

Hierarchical False Discovery Rate controlling methodology

Daniel Yekutieli *

September 25, 2007

Abstract

We discuss methodology for controlling the FDR in complex large-scale studies which involve testing multiple families of hypotheses: the tested hypotheses are arranged in a tree of disjoint subfamilies, and the subfamilies of hypotheses are hierarchically tested by the Benjamini and Hochberg FDR controlling (BH) procedure. We derive an approximation for the multiple family FDR for independently distributed test statistics: q – the level at which the BH procedure is applied, times the number of families tested plus the number of discoveries, divided by the number of discoveries plus 1. We provide a universal bound for the FDR of the discoveries in the new hierarchical testing approach, $2 \times 1.44 \times q$; and demonstrate in simulations that when the data has an hierarchical structure the new testing approach can be considerably more powerful than the BH procedure.

*This work was supported by a grant from the Israeli Science Foundation. The author thanks two anonymous referees, the associate editor and editor for comments and suggestions that greatly improved the quality of the paper.

1 Introduction

The Benjamini and Hochberg (1995) FDR controlling procedure (BH procedure) has been successfully applied in overcoming the multiplicity problem in many applications, but the task performed was limited – test a family of hypotheses, determined in advance, while controlling the FDR. Yekutieli et al. (2006) introduced hierarchical testing methodology for controlling the FDR in multiple families of hypotheses, and developed a general framework, based on the hierarchical approach, for controlling the FDR in complex experiments. In this paper we formally define the hierarchical FDR testing approach, study its performance in simulations, and derive bounds and approximations for the FDR of the various types of discoveries produced by the hierarchical testing methodology. We begin our exposition by demonstrating the use of this new testing approach in three different types of applications.

Analysis of microarray data Yekutieli et al. (2006) applied hierarchical testing to a microarray experiment which included expression levels of 25,600 genes in five mice brain regions, of ten inbred mice strains. A two-way ANOVA, with strain and brain region main effects, was fitted for each gene in order to identify genes with strain expression differences. The researchers were also interested in testing the interaction terms to locate areas in the brains and strains with abnormal expression levels. In the standard FDR approach the two questions are addressed separately: applied at level 0.05 to test the strain effect for the 25,600 genes the BH procedure yielded 957

discoveries; however, the 0.05 BH procedure applied to test the 1.2 million interaction terms yielded no discoveries. In the hierarchical approach the discovery of genes with significant strain effects is considered as the initial question for each gene, and localizing the effect to specific strains and brain regions are considered follow-up questions for genes with significant strain effects – separately applying the 0.05 BH procedure in each of the 957 families of interactions corresponding to a strain effect discovery yielded a total of 170 discoveries. In this case the researcher may be interested in four types of discoveries: (a) the entire set of 1127 discoveries; (b) the 957 strain effect discoveries; (c) the 170 interaction discoveries; (d) the most detailed information for each gene: either interaction discoveries, or strain effect discoveries – for genes for which no interaction discoveries are found. Reiner et al. (2007) applied hierarchical testing to the same data, but in that study the expression of each gene, with significant strain effect, was tested for association with a series of 17 behavioral traits.

Quantitative Trait Loci analysis QTL are genetic loci affecting quantitative traits. Weller et al. (1998) suggest searching for QTL by applying the BH procedure to p-values testing for linkage between a quantitative trait and a series of genetic markers. However, as linkage is not specific to a genetic marker, FDR control over marker discoveries does not imply control over the proportion of false QTL discoveries (see supplemental report in [http : //www.amstat.org/publications/jasa/supplemental_materials](http://www.amstat.org/publications/jasa/supplemental_materials)). To

control the occurrence of false QTL discoveries it is necessary to directly test for the existence of a QTL within a specified genomic region (Zheng, 1994), preferably, pinpointing the QTL to the smallest possible interval on the chromosome; however, due to decrease in the power to discover QTLs and the increase in the number of hypotheses considered – searching at too high resolution may result in failure to discover QTL.

The suggested solution is a hierarchical multi-resolution search for QTL. In the first level, apply the BH procedure to test all chromosome level hypotheses. On chromosomes in which a discovery has been made, proceed searching for QTL by testing the two half-chromosome level hypotheses. If a half chromosome discovery is made, proceed to the quarter-chromosome level. As long as a discovery is made continue searching at a higher resolution level. Adopting this testing strategy allows one to adaptively test hypotheses at the highest resolution possible. As the discovery of high resolution discoveries makes the initial low resolution discoveries irrelevant; in this case the researcher may only be interested in the highest resolution discovery in each genomic region.

Log-linear Analysis of mouse behavior data Kafkafi et al. (2006) present an algorithm for discovering behavior patterns that differentiate between mutant and wild-type rats: a filmed session of exploratory behavior is divided into a series of 54,000 frames; nine behavioral relevant endpoints are computed for each time frame, transformed into 3-5 level ordinal scales, and

summarized in a 9-way contingency table; the algorithm then scans these immense, sparse, contingency tables for patterns with significant frequency differences between mutant and wild-type rats. The main challenge is to work at the highest resolution level and still avoid over-fitting; thus, Kafkafi et al. (2006) only scan the 50,675 subsets of cells determined by combinations of up to four ordinal scales.

In the hierarchical approach log-linear models are fitted for the mutant rat and wild-type rat contingency tables. The log-linear models are constructed hierarchically: at the first stage the 0.05 BH procedure is applied to test the terms in the main effects model; at the second stage the 0.05 BH procedure is separately applied to test two-way interactions with each of the significant main effects from the first stage; in the third stage, the 0.05 BH procedure is separately applied to test three-way interactions with each of the significant two-way interaction; and so on. The tables are then scanned for the behavior patterns with the largest differences between the fitted values of the two log-linear models. In this case the researcher is interested in controlling the FDR for all the terms in the models, however the fact that hierarchical FDR model selection produces parsimonious models with small MSE may be a more relevant property.

1.1 Background

In their seminal paper Benjamini and Hochberg introduced the BH procedure and the FDR – a new measure for type I error in multiple testing. Since then

FDR methodology has become a very active field of research; but, essentially, all the methods only apply to a single family of hypotheses. Yekutieli and Benjamini (1999) expressed control of the FDR as an estimation problem: instead of comparing the sorted p-values to a series of critical values determined by the FDR level q , estimate the FDR of a fixed rejection region test. Storey (2002) and Storey (2003) discussed a Bayesian setting for the fixed rejection region FDR and introduced the positive FDR and the q-value with the following, very appealing, Bayesian interpretation: the conditional probability that a discovery is a false discovery given that its test statistic is in the rejection region. Efron et al. (2001) suggested empirical Bayes estimation of the FDR and even considered conditioning locally on the value of the test statistic, not just the rejection region. Another estimation effort is the estimation of the proportion of true null hypotheses (Benjamini and Hochberg, 2000; Storey, Taylor and Siegmund, 2004) in order to derive more powerful testing procedures. Genovese and Wasserman (2004) developed a framework in which the False Discovery Proportion, the number of false rejections divided by the number of rejections in a continuum of fixed rejection regions, is treated as a stochastic process. Benjamini and Yekutieli (2005a) generalized the FDR criterion to a measure for the validity of confidence intervals for parameters following selection. In this paper we discuss a general framework for applying the BH procedure and controlling the FDR in a wide variety of settings not considered before.

Hierarchical modelling is widely used in the construction of complex sta-

tistical models: for example the node splitting decisions in CART (Brieman et al., 1984), or in the treatment of interaction terms in linear models. Separately correcting for multiplicity in several families of hypotheses is also not new. When the Family Wise Error rate (FWE) is separately controlled in several families of hypotheses the FWE for the entire set of tested hypotheses is essentially the sum of the individual FWE levels. Thus the usual solution is splitting the nominal FWE level, α , between the families of hypotheses (for a QTL mapping example see Simonson and McIntyre, 2004). In the FDR literature it has been suggested to apply the BH procedure, at level q , in each family of hypotheses: Abramovich et al. (1998) discussed the asymptotic theoretical properties of this practice; in the context of QTL mapping, Lee et al. (2002) suggested a separate FDR controlled search for each quantitative trait; Yekutieli et al. (2006) suggest dividing the statistical analysis in complex studies into several main research directions, and controlling the FDR separately in each research direction. In general, FDR control separately applied in several families does not necessarily imply FDR control for the entire study (Benjamini and Yekutieli, 2005b). In this paper we compute a bound for the FDR when the BH is applied at level q to several families of hypotheses, and our main message is that hierarchical application of the BH procedure inherently implies global FDR control for the entire set of discoveries.

Screening the null hypotheses prior to testing is a related method of alleviating the multiplicity practiced in recent microarray analyses (Pavlidis,

2003; Letwin et al., 2006; Yekutieli et al., 2006; Reiner et al., 2007). Unlike hierarchical testing, any screening criterion can be applied to select the tested hypotheses, and all the screened hypotheses are tested simultaneously; similarly to hierarchical testing, screening must be independent of the hypotheses testing – Reiner et al. (2007) show that the FDR is not controlled when the BH procedure is applied to test pairwise strain expression differences of genes screened by a one-way ANOVA F-test. Reiner et al. (2007) also show in simulations that screening prior to testing offers less power than hierarchical testing.

In Section 2 we formally define the hierarchical testing approach, and summarize the theoretical results. Microarray analysis and signal denoising simulation studies are described in Section 3. Section 4 is devoted to the Discussion. We derive the FDR bounds in the Appendix. The technical issues needed to derive the theoretical results and detailed accounts of the simulations were deferred to the supplemental report.

2 Testing trees of hypotheses

In the hierarchical approach the set of tested hypotheses, $H_1 \cdots H_m$, is arranged in a tree with L levels. With the exception of hypotheses on the first level of the tree – which have no parent hypotheses; each hypothesis, H_i , on level $L(i) = 2 \cdots L$, is associated with a single parent hypothesis, indexed by $Par(i)$, on level $L(i) - 1$. Let $H_1 \cdots H_T$ denote the parent hypotheses, then

we can divide the m hypotheses into $T + 1$ families: $\mathcal{T}_0 = \{H_i : L(i) = 1\}$, and $\mathcal{T}_t = \{H_i : \text{Par}(i) = t\}$ for $t = 1 \cdots T$; m_t and m_t^0 are the total number of hypotheses and the number of true null hypotheses in \mathcal{T}_t .

The hierarchical test of the tree of hypotheses has two elements: (a) hypotheses in the same family are tested simultaneously; (b) testing begins with \mathcal{T}_0 and a family of hypotheses on higher levels of the tree is tested only if its parent hypothesis is rejected. Our concern in this work is with trees in which each family of hypotheses, \mathcal{T}_t , is tested by the BH procedure.

Definition 2.1 Level q BH procedure on \mathcal{T}_t .

1. Let $P_{(1)}^t \leq \cdots \leq P_{(m_t)}^t$ denote the set of ordered p-values corresponding to the hypotheses in \mathcal{T}_t .
2. Let $r_t = \max\{i : P_{(i)}^t \leq i \cdot q/m_t\}$.
3. If $r_t > 0$ reject the r_t hypotheses corresponding to $P_{(1)}^t \cdots P_{(r_t)}^t$.

We assume that the p-values are independently distributed; if H_j is a true null hypothesis then $P_j \sim U[0, 1]$, and for a false null hypothesis, H_j , P_j satisfies the following condition:

Condition 2.2 For all $0 < \alpha_1 \leq \alpha_2 \leq 1$

$$\alpha_1/\alpha_2 \leq \Pr(P_j \leq \alpha_1 \mid P_j \leq \alpha_2).$$

Condition 2.2 states that the conditional marginal distribution of all the p-values is uniform, or stochastically smaller than uniform. It is satisfied if the cdf of P_j is concave, e.g. under the monotone likelihood ratio condition, yet it is a somewhat stronger condition than stochastically smaller than $U[0, 1]$.

The final component of FDR trees is the specification of the set of discoveries that are of interest to the investigator. The FDR is then defined as the expected proportion of false "interesting" discoveries out of the total number of "interesting" discoveries made (this proportion is set to 0 if no discoveries are made). In this paper we discuss three types of FDR:

1. *Full tree FDR*: interest lies in the entire set of FDR tree discoveries.
2. *Level restricted FDR*: the investigator is only interested in the discoveries on a specific level of the tree. The level restricted FDR is also a model for the FDR of the entire study when the BH procedure is separately applied to several families of hypotheses (see Example A.5).
3. *Outer nodes FDR*: interest lies in discoveries which are not parents to other discoveries – e.g. in QTL analysis the highest resolution discovery in each genomic region.

A schematic drawing of a tree of hypotheses, and the results of the hierarchical test are presented in Figure 1. The tree includes 12 hypotheses in 6 families: $\mathcal{T}_0 = \{H_1, H_2\}$; $\mathcal{T}_1 = \{H_3, H_4\}$; $\mathcal{T}_2 = \{H_5, H_6\}$; $\mathcal{T}_3 = \{H_7\}$; $\mathcal{T}_4 = \{H_8, H_9\}$; $\mathcal{T}_5 = \{H_{10}, H_{11}, H_{12}\}$. Six discoveries are considered in the

full tree FDR. H_1 and H_2 are level-1 discoveries; H_3 and H_5 are level-2 discoveries; H_{10} and H_{12} are level-3 discoveries; H_3 , H_{10} and H_{12} are the outer nodes discoveries. \mathcal{T}_4 was not tested since H_4 was not rejected. Note that outer node discoveries are not necessarily in the leaves of the tree. For instance, H_3 is an outer node discovery because none of its children (only H_7 in this example) were rejected in the procedure.

2.1 Summary of the results

The theoretical results in this paper are derived under the assumption that the p-value are independently distributed, true null hypotheses p-values have $U[0, 1]$ distributions, and the false null hypotheses p-values satisfy Condition 2.2. The FDR bounds are sums over $t = 0 \dots T$ of FDR_t ,

$$FDR_t = E\{ I(\mathcal{T}_t \text{ is tested, } R > 0) \cdot \frac{V_t}{R} \},$$

where V_t is the number of false discoveries in \mathcal{T}_t and R is the total number of FDR tree discoveries.

To derive the bounds, FDR_t is expressed in terms of $R_t^{P_i=0}$ and $R^{P_i=0}$ – the number of discoveries in \mathcal{T}_t and total number of FDR tree discoveries given that a p-value corresponding to H_i , a true null hypothesis in \mathcal{T}_t , is set to 0 (setting $P_i = 0$ produces the conditional number of discoveries given that H_i is rejected). Employing this approach Benjamini and Yekutieli (2001) prove that the FDR of the level q BH procedure applied to a single family of independently distributed p-values is $q \cdot m^0/m$. Since it is not possible to aggregate the expression for FDR_t over multiple families of hypotheses we

substitute $R_t^{P_i=0}$ and $R^{P_i=0}$ with $R_t + 1$ and $R + 1$ — the unconditional number of discoveries in \mathcal{T}_t and total number of FDR tree discoveries, plus 1 (as true null hypotheses are rarely rejected in multiple testing procedures, the conditional number of discoveries given that H_i is rejected can be approximated by the unconditional number of discoveries plus 1); and the main technical issue addressed in this paper is finding a multiplicative factor, δ^* , as small as possible, yielding

$$E \frac{R_t^{P_i=0}}{R^{P_i=0}} / E \frac{R_t + 1}{R + 1} \leq \delta^*, \quad (1)$$

2.1.1 Assessment of δ^*

δ^* is assessed analytically and in simulations. At first, we consider the conditional distribution of $R_t^{P_i=1}$, the number of discoveries in \mathcal{T}_t for $P_i = 1$, given $R_t^{P_i=0}$. For $R_t^{P_i=0} = 1 \cdots m_t$, we examine the ratio in

$$R_t^{P_i=0} / E(R_t^{P_i=1} + 1 | R_t^{P_i=0}) \leq \delta^*. \quad (2)$$

We prove that for each value of $R_t^{P_i=0}$ the ratio is maximized when all the p-values have $U[0, 1]$ distributions, compute the ratio for each value of $R_t^{P_i=0}$ under this assumption, and verify that under Condition 2.2 the maximal value of the ratio is slightly less than 1.44 for $R_t^{P_i=0} = 4$. We then prove that (2) is a stronger condition than (1). Thus we prove that if the p-values are independently distributed and Condition 2.2 is kept then $\delta^* = 1.44$ satisfies Inequality (1).

In the first set of simulations assessing δ^* we study the distribution of

$R_t^{P_i=0}$ and the ratio in (2) for various p-value distributions; in the second set of simulations we directly assess the ratio in (1). The simulations reveal that for $q = 0.05$ the cases for which the ratio in (2) approaches its maximum, 1.44, are rare, and that in most cases $\delta^* \approx 1$ is sufficient to satisfy (1); while higher values of δ^* are only needed for testing several hundreds of hypotheses with a nearly uniform p-value distribution.

2.1.2 FDR bounds and approximations

With δ^* we can phrase the main result of the paper:

Proposition 2.3 *For the three types of tree discoveries and δ^* satisfying Inequality (1):*

$$FDR_t \leq \delta^* \cdot q \cdot \frac{m_t^0}{m} \cdot E\left\{ I(\mathcal{T}_t \text{ is tested}) \cdot \frac{R_t + 1}{R + 1} \right\} \quad (3)$$

The proof of Proposition 2.3, the lemmas leading to this result, and the assessment of δ^* were deferred to the supplemental report. The results given in the following paragraph are derived in the Appendix.

Summing (3) over $t = 0 \cdots T$ yields a bound for the FDR

$$FDR \leq q \cdot \delta^* \cdot E\left\{ \frac{\# \text{ of discoveries} + \# \text{ of families tested}}{\# \text{ of discoveries} + 1} \cdot \tilde{\pi}_0 \right\}, \quad (4)$$

where $\tilde{\pi}_0$ is mean of m_t^0/m_t , weighted proportionally to $R_t + 1$. Imposing the restriction of hierarchical testing yields universal bounds for the FDR:

1. The full tree FDR of any FDR tree is less than $q \cdot \delta^* \cdot 2$

2. The outer nodes FDR of any L level FDR tree is less than $L \cdot q \cdot \delta^* \cdot 2$.

Expression (4) also implies that if the number of discoveries greatly exceeds the number of families tested then the three types of hierarchical FDR are approximately $q \cdot \delta^* \cdot \tilde{\pi}_0$. We suggest using the observed number of families tested and the observed number of discoveries to approximate the FDR

$$FDR = q \cdot \delta^* \cdot \frac{\text{obs. \# of discoveries} + \text{obs. \# of families tested}}{\text{obs. \# of discoveries} + 1}. \quad (5)$$

While the theoretical properties of the term on the right side of (5) – the FDR multiplier – are not clear, the simulations in Section 3.1 reveal that it can be used to approximate the FDR, especially for the level restricted FDR which has no universal bound. Thus, in the analysis of microarray data example discussed in the introduction, the full tree FDR for the entire set of 1127 discoveries is approximately 0.092 ($= 0.05 \cdot (1127 + 957 + 1)/(1127 + 1)$), and the FDR for the 170 interaction discoveries is approximately 0.330 ($= 0.05 \cdot (170 + 957)/(170 + 1)$).

3 Simulation studies

3.1 Hierarchical FDR analysis of microarray data

The simulations model a 10,000 gene microarray experiment: a treatment is compared to a control on 8 mouse strains. The first research question is to find genes whose expression level is affected by the treatment; the follow-up questions are comparisons of the treatment effect between the 28 pairs of mice strains. Gene expression levels were modelled as treatment effect, plus

strain-specific treatment effect, plus correlated Normal noise. In the direct approach the 280,000 pairwise comparisons are tested simultaneously by the BH procedure. In the hierarchical approach the data is analyzed with a two level FDR tree tested by the 0.05 BH procedure: \mathcal{T}_0 includes the null hypotheses, for each gene, that the mean treatment effect (across all strains) is zero; and each hypothesis in \mathcal{T}_0 is parent to the 28 corresponding pairwise comparison hypotheses. In the hierarchical approach we consider four types of discoveries: (1) treatment effect discoveries are the level-1 discoveries; (2) pairwise discoveries are the level-2 discoveries; (3) full tree discoveries refers to the entire set of discoveries; (4) outer node discoveries are the pairwise differences plus the treatment effect discoveries for genes for which no pairwise differences were found. Note that the addition of correlated noise and the use of pairwise comparisons induces dependence across the tree, yet each pairwise comparison is independent of the initial treatment effect test.

The results of the simulations are presented in Table 1. The mean number of discoveries is listed in Column 5 and the FDR is listed in Column 6, their standard error is in parentheses. The mean observed value of the FDR multiplier, which is mean of the number of discoveries plus number of families tested divided by the number of discoveries plus 1 is in Column 7, its MAD is in parentheses.

Applied at level 0.05 the BH procedure yields 21 – 23 discoveries for sparse strain effect and dense strain effect configurations; in the sparse & small configuration the BH procedure loses its power: at level 0.05 the mean

number of discoveries goes down to 0.1, and even for $q = 0.50$ the mean number of discoveries is only 10. For zero treatment effect and sparse strain effect the power of the hierarchical approach was comparable to the power of the direct approach, but for the dense strain-specific effect the number of pairwise discoveries went up to 508. The performance of the hierarchical approach improved substantially for small and medium treatment effects: in the sparse & small strain-specific configuration with medium treatment effect the hierarchical procedure yielded approximately 627 treatment discoveries and 128 pairwise discoveries – over twelve times more discoveries than the level 0.50 BH procedure discoveries.

The level-1 restricted FDR was $0.05 \times m_0^0/m_0$. The full tree and outer nodes FDRs were less than the corresponding universal bounds, and even less than $0.05 \cdot \delta^*$ times the FDR multiplier ($\delta^* \leq 1.1$ in the simulations). In the first four parameter configurations the level-2 restricted FDR (which has no universal bound), was also less than $0.05 \cdot \delta^*$ times the FDR multiplier; but in the last configuration the level-2 FDR was 0.358 and exceeded its approximation $0.329 = 0.05 \cdot 1.1 \cdot 5.98$ – this indicates that hierarchical FDRs may increase in response to dependence across the tree. Examining the MAD of the FDR multiplier reveals that in many of the configurations the dispersion of the FDR multiplier is relatively small, this suggests that the observed FDR multiplier values (rather than the simulation mean FDR multiplier) can be used to approximate the FDR.

3.2 Hierarchical FDR signal denoising

We compare signal denoising via hard-thresholding based on two-level hierarchical testing schemes to the FDR wavelet thresholding approach introduced in Abramovich and Benjamini (1996), and demonstrate the effectiveness of incorporating prior knowledge in the design of hierarchical testing procedures. The signal consisted of K segments of 128 observations: a segment of 128 observation de-measured Doppler signal, followed by $K - 1$ segments of 128 zeroes. In each run of the simulation independent $N(0, 1)$ noise was added to the signal.

The Abramovich and Benjamini (1996) FDR wavelet thresholding was performed by applying the BH procedure at level 0.05 to the p-values corresponding to the $128 \times K - 1$ wavelet coefficient. For the hierarchical FDR approach a wavelet transform was separately applied to each of the K segments: for $k = 1 \cdots K$, \mathcal{T}_0 included a specific wavelet coefficient null hypothesis H_k , to test the null hypothesis that segment k is pure noise; and H_k , was parent to the family of 126 remaining wavelet coefficient null hypotheses corresponding to segment k . We experimented with two choices of wavelet coefficients in \mathcal{T}_0 : (1) The *default* choice, based on the assumption that in a non-null signal the effect size of the lowest resolution wavelet coefficient is large, was the lowest resolution wavelet coefficient. (2) The *optimal* choice, based on prior information on the type of signal used, was the largest wavelet coefficient in the wavelet transform of the Doppler signal.

The simulation study included 27 signal configurations: three levels of

Signal to Noise Ratio times nine values of K ranging from 1 to 256. Each configuration was run 3,000 times. For each run, we recorded the observed proportion of false discoveries for single stage FDR thresholding at $q = 0.05$ and for hierarchical FDR thresholding, also at $q = 0.05$; we then applied the inverse wavelet transform, and computed the SSE in relation to the true signal.

the simulation FDR levels for the two hierarchical procedures was approximately $0.05 \cdot \tilde{\pi}_0$. The SSE values of the three FDR denoising schemes are displayed in Figure 2. The optimal hierarchical procedure had the smallest SSE in all configurations – almost constant for all values of K . The SSE of the BH procedure was the same for $K = 1$ – but increased with K . For high SNR the SSE of the default hierarchical procedure was almost as small as the SSE of the optimal procedure; for small SNR, the poor choice of test statistic for H_1 made it difficult and recover the signal, resulting in high SSE.

4 Discussion

Hierarchical FDR methodology can be used to control the FDR in complex large-scale studies with multiple families of hypotheses. We have also shown that it is considerably more powerful than the BH procedure in sparse testing problems – providing that the data has a hierarchical structure. In the examples considered in this paper, the tested hypotheses were arranged in trees of homogenous families: either families consisting of true null hypotheses with a true null hypothesis parent, or families with a large proportion of

false null hypotheses and a false null hypothesis parent. Thus hierarchical testing passed over the families of true null hypotheses and adaptively applied the BH procedure to the families with high signal to noise ratio. As many of the statistical methods are hierarchical, these are intuitive and easily set-up testing schemes – the p-values tested in the examples discussed in the introduction and in the simulation studies were computed through a series of nested linear models.

Throughout the paper we make the assumption that the p-values are independently distributed. But it is important to make the distinction between dependence across the tree and dependence between a test statistic and its ancestors. The validity of hypotheses testing, in general, is based on the assumption the distribution of p-values corresponding to true null hypotheses is $U[0, 1]$ (or stochastically larger than $U[0, 1]$). In the hierarchical FDR approach it further implies that dependence between a p-value and any of its ancestors should not be allowed – thus when tested, i.e. given that all its ancestor hypotheses were rejected, the distribution of a p-value corresponding to a true null hypotheses is still $U[0, 1]$ (recall that this property was also needed for screening prior to testing). On the other hand, dependence across the tree can be allowed: it occurs in the examples presented in the introduction; and we study its effect on the FDR in the microarray analysis simulation study: the simulations indicate that while the universal FDR bounds apply for the dependent test statistics studied, dependence within the families of tested hypotheses and across the tree seems to result in higher

FDR values than expected for independent test statistics. We plan to study the effect of dependence across the tree on the hierarchical FDR in future work.

References

- [1] Abramovich F., Benjamini Y., (1996) “Adaptive thresholding of wavelet coefficients”, *Computational Statistics and Data Analysis*; **22**, 351-361.
- [2] Abramovich F., Benjamini Y., Donoho D. L., Johnstone I. M., (1998) “The amalgamation challenge to signal de-noising”, Research paper of the Department of Statistics and OR, Tel Aviv University, RP-SOR-98-03.
- [3] Benjamini Y., Hochberg Y., (1995) “Controlling the False Discovery Rate: a practical and powerful approach to multiple testing” *Journal of the Royal Statistical Society, Series B*; **57** (1): 289-300.
- [4] Benjamini Y., Hochberg Y., (2000) “On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics” *J. Educ Behav Stat*, **25**, 60-83.
- [5] Benjamini Y., Yekutieli D., (2001) “The control of the False Discovery Rate in multiple testing under dependency”, *The Annals of Statistics*; **29**, 1165-1188.

- [6] Benjamini Y., Yekutieli D., (2005a) “False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters” *Journal of the American Statistical Association*, **100**, 71-81.
- [7] Benjamini Y., Yekutieli D., (2005b) “Quantitative Trait Loci Analysis using the False Discovery Rate”, *Genetics*, **171**, 783-790.
- [8] Brieman L., Friedman J. H., Olshen R. A., Stone C. J., (1984) “Classification and regression trees”, the Wadsworth statistics/probability series, Wadsworth International Group, 1984.
- [9] Efron B., Tibshirani R., Storey J. D., Tusher V., (2001) “Empirical Bayes Analysis of a Microarray Experiment”, *Journal of the American Statistical Association*, **96**, 1151-1160.
- [10] Genovese C., Wasserman L., (2004) “A Stochastic Process Approach to False Discovery Control”, *Annals of Statistics*, **32** (3) 1035-1061.
- [11] Kafkafi N., Yekutieli D., Yarowsky P., Elmer G., (2006) “Pattern Array: a Novel Approach for Mining Raw Behavioral Data”, submitted to BMC Bioinformatics.
- [12] Lee H., Dekkers J. C. M., Soller M., Malek M., Fernando R. L., Rothschild M. F. (2002) “Application of the False Discovery Rate to Quantitative Trait Loci Interval Mapping With Multiple Traits”, *Genetics*, **161**, 905-914.

- [13] Letwin N.E., Kafkafi N., Benjamini Y., Mayo C., Frank B. C., Luu T., Lee N. H., Elmer G. I., (2006) “Combined application of behavior genetics and microarray analysis to identify regional expression themes and gene-behavior associations”, *J. Neurosci*, **26**, 5277-5287.
- [14] Pavlidis P., (2003) “Using ANOVA for gene selection from microarray studies of the nervous system”, *Methods*, **31**, 282-289.
- [15] Reiner A., Yekutieli D., Letwin N. E., Elmer G. I., Lee N. H., Kafkafi N., Benjamini Y., (2007) “Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay”, to appear in *Bioinformatics*.
- [16] Simonson K. L., McIntyre L. M., (2004) “Using Alpha Wisely: Improving Power to Detect Multiple QTL” *Statistical Applications in Genetics and Molecular Biology*, Volume 3, Issue 1, 2004 Article 1.
- [17] Storey J. D., (2002) “A direct approach to false discovery rates”, *Journal of the Royal Statistical Society: Series B*, **64** 479-498.
- [18] Storey J. D., (2003) “The positive false discovery rate: A Bayesian interpretation and the q-value”, *Annals of Statistics*, **31**, 2013-2035.
- [19] Storey J. D., Taylor J. E., Siegmund D., (2004) “Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach”, *Journal of the Royal Statistical Society: Series B*, **66**, 187-205(19).

- [20] Weller J. I., Song J. Z., Heyen D. W., Lewin H. A., Ron M., (1998) “A new approach to the problem of multiple comparisons in the genetic dissection of complex traits”, *Genetics*; **150** (4), 1699-1706.
- [21] Yekutieli D., Benjamini Y. (1999), “Resampling based false discovery rate controlling procedure for dependent test statistics”, *Journal of Statistical Planning and Inference*, **82** (1-2), 171-196.
- [22] Yekutieli D., Reiner A., Elmer G. I., Kafkafi N., Letwin N. E., Lee N. H., Benjamini Y. (2006) “Approaches to multiplicity issues in complex research in microarray analysis”, *Statistica Neerlandica*, **60** nr. 4, 414-437.
- [23] Zheng Z. B., (1994), “Precision mapping of quantitative trait loci”, *Genetics*, **136**, 1457-68.

A Appendix: FDR computations

Let D_j denote the event – H_j is rejected in the BH procedure on $\mathcal{T}_{Par(j)}$; and let D_j^{Par} denote the event – H_j and all of its ancestors are rejected in the BH procedure on their respective family, where $D_0^{Par} = \Omega$. We can now formally define the discovery of a null hypothesis, H_j , in the three tree testing schemes.

Definition A.1

1. H_j is a full tree discovery: D_j^{Par} .
2. H_j is an outer nodes discovery: $D_j^{Par} \cap (\cap_{k \in \mathcal{T}_j} \overline{D_k})$.
3. H_j is a level- l restricted discovery: $D_j^{Par} \cap L(j) = l$.

Let R_t and V_t denote the total number of full tree discoveries and the number of false full tree discoveries in \mathcal{T}_t , R is the total number of discoveries in any of the three tree testing schemes. Proposition 2.3 can be expressed as

$$FDR_t \leq \delta^* \cdot q \cdot \frac{m_t^0}{m_t} \cdot E\left\{ I(D_t^{Par}) \cdot \frac{R_t + 1}{R + 1} \right\}. \quad (6)$$

We partition the sample space into disjoint events indexed by subtrees of hypotheses K

$$D_K^{Par} = \left(\cap_{t \in K} D_t \right) \cap \left(\cap_{t \notin K} \overline{D_t} \right)$$

and for a given realization of the vector of m p-values, we denote the realized subtree by J . As the hypotheses are tested hierarchically, for $t = 0 \cdots T$

$$t \notin J \Rightarrow D_J^{Par} \cap D_t^{Par} = \emptyset, \quad \text{while } t \in J \Rightarrow D_J^{Par} \subseteq D_t^{Par} \quad (7)$$

A.1 Full tree FDR

Let $R^{Full} = \sum_{t \in J} R_t$ denote the total number of full tree discoveries, then full tree FDR can be expressed as

$$\begin{aligned}
FDR^{Full} &= E\left\{ \sum_{t=0}^T I(D_t^{Par}, R_0 > 0) \cdot \frac{V_t}{R^{Full}} \right\} \\
&= \sum_{t=0}^T E\left\{ I(D_t^{Par}, R_0 > 0) \cdot \frac{V_t}{R^{Full}} \right\} \\
&\leq \sum_{t=0}^T \delta^* \cdot q \cdot \frac{m_t^0}{m_t} \cdot E\left\{ I(D_t^{Par}) \cdot \frac{R_t + 1}{R^{Full} + 1} \right\} \tag{8}
\end{aligned}$$

$$\begin{aligned}
&= \delta^* \cdot q \cdot E\left\{ \sum_{t=0}^T I(D_t^{Par}) \cdot \frac{m_t^0}{m_t} \cdot \frac{R_t + 1}{R^{Full} + 1} \right\} \\
&= \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \sum_{t=0}^T I(D_t^{Par}, D_J^{Par}) \cdot \frac{m_t^0}{m_t} \cdot \frac{R_t + 1}{R^{Full} + 1} \right\} \\
&= \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{\sum_{t \in J} (R_t + 1) \cdot m_t^0 / m_t}{R^{Full} + 1} \right\} \tag{9}
\end{aligned}$$

$$= \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{R^{Full} + |J|}{R^{Full} + 1} \cdot \frac{\sum_{t \in J} (R_t + 1) \cdot m_t^0 / m_t}{\sum_{t \in J} (R_t + 1)} \right\} \tag{10}$$

where the inequality in (8) is due to (6) and the equality in (9) is due to (7).

Corollary A.2 *The full tree FDR is less than $\delta^* \cdot q \cdot 2$.*

Proof. With the exception of \mathcal{T}_0 , each $t \in J$ corresponds to the full tree discovery of H_t , thus $R^{Full} \geq |J| - 1 \geq 0$, and as $m_t^0 / m_t \leq 1$

$$\begin{aligned}
FDR^{Full} &\leq \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{R^{Full} + |J|}{R^{Full} + 1} \right\} \\
&\leq \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{|J| - 1 + |J|}{|J| - 1 + 1} \right\} < \delta^* \cdot q \cdot 2.
\end{aligned}$$

¶

Notice that R^{Full} equals $|J| - 1$ plus the number of leaf discoveries. Therefore the universal bound in Corollary A.2 can only be approached if there is a very small proportion of leaf discoveries and m_t^0/m_t in the families of tested hypotheses is close to 1. On the other hand, if $R^{Full} \gg |J|$ the expression in (10) is approximately $\delta^* \cdot q \cdot \tilde{\pi}_0$, where

$$\tilde{\pi}_0 = E\left\{ \sum_J I(D_J^{Par}) \frac{\sum_{t \in J} (R_t + 1) \cdot m_t^0/m_t}{\sum_{t \in J} (R_t + 1)} \right\}.$$

A.2 Outer nodes FDR

Let FDR^{Outer} , R^{Outer} and V_t^{Outer} denote the outer nodes FDR, the total number of outer node discoveries and the number of false outer node discoveries in \mathcal{T}_t . Per definition, $R^{Outer} \leq R^{Full}$ and $V_t^{Outer} \leq V_t$. In the simulations presented in this paper true null hypotheses are parents to families of true null hypotheses, as families of true null hypotheses rarely yield discoveries $V_t^{Outer} \approx V_t$, while R^{Outer} was usually less than R^{Full} , thus $FDR^{Outer} > FDR^{Full}$. The bounds for FDR^{Outer} are also greater than the bound for FDR^{Full} . But, in general, FDR^{Outer} can also be smaller than or equal to FDR^{Full} . To derive the bound for FDR^{Outer} , we substitute V_t^{Outer} with V_t and apply Proposition 2.3

$$\begin{aligned} FDR^{Outer} &= E\left\{ \sum_{t=0}^T I(D_t^{Par}, R_0 > 0) \cdot \frac{V_t^{Outer}}{R^{Outer}} \right\} \\ &\leq E\left\{ \sum_{t=0}^T I(D_t^{Par}, R_0 > 0) \cdot \frac{V_t}{R^{Outer}} \right\} \end{aligned}$$

$$\begin{aligned}
&\leq \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{\sum_{t \in J} (R_t + 1) \cdot m_t^0 / m_t}{R^{Outer} + 1} \right\} \\
&= \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{R^{Full} + |J|}{R^{Outer} + 1} \cdot \frac{\sum_{t \in J} (R_t + 1) \cdot m_t^0 / m_t}{\sum_{t \in J} (R_t + 1)} \right\}. \quad (11)
\end{aligned}$$

As J includes all indices of full tree discoveries which are not outer node discoveries $R^{Full} \leq R^{Outer} + |J|$; thus if $R^{Full} \gg |J|$ the bound in (11) is approximately $\delta^* \cdot q \cdot \tilde{\pi}_0$.

For the universal bound notice that R^{Outer} is always greater than or equal to the number of full tree discoveries at each level of the tree

$$R^{Full} + |J| \leq 2 \cdot R^{Full} \leq 2 \cdot L \cdot R^{Outer}. \quad (12)$$

As $m_t^0 / m_t \leq 1$, combining (12) and (11) yields:

Corollary A.3 *The outer nodes FDR is less than $\delta^* \cdot q \cdot 2 \cdot L$.*

In the following example we consider a very tall and narrow tree in which most discoveries are parent hypotheses. Thus we see that for large L , FDR^{Full} gets close to $\delta^* q \tilde{\pi}_0 2$, while FDR^{Outer} is less than half of $\delta^* q \tilde{\pi}_0 2L$.

Example A.4 $2 \cdot L$ null hypotheses are tested in a L level tree: for $l = 1 \cdots L$, H_l are false null hypotheses with $P_l = 0$; for $l = L + 1 \cdots 2L$, H_l are true null hypotheses with iid $U[0, 1]$ p-values; and for $l = 0 \cdots L - 1$, \mathcal{T}_l includes H_{l+1} and H_{L+l+1} . The tree of hypotheses is tested at level 0.05. Thus $J \equiv \{0 \cdots L - 1\}$, in each family $m_t^0 / m_t = 1/2$, and $\delta^* = 2/2.05$. In both testing schemes $V \sim \text{Binom}(L, 0.05)$, while $R^{Full} = L + V$ and $R^{Outer} = 1 + V$. For $L = 2, 3, 6$ and 10 : FDR^{Full} is 0.033, 0.036, 0.042

and 0.044, and FDR^{Outer} is 0.049, 0.073, 0.138 and 0.216. As $L \rightarrow \infty$, $FDR^{Full} \rightarrow 0.0476$ and $FDR^{Outer} \rightarrow 1$.

A.3 Level restricted FDR

For $1 \leq l \leq L$, let $J_l = \{t : t \in J, L(t) = l - 1\}$, and denote $R^l = \sum_{t \in J_l} R_t$ the number of level- l discoveries. The level- l restricted FDR is

$$\begin{aligned}
& FDR^{Level=l} \\
&= E\left\{ \sum_{t=0}^T I(D_t^{Par}, L(t) = l - 1, R_l > 0) \cdot \frac{V_t}{R^l} \right\} \\
&\leq \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{\sum_{t \in J_l} (R_t + 1) \cdot m_t^0 / m_t}{\sum_{t \in J_l} R_t + 1} \right\} \\
&= \delta^* \cdot q \cdot E\left\{ \sum_J I(D_J^{Par}) \cdot \frac{R^l + |J_l|}{R^l + 1} \cdot \frac{\sum_{t \in J_l} (R_t + 1) \cdot m_t^0 / m_t}{\sum_{t \in J_l} (R_t + 1)} \right\}. \quad (13)
\end{aligned}$$

There is no universal bound for $FDR^{Level=l}$, but if $R^l \gg |J_l|$ then the bound in (13) is approximately $\delta^* \cdot q \cdot \tilde{\pi}_0$ (in this case $\tilde{\pi}_0$ is the expected mean of m_t^0 / m_t over J_l).

Example A.5 A 2 level tree is tested. The level-1 hypotheses $- H_1 \cdots H_T$, are false null hypotheses with zero p-values. Each false null hypothesis H_t is parent to single true null hypothesis H_{T+t} . The number of false outer node discoveries and false full tree discoveries is $V \sim Binom(T, q)$; $R^{Full} = V + T$ hence $FDR^{Full} \approx q/(1 + q)$, and $R^{Outer} = T$ hence $FDR^{Outer} = q$. $FDR^{Level=1}$ is by definition 0. $FDR^{Level=2} = 1 - (1 - q)^T$ approaches 1 as $T \rightarrow \infty$

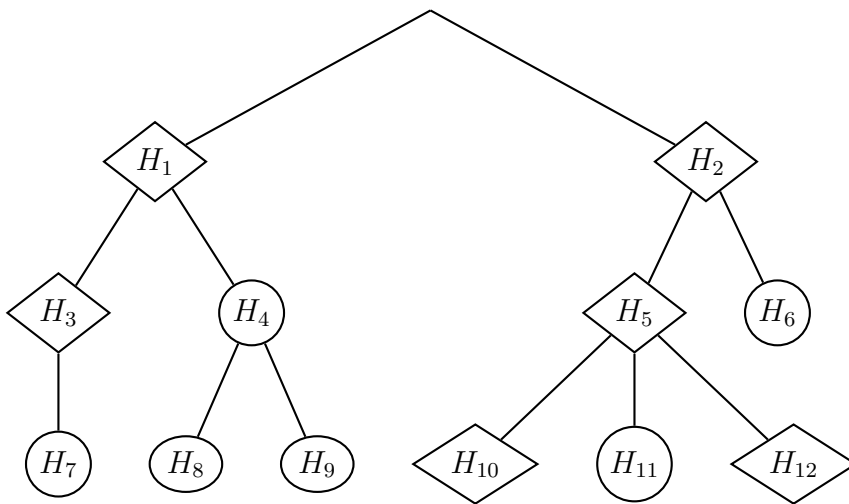


Figure 1: FDR tree schematic. Diamonds indicate null hypotheses rejected in the Hierarchical testing scheme; Circles are null hypotheses not rejected in the Hierarchical testing scheme; Ovals are null hypotheses in families not tested in the Hierarchical testing scheme.

Treat. effect	Strain effect	Analysis method	Discovery type	# of (s.e.) discoveries	FDR (s.e.)	FDR (MAD) multiplier
Zero	Sparse	BH – 0.05	Pairwise	21.1 (0.6)	0.044 (0.002)	1
		Hierar.	Treatment	9.6 (0.4)	0.044 (0.004)	1
			Pairwise	22.8 (0.9)	0.045 (0.004)	1.38 (0.21)
			Full tree	32.5 (1.3)	0.048 (0.003)	1.31 (0.11)
	Outer Node		27.6 (1.1)	0.055 (0.004)	1.36 (0.13)	
	Dense	BH – 0.05	Pairwise	21.7 (0.7)	0.044 (0.002)	1
		Hierar.	Treatment	404.5 (4.7)	0.039 (0.001)	1
			Pairwise	508.4 (5.7)	0.053 (0.001)	1.80 (0.08)
Full tree			912.9 (10.2)	0.047 (0)	1.44 (0.03)	
Outer Node	777.2 (8.7)		0.055 (0)	1.52 (0.03)		
Small	Sparse	BH – 0.05	Pairwise	22.1 (0.6)	0.048 (0.003)	1
		Hierar.	Treatment	157.0 (2.9)	0.04 (0.001)	1
			Pairwise	269.3 (4.4)	0.063 (0.001)	1.57 (0.09)
			Full tree	426.3 (7.2)	0.055 (0.001)	1.36 (0.03)
	Outer Node		362.3 (6.2)	0.064 (0.001)	1.43 (0.04)	
	Dense	BH – 0.05	Pairwise	22.9 (0.6)	0.044 (0.002)	1
		Hierar.	Treatment	1070 (5)	0.04 (0)	1
			Pairwise	1159 (6)	0.063 (0)	1.92 (0.07)
Full tree			2229 (11)	0.052 (0)	1.48 (0.02)	
Outer Node	1910 (10)		0.06 (0)	1.56 (0.02)		
Medium	Sparse & small	BH – 0.05	Pairwise	0.11 (0.02)	0.029 (0.007)	1
		BH – 0.35	Pairwise	2.73 (0.2)	0.314 (0.017)	1
		BH – 0.50	Pairwise	10.1 (0.8)	0.463 (0.016)	1
		Hierarch.	Treatment	626.8 (7.2)	0.039 (0)	1
			Pairwise	128.3 (1.7)	0.358 (0.003)	5.98 (0.91)
			Full tree	755.2 (8.6)	0.093 (0.001)	1.83 (0.02)
Outer Node	702.9 (8.0)		0.099 (0.001)	1.89 (0.02)		

Table 1: Results of hierarchical analysis of microarray data simulation study.

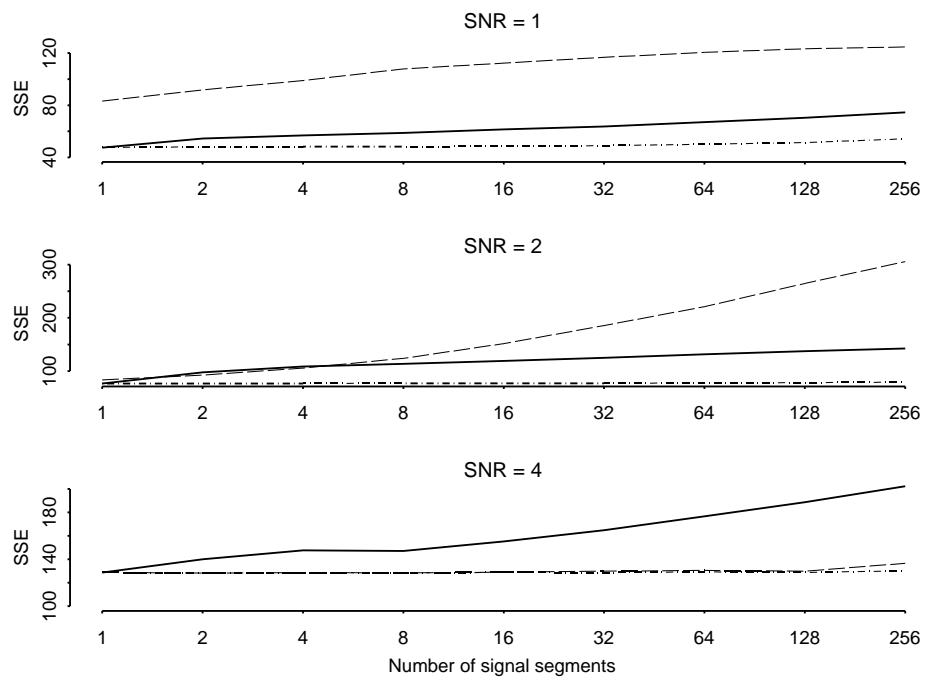


Figure 2: Simulation based mean SSE level from the signal denoising simulation study: BH level 0.05 FDR wavelet thresholding – solid lines; default hierarchical level 0.05 FDR wavelet thresholding – dashed lines; optimal hierarchical level 0.05 FDR wavelet thresholding – dotted-dashed lines.