Statistica Neerlandica (2006) Vol. 60, nr. 4, pp. 414–437

Approaches to multiplicity issues in complex research in microarray analysis

Daniel Yekutieli*, Anat Reiner-Benaim and Yoav Benjamini

Department of Statistics and Operations Research, Sackler Faculty of Exact Sciences, Tel-Aviv University, Ramat Aviv, P.O.B. 39040, Tel Aviv 61390, Israel

Gregory I. Elmer and Neri Kafkafi

Maryland Psychiatric Research Center, University of Maryland, Baltimore, MO, USA

Noah E. Letwin and Norman H. Lee

Department of Functional Genomics, The Institute for Genomic Research, Maryland & The George Washington University Medical Center, Washington. DC, USA

The multiplicity problem is evident in the simplest form of statistical analysis of gene expression data – the identification of differentially expressed genes. In more complex analysis, the problem is compounded by the multiplicity of hypotheses per gene. Thus, in some cases, it may be necessary to consider testing millions of hypotheses. We present three general approaches for addressing multiplicity in large research problems. (a) Use the scalability of false discovery rate (FDR) controlling procedures; (b) apply FDR-controlling procedures to a selected subset of hypotheses; (c) apply hierarchical FDR-controlling procedures. We also offer a general framework for ensuring reproducible results in complex research, where a researcher faces more than just one large research problem. We demonstrate these approaches by analyzing the results of a complex experiment involving the study of gene expression levels in different brain regions across multiple mouse strains.

Keywords and Phrases: false discovery rate, hierarchical testing, high throughput analysis.

*yekutiel@math.tau.ac.il

The results in this paper replace in part the results in the technical report: Benjamini, Y. and Yekutieli, D. (2002) *Hierarchical FDR* Testing of Trees of Hypotheses, Tel-Aviv University, Department of Statistics and Operations Research.

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

1 Introduction

1.1 Multiplicity issues in assessing differentially expressed genes

Identifying differentially expressed genes under two experimental conditions, possibly a treatment group versus a control, is the most direct form of utilizing expression data. Testing differential expression at each of the many genes, multiple testing is an immediate concern. When many hypotheses are tested, the probability that a type I error is committed increases sharply with the number of hypotheses. This problem of multiplicity is not unique to microarray technology, yet its magnitude here, where a single experiment may involve many thousands of genes, dramatically intensifies the problem.

While a control of the family-wise error rate (FWE) is needed in some cases, the multiplicity problem in microarray data analysis does not require protection against even a single type I error, so that the severe loss of power involved in such protection is not justified. Instead, it may be more appropriate to emphasize the proportion of errors among the identified differentially expressed genes. The expectation of this proportion is the false discovery rate (FDR) of Benjamini and Hochberg (1995, hereafter BH).

Advances in FDR methodology offer improved ways of incorporating FDR control in gene expression analysis. Reciprocally, the extensive studies by statisticians on microarray data have advanced our understanding of the FDR approach and enriched the methodology. We review the research related to the simultaneous testing of a family of hypotheses in Section 2 (a section that the reader more familiar with the area may choose to skip). Controlling the FDR in complex and large-scale studies of gene expression, discussed in later sections, is the focus of this paper.

1.2 Controlling FDR over a very large family

The FDR criterion is scalable in the sense that enough power can be retained even when testing a very large family, so simply using the linear step-up procedure in BH may accommodate a research direction that involves testing a very large family of hypotheses. In the following sections we present, and illustrate with real data examples, two additional approaches for controlling the FDR over a very large family. The second approach is to select a subset of hypotheses from the large pool of hypotheses considered, using test statistics that are independent from the those used to test the hypotheses, and apply the BH procedure to the subset of selected hypotheses. The third approach is to arrange the families of hypotheses in a hierarchical tree structure and apply a hierarchical testing procedure to the tree of hypotheses, where a hypothesis is tested only if its parent hypothesis is rejected, and where all hypotheses sharing the same parent are tested simultaneously. If the testing at each branch of the tree is conducted using the linear step-up procedure one can get bounds on the overall FDR, but it now matters what structure has the set of rejected

hypotheses that are of interest to the researchers. In Section 3 we introduce trees of hypotheses, hierarchical testing and three types of FDR that may be relevant. We discuss the theoretical properties of hierarchical FDR testing and demonstrate their use.

1.3 Complex research questions

The common concern of researchers of multiple inferences is the level of FDR (or FWE) in the set of rejected hypotheses when testing a family of null hypotheses. But when it comes to complex, large-scale studies, it is no longer clear that all hypotheses encountered form a single family, as demonstrated in the sequel. WESTFALL and YOUNG (1993) suggest three guidelines for determining whether a set of tests should be considered a 'family' of tests, thus tested in a multiple testing procedure. (a) The questions asked form a natural and coherent unit. (b) All tests are considered simultaneously. (c) It is considered *a priori* probable that many or all members of the family of null hypotheses are in fact true. It is important to note that the guidelines are subjective in the sense that a family of hypotheses is determined not only by the nature of the hypotheses tested but also by the goal of the study and by the prior knowledge of the researcher. The time may be ripe for a more serious discussion of the above criteria, but even without such a discussion, it becomes clear that the set of hypotheses that is of interest to the researcher in a single study does not necessarily form a single family of hypotheses, as demonstrated in the sequel. We call such a study, comprising of several families of hypotheses, a complex study. Complex studies are not limited to microarray analysis: even in two-way ANOVA all pairwise comparisons of one factor and all pairwise comparisons of the second factor are usually not treated jointly as a single family, but rather as two families, and multiplicity is controlled over each one separately. The novelty in microarray data is in the scale of the problem. For illustration, some of the statistical analyses described in this paper involve testing hypotheses in tens of thousands of separate families – a family of hypotheses for each of the $\sim$26,000 genes in the array. Therefore, a problem that could be overlooked in a single two-way ANOVA can no longer escape attention in complex microarray studies.

Some authors have already discussed the possibility of controlling the FDR separately in several families of hypotheses. ABRAMOVICH *et al.* (1998) discussed the asymptotic theoretical properties of this practice. LEE *et al.* (2002) suggested dividing a search for genetic determinant of multiple quantitative traits into separate FDR-controlled searches for each quantitative trait, and the practice was further discussed in BENJAMINI and YEKUTIELI (2005). In general, the relationship between the FDR in each of the families of tested hypotheses and the combined FDR is complicated. If all the hypotheses tested are true null hypotheses then the FDR will equal the FWE, and the combined FDR will increase with the number of families tested. On the other hand, if each family of hypotheses also consists of false null hypotheses with near-zero *P*-values then the proportion of false discoveries made by FDR

controlling procedure applied on each family will be approximately q, and the overall FDR of the entire study will also be approximately q.

In response to this complexity, our second offering in this paper is a comprehensive framework, based on the hierarchical FDR approach, for analyzing and addressing the multiplicity problem in complex studies. We suggest dividing the statistical analysis into several main research directions, or research questions, and controlling the FDR for each research direction separately. It is even not necessary to specify all the research directions simultaneously; additional research directions can be included at later stages of the analysis.

If the FDR is controlled locally in each research direction and the total number of discoveries greatly exceeds the number of research directions considered, then the FDR is also controlled globally in the entire study.

The specification of research directions and the choice of families of null hypotheses influence the discoveries made. Therefore, the interpretation of the results of a complex study is impossible without a detailed account of the choice of families of tested hypotheses, the considerations guiding the researcher in constructing the families of hypotheses, and the multiplicity corrections applied. We therefore describe in some detail the experimental setting and research questions involving gene expression-indifferent brain regions, which are of great biological interest, and do not merely serve as the vehicle for explaining and demonstrating our approaches.

2 The FDR

2.1 The FDR criterion

The traditional approach in simultaneous testing has been to construct a procedure that controls the probability of making one or more type I error – the FWE. Benjamini and Hochberg (1995) offered another measure for the erroneous rejection of a number of true null hypotheses, the FDR. The FDR is the expected proportion of erroneously rejected null hypotheses among the rejected ones. When some of the tested hypotheses are in fact false, FDR-control is less strict than FWE control, and thus FDR-controlling procedures are potentially more powerful. The analysis of gene expression data is a case where FDR control suffices, as its purpose is to extract genes that are potential candidates for further investigation. A single or even several erroneous rejections will not distort the conclusions at this stage of the investigation, as long as their proportion is small. Such errors do incur economical cost in that pursuing them at later stages will result in loss of time and money, so we would like to minimize their number. However, to ensure that the probability of even one such erroneous rejection will result in large loss of power, appears to be over-conservative. Controlling the FDR at some level q allows the researcher to control the proportion of effort invested in vain, on the average, at the next stage of the investigation.

In order to define the FDR precisely we need a few notations. Consider a family of m simultaneously tested null hypotheses $H_1, \ldots, H_m$, m_0 of which are true and

m_1, are false. For each hypothesis, H_i, a test statistic is calculated along with the corresponding P-value P_i. Let R denote the number of hypotheses rejected by a procedure, and V the number of true null hypotheses erroneously rejected. Let Q denote the proportion of false discovery V/R when $R>0$, and 0 otherwise. Then FDR is defined as $\mathrm{FDR}=E(Q)$.

2.2 The linear step-up procedure in (BH)

This procedure makes use of the ordered P-values $P_{(1)}\leq\cdots\leq P_{(m)}$. Denote the corresponding null hypotheses $H_{(1)},\ldots,H_{(m)}$. For a desired FDR level q, the ordered P-value $P_{(i)}$ is compared with the critical value $i\cdot q/m$. Let $k=\max\left\{i:P_{(i)}\leq i\cdot q/m\right\}$. Then reject $H_{(1)},\ldots,H_{(k)}$, if such a k exists.

Benjamini and Hochberg (1995) proved that the BH procedure controls the FDR at the level $q\cdot m_0/m\leq q$ for independently distributed test statistics. Benjamini and Yekutieli (2001) further show that the FDR is controlled for positively dependent test statistics as well. The type of positive dependency needed is a weak form of positive regression dependency – positive regression dependence of the entire set of test statistics on each of the test statistic corresponding to the true null hypotheses. In particular, the condition is satisfied by positively correlated normally distributed one-sided test statistics, and their Studentized t-tests.

For more general cases, in which the positive dependency conditions are not applicable, Benjamini and Yekutieli (2001) prove that replacing q with $q/\Sigma_{i=1}^{m}\left(1/i\right)$ in the BH procedure will provide FDR control for all types of test statistic distributions. However, this modification is needlessly conservative for the microarray problem. The simulation study presented in Reiner, Yekutieli and Benjamini (2003) demonstrates that applying the BH procedure at level q already controls the FDR for identifying differentially expressed genes.

2.3 Adaptive procedures

As the BH procedure controls the FDR at a level too low by a factor of m_0/m, it is natural to try to estimate m_0 and use $q^*=q\cdot m/\hat{m}_0$ instead of q to gain more power. In fact, in a well-defined asymptotic context, Genovese and Wasserman (2002) showed that the BH procedure at level $q\cdot m/m_0$ is the most powerful level q FDR-controlling procedure, in the sense that it minimizes the expected proportion of the hypotheses for which the alternatives hold among the non-rejected ones (minimizing the false non-discovery rate). Estimating m_0 from a set of P-values goes back to Schweder and Spjøvtoll (1982). Hochberg and Benjamini (1990) formalized their approach and synthesized an adaptive procedure that controls the FWE (see Turkheimer, 2001, for further progress). Benjamini and Hochberg (2000) suggest the adaptive procedure that combines the estimation of m_0 with the BH procedure. The procedure first uses the linear step-up procedure at level q, and stops if no hypotheses are rejected. Otherwise, m_0 is estimated by the following algorithm:

1. Compute $m_0[k] = (m + 1 - k)/(1 - P_{(k)})$.
2. Starting with $k = 2$, stop when for the first time $m_0[k] > m_0[k]$.
3. Estimate $\hat{m}_0 = \text{Ceiling}(\min(m_0[k], m))$.
4. Use the linear step-up procedure with $q^* = q \cdot m/\hat{m}_0$.

A related method of estimating m_0, actually $m_0/m = \pi_0$, appeared in Storey (2002) and implemented in Significance analysis microarray (SAM) (Storey and Tibshirani, 2003) in the context of the Bayesian FDR discussed below. Given λ, a tuning parameter, use $\hat{\pi}_0(\lambda) = (m - R(\lambda))/((1 - \lambda) \cdot m)$ as an estimator of π_0, where $R(\lambda)$ is the number of P-values less than λ. They further suggest choosing λ by bootstrapping.

It is quite impossible to review all methods suggested over the last 5 years to estimate the factor π_0 without devoting a full review to the topic. It is interesting to note that most of these methods were motivated by microarray applications where the many genes offer opportunities to estimate this factor either parametrically or non-parametrically. (For a partial list see Jiang, 2004.)

We would like to emphasize two adaptive procedures of proven properties. Storey, Taylor and Siegmund (2004) prove that adhering to fixed λ and adding 1 to the numerator of the above estimate of π_0, the above-the-linear step-up with q replaced by $q^* = q/\hat{\pi}_0(\lambda)$, controls the FDR at q for independent test statistics. Benjamini, Krieger and Yekutieli (2001) suggest a two-stage procedure with proven FDR-controlling properties under independence and show in simulations that their procedure offers FDR control under positive dependency. The procedure is as follows: apply the BH procedure at level $q' = q/(1 + q)$, let r_1 be the number of rejected hypotheses; if $r_1 = 0$ reject no hypotheses and stop; if $r_1 = m$ reject all m hypotheses and stop; otherwise, let $\hat{m}_0 = m - r_1$, and apply the BH procedure with $q^* = q' \cdot m/\hat{m}_0$.

Adaptive methods offer better performance only by utilizing the difference between m_0/m and 1. If the difference is small, i.e. when the potential proportion of differentially expressed genes is small, they offer little advantage in power while their properties are not well established under dependency. As more specific genes are pre-selected to the microarray experiments, such that the proportion of differentially expressed genes is not small, m_0/m gets smaller, and the adaptive procedures will offer a more detectable advantage.

2.4 Multiplicity-adjusted P-values

The results of a multiple testing procedure can be reported as multiplicity-adjusted P-values. As with the regular P-value, the adjusted P-value of each hypothesis is compared with the desired significance level of the test, and if smaller, the hypothesis is rejected. For an FWE-controlling procedure, the adjusted P-value of an individual hypothesis is the lowest level for which FWE $\leq \alpha$. For instance, the adjusted P-value for the Bonferroni procedure is simply $P_i \cdot m$.

For an FDR-controlling procedure, the adjusted P-value of an individual hypothesis is the lowest level of FDR for which the hypothesis is included in the set of rejected hypotheses. Thus the adjusted P-value using the linear step-up procedure is $P_{(j)}^{\mathrm{BH}} = \min_{j \le i}(P_i \cdot m/i)$.

2.5 *Bayesian formulation of the FDR*

A different line of development in FDR methodology is reflected in the evolution of the SAM software. Tusher, Tibshirani and Chu (2001) addressed the estimated FDR at a fixed threshold using resampling. Storey (2002, 2003) offered Bayesian interpretation to the FDR: by conditioning on at least one rejection being made, the positive FDR can be defined $\mathrm{pFDR} = E(V/R|R>0)$, with $\pi_0 = m_0/m$ as a prior in a mixture model, the pFDR has an immediate Bayesian interpretation. Furthermore, with an emphasis on the estimated FDR using a given threshold, Storey suggested that the results of the study be described by the levels at which each hypothesis will be rejected – the q-values.

With the use of estimate for π_0, Efron *et al.* (2001) gave an empirical Bayes interpretation of FDR. Given a statistic Z measuring the differential expression of a gene and having density f, define the local FDR as $\mathrm{fdr}(Z) = p_0 f_0(Z)/f(Z)$, where p_0 is the probability that a gene is unaffected and f_0 the density of Z for unaffected genes. Thus $\mathrm{fdr}(Z)$ is the *a posteriori* probability that a gene with score Z is unaffected. The FDR is the average of the local FDR over the rejection region.

Storey and Tibshirani (2003) presented FDR control by means of maximizing rejections while observing the estimated q-value, the smallest estimated FDR for which this P-value is still rejected. This suggestion completed the circle, in the sense that the latter is the same as working with the FDR adjusted P-values using an adaptive BH procedure with $\hat{\pi}_0(\lambda)$ as an estimator (see Section 2.3). Thus the two approaches are in fact similar.

3 The hierarchical FDR approach

3.1 *Trees of hypotheses*

The hierarchical FDR testing approach is a new way to address multiple testing of very large families of hypotheses. First, the set of m tested hypotheses is arranged in a tree structure of L levels, where $L(i)$ denotes the level of hypothesis H_i. With the exception of hypotheses on the first level – which have no parent hypotheses – each hypothesis H_i, on level $L(i) = 2, \ldots, L$ has a single-parent hypothesis, indexed as $\mathrm{Par}(i)$, on level $L(i) - 1$. The m hypotheses are divided into $T+1$ families: $\Im_0$ is the family of hypotheses at level 1, and for $t = 1, \ldots, T$ let $\Im_t = \{H_i : \mathrm{Par}(i) = t\}$. m_t and m_t^0 denote the total number and number of true null hypotheses in $\Im_t$.

3.2 Hierarchical testing

Hierarchical testing of a tree of hypotheses can be described easily by specifying two rules: hypotheses sharing the same parent are tested simultaneously as a family; a family of hypotheses is tested only if its parent hypothesis is rejected.

Testing begins with $\mathcal{H}_0$; at the next stage, each family of hypotheses corresponding to a second-stage discovery is tested; this testing process continues as long as additional rejections of parent hypotheses are made.

3.3 Types of FDR for hierarchical testing

The final component of the hierarchical FDR testing of a tree of hypotheses is the specification of the set of discoveries on which the inference is of importance to the investigator, and therefore for which the control of FDR is relevant. Once the set of interesting discoveries is determined, following Benjamini and Hochberg (1995), the FDR is 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). We present three types of FDRs suitable for different types of applications.

3.3.1 Full-tree FDR

Here interest lies in the entire set of discoveries, whether it is internal-node discovery or end node. This FDR is relevant when each discovery along the hierarchical testing path is important. It can also be used for high dimensional statistical modelling, in particular, applications where the model is specified hierarchically and it is important to extract all of the components of the model.

3.3.2 Level-restricted FDR

In this testing scheme, there is only interest in the set of discoveries on a specific level of the tree that is chosen in advance This FDR is relevant, for example, when the parent hypotheses serve only as screening devices leading to the inferences of importance.

3.3.3 Outer (end) node FDR

In the outer node scheme, the discoveries of interest are those on the outer nodes of the tree (end nodes) as reached by the hierarchical testing procedure. This FDR can be used for conducting multi-resolution searches, where higher resolution discoveries make the initial lower resolution discovery irrelevant. In fact, the application which motivated the development of this new testing approach was a genomic scan for genes affecting quantitative traits (Benjamini and Yekutieli, 2005).

3.4 FDR bounds for the hierarchical testing of trees of hypotheses

Yekutieli (2005) provides upper bounds for the three types of FDR under the assumptions that the test statistics are independently distributed and each family of hypotheses is tested by the BH procedure at level q. These results are reviewed in this section, and in Section 3.5 we consider, and present theoretical results for trees in which the families of hypotheses are tested by the subset-restricted FDR procedure.

The bounds for the FDR are sums over $t=0,\ldots,T$ of bounds for FDR_t, the expected proportion of false discoveries in $\Im_t$ out of the total number of discoveries made,

$$\mathrm{FDR}_t = E_{\mathbf{P}}\left\{ I(\Im_t \text{ is tested}) \cdot \frac{V_t}{R} \right\},$$

where V_t is the number of false discoveries in $\Im_t$ and R is the total number of FDR tree discoveries. To derive the bounds, FDR_t is expressed in terms of $R_t^{P=0}$ and $R^{P=0}$ – the number of hypotheses rejected in $\Im_t$ and the total number of FDR discoveries given that a specific true null hypotheses in $\Im_t$ is rejected. Employing this approach, Benjamini and Yekutieli (2001) prove that the FDR of the BH procedure applied to a single family of hypotheses is $m_0 q/m$. However, since the event conditioned on in each family of hypotheses is different, aggregating the bounds for FDR_t over the multiple families is not possible. To overcome this problem, Yekutieli (2005) substitutes $R_t^{P=0}$ and $R^{P=0}$ with R_t and R – the unconditional number of rejections in $\Im_t$ and finds a multiplicative factor δ^*, as small as possible satisfying,

$$\delta^* \geq E\frac{R_t^{P=0}}{R^{P=0}} \Big/ E\frac{R_t+1}{R+1}. \tag{1}$$

It is proven that if the P-values are independently distributed and their marginal cumulative density function (cdf) is star shaped in relation to 0, then setting $\delta^* = 1.44$ is sufficient to satisfy Expression (1). The condition, marginal cdf is star shaped in relation to 0, is somewhat a stronger condition than that of being stochastically smaller than $U[0,1]$. Yet, it holds if the distribution function of P_i is concave under the alternative, a condition satisfied under the often used condition of monotone likelihood ratio. Simulation results are also presented there, indicating that for most P-value configurations $\delta^* \approx 1$ is sufficient to satisfy (1). With inequality (1) Yekutieli (2005) proves that

$$\mathrm{FDR}_t \leq \delta^* \cdot q \cdot E_{\mathbf{P}}\left\{ I(\Im_t \text{ is tested}) \cdot \pi_t^0 \cdot \frac{R_t+1}{R+1} \right\}, \tag{2}$$

where π_t^0 is the proportion of true null hypotheses in $\Im_t$. The bound for FDR_t leads to a bound for the FDR:

$$\mathrm{FDR} = \sum_{t=0}^{T} \mathrm{FDR}_t \le \delta^* \cdot q \cdot \sum_{t=0}^{T} E_{\mathbf{P}} \left\{ I(\Im_t \text{ is tested}) \cdot \frac{(R_t+1)\pi_t^0}{R+1} \right\} \tag{3}$$

$$= \delta^* \cdot q \cdot E_{\mathbf{P}} \left\{ \sum_{t=0}^{T} I(\Im_t \text{ is tested}) \cdot \frac{(R_t+1)\pi_t^0}{R+1} \right\} \tag{4}$$

$$= \delta^* \cdot q \cdot E_{\mathbf{P}} \left\{ \frac{\text{no. of families tested} + \text{no. of discoveries}}{\text{no. of discoveries} + 1} \cdot \tilde{\pi}_0 \right\} \tag{5}$$

To derive (5), observe that summing $(R_t + 1)$ over all tested families yields the number of tested families plus R – the total number of discoveries, thus $\tilde{\pi}_0$ is a weighted mean of π_t^0.

The bound in (5) implies a universal bound for the full-tree FDR – $2 \cdot \delta^* \cdot q$ and the outer nodes FDR – $2 \cdot L \cdot \delta^* \cdot q$. But there is no universal bound for the level-restricted testing scheme, in some cases it may approach 1. Furthermore, Expression (5) implies that if the number of discoveries greatly exceeds the number of families tested then the three types of hierarchical FDR are approximately $q \cdot \delta^* \cdot \tilde{\pi}_0$. Thus, in spite of the lack of a universal bound in the case of level-restricted FDR, the implication of Expression (5) is straightforward: in a complex study with 10 research questions, each tested separately by applying the BH procedure at level 0.05, if the data only produce 10 discoveries then the combined FDR is approximately 0.10; the FDR is approximately 0.05 in large-scale studies with a few hundred discoveries.

3.5 The subset-selected BH procedure

The subset-selected BH is another method of alleviating the multiplicity problem. Each tested hypothesis H_i is associated with an indicator variable I_i, which determines whether the hypothesis is selected for testing or not; the BH procedure is applied to the subset $\{H_i : I_i = 1, i = 1, \ldots, m\}$ of the family.

We allow the vector **I** to be data dependent, but require it to be independent of the vector **P** of P-values corresponding that correspond to the tested hypotheses. This setup can also be applied to families of hypotheses – setting the same indicator, I_i, for all the members of a family of hypotheses means that the entire family of hypotheses is either selected or not based on a single test statistic (see Section 4.4); setting $I \equiv 1$, we are back at using the BH procedure on the entire family.

Note that I_i may be considered as a test of a parent hypothesis to H_i, and in that sense it is also a test of hierarchical tree of hypotheses. However, one distinction makes the situation much simpler here than in hierarchical testing discussed before: all hypotheses whose parents belong to the same 'grandparent' are tested jointly. This is the reason we can get the following result: If the components of **P** are positive regression dependent on the subset of P-values corresponding to true null hypotheses, then according to Benjamini and Yekutieli (2001), once condition-

ing on the values of **I**, the FDR of subset-selected BH procedure, conducted at level q, is less than or equal to $q \cdot m_0(I)/m(I)$, where $m(I)$ and $m_0(I)$ denote the number of selected hypotheses and number of selected true null hypotheses, respectively. Taking expectation over **I** yields the following obvious result: the subset-selected BH procedure controls the FDR at level $q \cdot E_I(m_0(I)/m(I))$.

3.6 Casting a complex study into the hierarchical testing framework

The hierarchical testing framework can be used to assess the global FDR of a complex study that consists of several research questions. We begin by considering the simple case of a complex study with multiple research directions where the family of null hypotheses corresponding to each of the research directions is separately tested by the BH procedure at level q. To evaluate the FDR of the entire study, we construct the following FDR tree: all the tested hypotheses are positioned on the second level of the tree, and each family of tested hypotheses is associated to a dummy parent P-value, which is set to 0, on the first level of the tree. Notice that the overall FDR for the entire study can be expressed as the level-2 restricted FDR. This implies that if the number of discoveries greatly exceeds the number of research directions, then the FDR of the entire study is approximately $q \cdot \delta^* \cdot \tilde{\pi}_0$.

We now consider the more general case of a complex study with multiple research directions, where each research direction is tested by a FDR tree and each family of null hypotheses is separately tested by the subset-restricted BH procedure. Let $\Im_{t,d}$ denote the hypotheses in family $t = 0, \ldots, T_d$ of research direction $d = 1, \ldots, D$; $V_{t,d}$ is the number of false discoveries made in $\Im_{t,d}$ by the FDR tree applied to research direction d. To express the FDR of the entire study as the FDR of a tree of hypotheses, we combine the FDR trees from each research direction by lifting all the hypotheses one level higher and associating the hypotheses in $\Im_{0,d}$ with a dummy parent P-value, $P_{0,d} = 0$, positioned on the first level of the tree. To accommodate the subset-restricted BH procedure we define an indicator variable, $I_{i,d}$, for each tested hypothesis $H_{i,d}$, such that for $i \in \Im_{t,d}$ the hypothesis $H_{i,d}$ is tested in the subset-restricted BH procedure of $\Im_{t,d}$ if and only if $I_{i,d} = 1$. Notice that the FDR corresponding to the set of discoveries of interest is unlike any of the FDRs described in Section 3.3: unlike the full-tree FDR and outer nodes FDR we are not interested in any of the level-1 discoveries; unlike the level-restricted FDR we may be interested in a discoveries on any level greater than 1. As a result of the independence of **I** and **P**, conditioning on **I**, the hierarchical test of the FDR tree for the entire study only involves repeated hierarchical applications of the BH procedure, hence we can apply the methods presented in Section 3.4 to bound the FDR.

If we denote the total number of discoveries in the study by R, then the FDR of the study is

$$\mathrm{FDR} = E_{\mathbf{I}} E_{\mathbf{P}} \left\{ \sum_{d=1}^{D} \sum_{t=0}^{T_d} I(\Im_{t,d} \text{ is tested}) \cdot \frac{V_{t,d}}{R} \right\} = E_{\mathbf{I}} \sum_{d=1}^{D} \sum_{t=0}^{T_d} \mathrm{FDR}_{t,d}. \tag{6}$$

Applying the bound in (2) for each value of **I** we get,

$$\mathrm{FDR}_{t,d} \le \delta^* \cdot q \cdot E_{\mathbf{P}} \left\{ I(\Im_{t,d} \text{ is tested}) \cdot \frac{m_0(I)}{m(I)} \pi_{t,d} \cdot \frac{R_{t,d}+1}{R+1} \right\}. \tag{7}$$

Incorporating Expression (7) into Expression (6) we can phrase the FDR for the entire study as,

$$\begin{aligned} \mathrm{FDR} &\le \delta^* \cdot q \cdot E_{\vec{I}} \left\{ \sum_{d=1}^{D} \sum_{t=0}^{T_d} \cdot E_{\vec{P}} \left[I(\Im_{t,d} \text{ is tested}) \cdot \frac{m^0_{t,d}(\vec{I})}{m_{t,d}(\vec{I})} \cdot \frac{R_{t,d}+1}{R+1} \right] \right\} \\ &= \delta^* \cdot q \cdot E_{\vec{I}} E_{\vec{P}} \left[\frac{\sum_{d=1}^{D} \sum_{t=0}^{T_d} I(\Im_{t,d} \text{ is tested}) \cdot m^0_{t,d}(\vec{I}) / m_{t,d}(\vec{I}) \cdot (R_{t,d}+1)}{R+1} \right] \end{aligned} \tag{8}$$

Expression (8) is very similar to the bound for the simple case. The main difference is that the total number of FDR tree discoveries is compared with the total number of families considered in the entire study, which can be much greater than the number of research directions. Still, if the number of discoveries greatly exceeds the number of families considered, the FDR of the entire study is approximately $q \cdot \delta^* \cdot \tilde{\pi}_0$.

4 The experiment

The experiment involved 10 adult males of 10 strains of inbred mice, which included the major inbred strains traditionally used for research, among them inbred strains derived from wild mice. Note that inbred strains are homogeneous in terms of genetic background, as all animals of the same strains are homozygous in all genes. The exploratory behaviour of the mice was tracked as part of a large high throughput phenotyping experiment (Kafkafi *et al.*, 2005), and the different strains exhibited a wide range of behaviour.

The gene expression part of the experiment involved harvesting tissue from the mice used in the behavioural assessment protocol 7–12 days following the behavioural experiment. Five tissue areas were dissected (prefrontal cortex, ventral striatum, temporal lobe, periaquaductal grey and cerebellum). The mice of same strain were divided into two groups, and tissue within group pooled, providing two biological replications per strain. The level of gene expression was measured versus a pooled control, swapping the dyes between the two replicates. In total, ~27,000 genes were analyzed.

We denote the expected expression level of gene in the brain region of mouse strain by μ_{gsr}; the average expression level over the brain by μ_{gs+}; the average expression level over the strains in one brain region by μ_{g+r}; and the average expression level over the strains and over the brain by μ_{g++}.

4.1 Question 1: strain ANOVA per gene

The first question addressed in this study is the discovery of genes with strain differences in their expression levels in the brain. This is an immediate extension of the discovery of two groups of differentially expressed genes. This question can be easily posed as a testing problem for the intersection null hypothesis for each gene:

$$H_S^0(g): \mu_{g1+} = \mu_{g2+} = \cdots = \mu_{g10+}.$$

These hypotheses can be tested by one-way ANOVA, using F-test, or Tukey's test. A discovery in this analysis is that there is some difference in brain expression levels of gene g between the strains. The multiplicity issue that stems from testing many genes is typically addressed by controlling the FDR over all the genes.

Unlike the usual application of ANOVA, microarray analysis may end with the rejection of the intersection null hypothesis. As an important example take the case where the analysis is performed as a dimension-reduction preliminary step, to be followed by clustering analysis or discriminant analysis. More commonly in ANOVA, the rejection of the intersection hypothesis is followed by an analysis of the pairwise comparisons. This question is discussed next.

In the current study, the emphasis is on differences in expression levels related to the studied strains. In another phase of this study, the parallel question regarding the testing across brain regions for differences in expression levels of genes, as reflected by $H_R^0(g): \mu_{g+1} = \mu_{g+2} = \cdots = \mu_{g+5}$ (in obvious notation), may be of similar interest. This certainly constitutes a different research question as well.

4.2 Question 2: strainwise comparisons in brain analysis per gene

A question of immediate interest to researchers is the identification of which specific pairs of strains differ in their expression level in a gene where difference exists. This is the classical pairwise comparison problem, where the strainwise comparison of strains s'' and s'', in terms of their brain expression level in gene g is expressed in the hypothesis

$$H_S^0(s', s''; g): \mu_{gs'1+} = \mu_{gs''+}.$$

Pairs of strains identified to differ in the expression level of a gene of importance may serve as the breeding source for a backcross experiment that tries to identify the genomic locations of other genes that caused the difference in their expression. Clearly discovering such pairs is of scientific importance.

Note that there are $10 * 9/2 = 45$ such pairwise comparisons per gene, and with the $\sim$27k genes analyzed we reach the multiplicity problem of about 1.2 million pairwise comparisons.

4.3 Question 3: strain $*$ region interactions in brain analysis per gene

While identifying strains that differ in their expression level of a certain gene over the brain is of obvious interest, the availability of data on five brain regions offers

the possibility to search the data for interesting findings in another direction as well: identifying specific strain $*$ region interactions in the expression level of a gene. The relevant hypothesis is

$$H^0_{\mathrm{S}\times\mathrm{R}}(s,r;g):\mu_{gsr}-\mu_{gs+}-\mu_{g+r}+\mu_{g++}=0$$

where finding a significant interaction means that the expression level of gene g, in strain s, is different than what would be expected under a typical additive pattern for that gene. This is clearly a data-mining operation, looking for interesting clues to pursue.

Two approaches can be taken. The first argues that such an interaction is interesting only after brain analysis showed strain differences in that gene. The second argues that any identification of interaction is of interest, but only after the interaction was shown to exist in the gene at large. Taking the first approach, any finding of strainwise difference in a gene is of interest, and then finding the interaction is another discovery. In the second approach genes are first screened for significant interaction, and then the specific interactions become the only interesting discoveries. This is actually parallel to the relationship between the testing of the strain factor in the ANOVA that is followed by the identification of specific pairwise comparisons. Both approaches are legitimate, but in the current study we take the first approach.

4.4 FDR-controlled testing in the experiment

4.4.1 The direct approach using the procedure in BH

In the direct approach we consider the entire set of tested hypotheses, even if it consists of several separate families, as a single family, and use the BH procedure on the P-values. FDR is controlled if the two conditions in BENJAMINI and YEKUTIELI (2001) are satisfied: (i) the test statistics are positive regression dependent on the subset of true null hypotheses; (ii) the distribution of each P-value corresponding to a true null hypothesis is either $U[0,1]$ or stochastically greater than $U[0,1]$. As a result of (ii), if the entire family of tests is not at our disposal and the missing is at random, we may put the unknown P-values at 1, and get valid (if somewhat conservative) inference.

As we noted in Section 1, the FDR criterion is quite accommodating when families of hypotheses are combined. It is important to note that the BH procedure has some asymptotic properties that make it very effective.

4.4.2 The direct approach for selected subset

In the direct approach the family of tested hypotheses is very large. In some cases, this multiplicity problem can be alleviated if there is a way to exclude a large proportion of the families of hypotheses in a way that little information is lost and the

FDR is still controlled when the BH procedure is applied to the hypotheses in the remaining families of hypotheses. One way to achieve both these goals is to choose families of the null hypotheses according to prior knowledge.

Extra care must be taken when the choice of families is data dependent. Notice for example that in research question 3, a strain $*$ region interaction in gene expression levels is not possible if there is no strain effect in the expression level. We therefore test each gene for a significant strain effect, and then apply the direct approach to the hypotheses belonging to families of hypotheses with significant strain effects. The test statistics we use to test the strain effect is independent of the interaction term test statistics. This means that any selection criterion applied to strain effect test statistics does not violate the conditions needed for FDR control. In Section 5 the selection criterion used is the BH procedure at level 0.05. Note that this approach cannot be used for research question 2. The dependency between the strain effect test statistics and the individual strain pairwise test statistics rules out the possibility of using the strain effects to choose the set of families used in the direct approach.

Note also that this is in sharp contrast to the case where the 'same hypotheses are tested again' using independent data in both stages. Under such circumstances, if FDR is first controlled at level q_1 to select a subset of hypotheses, and then controlled at q_2 on the selected subset, the joint procedure enjoys an FDR level of $q_1 * q_2$ (Reiner *et al.*, 2003). The difference lies in the fact that using that method, if a hypothesis is erroneously selected in stage II it was also at error when selected at stage I, the probability for an error being smaller. This is not the case in the general hierarchical selection outlined above, where the hypotheses are different from one stage to the next.

5 Analysis and results

5.1 Strain by brain ANOVA per gene: repeated median removal of chip layout effects

Our approach is to use an F-test at each gene to test the hypothesis

$$H_S^0(g): \mu_{g1+} = \mu_{g2+} = \cdots = \mu_{g10+}.$$

No pooling across genes was used (with no fudge factor as in SAM), so the distribution of the test statistics under the null hypothesis should be F-distribution with 9 and 40 degrees of freedom. Figure 1a presents the results of close to 25,600 P-values using their quantile plot, where the ordered P-values are plotted versus their rank. This is equivalent to a *P–P plot* under the assumed F-distribution. We expect to see in the upper right part of the plot a straight line, corresponding to genes where the hypothesis of no strain differences is true. Instead the line curves concavely, indicating that the P-values are bigger than they should be.

We therefore revisited the preprocessing stage and studied the possible effects of the physical layout of the chips on individual expression levels. The structure of an array is schematically described in Figure 2. Effects of block row and column and

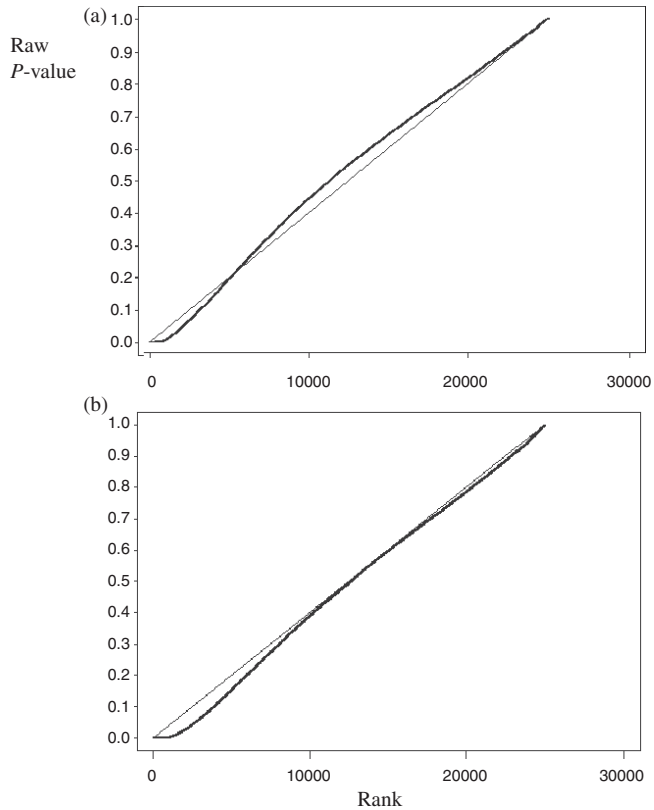

Fig. 1. Significance of strain effects before and after removal of spatial effects (a) Before spatial effects removal – raw P-values distribution is clearly not $U(0, 1)$ even for ranks from 1000 and higher, as indicated by the deviation of the ranked raw P-values from a straight line; 730 genes are identified as differentially expressed. (b) After spatial effects removal – raw P-values distribution is closer to $U(0, 1)$, deviating substantially from a straight line only for low ranks. Now nearly 1000 genes are identified. Thus effect removal improves the sensitivity of the testing procedure.

within-block row and column were estimated for log-transformed Cy3 signal and Cy5 signal. A four-way median polish model, which is an extension of the two-way median polish (TUKEY, 1977; EMERSON and WONG, 1985), was applied for the purpose of identifying block and slide spatial effects. This algorithm iteratively subtracts the median signal for each category of each effect (block row, block column, within-block row, within-block column), until the residuals no longer change. The 'polished' signals are then used for data analysis. It is essential that this resistant procedure be used, and not the usual ANOVA that is based on mean values, so that real differences in expression levels will not be smoothed away.

Figure 1b presents the results of the same ANOVA tests as those displayed in Figure 1a, except after the four-way median polishing. The improvement is noticeable, with the P-values on the right side lying closer to a straight line than before. The

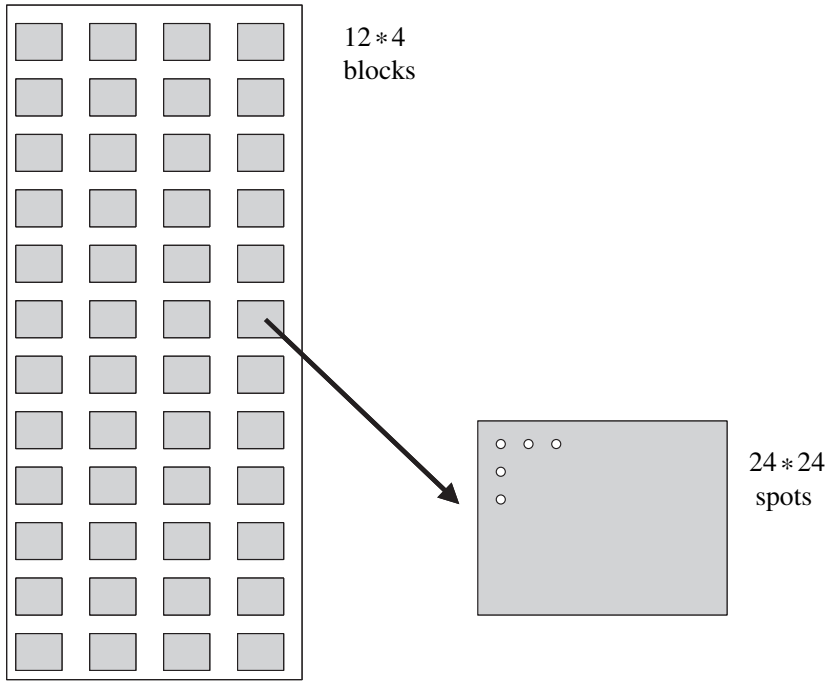

Fig. 2. A schematic structure of a microarray: the printing structure of the array comprises the effects of block row, block column, within block row, and within block column.

large P-values are still somewhat larger than expected under the Uniform. This may stem from the lack of normality of observations where the F-test is known to be conservative under such circumstances. Fortunately, this fact does not harm the validity of the conclusions as the BH procedure still guarantees FDR control when the P-values under the null are stochastically larger than Uniform.

5.2 Strain ANOVA per genes

A two-way ANOVA was applied to identify genes that could distinguish between strains, while adjusting for the effect of brain region. Each gene was analyzed separately, resulting in a P-value for the main factor of strain for the gene.

Among the 25,600 P-values, 2676 were lower than 0.05 (10.5%). Adjusting the P-values for multiplicity using the linear step-up procedure in BH, 957 genes were found statistically significant at 0.05 FDR level. The procedure simply amounts to passing a line through the points $(1, q/m)$ and (m, q), and looking for the largest value under this line in Figure 2b. The FDR-adjusted P-values using the BH procedure for the most significant genes are presented in Figure 3. We note that applying the model to non-polished logs difference produced around 700 adjusted P-values lower than 0.05, being a clear evidence that removing the spatial effects can substantially improve the sensitivity of the analysis.

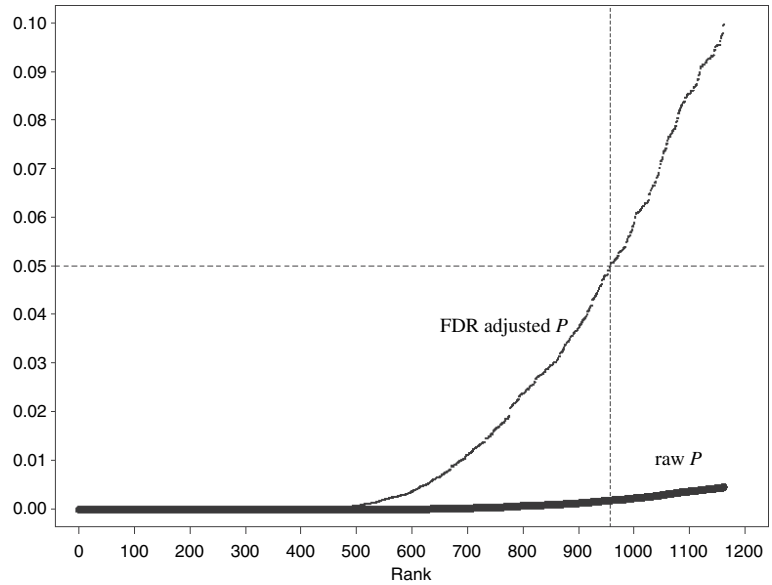

Fig. 3. Raw and FDR adjusted *P*-values – strain effect. Zooming in for FDR less or equal to 0.1, the multiplicity effect on the significance threshold is visible. After FDR adjustment at the 0.05 level, nearly 1000 genes are found to have a differential expression across strains.

5.3 Identifying specific strainwise differences in the brain analysis per gene

Comparing strains by pairs through the two-way ANOVA model described above involves approximately $25,600 * 45 = 1.15$ million comparisons. The hypotheses have a natural and intuitive tree structure where the parent of all pairwise hypotheses in a gene is the intersection hypothesis for that gene residing at level 1. It makes sense to follow hierarchical testing at this stage, but the test statistics will clearly not be independent. As we do not know how the FDR is inflated under such circumstances, we turn to direct testing of all 1.15 million hypotheses, relying on the many studies indicating that the BH procedure controls the FDR under the pairwise correlation structure (YEKUTIELI, 2002a,b; KESSELMAN, CRIBBIE and HOLLAND, 1999; BENJAMINI, HOCBBERG and KLING, 1999). Controlling the FDR at the 0.05 level yields 7771 significant pairs. Figure 4 shows the frequency of significant pairwise comparisons for each pair of strains. Although each of the strains was found to have at least 10 genes at which it differs from any other strain, the pattern of significant pairs indicated that two of the strains, CAST/Ei and SPRET/Ei, are different in a substantial way from all other strains as well as from each other (as reflected by hundreds of gene expression level differences).

Interestingly, the SPRET/Ei belongs to the Mus spretus species, the CAST genotype belongs to the subspecies *Mus musculus castaneus* while the eight other inbred strains belong to the subspecies *Mus musculus musculus* (BONHOMME and GUENET, 1996). It is thus interesting to find that the divergent speciation of SPRET and

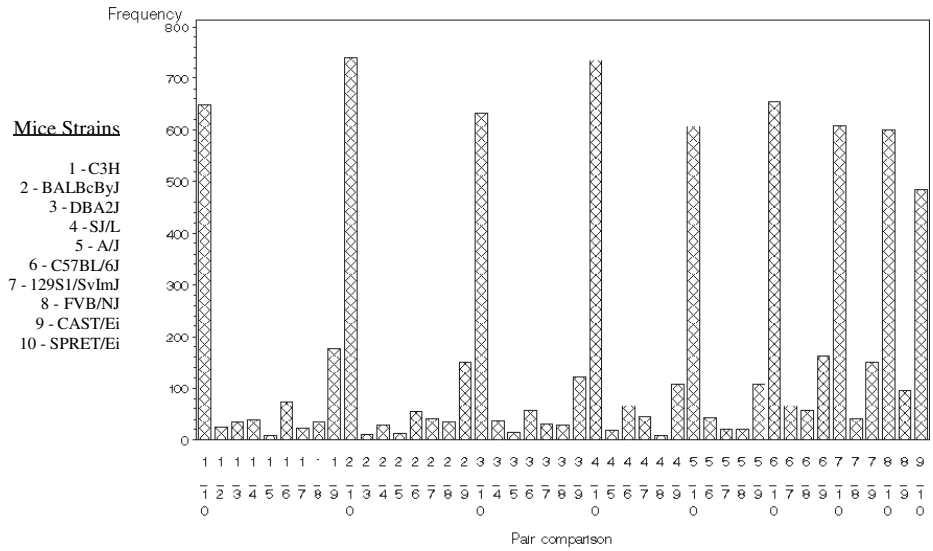

Fig. 4. Frequency of genes with significant pairwise difference. The number of genes found to have a different expression between strains is visibly higher when comparing to strain 10 (SPRET/Ei), implying a large-scale change in genetic activity in the brain for this strain.

CAST is reflected in the large gene-expression distinction consistently recognized in the pair-wise comparisons in the brain. This raises the possibility that differences in behaviour that are known to exist between the strains will be associated with differences in levels of gene expression in the brain.

5.4 Brain region interactions with strain per gene

Standardized residuals from the above analysis of variance model were used to calculate a test statistic for each combination of strain and brain region, for each gene. Testing those interactions only in those 957 genes that were identified at the first stage of the analysis still involves the testing of about 50,000 interactions. We can restrict our testing to these interactions only by the subset selection method, because the tests of the interactions are independent of the tests of strain effects in ANOVA. Only 13 significant interactions of strain and brain region were found this way at FDR level 0.05.

Turning to hierarchical testing, it is natural to arrange all hypotheses of strain effects for the genes at level 1, with the family of interactions per gene being its progeny at level 2. If any such discovery amounts to a meaningful discovery, it calls for a full-tree testing scheme. Thus, if each one of the first-level analysis is performed at the FDR level of 0.017, and at each rejected gene separately the testing of the 50 relevant hypotheses is performed at FDR level of 0.017, the overall FDR level is bounded by $2 * 1.44 * 0.017 - 0.05$.

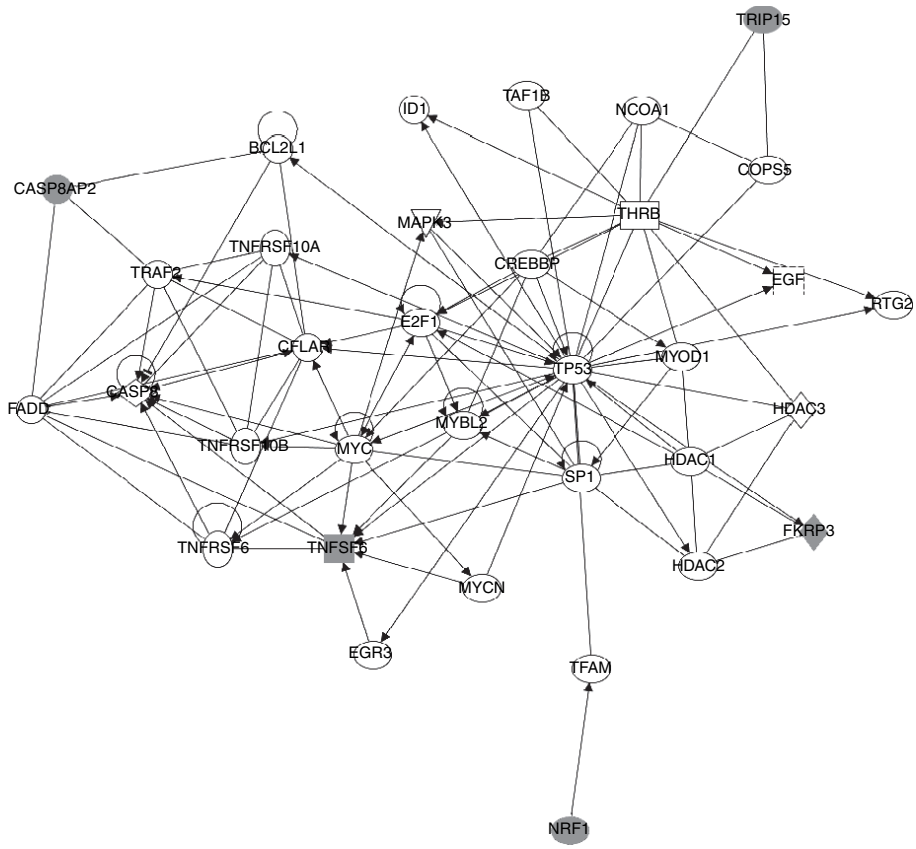

Fig. 5. Genes mapped into a network linked with seizure activity. Five genes in the seizure activity network that exhibit significant interactions in the CAST strain and various brain regions are marked in dark.

Some 758 genes and 76 interactions are found significant using the hierarchical testing approach. The interaction component of the hierarchical analysis was sufficiently powerful to detect the brain region by strain interactions. Of these 76 interactions, two-thirds (47) were associated with the CAST genotype indicating a non-additive pattern against the model suggested by the parent species *Mus musculus musculus*. A possible explanation for this disproportionate representation, which reflects the deviation from the additive pattern formed by the eight *Mus musculus musculus* genotypes, lies in either a *cis*- or *trans*-acting regulatory polymorphism whose consequence affects a gene network that is tissue-specific (Cowles *et al*., 2002). An evidence for such a functional explanation of the brain region by strain interactions can be found in the mapping of these genes into networks (Figure 5). For example, five of the 47 genes (Casp8AP2, Tnfsf6, Trip15, Fkbp3, Nrf1) exhibiting significant interactions associated with CAST are part of a gene network that has

been linked with seizure activity. Interestingly, CAST is particularly resistant to beta-carboline-3-carboxylate induced seizures (Le Roy *et al.*, 1998). A possible explanation to the fact that the CAST is involved in such interactions more than other strains lie in the increased incidence in which such wild-derived genotypes naturally carry Robertsonian chromosomes, but we shall not dwell into this topic here.

5.5 Assessing the FDR level of the entire study

In this paper, we presented two methods for addressing the multiplicity of interactions in research question 3, which lead to two FDR trees into which the complex study is imbedded. We describe the structure of the two FDR trees and evaluate the realized value of the bound FDR for the entire study – Expression (8). As the statistical analysis included a very large number of hypotheses, the bound for the FDR is probably close to its realized value.

In the first framework, the complex study comprises of three research directions. In the first research direction, strain ANOVA per gene, the BH procedure was applied at level 0.05 on a single family of 25,600 hypotheses, and 957 genes were found statistically significant. The strainwise differences research direction, a single family of 1.15 million hypotheses, was tested by the BH procedure at level 0.05, yielding 7771 discoveries. In the third research direction, the initial family consisted of about 1.325 million interaction null hypotheses; subset selection of 957 genes with significant strain differences reduced this number to approximately $957 * 50$ interaction null hypotheses, and applying the BH procedure at level 0.05 on the selected subset of hypotheses yielded 13 discoveries. In (8) a bound for the FDR of the entire complex study is given in the form of the $\delta^* q$ times the expected ratio of a weighted mean of the proportion of true null hypotheses multiplied by the number of discoveries plus the number of families tested divided by the number of discoveries plus 1. In this case the realized value of (8) is less than $\delta^* q$ multiplied by $(957+7771+13+3)/(957+7771+13+1) \approx 1$, thus the FDR for the entire study is at most 0.072.

In the second framework, the complex study comprises of only two research directions. The first research direction is a FDR tree in which in the first level included the 25,600 strain ANOVA per gene hypotheses, and the families in the second level of the tree include the 50 strain $*$ brain interactions for each gene exhibiting a significant strain effect. Applying the BH procedure at level $q = 0.017$ in each family, yielded 758 level-1 discoveries; the 758 families tested in level-2 yielded 76 additional discoveries. The second research direction regarding strainwise differences involves the same analysis used in the first framework. In the first research direction $q < 0.05/(2 \cdot 1.44)$ ensures that the universal bound for the full-tree FDR is less than 0.05. For evaluating the total FDR notice that the total number of discoveries was $8605 = 758 + 76 + 7771$ and the number of families tested is $759 = 758 + 1$, thus the realized value of (8) is now at most $\delta^* q$ multiplied by $(8605+759)/(8605+1) \approx 1.09$, which is less than 0.078.

6 Discussion

The purpose of this paper was to demonstrate how our understanding the way by which the FDR progresses when testing hypotheses hierarchically, can enable us to test complex and extremely large families of hypotheses. Generally, we used the theory developed for independent test statistics in the dependent situations faced in microarray analysis. This is based on our experience in using the BH procedure in microarray analysis (Reiner *et al.*, 2003), where under the type of dependency encountered in practice the FDR is controlled. Further verification of the validity of the procedure is being planned. Moreover, experience at quantitative trait locus (QTL) analysis using hierarchical testing revealed that the upper bounds offered by the current theory are somewhat high. Further theoretical and simulation work may allow us to lower them. The multiple-testing FDR-controlling procedures discussed in this paper, were implemented using the MULTTEST procedure of SAS software. The input to this procedure is the set of *P*-values derived from an ANOVA model, or from tests of correlations. The flexibility of this approach lies in the fact other procedures, such as those based on mixed model, repeated measurements and the like, can all be easily implemented. Resampling-based procedures have also been implemented using R programming language (and Splus), so the above approach should be easy to implement for other settings.

The biological implications of the results obtained here will be detailed elsewhere, but some conclusions of importance were discussed here. In particular, it was found that even though we started with two different research questions, they turn out to be related. Moreover, the results reported here are the first stage of an even more complex study, where interest further lies in the relationship between measured behavioural aspects of the mouse strains and the gene expression levels in the brain of those strains, in those genes where strain differences are evident. In particular, we have interest in correlating gene expression levels in the brain with the 17 measures of behaviour, which quantify exploratory behaviour giving rise to a multiplicity problem of about 2.2 million tests of hypotheses. While the approach taken here is applicable there as well, the solutions developed will be discussed in a later study.

Acknowledgements

This research is supported by a United States National Institute of Health grant, and by a FIRST grant of the Israeli Academy of Sciences.

References

Abromovich, F., Y. Benjamini, D. L. Donoho and I. M. Johnstone (1998), *The amalgamation challenge to signal de-noising*, Research paper of the Department of Statistics and OR, Tel Aviv University, RP-SOR-98-03.

Benjamini, Y. and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society* B **57**, 289–300.

Benjamini, Y. and Y. Hochberg (2000), On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics* **25**, 60–83.

Benjamini, Y. and D. Yekutieli (2001), The control of the false discovery rate under dependency, *Annals of Statistics* **29**, 1165–1188.

Benjamini, Y. and D. Yekutieli (2005), Quantitative trait loci analysis using the false discovery rate, *Genetics* **171**, 783–790.

Benjamini, Y., Y. Hochberg and Y. Kling (1999), *False discovery rate controlling procedures for pairwise comparisons*, Technical Report, Department of Statistics and Operations Research, University of Tel-Aviv.

Benjamini, Y., A. Krieger and D. Yekutieli (2001), *Two-staged linear step-up FDR controlling procedure* (Revised, 2003). Department of Statistics and Operation Research, Tel-Aviv University, and Department of Statistics, Wharton School, University of Pennsylvania, Technical Report.

Bonhomme, F. and J. L. Guenet (1996), The laboratory mouse and its wild relatives, in: M. F. Lyon, S. Rastan and S. D. M. Brown (eds.), *Genetic variants and strains of the laboratory mouse*, Oxford University Press, Oxford, 1577–1596.

Cowles C. R., J. N. Hirschhorn, D. Altshuler and E. S. Lander (2002), Detection of regulatory variation in mouse genes, *Nature Genetics* **32**, 432–437.

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

Emerson, J. D. and G. Y. Wong (1985), Resistant nonadditive fits for two-way tables, in: D. C. Hoaglin, D. Mosteller and J. W. Tukey (eds.), *Exploring data tables, trends and shapes*, Wiley, New York, 67–124.

Hochberg, Y. and Y. Benjamini (1990), More powerful procedures for multiple significance testing, *Statistics in Medicine* **9**, 811–818.

Genovese, C. and L. Wasserman (2002), Operating characteristics and extensions of the false discovery rate procedure, *Journal of the Royal Statistical Society* B **64**, 499–517.

Jiang, H. (2004), *A two step procedure for multiple pairwise comparisons in microarray experiments*, PhD thesis, Department of Statistics, Purdue University.

Kafkafi, N., Y. Benjamini, A. Sakov, G. I. Elmer and I. Golani (2005), Genotype-environment interactions in mouse behavior: a way out of the problem, *Proceedings of the National Academy of Sciences USA* **102**, 4619–4624.

Kesselman, H. J., R. Cribbie and B. Holland (1999), The pairwise multiple comparison multiplicity problem: an alternative approach to familywise and comparisonwise type I error control, *Psychological Methods* **4**, 58–69.

Lee, H., J. C. M. Dekkers, M. Soller, M. Malek, R. I. Fernando and M. F. Rothschild (2002), Application of the false discovery rate to quantitative trait loci interval mapping with multiple traits, *Genetics* **161**, 905–914.

Le Roy, I., P. L. Roubertoux, L. Jamot, F. Maarouf, S. Tordjman, S. Mortaud, C. Blanchard, B. Martin, P.-V. Guillot and V. Duquenne (1998), Neuronal and behavioural differences between mus musculus domesticus (C57BL/6JBy) and *Mus musculus castaneus* (CAST/Ei), *Behavioral Brain Research* **95**, 135–142.

Reiner, A., D. Yekutieli and Y. Benjamini (2003), Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* **19**, 368–75.

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

Storey, J. D. (2002), A direct approach to false discovery rates, *Journal of the Royal Statistical Society Series* B **64**, 479–498.

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

Storey, J. D. and R. Tibshirani (2003), SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays, in: G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger (eds.), *The analysis of gene expression data: methods and software*, Springer, New York, 272–290.

Storey J. D., J. E. Taylor and D. Siegmund (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.

Tukey, J. W. (1977), *Exploratory data analysis*, Addison-Wesley, Reading, MA.

Turkheimer, F. E., C. B. Smith, and K. Schmidt (2001), Estimation of the "true" null hypotheses in multivariate analysis of neuroimaging data. *Neuroimage* **13**, 920–930.

Tusher, V., R. Tibshirani and G. Chu (2001), Significance analysis of microarrays applied to transcriptional responses to ionizing radiation, *Proceedings of the National Academy of Science* **98**, 5116–5121.

Westfall, P. H. and S. S. Young (1993), *Resampling based multiple testing*, Wiley, New York.

Yekutieli, D. (2002a), *Theoretical results needed for applying the false discovery rate in statistical problems*, PhD Thesis, Department of Statistics and Operations Research, Tel-Aviv University.

Yekutieli, D. (2002b), *Elkund-Seeger-Simes is conservative for testing all pairwise comparisons*, Technical Report, Department of Statistics and Operations Research, University of Tel-Aviv.

Received: December 2004. Revised: December 2005.