and $c \le m F_i(\alpha_m) \le C$.

3 Statistical methods

In this section we derive the maximum likelihood estimator for $F_0(x)$ and compute its efficiency relative to the empirical distribution function. The distribution $F(x)$ may be estimated using maximum likelihood methods for mixture distributions. Alternatively, an approximation of the extreme tail mixture model makes it possible to use same estimator as for $F_0(x)$. We assume that p-values are independent, except in a final general theorem.

To estimate $F_0(x)$ we assume that it has been possible to obtain a sample of m_0 p-values, $p_1^0, \ldots, p_{m_0}^0$, from the true null distribution. Then, one chooses a small threshold $u_0 > 0$ and in the analysis one only uses the p_i^0-s which are less than u_0, and only estimates $F_0(x)$ for $x \leq u_0$. The important point is that one only trusts the model $F_0(x) = c_0 x^{1/\gamma_0}$ to be sufficiently accurate for $x \leq u_0$, but that u_0 can often be chosen much larger than α, so that the model-based estimate of $F_0(\alpha)$ uses many more observations and thus has much smaller variance than the empirical distribution function. The choice of u_0 is a compromise between bias and variance, guided by goodness-of-fit test and plots: a small u_0 leads to less model error, and hence less bias, but also to fewer observations to base estimation on, and hence more variance, see e.g. Coles (2001), Section 4.3.1.

Asume P^0 has the true conditional null distribution, $\mathrm{pr}\big(P^0/u_0 \leq x \mid P^0 \leq u_0\big) = (x/u_0)^{1/\gamma_0}$ for $0 \leq x \leq u_0$. Then, differentiating the log likelihood function shows that the maximum likelihood estimate of γ_0, based on the p_i^0-s which are less than u_0, is

$$\hat{\gamma}_0 = \frac{1}{N_0(u_0)} \sum\nolimits_{p_i^0 \leq u_0} - \log(p_i^0/u_0), \tag{6}$$

with $N_0(u_0) = \#\{p_i^0 \leq u_0, 1 = i, ..., m_0\}$. Since $F_0(x) = \mathrm{pr}\big(P^0 \leq u_0\big)\mathrm{pr}\big(P^0 \leq x \mid P^0 \leq u_0\big)$, and estimating $\mathrm{pr}\big(P^0 \leq u_0\big)$ by $N_0(u_0)/m_0$ we estimate $F_0(x)$ by

$$\hat{F}_0(x) = N_0(u_0)/m_0 \, (x/u_0)^{1/\hat{\gamma}_0}, \quad x \leq u_0. \tag{7}$$

The variance of $N_0(u_0)/m_0$ is estimated by $\{N_0(u_0)/m_0\}\{1 - N_0(u_0)/m_0\}/m_0$. Conditionally on $P_0 < u_0$ the summands in (6) have a mean γ_0 exponential distribution and hence, conditionally on $N_0(u_0)$, the variance of $\hat{\gamma}_0$ is estimated by $\hat{\gamma}_0^2/N_0(u_0)$. Using that $N_0(u_0)/m_0$ and $\hat{\gamma}_0$ are asymptotically uncorrelated and normally distributed, see the Supplementary Material, confidence intervals can be computed using the delta method.

By (2), $\mathrm{pr}(P/u \leq x \mid P \leq u) = px^{1/\gamma_0} + (1-p)x^{1/\gamma_1}$ for $p = \pi_0 c_0 u^{1/\gamma_0}/(\pi_0 c_0 u^{1/\gamma_0} + \pi_1 c_1 u^{1/\gamma_1})$ and $0 \leq x \leq 1$. Thus the conditional distribution of $\{p_i/u\}$ is a mixture distribution with parameters p, γ_0, and γ_1. These may be estimated using numerical maximum likelihood, and for $N(u) = \#\{p_i \leq u, 1 \leq i \leq m\}$, we may estimate $F(x)$ by

$$\hat{F}(x) = N(u)/m \left\{ \hat{p}\,(x/u)^{1/\hat{\gamma}_0} + (1-\hat{p})\,(x/u)^{1/\hat{\gamma}_1} \right\}, \quad x \leq u. \tag{8}$$

This can be done using three methods: (i) if a sample from the null distribution is available, maximize the product of the likelihoods for the p_i^0-s which are smaller than u_0 and the p_i-s which are smaller than u, (ii) for the cases when the null hypothesis holds, or when the tail asymptotics of Zholud (2014) apply, maximize the likelihood for the p_i-s which are smaller than u, with γ_0 set to 1, or (iii) maximize the likelihood for the p_i-s which are smaller than u. Confidence intervals are obtained using standard techniques.

The approach (iii) provides an estimate of F_0 without a null sample, but since it requires estimating three parameters, estimation uncertainty will often be large. Additionally, often $\gamma_0 \approx \gamma_1 \approx 1$ so that (8) is close to non-identifiability, and then none of the methods will work. This was the case for the yeast genome screening data considered below, where all three methods gave estimates of γ_0 and γ_1 which were quite close to 1, but where the confidence intervals were too wide to make the methods practically useful.

Also generally, estimation uncertainty for the maximum likelihood estimates in the mixture model is often large. However, if $\gamma_0 \approx \gamma_1$ then (2) reduces to (9) below, with $c = \pi_0 c_0 + \pi_1 c_1$, and with γ the common value of γ_0 and γ_1. Thus a widely useful shortcut method to obtain more accurate estimates is to approximate (2) by

$$F(x) = cx^{1/\gamma}, \tag{9}$$

and then to estimate $F(x)$ in the same way as for $F_0(x)$. Whether (9) is reasonable may be checked by comparing estimates of γ_0 with estimates of γ.

Storey (2002) proposed the conservative estimator $\hat{\pi}_0 = (\#\{p_i > \lambda\}/m)/(1-\lambda)$ of π_0, with $\lambda \in (0,1)$ a suitably chosen, not small, number. In the present situation where the true null distribution may be non-uniform one can instead use the estimator

$$\hat{\pi}_0 = \frac{\#\{p_i > \lambda\}/m}{\#\{p_i^0 > \lambda\}/m_0}.$$

Now let $F_E(x)$ be the empirical distribution function estimator of $F(x)$ in (9) and let $\hat{F}(x)$ be the estimator provided by (7). It can be shown, see the Supplementary Material, that for large m, small $F(u)$, $x \leq u$, and, say, $mF(u) \geq 35$, the efficiency is

$$e = \frac{\mathrm{var}\{F_E(x)\}}{\mathrm{var}\{\hat{F}(x)\}} \approx \left(\frac{u}{x}\right)^{1/\gamma} \left[1 + \frac{1}{\gamma^2}\left\{\log\left(\frac{u}{x}\right)\right\}^2\right]^{-1}. \qquad (10)$$

The efficiency for typical values of u/x and γ is shown in Figure 1. Further, if one uses the assumption $\gamma = 1$, the efficiency can be shown to be $e = u/x$, and is thus higher.

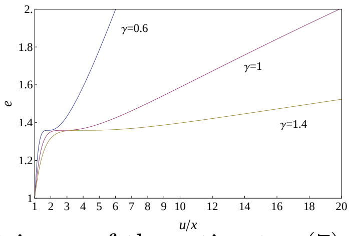

Figure 1. Efficiency of the estimator (7) of $F(x)$ in (9).

The same results hold for $\hat{F}_0(x)$. Finally, for very small values of x, which sometimes are of interest, the empirical distribution function cannot be used as estimator.

Efron (2004, 2008) uses the theoretical null distribution to transform the test statistic to a N(0, 1) distribution and then accounts for deviations from it by fitting a N(μ, σ) distribution. More complex procedures are proposed by, e.g., Schwartzman (2008). However, these methods can lead to large bias and this can be checked by looking carefully into the tails.

The positive false discovery rate is of central interest in this paper. We also use the Efron et al. (2001) local false discovery rate which measures the a posteriori likelihood of false rejection of a hypothesis with p-value equal to x, and is defined by

$$\mathrm{fdr}(x) = \mathrm{pr}(H_0 \text{ true} \mid P = x) = \frac{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0}}{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0} + \pi_1 c_1 \gamma_1^{-1} x^{1/\gamma_1}} = \frac{\pi_0 dF_0(x)/dx}{dF(x)/dx}.$$

Our methods directly give estimators of these parameters and of other false discovery rate parameters: see Table 1, in which V and R are the number of false positives, and the total number of positives, respectively.

Table 1. Error control parameters and estimators

Parameter	**Definition**	**Estimator**
false discovery rate	$E(V/R \mid R > 0)\,\mathrm{pr}(R > 0)$	$\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}\{1 - e^{-m\hat{F}(\alpha)}\}$
positive false discovery rate	$E(V/R \mid R > 0)$	$\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}$
local false discovery rate	$\frac{\pi_0 dF_0(x)/dx}{dF(x)/dx}$	$\frac{\hat{\pi}_0 d\hat{F}_0(x)/dx}{d\hat{F}(x)/dx}$
familywise error	$\mathrm{pr}(V \neq 0)$	$1 - e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$
k-familywise error	$\mathrm{pr}(V \geq k)$	$\sum_{i=k}^{\infty} \frac{\{m\hat{\pi}_0 \hat{F}_0(\alpha)\}^i}{i!} e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$

Conservative estimates are obtained by setting $\hat{\pi}_0 = 1$. The estimators follow from the conditional binomial distribution of the number of false positives together with the asymptotic Poisson distribution of the number of false positives. Our estimators of the positive false discovery rate and the false discovery rate differ from the Storey (2002) estimators by factors $1 - \exp\{-m\hat{F}(\alpha)\}$.

The estimators are consistent and asymptotically normal, also when the p-values, $\{P_i\}$, form a stationary dependent sequence. Here and below we omit the subscripts $i = 0, 1$. Assume (3) holds, so that $F(x) = \ell(x)x^{1/\gamma}$, where $\ell(x)$ is slowly varying as $x \to 0$. Split the sequence $1, 2, \dots n$ up into $k_m = [m/r_m]$ blocks $B_{m,i} = ((i-1)r_m, ir_m]$, $1 \le i \le k_m$, of length r_m. Next, consider levels $u_m \to 0$ as $m \to \infty$, and, writing 1_i for the indicator function of the event that $P_i \le u_m$, define $F_E(u_m) = m^{-1}\sum_{i=1}^m 1_i$, $\hat{\gamma}(u_m) = \sum_{i=1}^m \{-\log(P_i/u_m)\, 1_i\}/\sum_{i=1}^m 1_i$ and $\gamma_m = \frac{1}{\ell(u_m)u_m^{1/\gamma}}\int_0^{u_m} \ell(x)x^{1/\gamma-1}dx$. Let

$$Z_{m,i} = \sum_{j\in B_i}\{-\log(P_j/u_m) + C\}F(u_m)^{-1}1_j, \quad C = -\gamma_m\{1 + \gamma_m/\log(x/u_m)\}$$

and

$$Z_{m,i}^{(1)} = \sum_{j\in B_i} -\log(P_j/u_m)F(u_m)^{-1}1_j, \quad Z_{m,i}^{(2)} = \sum_{j\in B_i} F(u_m)^{-1}1_j.$$

Set $\sigma_m^2 = k_m \mathrm{var}(Z_{m,1})$, $\sigma_{m,i}^2 = k_m \mathrm{var}(Z_{m,1}^{(i)})$, $i = 1, 2$, introduce sample block sums

$$\hat{Z}_i = \hat{Z}_{m,i} = \hat{D}\sum_{j\in B_i}\{-\log(P_j/u_m) + \hat{C}\}F_E(u_m)^{-1}1_j,$$

with $\hat{D} = -m^{-1}\log(x/u_m)\,(x/u_m)^{1/\hat{\gamma}}\,F_E(u_m)\hat{\gamma}^{-2}$, $\hat{C} = -\hat{\gamma}\{1 + \hat{\gamma}/\log(x/u_m)\}$, and set

$$s_m^2 = \sum_{i=1}^{k_m}(\hat{Z}_i - \bar{Z})^2, \quad \bar{Z} = k_m^{-1}\sum_{i=1}^{k_m}\hat{Z}_i.$$

Let $\{\mathcal{B}_{i,j}\}$ be the σ-algebra generated by $P_i, \dots, P_j$, define the strong mixing coefficients $\alpha_{m,\ell} = \sup\{|\mathrm{pr}(AB) - \mathrm{pr}(A)\mathrm{pr}(B)| : A \in \mathcal{B}_{1,k}, B \in \mathcal{B}_{k+\ell,m}, 1 \le k \le m-\ell\}$, and introduce the following conditions:

C1: There exist integers $\ell_m < r_m \to \infty$ with $r_m = o(m)$ such that, for $k_m = [m/r_m]$,

$$k_m(\alpha_{m,\ell_m} + \ell_m/m) \to 0 \;\text{ and }\; k_m^{-1}mF(u_m) \to 0.$$

C2: There exist integers $w_m > 1$ such that

$$r_m\{F(u_m)\sigma_m\}^{-1}w_m\{mF(u_m)\sigma_m^{-1}e^{-w_m} + 1\} \to 0 \;\text{ and }\; \sigma_m\{mF(u_m)\}^{-1} \to 0.$$

Under these conditions the estimator (7), i.e. $\hat{F}(x) = F_E(u_m)\,(x/u_m)^{1/\hat{\gamma}(u_m)}$, of $F(x)$, $x \le u_m$, asymptotically has a normal distribution.

Theorem 3.1. *(i) Suppose C1 and C2 hold, and that there exist constants $0 < k < K$ such that $k \le \sigma_{m,i}/\sigma_m \le K$ for $i = 1, 2$. Then for any fixed $y \in (0,1)$, as $m \to \infty$,*

$$\frac{1}{D(y,u_m)\sigma_m}\left\{\hat{F}(yu_m) - F(u_m)y^{1/\gamma_m}\right\} \to_d N(0,1),$$

for $D(y,u_m) = -m^{-1}\log(y)\,y^{1/\gamma}F(u_m)\gamma^{-2}$, and

$$\frac{m}{\sigma_{m,1}}(\hat{\gamma} - \gamma_m) \to_d N(0,1), \quad \frac{m}{\sigma_{m,2}}\{F_E(u_m) - F(u_m)\} \to_d N(0,1).$$

In particular $\hat{\gamma} \to_{pr} \gamma$ *and* $F_E(u_m)/F(u_m) \to_{pr} 1$.

(ii) If, in addition, $k_m var(Z_{m,1}^2) \to 0$, *then*

$$\frac{1}{s_m}\left\{\hat{F}(yu_m) - F(u_m)y^{1/\gamma_m}\right\} \to_d N(0,1).$$

The proof, explanation of the conditions, extensions to more complex dependence structures, discussion of sandwich estimators for construction of confidence intervals, and extensions of Theorem 2.2 to dependent cases are given in the Supplementary Material.

4 Examples

Warringer et al. (2003) performed genome-wide screening experiments for detecting differential growth in *Saccharomyces Cerevisiae*, baker's yeast. In the experiments different yeast strains were grown on two 100-well honeycomb agar plates. We consider a growth parameter, logarithmic doubling time, extracted from the resulting 200 growth curves.

In the experiments, 96 mutant yeast strains were grown in the same positions in each of the two plates. Reference wild type strains were grown in four wells in each plate, one in each quadrant. For each of the mutant strains, differential growth was measured by subtracting the average of the logarithmic doubling times of the four reference strains from the logarithmic doubling time of the mutant strain. This gives one value per mutant for each plate. High differential growth for the mutant was then tested by comparing the two measured values with zero in a one-sample t-test with 1 degree of freedom.

We considered three data sets from PROPHECY collection of deletion strains in yeast: the wild type data with 1,728 observed p-values, the genome-wide data with 4,896 observed p-values, where the mutants and reference wild type strains were grown under normal conditions, and the salt stress data with 5,280 observed p-values, where the strains were grown under salt stress. The wild type data were obtained for quality control purposes, and were analyzed in the same way as the genome-wide data, and hence were a sample from the true null distribution. As discussed above, asymptotically one expects that $\gamma_0 = \gamma_1 = 1$, but non-asymptotically other values might give a better fit.

Figure 2 shows that the true null distribution is non-uniform, and that the model (1) fits quite well. A Kolmogorov-Smirnov test, after a log transformation to get exponentially distributed variables (Schafer et al., 1972), gave the p-value 0.31 and hence did not reject (1), and the p-value for likelihood ratio test of $\gamma_0 = 1$ was 0.8. Plots which guided threshold choice are given in the Supplementary Material.

For the genome-wide data, the maximum likelihood estimates of γ_0 and γ_1 obtained from (8), with $u_0 = 0.05$ and $u = 0.01$, were both close to 1 for all methods (i) - (iii) described in Section 3. The estimates of p varied much more. Method (ii), where γ_0 is set to 1, gave the shortest confidence intervals, but they still were too wide for the estimates to be useful: the estimates were $\hat{p} = 0.0$ and $\hat{\gamma}_1 = 1.05$, with confidence intervals $(0, 1)$ and $(0.73, 1.36)$, respectively.

The model (9) estimate was $\hat{\gamma} = 1.05$, and comparing it with $\hat{\gamma}_0 = 0.98$ obtained from the wild type data, this also indicates that (9) is appropriate. Additionally, Figure 3, with Kolmogorov-Smirnov p-value 0.21, shows that (9) fits the genome-wide data well.

Using (9) the estimate of the positive false discovery rate at $\alpha = 0.001$, with 95% confidence interval, was 0.25 ± 0.14. Since there were 44 p-values less than 0.001, the expected number of false positives was hence estimated to be 11. Using the binomial approximation, the number of false positives was estimated to be at most 15, with probability greater than 95%. If one instead uses a uniform null distribution, the number of false positives is estimated to be at most 7 with probability greater than 95%, a much too positive picture of experimental precision.

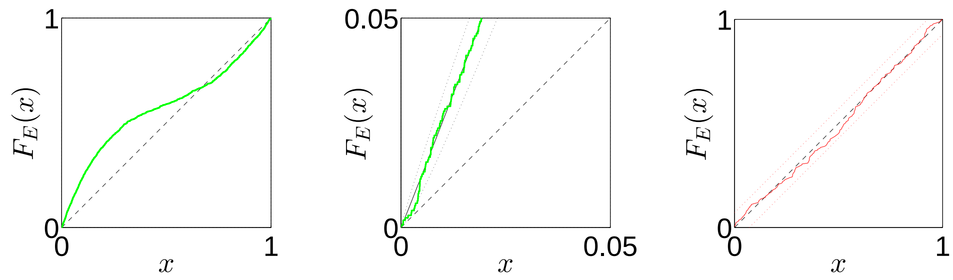

Figure 2. *Goodness of fit plots for the wild type data*
Left: Empirical distribution function. Dashed line is uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.05$ (226 values). Solid line is (1) estimated using $u = 0.05$; Dotted lines are 95% pointwise confidence intervals. *Right:* Empirical conditional distribution function of $-\log(p/0.05)$ for $p \leq 0.05$, transformed to uniform scale, and Kolmogorov-Smirnov 95% goodness of fit limits.

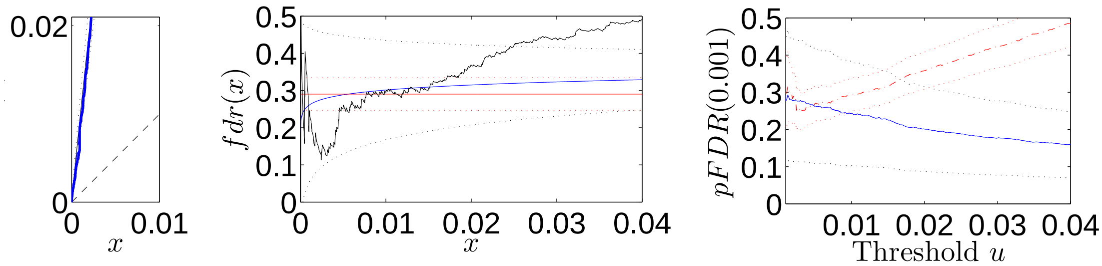

Figure 3. *Goodness of fit plots for the genome-wide data*
Left: Empirical distribution function for $p \leq 0.01$ (441 values); dashed line is uniform distribution. Solid line is (9) estimated using $u = 0.01$. *Middle:* Estimated false discovery rate (smooth curve) and empirical false discovery rate (jagged curve) for $u = 0.01, \pi_0 = 1$. Straight line is the local false discovery rate with γ_0, γ_1 set to 1. *Right:* Solid line is the positive false discovery rate at $\alpha = 0.001$ as function of u, for $\pi_0 = 1$. Dot-dashed line is with γ_0, γ_1 set to 1. Dotted lines are 95% pointwise confidence intervals.

The local false discovery plot in Figure 3 indicates that it is slightly more probable that it is the rejections with the smallest p-values which are the true positives. Still, it is quite likely that some of the smallest p-values are false positives. The salt stress data was analyzed in the same way as the genome-wide data and also showed good model fit.

For the wild type data, the genome-wide data, and the salt stress data $\mathrm{var}\{F_E(0.001)\}/\mathrm{var}\{\hat{F}(0.001)\}$ was estimated to be 2.3, 1.4, and 1.5, respectively, with $u = 0.01$ also for the salt stress data. Variance estimates for the positive false discovery rate based on the empirical distribution functions do not seem to be available. The estimates above are biased upwards as we have set $\pi_0 = 1$. This bias ought to be small.

These data sets dramatically illustrate the danger of letting parametric models for centers of data determine tails: in the spirit of Efron (2004, 2008) we transformed the p-values in the wild type data to z-values using the inverse of the standard normal distribution function, fitted a $N(\mu, \sigma)$ distribution to these z-values, and then used the fitted distribution to estimate $F_0(0.001)$ to be 0.0116. Instead the empirical estimate of $F_0(0.001)$ was 0.0017 and our estimate was 0.0025. Thus, this normality based parametric estimate seemed severely wrong. Continuing, the empirical estimate of $F(0.001)$ for the genome-wide data was 0.0090, and hence the normality based estimate led to the estimate $\pi_0 \times 0116/0.0090 \approx 1.3 > 1$ for the positive false discovery rate!

Acknowledgement

We thank Simon Tavaré, Anthony Davison, Olle Nerman, Johan Segers, Anders Blomberg, and Jonathan Taylor. Research supported by the Wallenberg foundation.

Supplementary materials

Contains short introduction to SmartTail - software implementation of the methods considered in the paper; additional plots; proofs; and analysis of an *Arabidopsis* microarray experiment and an fMRI brain imaging experiment [PDF].

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful aproach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995.

S.G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001.

S. Dudoit and M.J. van der Laan. *Mulitple Testing Procedures with Applications in Genomics*. Wiley, New York, 2008.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Ass.*, 99(465):96–104, 2004.

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008.

B. Efron, R.J. Tibshirani, J.D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96(456):1151–1160, 2001.

J. Fan, P. Hall, and Q. Yao. To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.*, 102(480):1282–1288, 2007.

J. Jin and T.T. Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.*, 102(478):495–506, 2007.

K.F. Kerr. Comments on the analysis of unbalanced microarray data. *Bioinform.*, 25(16):2035–2041, 2009.

T.A. Knijnenburg, L.F.A. Wessels, J.T.M. Reinders, and I. Shmulevich. Fewer permutations, more accurate P-values. *Bioinformatics*, 25(12):161–168, 2009.

D. Ruppert, D. Nettleton, and J.T.G. Hwang. Exploring the information in p-values for the analysis and planning of multiple-test experiments. *Biometrics*, 63(2):483–495, 2007.

R.E. Schafer, J.M. Finkelstein, and J. Collins. On a goodness-of-fit test for the exponential distribution with mean unknown. *Biometrika*, 59(1):222–224, 1972.

A. Schwartzman. Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Statist.*, 2:1332–1359, 2008.

J.D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3):479–498, 2002.

J.D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, 31(6):2013–2035, 2003.

J. Warringer, E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc. Natl. Acad. Sci. USA*, 100(26): 15724–15729, 2003.

D. Zholud. Tail approximations for the student t-, F-, and Welch statistics for non-normal and not necessarily i.i.d. random variables. *Bernoulli*, 20(4):2102–2130, 2014.