Tail estimation methods for the number of false positives in high-throughput testing

Holger Rootzén and Dmitrii Zholud [*]

Department of Mathematical Statistics, Chalmers University of Technology and University of Gothenburg, Sweden.

Abstract This paper develops methods to handle false rejections in high-throughput screening experiments. The setting is very highly 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 conditional distribution of the number of false positives, given that there is in all r positives, approximately has a binomial distribution, and develop efficient and accurate methods to estimate its success probability parameter. Furthermore we provide efficient and accurate methods for estimation of the true null distribution resulting from a preprocessing method, and techniques to compare it with the theoretical null distribution. Extreme Value Statistics provides the natural analysis tools, a simple polynomial model for the tail of the distribution of p-values. We provide asymptotics which motivate this model, exhibit properties of estimators of the parameters of the model, and point to model checking tools, both for independent data and for dependent data. The methods are tried out on two large scale genomic studies and on an fMRI brain scan experiment. A software implementation, SmartTail, may be downloaded from the web.

1 Introduction

Setting: High-throughput measurements and screenings in modern bioscience differ from classical statistical testing in several ways. First, it involves testing thousands - or hundreds of thousands - of hypotheses. Second, to get a manageable amount of rejected null hypotheses, testing is typically done at extreme significance levels, e.g. with $\alpha = 0.001$ or smaller. Third, each of the individual tests are often based on very few observations, so that the degrees of freedom for t- or F-tests may be as low as 1 or 2, and degrees of freedom less than 10 are common. Fourth, in such

---

[*]E-mails: hrootzen@chalmers.se and dmitrii@zholud.com

large and complicated experiments the real null distribution of test statistics and p-values frequently deviates from the theoretical one.

In Section 5 we consider three such high throughput investigations: testing of gene interaction in yeast using a Bioscreen C Analyzer robot Warringer et al. (2003), Zholud et al. (2011); a genome-wide association scan *Arabidopsis* microarray experiment Zhao et al. (2007); and a fMRI brain imaging experiment (Dehaene-Lambertz et al. (2006), Taylor and Worsley (2006)). The number of tests for the different data sets in these investigations varied between 1,700 and 35,000 and the typical significance levels were 0.001 or less. The degrees of freedom was 1, the lowest possible, in the Bioscreen experiment. It seemed clear that for all three investigations the real null distribution differed from the theoretical one.

Property four, that the real null distribution often is different from the hypothetical one has been widely observed and discussed in the literature. For a few examples see Efron et al. (2001), Efron (2004, 2008), Jin and Cai (2007), Zhao et al. (2007), and Schwartzman (2008). Cope et al. (2004) developed a standardized set of graphical tools to evaluate high-throughput testing data.

False positives: The aim of this paper is to develop methods to understand and handle false rejections in high-throughput screening experiments. Thus we study very highly multiple testing problems where testing is performed at extreme significance levels and with low degrees of freedom, and where the true null distribution may differ from the theoretical one. Our illustrations come from biology, but the same problems appear in many other areas, too.

Our point of view is the following: the tests use a very small α and hence false rejections are determined by the extreme tails of the distributions of test statistics and p-values, and the central parts of the distributions are largely irrelevant. For this reason we throughout use the powerful methods for tail estimation which have been developed in Extreme Value Statistics.

Our main results are answers to the questions *"how many of the positive test results are false?"* and *"how should one judge if one preprocessing method makes the true null distribution closer to the theoretical one than another method?"* for such testing problems.

Our answer to the first question is i) the conditional distribution of the number of false positives given that there are in all r positives is approximately binomial, and ii) efficient and accurate methods to estimate the success probability parameter of this binomial distribution.

As answer to the second question we provide efficient and accurate methods for estimation of the true null distribution resulting from a preprocessing method, and techniques to compare it with the theoretical null distribution.

Perhaps the words *efficient* and *accurate* above should be emphasized. Existing approaches use either fully parametric models for the distributions of test quantities or p-values, or else use the empirical distribution function as estimator. Our approach instead is semi-parametric: we use a parametric model, but only for the tails of the distributions. The meaning of "efficient" then is that the random variation in our estimates is substantially smaller than for the empirical distribution function.

With "accurate" we mean that we do not make the very strong assumptions that models like normal or beta distributions can be trusted far out in the tails of the distribution. Instead we only model the tail, and let data determine the part of the tail for which the model fits well. For the details of this, see Sections 2 below, and for some concrete numerical results see Section 5.

A third contribution of this paper is that it provides an accurate estimator of Efron's local false discovery rate $fdr(x)$, see later in this section. Note that the empirical distribution does not provide any estimate of fdr(x) at all, and hence one can not talk about our estimator's "efficiency" relative to the empirical estimator (this is why here we write "accurate", not "efficient and accurate").

There is an enormous and rapidly expanding literature on multiple testing. The monograph of Dudoit and van der Laan (2008) is one entrance point into this literature. Kerr (2009) gives a recent useful review of the area. Noble (2009) is directed at practitioners. Below we discuss some specific parts of the recent literature on multiple testing more in detail. The recent paper Knijnenburg et al. (2009) suggests using Generalized Pareto approximations to improve efficiency of permutation test, in particular in bioinformatics. We are not aware of any other papers which connect EVS and high-throughput testing.

Estimation, not error control: The aim of this paper is estimation of the distribution of the number of false positives and of related quantities, and not "error control". The False Discovery Rate (FDR) error control procedure introduced by Benjamini and Hochberg (1995) has had an enormous influence on the field of multiple testing, and has seen extensive further development. However, in screening studies the aim isn't final confirmation of an effect. It instead is to select a number of interesting cases for further study. In such situations, error control may be less natural. The estimation approach to multiple testing of course already has attracted significant interest in the literature. E.g., this is the point of view in Storey (2002, 2003, 2004), Efron et al. (2001), Efron (2004, 2008), Ruppert et al. (2007), and Jin and Cai (2007).

By way of further comment, in high throughput testing and for many standard tests, such as t- or F-tests, the error control provided by the Benjamini-Hochberg method is different from what one naively could be led to expect. The reason is that for such tests the ratio {probability of false rejection}/{probability of rejection} converges to a constant as the significance level tends to zero (see e.g. Zholud (2011a)). Then, if the desired FDR is less than this constant, the Benjamini-Hochberg method simply makes it highly probable that there are no rejections at all. Since FDR is defined to be zero if there are no rejections this in turn makes the achieved FDR small. However, this may be more a formality than anything else. For high throughput screening it seems to be more useful to have good estimates of the probability of false rejection and of the distribution of the number of false rejections, rather than FDR control. These issues are discussed in further in e.g. Storey (2002, 2003), Kerr (2009), and Zholud (2011a).

Tail model: Our methods can equivalently be presented in terms of test statistics or in terms of p-values. We have found the latter formulation convenient and use

it throughout.

Let P denote a generic p-value, and, as usual, write H_0 for the null hypothesis and H_1 for the alternative hypothesis. Our basic model is that there are positive constants $c_0, \gamma_0, c_1, \gamma_1$ such that

$$F_0(x) = \Pr(P \leq x \mid H_0) = c_0 x^{1/\gamma_0}(1 + o(1)) \tag{1}$$

and

$$F_1(x) = \Pr(P \leq x \mid H_1) = c_1 x^{1/\gamma_1}(1 + o(1)), \tag{2}$$

as $x \to 0$. This model is motivated by general asymptotic arguments from extreme value theory, and also by extensive practical experience from extreme value statistics. The theoretical motivation and references are given in Section 4. Section 5 gives a data-based motivation using three studies from biology, and also introduces and illustrates a number of model checking tools, cf. Figure 2 in Section 5.

If one further assumes that tests are independent, that H_0 is true with probability π_0, and H_1 is true with probability $\pi_1 = 1 - \pi_0$, then we get the following extreme tail mixture form for the distribution of p-values,

$$F(x) = \Pr(P \leq x) = \big(\pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}\big)(1 + o(1)), \tag{3}$$

for the observed p-values. If the testing procedure is reasonable, then small p-values should be more likely under H_1 than under H_0, so it is typically reasonable to expect that $\gamma_0 \leq \gamma_1$.

Mixture models are of course very widely used in many areas, and in particular in multiple testing in bioinformatics. For a long list of references to this, see Kerr (2009); in particular Allison et al. (2002) discusses such models under the heading Mixture Model Methods, or MMM-s. We find these models natural and useful. However the results of this paper continue to hold also for models where the numbers of true and false null hypotheses are considered as fixed numbers, see below.

Asymptotics and the tail model: The situation described above is in an "asymptotics formulation" described as

$$n \text{ fixed}, m \to \infty, \ \alpha \to 0, \tag{4}$$

where n is the number of observations used in the individual tests, m is the number of tests, and α is the significance level.

The first two assumptions, n fixed, $m \to \infty$, delineate the class of high through-put testing situations which are studied in this paper. As for the third one, $\alpha \to 0$, suppose one chooses to reject the null hypotheses for all tests that give a p-value less than some critical value α. For example, in a Bonferroni procedure one would choose $\alpha = \eta/m$ (cf. Gordon et al. (2007)), with η "a fixed number". In theory the choice of η, and hence of α, would be guided by the fact that $\pi_0 \eta$ is the expected number of false rejections (provided the null distribution one uses in fact is the true one). In practice the choice of α is often based on beliefs of how often the null hypothesis is violated, and on available capacity for further study of rejected hypotheses. Since the number of p-values, m, is assumed to be very large, α in the

end typically in the situations we consider is chosen quite small, also because one wants to get a manageable number of rejections. Hence, the assumption $\alpha \to 0$.

There of course exists very large literature on central limit type approximations for the case $n \to \infty$. E.g., sharp results for uniformity in approximate t-distributions when $n \to \infty$ and $m \to \infty$ simultaneously and a literature review is given by Fan et al. (2007). However, this is not the case of interest here, and it is of course well known that for low values of n approximation by t or F-distributions can be quite inaccurate. In a another set of literature it is instead proven that if the underlying observations deviate from normality or independence then, under very general conditions, tails of one and two-sample t-statistics and of F-statistics are not the same as if the observations really were normal, see Hotelling (1961), Zholud (2011a), and references in the latter paper. However this literature also shows that the deviation is of a simple kind: under the asymptotics (4) the tail probabilities under non-normality are proportional to the tails of the relevant t or F-distributions. It in particular follows that (1) - (3) are satisfied, since these equations are known to hold for t and F-distributions.

The Extreme Tail Mixture Model: It is reasonable to neglect the $o(1)$-terms in (1) - (3) if x is not too far from α, and α is small. This leads to the following Extreme Tail Mixture Model

$$F_0(x) = \Pr(P \le x \mid H_0) = c_0 x^{1/\gamma_0}, \tag{5}$$

$$F_1(x) = \Pr(P \le x \mid H_1) = c_1 x^{1/\gamma_1}, \tag{6}$$

and

$$F(x) = \Pr(P \le x) = \pi_0 c_0 x^{1/\gamma_0} + \pi_1 c_1 x^{1/\gamma_1}. \tag{7}$$

If this model holds, tests are made at the "fixed" level α, and assuming independence between tests, then (see Section 3 below) the conditional distribution of the number false rejections, given that there has been $r > 0$ rejections in total, has approximately a binomial distribution with "number of trials" parameter r and success probability parameter

$$\text{pFDR} = \frac{\pi_0 c_0 \alpha^{1/\gamma_0}}{\pi_0 c_0 \alpha^{1/\gamma_0} + \pi_1 c_1 \alpha^{1/\gamma_1}} = \frac{\pi_0 F_0(\alpha)}{F(\alpha)}. \tag{8}$$

Here we use the notation pFDR for this parameter because it coincides with the pFDR of Storey (2002), see Section 3. (The leftover case $r = 0$ is not interesting - if there are no rejections, then one knows for sure that there are no false rejections either!)

These results apply also for the more ad hoc method, which is presumably often used in practice, where one chooses to reject the null hypothesis for the r tests which gave the smallest p-values, with "r a fixed number". Again, the choice of r is typically influenced by beliefs about the frequency of null hypothesis violations, and on available capacity for study of rejected hypotheses.

Further, the local false discovery rate (Efron et al. (2001)), which measures the "a posteriori likelihood of false rejection" of a hypothesis with p-value $x \leq \alpha$ is then

$$
\begin{aligned}
\mathrm{fdr}(x) = \Pr(H_0 \text{ true} \mid P = x) &= \frac{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0 - 1}}{\pi_0 c_0 \gamma_0^{-1} x^{1/\gamma_0 - 1} + \pi_1 c_1 \gamma_1^{-1} x^{1/\gamma_1 - 1}} \\
&= \frac{\pi_0 \frac{d}{dx} F_0(x)}{\frac{d}{dx} F(x)}. \qquad (9)
\end{aligned}
$$

The empirical distribution function doesn't provide any estimate of fdr. An alternative nonparametric possibility could be to use some kernel type estimator. However such estimators are well known to provide erratic tail estimators.

The simplest situation is when the theoretical null distribution is in fact the true null distribution, so that F_0 is the uniform distribution and $c_0 = \gamma_0 = 1$. However our focus (cf. discussion above) is situations where this doesn't hold, but we instead have an additional sample where the null hypothesis is known to be true. It is typically quite worth the effort to acquire such a sample. It could be achieved by performing an extra experiment, as in the Bioscreen example in Section 5, or, in the brain imaging example, by finding regions where it seems likely there are no effects, or by randomly changing signs in contrasts, or in other ways, see e.g. in Efron et al. (2001) and Taylor and Worsley (2006).

In this paper we show how Extreme Value Statistics can be used to get *efficient* and *accurate* estimates of the distributions (5) and (7), and of their derivatives, for the small x which are the values of interest in the present situation. EVS in addition provides confidence intervals and goodness-of-fit tests and focusses analysis and graphics on what is at the heart of the problem. The remaining parameter π_0 in (8) and (9) has to be estimated through other means, e.g. by a variant of the method in Storey (2002) – this is the only point where the entire range of p-values comes into play. Alternatively, π_0 can be conservatively estimated by setting it equal to 1. The loss of accuracy in the conservative approach is small in the situations we consider, and is quantified by (8) and (9). The estimates of $F_0(x)$, $F(x)$, and π_0 directly lead to estimates of the success probability parameter pFDR in the binomial distribution, and of fdr(x).

Dependence: Complexity and preprocessing in high throughput testing can introduce dependence between tests. The effects of time series dependence on extremes has been extensively studied in the extreme value literature, and many of the issues are well understood. In particular, extremes may be asymptotically independent even if typical observations are dependent. For one instance of this, in a paper inspired by high throughput testing in biology, see Clarke and Hall (2009). However, also for the opposite case when extremes are "asymptotically dependent", there exist good methods to deal with time dependence. Further, even if less is rigourously proven for the more complicated "spatial" dependence which often is of interest, e.g. in gene expression experiments, it is typically relatively clear how the known time series results extend to such situations. In the following sections we provide some more details on this.

Overview: In Section 2 we develop the estimation methods, and discuss how dependence influences estimation. Section 3 derives the conditional binomial distribution for the number of false positives. It also provides estimates of error control parameters such as the Benjamini-Hochberg False Detection Rate and the FamilyWise ERror. Section 4 gives the motivation for the models (1) and (2). In Section 5 we use our methods to analyze the two data sets from genomics, and the brain imaging data set discussed above. Section 6 contains a concluding discussion.

A statistical software tool, SmartTail, for performing the analyzes described in this paper may be downloaded from www.smarttail.se, and details on technical implementation of the methods can be found in Zholud (2011c,b).

2 Statistical methods

In this section we first discuss how to estimate $F_0(x)$, and how to test and produce confidence intervals. This discussion is for the case of independent p-values. A natural simplification of the mixture model (3) then makes it possible to use the same methods to estimate $F(x)$. For t- and F-tests with low degrees of freedom, it may sometimes be reasonable to replace γ_0 and γ_1 by 1, and, if the theoretical null distribution also is the true one, then $c_0 = \gamma_0 = 1$.

We further discuss how the method of Storey (2002) to estimate π_0 translates to the present setting. Together this provides all the ingredients needed for estimation of (8) and (9). Finally, if there is dependence between observations the estimators still are consistent and asymptotically normal, but if there is clustering of extremely small p-values, then the standard deviations of the estimators may be inflated.

Estimation of $F_0(x)$: To estimate F_0 we assume that it somehow has been possible to obtain a (possibly approximate) sample of m_0 p-values, $p_1^0, \ldots, p_{m_0}^0$, from the true null distribution, cf. the discussion in the introduction. Our EVS procedure for estimation of the parameters of (1) is then as follows:

The first step is to choose a threshold $u > 0$ which is small enough to make it possible to neglect the $o(1)$ term in (1) for $x \leq u$ and hence to use the Extreme Tail Model (5). This u then plays the central role that the statistical analysis only uses those of the observations (i.e. the p_i^0-s) which are less than u, and that the analysis only tries to estimate values of $F_0(x)$ for $x \leq u$.

This choice of the threshold u is a compromise between bias and variance (or, in the terminology of the introduction, between accuracy and efficiency) and is similar to the choice of bandwidth in kernel density estimation: a small u leads to less "model error", and hence less bias, but also to fewer observations to base estimation on, and hence more variance. In practice, the choice of u is guided by goodness-of-fit test and plots; see Coles (2001), Beirlant et al. (2004), and the analysis of the examples in Section 5.

For the next step, write P^0 for a random variable which has the (true) null

distribution of the p-values. From (5) it follows that

$$
\Pr\left(-\log(P^0/u) \geq x \mid P^0 \leq u\right) = \frac{c_0(ue^{-x})^{1/\gamma_0}}{c_0 u^{1/\gamma_0}} = e^{-x/\gamma_0}, \tag{10}
$$

for x positive. Thus, conditionally on $P^0 \leq u$, the variable $-\log(P^0/u)$ has an exponential distribution with mean γ_0. Let $N = \#\{1 \leq i \leq m_0;\ p_i^0 \leq u\}$ be the number of the $p_1^0, \ldots, p_{m_0}^0$ that are less than u. Since the mean of the observations is the natural estimator of the mean of an exponential distribution, the natural estimator of γ_0 is

$$
\hat{\gamma}_0 := \frac{1}{N} \sum_{1 \leq i \leq m_0;\ p_i^0 \leq u} -\log(p_i^0/u). \tag{11}
$$

This is just the ubiquitous Hill estimator in a somewhat different guise, cf. Beirlant et al. (2004), Section 4.2. Further, for $0 \leq x \leq u$, we have that $F_0(x) = \Pr\left(P^0 \leq x\right) = \Pr\left(P^0 \leq u\right)\Pr\left(P^0 \leq x \mid P^0 \leq u\right) = \Pr\left(P^0 \leq u\right) c_0 x^{1/\gamma_0}/(c_0 u^{1/\gamma_0})$. Since N/m_0 is the non-parametric estimator of $\Pr\left(P^0 \leq u\right)$ we get the semiparametric estimator

$$
\hat{F}_0(x) = \frac{N}{m_0} \left(\frac{x}{u}\right)^{1/\hat{\gamma}_0} \tag{12}
$$

of $F_0(x)$, for $0 \leq x \leq u$. An estimate of $\frac{d}{dx}F_0(x)$ is obtained by differentiating (12).

The important point here is the following. We only trust the model $F_0(x) = c_0 x^{1/\gamma_0}$ to be sufficiently accurate for "small" values of x, i.e for $x \leq u$, where u is "small". However, still this threshold u often can be chosen much larger than the critical value $x = \alpha$ used to decide if a test rejects or not, and hence the estimate $\hat{F}_0(\alpha)$ is based on many more observations - and accordingly is much more efficient - than the standard empirical distribution function estimator of $F_0(\alpha)$. Quantitative examples of this are given below, and for any specific data set the efficiency gain can be obtained from our SmartTail software.

If the observations $p_1^0, \ldots, p_{m_0}^0$ are independent, then the variance of N/m_0 is estimated by $\frac{N}{m_0}\left(1 - \frac{N}{m_0}\right)/m_0$. Since the variance of an exponential distribution is equal to the mean we have that conditionally on N the variance of $\hat{\gamma}_0$ is γ_0/N. Hence the variance of $\hat{\gamma}_0$ may be estimated by $\hat{\gamma}_0/N$. Further the parameter estimators N/m_0 and $\hat{\gamma}_0$ are asymptotically uncorrelated and asymptotically normally distributed. Thus, asymptotic confidence intervals, e.g. for $\hat{F}_0(\alpha)$, can be computed using the delta method, see Zholud (2011b).

Estimation of $F(x)$: The straightforward way to estimate parameters in the mixture density (3) would be to write down the joint conditional likelihoods of the sample from the null distribution and of the observed p-values $p_1, \ldots, p_m$ that are less than the threshold u and then maximize numerically to find the parameters $c_0, c_1, \gamma_0, \gamma_1, \pi_0$. However, if $\gamma_0 = \gamma_1 =: \gamma$ then the model collapses to

$$
F(x) = cx^{1/\gamma}, \tag{13}
$$

with $c = \pi_0 c_0 + \pi_1 c_1$, and the parameters become unidentifiable. This would presumably also make it difficult to estimate parameters if γ_0 and γ_1 are similar, even if they are not exactly equal. This identifiability problem is further compounded by the fact that in typical situations where the test works as desired, the first term in (7) would be substantially smaller than the second one – if not there would be too many false rejections.

However, turning this around, one can often expect that (13) in fact would model the observed p-values quite well. We hence propose the following procedure: First estimate γ in the model (13) from the observed p-values $p_1, \ldots, p_m$ in precisely the same way as γ_0 was estimated from $p_1^0, \ldots, p_{m_0}^0$. If this estimate is reasonably close to the estimate of γ_0 just use the model (13) for the distribution of p-values in the experiment and estimate $F(x)$ in the same way as $F_0(x)$. Confidence intervals for $F(x)$ are also obtained in the same way as for $F_0(x)$.

If the estimated γ_0 and γ are substantially different, then one might try to complement with the maximum likelihood approach outlined above, perhaps with π_0 estimated "externally" by just guessing, or by the Storey (2002) method which we discuss below. The final decision on whether to use (7) or (13) can then be based on the extent to which the fitted distributions differ, and on the relative sizes of the two terms in (3).

As a further comment, the preceding results of course are valid not only for the mixture model, but also if one assumes fixed numbers of true and false null hypotheses.

As mentioned above, there are important cases when some of the parameters are known from basic theory. In particular, if the p-values have been produced by one- or two-sample t-tests or F-tests, then $\gamma_0 = \gamma_1 = 1$, see Zholud (2011a), and in particular (13) is satisfied. If the theoretical null distribution is in fact equal to the true one, then $\gamma_0 = c_0 = 1$. In such cases one may of course use these known values instead of the estimates – but it still may be a good idea to check if they agree with the estimates.

Estimation of π_0: Storey (2002) proposed the following conservative estimator $\hat{\pi}_0 = (\#\{p_i > \lambda\}/m)/(1-\lambda)$ for the proportion of cases where the null hypothesis is true. Here $\lambda \in (0,1)$ is a suitably chosen (not "small") number. The idea behind the estimator is that "most" of the large p-values come from the null distribution, so that the numerator is an estimate of $\pi_0 \Pr(P^0 > \lambda)$, while the denominator is an estimate of $\Pr(P^0 > \lambda)$, provided the p-values in fact are uniformly distributed under the null hypothesis. In the present situation where this last assumption may not be true one can instead use the estimator

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

The choice of λ is discussed in Storey's paper.

Efficiency and accuracy : As one quantification of the gain in efficiency from using the Extreme Tail Model instead of the empirical distribution function, assume that the model (13) holds exactly for $x \leq u$, and let $\hat{F}_E(x)$ be the empirical distribution
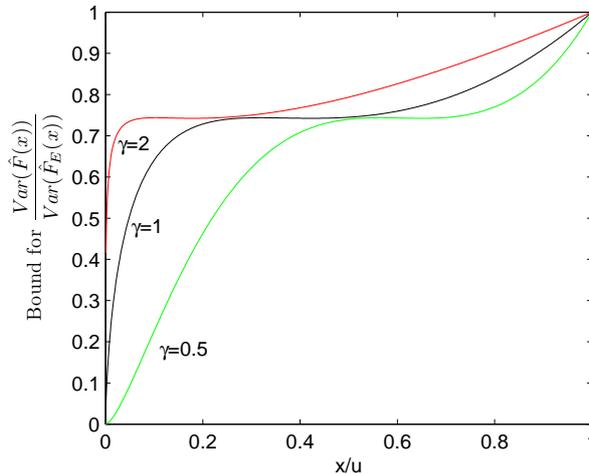
Figure 1: Efficiency gain from using the Extreme Tail Model as compared to the standard empirical CDF.

function estimator of $F(x)$. Straightforward but lengthy calculations show that then, for $x \leq u$ and $F(x) \leq 0.01$,

$$\frac{\text{Var}(\hat{F}(x))}{\text{Var}(\hat{F}_E(x))} \leq \left(\frac{x}{u}\right)^{1/\gamma} \left(1 + \frac{1.02}{\gamma^2}\left(\log\left(\frac{x}{u}\right)\right)^2\right),$$

and that this bound is quite precise. For details see Zholud (2011b). In practical use often x/u would be of the order of 0.1 - 0.01 and the value of γ would be around 1. The resulting efficiency gain is illustrated in Figure 1 for three values of γ. The same results of course apply to $\hat{F}_0(x)$. The examples in Section 5 contain some further numerical examples of the increased efficiency which can be obtained by using the Extreme Tail Model. As a final comment, for very small values of x, which sometimes can be of interest, the empirical distribution function is just not possible to use as an estimator.

As for accuracy, e.g. the papers Efron (2004, 2008, 2010) first uses the theoretical null distribution to transform test statistics to a N(0, 1) distribution and then accounts for deviations from this theoretical null distribution by fitting a N(μ, σ) distribution to these values, and finally uses the result as a tail approximation. Schwartzman (2008) instead assumes an exponential family of distributions. These papers emphasize the risks for wrong inferences which can result from the theoretical null distribution being different from the theoretical one, but don't seem to account for the risk that the assumed distribution doesn't fit in the tails. In Allison

et al. (2002) tail approximations are instead obtained by fitting a beta distribution to observed p-values. Tang et al. (2007) increases the sophistication of this approach by adding a Dirichlet prior. In these approaches, data from the center of the distributions determine the fit, and thus the statistical variability of the resulting estimators is small. Instead the fitted models are assumed to hold in the extreme tails. However, for the case considered here where each test is based on very few observations, such an assumption is not backed by theory, and in many similar situations has been observed to lead to bad modeling of tails (also in cases where the parametric model fitted very well in the center of the data). Thus there is substantial risk that these methods could lead to bad accuracy and, in our opinion, if they nevertheless are used they should at least be checked through comparison with methods such a those in this paper. In particular, it is wise to employ appropriate graphical tools which concentrate on tail fit, rather than on overall fit of distributions.

Dependence: The estimators discussed above are consistent and asymptotically normal also for dependent observations, in quite general circumstances. For the standard Hill estimator and for the "time series case" Rootzén et al. (1991) and Hsing (1991) give the precise conditions for this, using strong mixing as the dependence condition. Via (10) and (11) this directly provides corresponding results for the present situation.

Dependence can potentially inflate the variances of the estimators. However this only happens if there is clustering of small p-values, i.e. if the small p-values appear in clusters located closely together. Thus, if there isn't any clustering one can just ignore dependence and proceed as if the p-values were independent. How to check for clustering is discussed in the next section.

If extreme values in fact do cluster, it is nevertheless still possible to estimate the asymptotic variance by "blocking", see Rootzén et al. (1991). In this method the p_i^0-s are first grouped in equal sized blocks of neighboring observations, and then estimates of γ in the blocks are used to estimate the variance, in the following way: Let ℓ be the lengths of the blocks, and for simplicity assume that $n = \ell \times k$ for k integer (if this is not satisfied, one may discard the last few observations). The standard deviation of $\hat{\gamma}_0$ is then estimated by the standard deviation of the k γ-estimates computed in the blocks, divided by $\sqrt{k}$. The standard deviation of $\hat{\gamma}$ is estimated in the same way.

The choice of ℓ is again a compromise between bias and variance: too small an ℓ might cut to many clusters into two and lead to an underestimated standard deviation, while too large an ℓ gives too few block averages to use for estimation, and hence very variable estimates of the standard deviation. In practice one typically would compute the estimates for a range of values of ℓ and use this together with visual information about clustering to make the final choice of ℓ. This method of course is closely related to the block bootstrap methods used in time series analysis.

In addition to γ_0 the estimator (12) of $\hat{F}_0(x)$ also contains the factor N/m_0 where $N = \#\{1 \leq i \leq m_0; \ p_i^0 \leq u\}$. The discussion above of the influence of dependence on γ_0 carries directly over to the variance of N and the covariance

between γ_0 and N, see Rootzén et al. (1991).

In biological applications it is common that there is no natural linear ordering of the observations. However, it may still be possible to group the observations into equal-sized blocks such that there may be considerable dependence inside the blocks, while block averages are substantially independent. If this is possible, the method described above can still be used to estimate the standard deviations of $\hat{\gamma}_0$ and $\hat{\gamma}$.

3   Basic theory

The basic binomial conditional distribution for the number of false positives in the introduction follows from completely elementary reasoning. However, for completeness we still give a derivation of it for three cases: the basic model (3) where H_0 is true with probability π_0 and H_1 is true with probability π_1; the case when $m_0 = \#\{$false null hypotheses$\}$ is thought of as non-random; and a case when the critical level α is a random variable, e.g. when it is equal to the k-th largest p-value for "k non-random".

Although this is not the main thrust of this paper we also briefly illustrate how the estimates from the previous section may be used to give efficient and accurate estimates of some other standard error control parameters.

Approximate binomial distribution of the number of false positives: With notation from above we have that

$$
\begin{aligned}
m_0 &= \# \text{ true null hypotheses} \\
m_1 &= \# \text{ false null hypotheses} \\
m &= m_0 + m_1 = \text{total number of tests} \\
r &= \# \text{ rejections} \\
\alpha &= \alpha_m = \text{critical level of the tests.}
\end{aligned}
$$

Further, in the derivations below, if nothing is said to the contrary, we assume that the observed p-values are mutually independent.

Now, suppose m_0 and m_1 are non-random and tend to infinity and that α_m tends to zero, in a coordinated way so that $m_0 F_0(\alpha_m)$, and $m_1 F_1(\alpha_m)$ are bounded. Accordingly, also the expected total number of rejections is bounded. Then, by the standard Poisson limit theorem for the binomial distribution, the number of false rejections is approximately Poisson distributed with parameter $m_0 F_0(\alpha_m)$ and the number of correct rejections is approximately Poisson distributed with parameter $m_1 F_1(\alpha_m)$. It then follows at once that the conditional distribution of the number of false rejections, given that there are in total r rejections, is approximately Binomial with number of trials parameter r and success probability $m_0 F_0(\alpha_m)/(m_0 F_0(\alpha_m) + m_1 F_1(\alpha_m))$. This is the same as (8) if $\pi_0 = m_0/m, \pi_1 = m_1/m$.

If instead the mixture model (3) is assumed to hold then $m_0/m \xrightarrow{P} \pi_0$ as $m \to \infty$. Thus, for any $\epsilon > 0$ we have that for sufficiently large m the number of

false rejections is less than the number of rejections of a sample of size $(\pi_0 + \epsilon)m$ so that the number of false rejections is stochastically smaller than a binomial variable with $m_0 + \epsilon m$ trials and success probability $F_0(\alpha_m)$. Similarly one gets an upper bound for the number of correct rejections, and also corresponding lower bounds. Using monotonicity and the Poisson limit distribution of binomial distributions, as above one obtains an approximate conditional binomial distribution with success probability (8) for the number of false rejections.

Next, if instead α_m is random and it is assumed that there exists a non-random sequence $\tilde{\alpha}_m$ such that $m|F(\alpha_m) - F(\tilde{\alpha}_m)| \xrightarrow{P} 0$, then a simple argument similar to the previous one shows that asymptotically there are no p-values in the interval with endpoints α_m and $\tilde{\alpha}_m$. It then follows that the conditional distribution of the number of false rejections is the same as if the rejection level was $\tilde{\alpha}_m$, and hence the results above again apply. In particular this is the case if α_m is the r-th smallest of the p-values, for some fixed r.

Dependence and clustering of small p-values: So far we have assumed that the p-values were independent. If they instead are dependent, then the results above continue to hold if Leadbetter's conditions $D'(\alpha_m)$ and $D(\alpha_m)$ are satisfied; see Leadbetter (1974) and Leadbetter and Rootzén (1998). Here $D(\alpha_m)$ is a quite weak restriction on dependence at long distances, and can be expected to hold very generally. Instead $D'(\alpha_m)$ restricts dependence between neighboring variables. It may be violated in circumstances where small p-values occur in clusters, and typically holds otherwise.

Clustering of p-values which could make $D'(\alpha_m)$ invalid can be investigated informally by inspection of the samples, and there is also a large literature on formal estimation of the amount of clustering, as measured by the so-called Extremal Index, see e.g. Beirlant et al. (2004), Section 10.3.2. However, the issue is somewhat delicate: clustering caused by local dependence will violate the asymptotic Poisson distribution, but clusters of very small p-values may also be caused by non-null experiments occurring at neighboring locations, and this would then not contradict an asymptotic Poisson distribution. The latter situation, for example, is expected to occur in the brain scan experiment discussed in Section 5 below.

Estimation of error control parameters: With standard notation in multiple testing, let the random variables V and R be the number of false positives, and the total number of rejections, respectively. We now list number of common error control quantities, and how they may be estimated using the results from Section 2 (it is assumed that α and x are less than the threshold u). The second and third one have already been discussed above, but are included for completeness. For comprehensive listing and discussion of such parameters we refer to Dudoit and van der Laan (2008). Motivation of the estimators comes after the table. In each case a conservative estimate is obtained by setting $\hat{\pi}_0 = 1$, and the degree of conservatism can be judged directly from the formulas for the estimators.

Parameter	Estimate

The Benjamini and Hochberg (1995) False Detection Rate:

$\quad$ FDR:$= E\left(\frac{V}{R} \mid R > 0\right) \Pr(R > 0)$ $\qquad$ $\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}(1 - e^{-m\hat{F}(\alpha)})$.

The Storey (2002) False Detection Rate:

$\quad$ pFDR:$= E\left(\frac{V}{R} \mid R > 0\right)$ $\qquad$ $\frac{\hat{\pi}_0 \hat{F}_0(\alpha)}{\hat{F}(\alpha)}$.

The Efron et al. (2001) Local False Detection Rate:

$\quad$ fdr(x):$= \Pr(H_0 \text{ true} \mid P = x) = \frac{\pi_0 \frac{d}{dx} F_0(x)}{\frac{d}{dx} F(x)}$ $\qquad$ $\frac{\hat{\pi}_0 \frac{d}{dx} \hat{F}_0(x)}{\frac{d}{dx} \hat{F}(x)}$.

The FamilyWise ERror:

$\quad$ FWER:$= \Pr(V \neq 0)$ $\qquad$ $1 - e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$.

The k-FamilyWise ERror:

$\quad$ k-FWER:$= \Pr(V \geq k)$ $\qquad$ $\sum_{i=k}^{\infty} \frac{(m\hat{\pi}_0 \hat{F}_0(\alpha))^i}{i!} e^{-m\hat{\pi}_0 \hat{F}_0(\alpha)}$.

The estimate of pFDR is a consequence of the conditional binomial distribution of the number of false positives. Specifically, conditional on $R = r > 0$ the number of false positives has a binomial distribution with parameters r and $\pi_0 F_0(\alpha)/F(\alpha)$ so that $E\big(V/R \mid R = r\big) = \pi_0 F_0(\alpha)/F(\alpha)$. Hence we also have that $E\big(V/R \mid R > 0\big) = \pi_0 F_0(\alpha)/F(\alpha)$. The estimate of FDR is obtained by using the pFDR estimate for the first factor and the asymptotic Poisson distribution of the number of false positives to estimate the second factor. The FWER and k-FWER estimates use the asymptotic Poisson distribution of the number of false positives. Since they also involve $\hat{\pi}_0$ they may be harder to use in practice.

It may be noted that our estimates of pFDR and FDR are slightly different from the Storey (2002) estimates: translating Storey's estimates to the present situation, Storey's estimate of pFDR is the same as our estimate of FDR, and Storey's estimate of pFDR is obtained by dividing our pFDR estimate by $1 - e^{-m\tilde{F}(\alpha)}$.

The pFDR estimate also works quite generally for dependent p-values, see Section 2. The other three estimates require that "local dependence" of p-values is negligible for small p-values so that there is no clustering of extreme values, cf. the discussion of the condition $D'(\alpha_n)$ at the end of Section 3.

4 Motivation

In this section we give the theoretical motivation for the models (1) and (2) by showing that, very generally, they are asymptotically valid when m is large and one is far out into the tails – basically the models apply if tails of test statistics

have a natural "asymptotic stability property", or, equivalently, if the distribution of test statistics is in the domain of attraction of an extreme value distribution. This motivation of course comes from general mathematical arguments (as does the motivation for the use of normal distribution), and not from, say, specifics of biology.

Practical motivation is given by the analysis of three examples in Section 5 below and, of course, more generally from very extensive experience in using extreme value statistics in many areas of science.

For simplicity, in this section we phrase the discussion in terms of large values of the test statistic being "significant" i.e. leading to small p-values. Let T_h, T_0, and T_1 be random variables which have the distribution of the test statistic under the hypothetical (=theoretical) null distribution, the true null distribution, and the distribution under the alternative hypothesis, respectively, and let G_h, G_0, and G_1 be the corresponding distribution functions. Further, throughout let $\bar{G} = 1 - G$ denote the tail (or "survival") function associated with a distribution function (d.f.) G.

The simplest motivation is as follows. Suppose that there are constants $C_h > 0$ and $\tilde{\gamma}_h > 0$ such that

$$\bar{G}_h(x) \sim C_h \frac{1}{x^{1/\tilde{\gamma}_h}}, \quad \text{as } x \to \infty, \tag{14}$$

which e.g. holds for one-and two-sample t-statistics, and for F-statistics. Further suppose that $\bar{G}_0$ and $\bar{G}_1$ satisfy corresponding expressions. For one- and two-sample t-statistics this again holds very generally indeed, with $\tilde{\gamma}_h = \tilde{\gamma}_0 = \tilde{\gamma}_1 = f$, where f is the degrees of freedom; see Zholud (2011a). The distribution F_0 of the p-values under the true null distribution is $G_0(G_h^\leftarrow)$, where $G^\leftarrow$ denotes is the right continuous inverse of a d.f. G. Thus,

$$F_0(x) = G_0(G_h^\leftarrow(x)) \sim C_0 \frac{1}{\left((x/C_h)^{-\tilde{\gamma}_h}\right)^{1/\tilde{\gamma}_0}} = c_0 x^{1/\gamma_0}, \quad \text{as } x \to 0,$$

with $c_0 = C_0/C_h^{\tilde{\gamma}_h/\tilde{\gamma}_0}$ and $\gamma_0 = \tilde{\gamma}_h/\tilde{\gamma}_0$, so that (1) holds. Similarly it follows that (2) holds with $c_1 = C_1/C_h^{\tilde{\gamma}_h/\tilde{\gamma}_1}$ and $\gamma_1 = \tilde{\gamma}_h/\tilde{\gamma}_1$.

However, (1) and (2) hold much more generally. Let T be a random variable with d.f. G and suppose that G satisfies either one of the following two equivalent conditions: a) G belongs to the domain of attraction of an extreme value distribution, i.e. the distribution of linearly normalized maxima of i.i.d. variables with d.f. G converges, or b) the tail of G is asymptotically stable, i.e. the distribution of a scale normalized exceedance of a level u converges as u tends to the right hand endpoint of the distribution G. Then there are constants $\sigma = \sigma_u > 0$ and γ such that

$$\Pr\left(\frac{T - u}{\sigma} > x \mid T > u\right) \approx \left(1 + \frac{\gamma}{\sigma}x\right)_+^{-1/\gamma}, \tag{15}$$

for u close to the right endpoint of G. Here the + signifies that the expression in parentheses should be replaced by zero if it is negative, and the right hand side

is the tail function of a Generalized Pareto distribution. The parameter γ can be positive, zero, or negative. For $\gamma = 0$ the last term in (15) is interpreted as its limit as $\gamma \to 0$, i.e. it is $e^{-x/\sigma}$. Writing $v = \Pr(T > u)$ we get that

$$\bar{G}(x) \approx v \left(1 + \frac{\gamma}{\sigma}(x - u) \right)_+^{-1/\gamma}, \ \text{ for } \ x > u,$$

and

$$\bar{G}^{\leftarrow}(y) \approx u + \frac{\sigma}{\gamma} \left(\left(\frac{v}{y} \right)^{\gamma} - 1 \right), \ \text{ for } \ y \leq v.$$

Suppose now that G_h and G_0 satisfy (15). Then, repeating the calculations above (with the same u for both distributions) we get, with self-explanatory notation, that for $\gamma_h, \gamma_0 > 0$

$$F_0(x) \approx v_0 \left(1 - \frac{\gamma_0 \sigma_h}{\gamma_h \sigma_0} + \frac{\gamma_0 \sigma_h}{\gamma_h \sigma_0} \left(\frac{v_h}{x} \right)^{\gamma_h} \right)^{-1/\gamma_0} \approx v_0 \left(\frac{\gamma_h \sigma_0}{\gamma_0 \sigma_h} \right)^{\gamma_0} v_h^{-\gamma_h/\gamma_0} x^{\gamma_h/\gamma_0},$$

for small x, so that (1) again holds.

If instead $\gamma_h = \gamma_0 = 0$ one obtains

$$F_0(x) \approx v_0 x^{\sigma_h/\sigma_0},$$

which again is of the form (1). Cases where one γ is > 0 and the other is ≤ 0, or where one γ is ≥ 0 and the other is < 0 are not interesting, since one of the tails then completely dominates the other. Calculations become more complex if both γ-s are negative, so that the corresponding generalized Pareto distributions have a finite upper endpoint. However such cases are not expected to occur in practice, either. The motivation for (2) is the same as for (1).

5 Examples

In this section the methods introduced here are illustrated by analyses of three different data sets. All analyses were made using the SmartTail tool, see also Zholud (2011c) and Zholud (2011b).

Example *(Yeast genome screening, Warringer et al. (2003), Zholud et al. (2011)):* The data sets in this example come from a long sequence of genome-wide screening experiments for detecting differential growth under different conditions. The experiments use *Saccharomyces cerevisiae*, baker's yeast, a model organism for advancing understanding of genetics. The experiments were run on a Bioscreen Microbiology Reader (also known as Bioscreen C Analyzer). In an experiment different yeast strains are grown on two 100-well (10×10) honeycomb agar plates. The output is 200 growth curves, each representing a time series of optical density measurements from a well. Here we only consider one of the parameters extracted from these curves, the so-called *logarithmic doubling time*.

In a typical experiment a mutant yeast strain with one gene knocked out is grown in the same position in each of the two plates. A reference wild type strain without any gene knockout is grown in four wells in each plate, one well in each quadrant of the plate. Differential growth caused by a mutant is measured separately for each of the two plates by subtracting the average of the logarithmic doubling times of the four reference strains from the logarithmic doubling times for the mutant strain. This gives one value for each plate. Differential growth is then tested by comparing these two values with zero in a one-sample t-test with 1 degree of freedom.

We consider three data sets: A *Wild Type Data Set* with 1,728 observed p-values, a *Genome Wide Data Set* with 4,896 observed p-values, where the single knockout mutants and reference strains were grown under normal conditions, and a *Salt Stress Data Set* with 5,280 observed p-values, where all single knockout mutants and reference strains were grown in a salt stress environment. The Wild Type data set was obtained for quality control purposes, and hence was analyzed in exactly the same way as the genome-wide scans. However for this data set one knows that there are no real effects, so it in fact is a sample from the true null distribution. These data sets are available from the PROPHECY database (see PROPHECY in the list of references, and Fernandez-Ricaud et al. (2006)).

As a theoretical background, from Zholud (2011a) follows that for one sample t-tests the models (1) and (2) are expected to hold, with $\gamma_0 = \gamma_1 = 1$. It then follows that also (13) holds, with $\gamma = 1$ and $c = \pi_0 c_0 + \pi_1 c_1$. However, non-asymptotically, other values of the γ-s may give a better fit.

Figure 2 illustrate the results of the analysis of the wildtype data set (recall that this data set is a sample from the true null distribution). From the top right and middle panels it is clear that the true null distribution is different from the uniform distribution, and also that the model (5) fits quite well. A formal likelihood ratio test of the hypothesis $\gamma_0 = 1$ gave the p-value 0.67, and setting $\gamma_0 = 1$ doesn't change the estimates much (the estimated value of γ_0 is 0.96). The remaining three plots in Figure 2 illustrate different ways of checking the model assumption (5).

The top right panel shows the Kolmogorov-Smirnov 95% simultaneous confidence limits for the exponential distribution of $-\log(p/u)|p \leq u$ (see Lilliefors (1969) and Schafer et al. (1972)), transformed to the uniform scale. The test doesn't reject the model (5) (p-value 0.49). The bottom right panel shows that the estimate of $\hat{F}_0$ is quite insensitive to the choice of the threshold u. The individual estimates of $1/\gamma_0$ (bottom left panel) and c_0 (not shown) change slightly more when u is changed, but are still quite stable. The two bottom plots are customarily used to guide the choice of u and to check model fit.

The left panel of Figure 3 indicates that the model (13) fits the Genome Wide Data Set well (the Kolmogorov-Smirnov p-value was 0.38). The estimates of pFDR at $\alpha = .001$ are somewhat higher for small values of the threshold u. This behavior is reversed if γ is set to 1. The model checking plots corresponding to the four last panels in Figure 2 where slightly less stable than those for the wildtype data set.

17

The estimate of pFDR at $\alpha = .001$ is 0.34 (here $\gamma_0 = 0.96$ and $\gamma = 1.03$ are estimated from the data using $u = 0.05$ and $u = 0.035$ accordingly). Since there were 14 p-values less than 0.001 we hence estimate the expected number of false positives at this level to be 4.76. Using the binomial approximation to the number of false positives, we further estimate that with probability greater than 95% the number of false positives was at most 8.
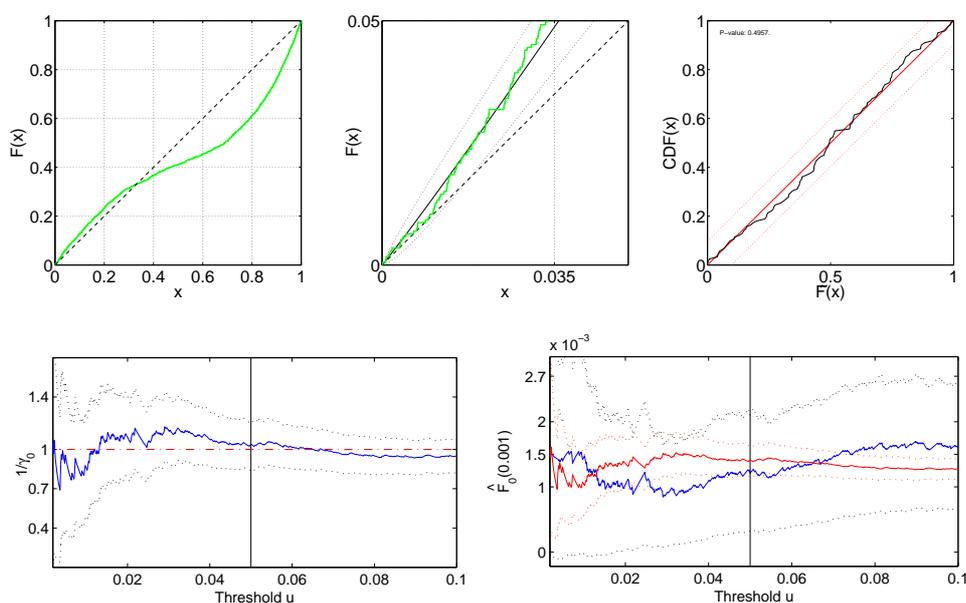

Figure 2: *The Wild Type Data Set*

Top left: Empirical distribution function. Dashed line is uniform distribution. *Top middle:* Empirical distribution function for $p \leq 0.05$ (122 values). Solid line is (5) estimated using $u = 0.05$; dashed line is uniform distribution. Dotted lines are 95% pointwise confidence intervals. *Top right:* Empirical conditional distribution function of $-\log(p/0.05)|p \leq 0.05$ transformed to the uniform scale and Kolmogorov-Smirnov 95% goodness of fit limits. *Bottom left/right:* Estimated $1/\gamma_0$ and $\hat{F}_0(0.001)$ as function of the threshold u. Red line is the same function but with γ_0 set to 1. Dotted lines are 95% pointwise confidence intervals.

If we instead had believed in a uniform distribution under the null hypothesis, we would have estimated the mean number of false positives to be 4 and that the number of false positives with probability greater than 95% was less than 7 - a somewhat too positive picture of experimental precision.

18

A further practically important question is "Which out of the 14 rejections are the true positives?". Sometimes one meets the idea that one should make an ordered list of the p-values corresponding to rejected null hypotheses and make further investigation starting with the smallest p-value, then go to the next smallest one, and so on, in the hope that the smaller the p-value, the more likely it is to correspond to true positives, see e.g. Noble (2009). For t- and F-tests with low degrees of freedom, and far out in tails theory suggests that this hope often is unfounded, see Zholud (2011a).
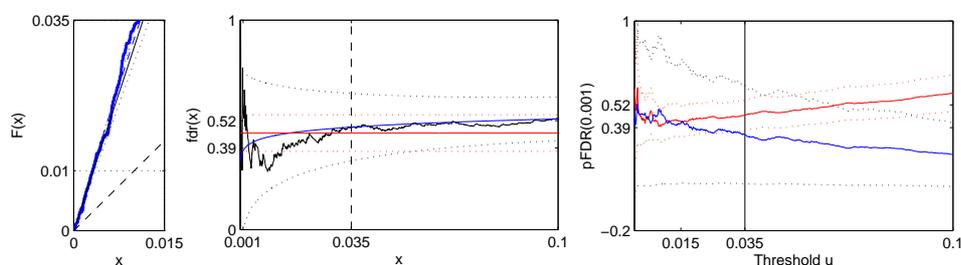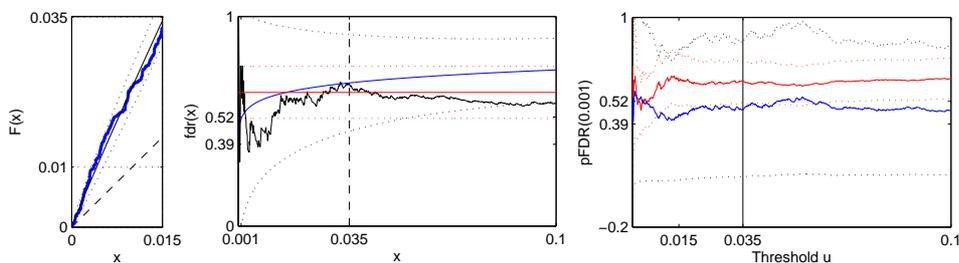

Figure 3: *The Genome Wide Data Set*
Left: Empirical distribution function for $p \leq 0.035$ (525 values); dashed line is uniform distribution. Solid line is (13) estimated using $u = 0.035$; dot-dashed line is same function for γ_0 set to 1. *Middle:* fdr(x) (smooth curve) and empirical FDR (edgy curve) for $u = .035, \pi_0 = 1$. Horizontal line is fdr(x) with γ_0 set to 1. *Right:* pFDR at $\alpha = .001$ as function of the threshold u, for $\pi_0 = 1$. Dot-dashed line is the same function with γ_0 set to 1. Dotted lines are the corresponding 95% pointwise confidence intervals.

Nevertheless, for less extreme situation Efron's fdr(x) can be used to measure how likely it is that a rejection is a false positive. The fdr(x) plot in middle panel of Figure 3 decreases as x tends to zero, but the decrease is small. This indicates that it is slightly (but only slightly) more probable that it is the rejections with the smallest p-values which are the true positives. However, it is still quite likely that also some of the tests with the smallest p-values are false positives.

Theoretically, that fdr(x) is almost constant for small x of course is a consequence of the asymptotic tail behavior of t-statistics discussed above. For the present data set this theory is also borne out by the empirical results.

The Salt Stress Data Set by and large behaved in the same way as the Genome-Wide Data Set, see Figure 4. A difference was that model checking plots were more stable, and in fact model fit seemed even better than for the wild type data.

To illustrate the gain in efficiency from using the estimates from Section 2 instead of the empirical distribution function estimator, SmartTail for the Wild Type Data Set and $u = 0.05$ estimated that $\mathrm{Var}(\hat{F}_0(0.001))/\mathrm{Var}(\hat{F}_E(0.001)) = 0.46$ and for $u = 0.035$ that $\mathrm{Var}(\hat{F}(0.001))/\mathrm{Var}(\hat{F}_E(0.001))$ was 0.7 for the Genome Wide Data Set and 0.69 for the Salt Stress Data Set. Variance estimates for the version of pFDR which uses the empirical distribution functions don't seem to be available, and hence we haven't compared the variance of the pFDR estimates from Section 2 with the variance of the empirically estimated pFDR.

Figure 4: *The Salt Stress Data Set*

Left: Empirical distribution function for $p \leq 0.035$ (409 values); dashed line is uniform distribution. Solid line is (13) estimated using $u = 0.035$; dot-dashed line is same function for γ_0 set to 1. *Middle:* fdr(x) (smooth curve) and empirical FDR (edgy curve) for $u = .035, \pi_0 = 1$. Horizontal line is fdr(x) with γ_0 set to 1. *Right:* pFDR at $\alpha = .001$ as function of the threshold u, for $\pi_0 = 1$. Red line is the same function with γ_0 set to 1. Dotted lines are the corresponding 95% pointwise confidence intervals.

At this point it should perhaps be recalled that the estimates of fdr and pFDR are somewhat biased upwards as we have set $\pi_0 = 1$. However, for the situations we are interested in, the true π_0 should be close to 1, and this bias accordingly is insignificant. The same comment applies also to all the plots which follow.

Example(*Association mapping in Arabidopsis,* Zhao et al. (2007)): This data set comes from 95 *Arabidopsis Thaliana* samples, with measurements of flowering-related phenotypes together with genotypes in the form of over 900 short sequenced fragments, distributed throughout the genome. The goal was association mapping, i.e. identification of regions of the genome where individuals who are phenotypically similar are also unusually closely genetically related. A problem is that spurious correlations may arise if the population is structured so that members of a sub-group, say samples obtained from a specific geographical area, tend to be closely related. One of the main thrusts of the paper was to evaluate 9 different statistical

20

methods to remove such spurious correlations. But of course an ultimate aim is to identify interesting genes.

Here we only consider the SNP (Single Nucleotide Polymorphism) data, and one phenotype, the one called JIC4W, which we choose since it was of special interest in the paper. Further, we only display results for two of the statistical methods, the KW method which just consisted in making Kruskal-Wallis tests without correction for population structure, and a method called Q+K which may have been the most successful of the 9 methods studied. The number of tests was 3745.

Figures 5 and 6 show that the model (13) fits both the Kruskal-Wallis and the Q+K p-values well for the values $u = 0.001$ and $u = 0.01$ of the threshold accordingly. The p-values for the Kolmogorov-Smirnov test were 0.43 and 0.38, respectively. The estimate of pFDR at $\alpha = 0.0001$ for the Kruskal-Wallis test was 0.013, and for the Q+K the estimate 0.1, i.e. more than 7 times bigger. Both these numbers assume that the true null distribution is the uniform distribution. Zhao et al. (2007) argue that most of the Kruskal-Wallis p-values are spurious. We also performed the same analysis for the other test methods proposed in their paper. For most, but not all, of them, the model (13) gave a good fit. Of course the quality of the fit also depended on the choice of u.
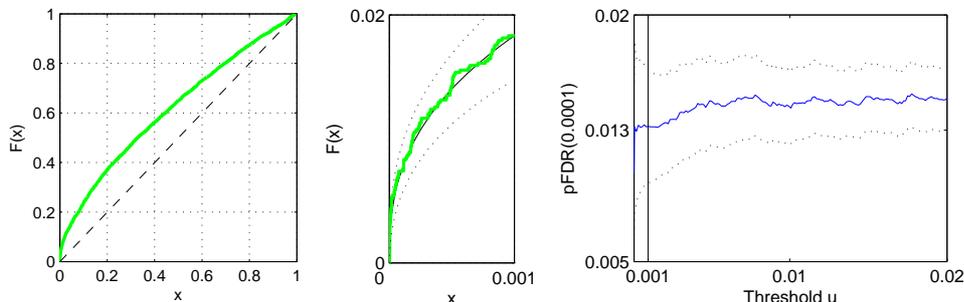

Figure 5: *The KW analysis of the Arabidopsis JIC4W data set*
Left: Empirical distribution function. Dashed line is uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.001$ (99 values). Solid line is (13) estimated using $u = 0.001$. Dotted lines are 95% pointwise confidence intervals. (OBS Scale: x-axis is stretched 10 times) *Right:* pFDR at $\alpha = .0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

Led by some speculation in the paper, we tried to use chromosomes 2 and 3 as a surrogate null distribution sample. However, the tail distribution of p-values in those chromosomes were in fact, if anything, heavier than for those in chromosomes

1, 4, 5, although the differences were well within the range of random statistical variation. Thus if there indeed were no effects present in chromosomes 2 and 3, then also most of the positives in the other genes might be false, as also is discussed in Zhao et al. (2007).

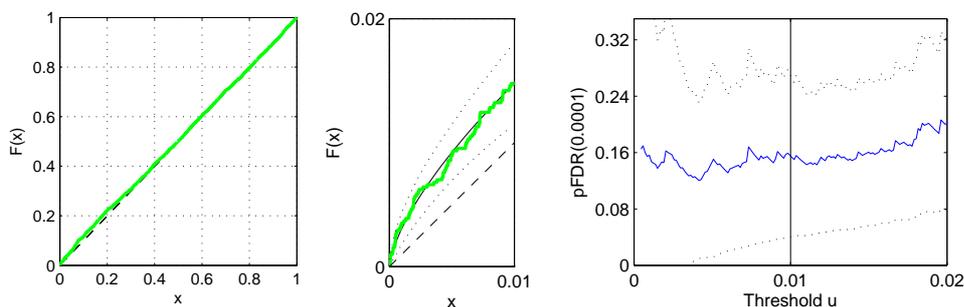Figure 6: *The K+Q analysis of the Arabidopsis JIC4W data set*
Left: Empirical distribution function. Dashed line is uniform distribution. *Middle:* Empirical distribution function for $p \leq 0.01$ (76 values). Solid line is (13) estimated using $u = 0.01$. Dotted lines are 95% pointwise confidence intervals. *Right:* p-FDR at $\alpha = .0001$ as function of the threshold u, for $\pi_0 = 1$. Dotted lines are 95% pointwise confidence intervals.

Again to illustrate the gain in efficiency from using the estimates from Section 2, SmartTail for $u = 0.001$ estimated that $\mathrm{Var}(\hat{F}(0.0001))/\mathrm{Var}(\hat{F}_E(0.0001))$ was 0.53 for the KW method and 0.85 for the Q+K method and $u = 0.01$. Note that the values of γ for these two sets of p-values were 2.6 and 1.5 accordingly.

Example*(fMRI brain scans, Taylor and Worsley (2006)):* The Functional Image Analysis Contest (FIAC) data set contains results from an fMRI experiment aimed at exploring functional organization of the language network in the human brain, see Dehaene-Lambertz et al. (2006). The part we use here is "the Block Experiment". In this experiment 16 subjects were instructed to lie still in a scanner with eyes closed and to attentively listen to blocks of 6 sentences, either different ones or the same sentence, and either read by the same speaker or by different speakers. Each subject was asked to participate in two "runs", with 16 blocks presented in each run. In Taylor and Worsley (2006), for each run and each voxel in the brain scans, the data was used to study the significance of two contrasts, "different minus same sentence" and "different minus same speaker" and the interaction between these two. Roughly 35,000 voxels per subject were used. For each voxel in each subject and each run quite sophisticated preprocessing was used to construct the corresponding 3 t-test quantities. One subject dropped out of the experiment, and

one only completed one run, so the end results was $(15 \times 2 + 1) \times 3 = 93$ sets of roughly 35,000 t-test quantities.

To study the fit of Equation (13) we transformed these t-values to p-values using a t-distribution with 40 degrees of freedom (d.f.). (This was the approximate d.f.-s according to Taylor and Worsley (2006) - it can in fact be seen that to check model fit the precise number is not important.) For each of the 93 resulting data sets of about 35,000 p-values we performed a Kolmogorov-Smirnov goodness-of-fit test of the fit of the model (13) for the p-values which were smaller than the threshold $u = 0.01$. In these 93 data sets the smallest number of p-values less than 0.01 was 117, and the largest number was 973. Figure 7 shows that the distribution of the 93 goodness-of-fit p-values are somewhat skewed towards smaller values, as compared with the uniform distribution. However, this deviation from uniformity is small, and the overall impression is that Equation (13) fits the Block Experiment FIAC data well. In fact, even for the two data sets where (13) was clearly rejected (the Kolmogorov-Smirnov p-values were 0.005 and 0.007), the Kolmogorov-Smirnov plots showed that the deviations from the model were still quite moderate, and, as expected, even smaller for thresholds u lower than 0.01 (Kolmogorov-Smirnov p-values 0.32 and 0.29 accordingly for $u = 0.005$).
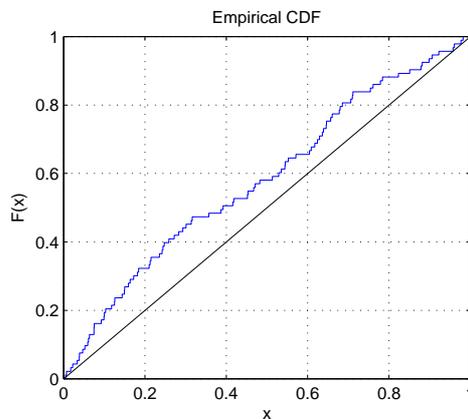

Figure 7: Empirical d.f. of the Kolmogorov-Smirnov goodness-of-fit p-values from the 93 sets of p-values in the fMRI brain scan data set.

This FIAC experiment opens possibilities for substantial further analysis. For example, Taylor and Worsley (2006) suggest randomly changing signs in contrasts to get surrogate observations from the true null distribution. A null sample might also be obtained from areas in the brain which are not involved in language processing. However, we stop at this point.

6   Discussion and conclusions

This paper is about high-throughput screening – experiments where very many individual statistical tests are performed, but where one expects that it is only possible to detect a real effect in a small or moderate number of the tests, so that testing is done at quite extreme significance levels. High-throughput testing typically involves considerable preprocessing of data before the tests are made. This, and the complexity of the experiments often cause the true null distribution to be different from the theoretical null distribution. We believe that if one suspects this is the case it may be well worth the effort to try to obtain a sample from the true null distribution, both to get a better grip on risks for false positives and for general quality control purposes. Examples of how this can be done are mentioned above.

This paper gives answers to the two questions from the introduction: "How many of the positive test results are false?" and "How should one judge if one preprocessing method makes the true null distribution closer to the theoretical one than another method?". The questions concern tails of distributions, with the central part of the distributions being largely irrelevant. We accordingly use Extreme Value Statistics in the answers. Our answer to the first question is that the conditional distribution of the number of false positives is approximately binomial, and efficient and accurate methods to estimate the success probability parameter of this binomial distribution. The answer rests on assuming a simple polynomial model for the lower tail of the distribution of p-values (cf. (5) and (6)). In Section 4 this assumption is shown to be quite generally asymptotically valid. However, of course, whether these asymptotics are relevant in a concrete testing problem has to be checked from data. We also provide methods for such model checking (see the analyzes in Section 5, in particular Figure 2).

Our answer to the second question is to compare the estimates of the true null distribution with the theoretical uniform distribution. This can be done informally from plots, or by a formal test of the hypothesis that the parameters in the null distribution (1) satisfy $c_0 = \gamma_0 = 1$. Again it is useful to complement this analysis with model checking.

A third basic question is "Which of the rejections are caused by real effects?". The answer one might hope for is that the smallest of the p-values which lead to rejections are those which correspond to real effects. Our $\mathrm{fdr}(x)$ plots can be used to judge if this in fact is the case. However, both from asymptotic theory and from our experience with data analysis, the answer might be disappointing: often the real effects are fairly randomly spread out amongst the rejections.

The p-values obtained from high throughput screening sometimes are dependent. However, not unusually this dependence affects the extreme tails less than the centers of distributions - whether this is the case or not depends on the amount of clustering of small p-values. This is discussed in Sections 2 and 3. A comforting message is that even in cases where dependence persists into the extreme tails, the estimates of basic quantities, such as pFDR, still under very wide conditions are consistent and asymptotically normal. There exists a very extensive literature

about dependent extremes for the case when observations are ordered "in time". However less is proven for the much more complicate "spatial" dependence patterns which may occur in high throughput testing, and more research is needed.

We have applied the methods developed in this paper to data from two genomics experiments, a Bioscreen yeast experiment, and an *Arabidopsis* study, and to a fMRI brain scan experiment. For all three data sets our analysis methods seem to fit the data well, and to provide useful information. In particular, they proved that for the yeast data the real null distribution was different from the uniform distribution, and quantified the rather low specificity of the tests. For the *Arabidopsis* data the methods put numbers on the differences between alternative statistical processing methods and indicated that even for the best test method, specificity may not have been all that good.

Finally, the aim of this paper is not just technical development. It is also to deliver a message: *If you are concerned with false positives in high-throughput testing, then it is the tails (and not the centers) of distributions which matter!* And, Extreme Value Statistics is *the* instrument for looking at tails. Further, already in the near future, screening experiments will become even much larger, and testing will be done at even more extreme significance levels - so the issues raised in this paper will become even more important than they now are.

References

D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernndez, C.-K. Lee, T. A. Prolla, and R. Weindruch. A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.*, 39(1):1–20, 2002. 4, 10

J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes, Theory and Applications*. Wiley, Chichester, 2004. 7, 13

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. 3, 14

S. Clarke and P. Hall. Robustness of multiple testing procedures against dependence. *Ann. Statist.*, 37(1):332–358, 2009. 6

S. G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001. 7

L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, and T. P. Speed. A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, 20(3):323–331, 2004. 2

G. Dehaene-Lambertz, S. Dehaene, J.-L. Anton, A. Campagne, A. Jobert, D. LeBihan, M. Sigman, C. Pallier, and J.-B. Poline. Functional segregation of cortical language areas by sentence repetition. *Human Brain Mapping*, 27(5):360–371, 2006. 2, 22

S. Dudoit and M.J. van der Laan. *Mulitple Testing Procedures with Applications in Genomics*. Wiley, New York, 2008. 3, 13

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Ass.*, 99(465):96–104, 2004. 2, 3, 10

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23 (1):1–22, 2008. 2, 3, 10

B. Efron. Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.*, 105(491):1042–1055, 2010. 10

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96 (456):1151–1160, 2001. 2, 3, 6, 14

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. 5

L. Fernandez-Ricaud, J. Warringer, E. Ericson, K. Glaab, P. Davidsson, F. Nilsson, G. J. Kemp, O. Nerman, and A. Blomberg. PROPHECY-a yeast phenome database, update 2006. *Nucleic Acids Res*, 35:D463–D467, 2006. 17

A. Gordon, G. Glazko, X. Qui, and A. Yakovlev. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Ann. Appl. Statist.*, 1(1):179–190, 2007. 4

H. Hotelling. The behavior of some standard statistical tests under non-standard conditions. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:319–360, 1961. 5

T. Hsing. On tail index estimation using dependent data. *Ann. Statist.*, 19(3): 1547–1569, 1991. 11

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. 2, 3

K. F. Kerr. Comments on the analysis of unbalanced microarray data. *Bioinformatics*, 25(16):2035–2041, 2009. 3, 4

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. 3

M. R. Leadbetter. On extreme values in stationary sequences. *Probability Theory and Related Fields*, 28(4):289–303, 1974. 13

M.R. Leadbetter and H. Rootzén. On extreme values in stationary random fields. In I. Karatzas, B.S. Rajput, and M.S. Taqqu, editors, *Stochastic Processes and Related Topics, in Memory of Stamatis Cambanis, 1943–1995*, pages 275–285. Birkhäuser, Boston, 1998. 13

H. W. Lilliefors. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *J. Amer. Statist. Assoc.*, 63(325):387–389, 1969. 17

W. N. Noble. How does multiple testing correction work? *Nature Biotechnology*, 27:1135 –1137, 2009. 3, 19

PROPHECY. *url:* prophecy.lundberg.gu.se - quantitative information about phenotypes for the complete collection of deletion strains in yeast (*Saccharomyces cerevisiae*). 17

H. Rootzén, L. de Haan, and M.R. Leadbetter. Tail and quantile estimation for strongly mixing stationary sequences. Technical report, dept of Statistics,University of North Carolina, 1991. 11, 12

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. 3

R. E. Schafer, J. M. Finkelstein, and John Collins. On a goodness-of-fit test for the exponential distribution with mean unknown. *Biometrika*, 59(1):222–224, 1972. 17

A. Schwartzman. Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Statist.*, 2:1332–1359, 2008. 2, 10

SmartTail. *url:* www.smarttail.se - software for correction of theoretical p-values and analysis of false discovery rates in high-throughput screening experiments. 7, 16, 22

J.D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3): 479–498, 2002. 3, 5, 6, 7, 9, 14

J.D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. 3

J.D. Storey. Srong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66(1):187–205, 2004. 3

Y. Tang, S. Ghosal, and A. Roy. Nonparametric Bayesian Estimation of Positive False Discovery Rates. *Biometrics*, 63(4):1126–1134, 2007. 11

J. E. Taylor and K. J. Worsley. Inference for magnitudes and delays of response in the FIAC data using BRAINSTAT/FMRISTAT. *Human Brain Mapping*, 27: 434–441, 2006. 2, 6, 22, 23

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. 2, 16

K. Zhao, M. J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet*, 3(1):71–82, 2007. 2, 20, 21, 22

D.S. Zholud. Extremes of Student's one- and two-sample T-test, Welch's test, and F-test for non-normal and not necessarily i.i.d. random variables. *Submitted*, 2011a. 3, 5, 9, 15, 17, 19

D.S. Zholud. On confidence intervals for SmartTail estimator. *Work in progress*, 2011b. 7, 8, 10, 16

D.S. Zholud. SmartTail - software for analysis of False Discovery Rates in high--throuput screening experiments. *Work in progress*, 2011c. 7, 16

D.S. Zholud, H. Rootzén, O. Nerman, and A. Blomberg. Positional effects in biological array experiments and their impact on False Discovery Rate. *Work in progress*, 2011. 2, 16