BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 17 2007, pages 2239–2246
doi:10.1093/bioinformatics/btm300

Gene expression

Associating quantitative behavioral traits with gene expression in the brain: searching for diamonds in the hay

Anat Reiner-Benaim[1], Daniel Yekutieli[2], Noah E. Letwin[3], Gregory I. Elmer[4], Norman H. Lee[3], Neri Kafkafi[4] and Yoav Benjamini[2,*]

[1]Department of Statistics and Operation Research, Tel-Aviv University and Stanford University, Stanford, [2]Department of Statistics and Operation Research, Tel-Aviv University, [3]Department of Functional Genomics, The Institute for Genomic Research, Maryland and The George Washington University Medical Center, Washington D.C. and [4]Maryland Psychiatric Research Center, University of Maryland, Baltimore

Received on January 1, 2007; revised on May 10, 2007; accepted on May 28, 2007

Advance Access publication September 7, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Gene expression and phenotypic functionality can best be associated when they are measured quantitatively within the same experiment. The analysis of such a complex experiment is presented, searching for associations between measures of exploratory behavior in mice and gene expression in brain regions. The analysis of such experiments raises several methodological problems. First and foremost, the size of the pool of potential discoveries being screened is enormous yet only few biologically relevant findings are expected, making the problem of multiple testing especially severe. We present solutions based on screening by testing related hypotheses, then testing the hypotheses of interest. In one variant the subset is selected directly, in the other one a tree of hypotheses is tested hierarchical; both variants control the False Discovery Rate (FDR). Other problems in such experiments are in the fact that the level of data aggregation may be different for the quantitative traits (one per animal) and gene expression measurements (pooled across animals); in that the association may not be linear; and in the resolution of interest only few replications exist. We offer solutions to these problems as well. The hierarchical FDR testing strategies presented here can serve beyond the structure of our motivating example study to any complex microarray study.

Contact: ybenja@post.tau.ac.il

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

1 INTRODUCTION

The initial and simplest statistical question in functional genomics is the identification of genes differentially expressed between two classes that differ in functionality. The two classes may involve healthy versus sick cells, one region in the brain versus another, hearing versus deaf animals, or a high activity strain of mice versus a low one. The development of microarray technology has opened a new era for functional genomics, in that tens of thousands of such experiments can be conducted concurrently (for review see Lee and Saeed, 2006). At the same time it has generated a concern regarding the effect of conducting multiple statistical tests, a concern that may cast doubt on the validity of the statistical discoveries (e.g. Dudoit *et al*., 2002, Efron *et al*., 2001, Reiner *et al*., 2003, Storey and Tibshirani, 2003).

When more than two classes of function exist, that is the functionality is classified into a few categories, the resolution offered by the experiment increases yet the multiplicity problem is compounded. Here, the effort may initially involve the identification of genes that are not similarly expressed across all functional categories, leading towards the use of one-way analysis of variance (ANOVA). Next, one may be interested in comparisons among pairs of categories in order to specifically identify where the difference lies (Smyth *et al*., 2005, Yekutieli *et al*., 2006). More generally, the interest is in linear contrasts that compare the average expression in one subset of levels with the average in another, as when comparing the expression pattern of genes in active mouse strains versus low activity strains (Pavlidis and Noble, 2001).

A better opportunity to study the connection between the gene expression level and functional outcome is when the information on individuals can be measured quantitatively. For example, measuring some quantitative traits that reflect the activity of the animal, such as the distance traveled, are preferred for associating expression levels with activity, than merely classifying activity to high and low. The approach is gaining popularity for example in correlating gene expression and/or expression QTLs (eQTLs) with a particular phenotypic trait (Chesler *et al*., 2005, Kerns *et al*., 2005, Letwin *et al*., 2006), although only trait averages on other animals, taken from data bases, are used in the above studies.

1.1 Multiple testing

As the above-mentioned studies exemplify, the number of hypotheses being tested in microarray studies has increased dramatically. Therefore, the concern about controlling the increased type I errors (false discoveries) resulting from

*To whom correspondence should be addressed.

multiple testing, while not loosing too much power, becomes most important. It has become quite common to confront this concern about multiplicity by using the False Discovery Rate (FDR) methodology in one form or another (e.g. Efron *et al*., 2001, Reiner *et al*., 2003, Storey and Tibshirani, 2003). The linear step-up procedure offered by Benjamini and Hochberg (1995, hereafter BH) was shown to control the FDR under independence, and under certain type of positive dependence (Benjamini and Yekutieli, 2001), as demonstrated on gene expression data by Reiner et al (2003). More recently, Reiner (2007) proved that the BH procedure controls the FDR for two sided normally distributed tests with any correlation structure at the desired level q. Many adaptive procedures offer a way to estimate the proportion of true hypotheses $\pi_0 = m_0/m$, and then use the BH procedure at level q/π_0 thus gaining power when m_0/m is substantially smaller than 1 (e.g. Benjamini and Hochberg, 2000, Jiang, 2004, Storey 2002, Storey and Tibshirani, 2003, Benjamini *et al*., 2006). The results of Reiner (2007) do not carry over to this dependent case (as the FDR is higher than the bound when independent) and the sensitivity of the adaptive FDR procedures to the dependency remains an issue.

Even within the FDR framework, however, the more complex studies, such as the ones described above, pose new multiplicity challenges, and the problems involved have not been adequately addressed (see comment in Letwin *et al*., 2006). In this article we aim to use recent advances in FDR methodology (Yekutieli, 2007, Yekutieli *et al*., 2006) in order to make well-founded inference on gene-level functionality. The underlying idea is that the problem of testing a potentially very large family of hypotheses can be alleviated by initial screening, that is excluding a large proportion of the hypotheses where findings are not likely to occur. The remaining ones are tested using a multiple testing procedure either (a) jointly as a single family, or (b) separately within each emerging subfamily. In the latter case the hypotheses are organized in a tree structure and tested hierarchically. We take the second strategy a step further, by showing that the demands raised by existing theoretical bounds can be lowered. At the same time we warn against the indiscriminate use of test statistics that are dependent across stages of the analysis.

Rather than continuing in generality, we describe the experiment that motivated the methodological work presented here, and discuss the above issues in this context.

1.2 The experiment

The goal of the current study is to find associations between open field exploratory behavior of mice and the level of gene expression in different brain regions, some of which will then be followed by biological verification. On the behavioral side, the purpose was to use the highly informative and quantitative characterization of open-field exploratory behavior, as encompassed by the recently developed SEE (Software based strategy for Exploring Exploration, Benjamini *et al*., 2001, Drai *et al*., 2000, Kafkafi *et al*., 2005). This strategy attempts to capture highly structured behavioral patterns using ethologically relevant measures (so-called tests, behavioral traits or behavioral endpoints). For that purpose, the path of the animal in a large open field is automatically tracked and digitized for 30 min. It is subsequently smoothed robustly and statistically segmented into discrete behavioral units of stops (lingering episodes) and progression segments. The quantitative properties of the segments, such as their length, duration, maximal speed and spatial spread constitute a large number of the traits studied, that otherwise include more traditional traits such as total activity and time spent at center. The traits studied at three laboratories exhibited high broad sense heritability with significant strain differences (Kafkafi *et al*., 2005). In the current study, we used the tracking data of 10 males from each of 8 traditionally used inbred mouse strains measured in the Maryland Psychiatric Research Center laboratory of the Kafkafi *et al*. (2005) study.

The gene expression part of the experiment involved harvesting tissue from the same mice 7–12 days following the behavioral assessment. Five brain regions were dissected (prefrontal cortex; ventral striatum; temporal lobe; periaqua-ductal gray and cerebellum). Each region from mice of the same strain was pooled into two groups in order to have sufficient quantities of RNA for measuring expression levels, thereby providing two biological replications per strain per region, for some ~27 000 genes (such pooling strategy is often employed, see Lee and Saeed, 2006). For experimental details related to the microarray hybridizations and data pre-processing see Letwin *et al*. (2006).

1.3 The problem of testing following screening

The search for gene-behavior associations involves testing over 17 behavioral traits, 5 regions and ~27 000 genes—representing more than 2 millions potential hypotheses. We approach this mega-family of hypotheses by first screening for potential families of hypotheses for which the correlation is non-zero. For that purpose, we make use of general evidence from the previous analysis in other labs that the 17 behavioral traits vary between the strains. Therefore, if the expression of a particular gene in a particular brain region is the same for all strains, then there is no reason to believe it can be correlated with behavior. Hence, an informative screening question is whether genes are differently expressed between the strains in each region of the brain (a question of scientific value by itself). This strategy is intuitive, and has been practiced in recent microarray analyses (Letwin *et al*., 2006, Pavlidis, 2003) and implemented in software (LIMMA at Bioconductor, see Smyth, 2005).

Still, when (multiple) testing is preceded by screening the same data, the distribution of the *P*-values corresponding to a true null hypothesis may no longer remain Uniform (0,1) or stochastically larger, as needed for valid testing. For example, in a small simulation study we investigated all pairwise differences between gene expressions of 8 strains for 10 000 genes. When the procedure in Benjamini and Hochberg (1995), at level 0.05, is used to test jointly all 280 000 *P*-values from the pairwise t-statistics, the FDR was 0.036; When the same procedure was applied first to the 10 000 one-way ANOVA F-test *P*-values in order to screen genes, and then applied to simultaneously test all screened hypotheses—the FDR was 0.272 (all simulation SE < 0.006).

The phenomenon is well known at the level of a single ANOVA, where it is termed *post-hoc* analysis. Such questions have not received much theoretical or practical consideration when the *post hoc* analysis is in the context of screening and with less obvious connection between the screening test and the follow-up one.

1.4 Additional considerations

The joint analysis of behavior and expression in a single experiment raises additional concerns that require methodological attention. First, screening with ANOVA at each combination of gene and brain region, provides less than ideal number of biological repetitions at each combination, because of tissue pooling. Second, how should the association between the expression level and the behavioral trait be measured? Third, the implications of measuring behavior at the individual animal level while measuring expression level from pooled tissues from the same brain region because of the scarcity of biological material, may be an exaggerated correlation, a phenomenon sometimes termed 'ecological correlation'. These problems are known, even though not always recognized, and have recently come to light in a number of gene-phenotype association studies toxicogenomics and environmental genomics. The first two problems have known answers to be merely highlighted; the solution for the last problem has to be tailored for the situation at hand.

2 METHODS

Let us first introduce some notations for the parameters of the model describing the above experiment, and for the measured values. The behavioral endpoint b of mouse m from strain s is denoted by B_{bsm}, and its expectation given the strain is β_{bs}. M_{grsm} is the level of the expression of gene g (averaged over dye-swap hybridizations), in mouse m, in strain s, in region r and its expectation given the strain is μ_{grs}. We denote the average expression level across all the strains in brain region r by μ_{gr+}, and the average expression level across all strains and brain regions by μ_{g++}. With these notations the hypothesis $H_0^{\mathrm{ASSOC}}(b,g,r)$ states B_{bsm} and M_{grsm} are not associated; the hypothesis for each gene g and each brain-region used for screening of no strains difference is $H_0^{\mathrm{STR}}(g,r)$: $\mu_{gr1}=\mu_{gr2}=\ldots=\mu_{gr8}$.

2.1 Testing following screening

2.1.1 Selected subset testing The procedure involves screening for potentially successful hypotheses while controlling the FDR at level q_1 in the first stage, and then testing the set of identified hypotheses as a single family while controlling the FDR at level q_2 in the second stage. Independence between the tests in the first stage and those in the second stage is crucial. One important example of data dependent choice where the condition is satisfied is the case where the same hypotheses are tested again using independent data in both stages. This approach can be much more powerful than testing indiscriminately the original family of hypotheses (Benjamini and Yekutieli, 2005, Zehetmayer *et al.*, 2005, Reiner *et al.*, 2003).

In the case we address here: (1) The same data is used in both stages; (2) Different hypotheses are tested at the screening stage and at the second stage. However, if we still can assure the test statistics for the true null hypotheses at the second stage are independent of those in the first stage—the procedure will obviously control the FDR at level q_2.

In our case, we first test $H_0^{\mathrm{STR}}(g,r)$ using ANOVA for strain differences in expression levels in a brain region and then test for association using the Spearman's test, as displayed in Figure 1. Under the null hypothesis the Spearman's test is not dependent on whether the distribution of the expression data is more dispersed between the strains or less so, which is the hypothesis tested at the first stage. Thus using the procedure in BH at level q_2 at the second level controls the FDR at q_2. In fact one may even use an adaptive procedure at stage two and similarly control the FDR at that level.

2.1.2 Hierarchical FDR testing In the subset selection method discussed above, all hypotheses regarding the correlations that passed the first screening are tested simultaneously as a single family. However, there is more structure than that to our problem. Every rejected hypothesis at the screening stage suggests that for this pair of gene and brain region the expression level may be associated with some or all of the 17 behavioral traits. Can we test each such subfamily of 17 hypotheses by itself while controlling the FDR of the entire process of inference? The question can be answered in the positive but elaboration is needed. Our problem can be imbedded in the general scheme for hierarchical testing of trees of hypotheses described in Yekutieli *et al.* (2006). In our case the tree has two levels, as displayed in Figure 1. In the first level, the family of hypotheses tested are the hypotheses regarding the question of strain differences within brain regions for each gene $\{H_0^{\mathrm{STR}}(g,r):g=1,2,\ldots 27000;\ r=1,2,\ldots,5\}$; each hypothesis at the first level is parent to the subfamily of hypotheses regarding the association between the expression of the gene within the brain region and the 17 behavioral traits, namely $\{H_0^{\mathrm{ASSOC}}(b,g,r):b=1,2,\ldots,17\}$. These hypotheses are referred to as Level-2 hypotheses. Now, testing begins by applying the BH at level q procedure to the family of hypotheses at level 1; Level-1 discoveries only are followed to Level 2, where the BH procedure is separately applied to each subfamily of association hypotheses, again at level q.

Yekutieli (2007) proved a bound for the FDR of the hierarchical testing procedure under the assumption that the P-values corresponding to the hypotheses are independently distributed:

$$FDR \le E\left(\frac{R_t+J}{R_t+1}\right)\cdot q\delta^*\cdot\tilde{\pi}_0, \tag{1}$$

where J is the number of families tested, R_t is the total number of discoveries, $\tilde{\pi}_0$ is a weighted mean of the proportion of true null hypotheses in the J families of hypotheses, and δ^* is family-specific multiplicative factor (its upper bound is shown to be smaller than 1.44, but typically $\delta^*\sim 1$). In some cases it is possible to derive universal bounds for the FDR. (a) If the researcher is interested in the entire set of discoveries (both strain differences and association discoveries), namely the full-tree FDR, then it is easy to see that the bound for the FDR in expression (1) is less than is $2q\delta^*$. (b) It can also be argued that once expression is found to be associated with behavior, the initial strain difference discovery is no longer of interest—thus the 'interesting' discoveries are all the association discoveries, and strain difference discoveries with no subsequent association discoveries, namely end-node FDR. Note that in this case R_t is greater than R_t in case (a) divided by the number of levels in the FDR tree—therefore the FDR is less than $4q\delta^*$ (the bound in (a) times L). (c) If the researcher is only interested in the association discoveries, then there is no universal upper bound for this level-2 FDR.

Yekutieli (2007) addresses both theoretically and via simulations the power and FDR control of hierarchical testing in cases (a), (b) and (c). He finds that settings like the one confronted here, where the number of sub-families is large and each one has few hypotheses, are the most demanding in terms of FDR control. We therefore choose to verify using a simulation study settings somewhat similar to the

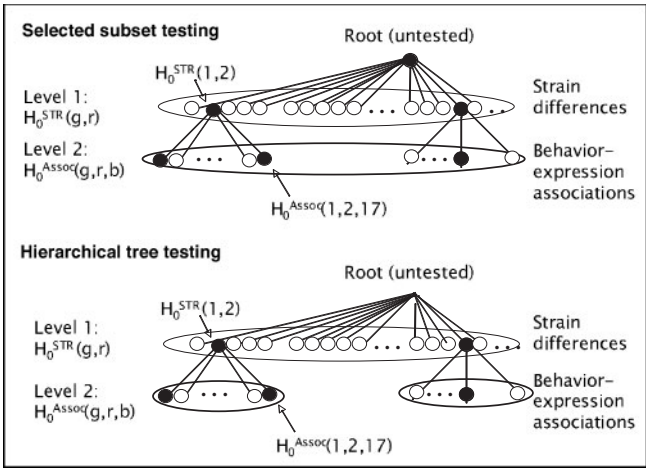

Fig. 1. Schematic display of the selected subset testing and the hierarchical tree testing for testing for association after screening for strain differences in expression levels in specific brain region. The set of discoveries is marked by black circle; Hypotheses tested together as a family, using the BH procedure, are enclosed in an ellipse.

problem we face, and study the behavior of

$$B(R_1,J) = E\left(\frac{R_1+J}{R_1+1}\right).$$

Moreover, let $\hat{B}(R_1,J)$ be its observed value, namely the ratio between the total number of association discoveries plus the number of families and the total number of discoveries plus 1. We use it as to estimator in the FDR for the association discoveries: $\hat{B}(R_1,J)\cdot q\delta^*$. Finally, note that if the number discoveries greatly exceeds the number of tested families (i.e. most of discoveries are association discoveries) then the FDR is also $\sim q\delta^*\cdot\tilde{\pi}_0$ in testing scenarios (a) and (b).

2.2 Screening using ANOVA

As discussed before, analysis of variance will be used to test the difference between strains at the gene-brain region level. When strain differences are assessed during the first stage of research we face the problem of a small sample size since the data consists of only two biological replicates for each gene in each brain-region, resulting in $8(2-1)$ degrees of freedom for estimating the error term. This problem is quite common in microarray experiments that include more than one factor, for the mere reason of high economical cost. Pavlidis *et al.* (2003) show in this context that statistical analysis of experiments containing less than five biological replicates may result in poor power and reproducibility Thus, rather than using the conventional one-way ANOVA F-test, we will test the simple effect of strain within brain-region, as proposed by Winer (1971). Namely, for each brain-region, the test is based on the between-strain variation estimated from the one-way model only for that specific brain region (numerator), and the within-strain variation is estimated from the full two-way model with interaction (denominator). Since expression levels from the same biological replicate in different brain regions may be correlated, we use the simple effect analysis within a mixed model framework, where observations from the same biological replicate across the brain region are treated as repeated measurements and the model errors having compound symmetry covariance structure. For details for implementation see Supplementary Material.

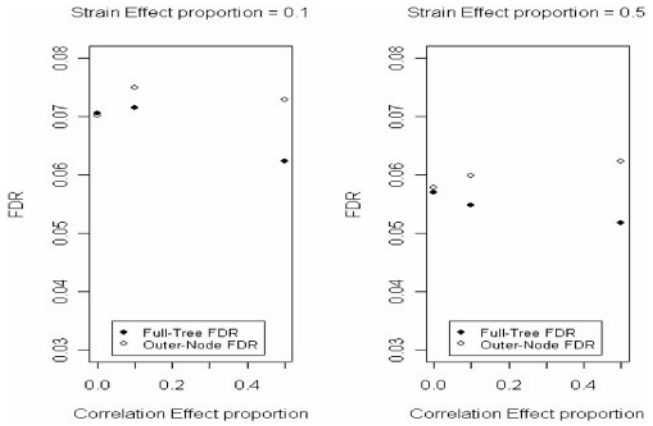

Fig. 2. FDR for full-tree and outer node schemes, as estimated from the simulation study. The FDR for full tree is lower than that of the outer-node; both are much lower than the theoretical bound 0.144.

2.3 Testing for association

In order to assess the association between gene expression and the quantitative trait we use Spearman's correlation rather than Pearson's. Our choice is motivated by the fact that the rank-based Spearman's correlation assesses the strength of a monotone relationship between the two, while Pearson's assesses the strength of the linear relationship. It is true that through the use of suitably chosen transformations for each combination of gene, region and trait, Pearson's correlation may be almost as useful. However, with little-monitored screening of many correlations this becomes a formidable task. The use of Pearson correlation and the effort involved cannot be avoided when the number of observations available for the correlation analysis is very small, in this case Spearman's correlation cannot achieve low enough *P*-values even when the association is strong due to the granularity of rank correlation distribution. This is not the case in our example.

2.4 Jittering for estimating correlation when pooling samples for gene expression

Recall that within each strain we have individual data for the behavioral trait but we have only two means for the gene expression data due to pooling. One option was thus to reduce all information to strain averages, producing only 8 points, which amounts to using strains as the unit of correlation analysis. This type of correlation has been termed ecological correlation (e.g. Greenland and Robins, 1994, Robinson, 1950) that has limitations when it comes to interpretation, since it is the individual animal for which we wish to study the behavior and gene expression. Furthermore, this approach yields high correlations but also high *P*-values (not very significant) because of the small number of strains available.

It may still be argued that since strains are genetically identical it is reasonable to assume that the individual gene expressions for each strain are all at the same (mean) level while still making use of the individual behavioral information. Such an approach leads to a correlation that is always too high due to the artificially reduced variation of the expression level: In the calculation of Spearman's correlation, the ranks of expression levels are being used. Thus all expression levels of animals from the same strain will receive the same rank, artificially reducing the real variability of expression levels.

We may simulate individual values of gene expression for mice, under the null hypothesis of no correlation between expression level

and behavior, by jittering the mean expression values: from the estimated variance of the biological replicates $\hat{\sigma}^2_{b,gr}$ we can estimate for gene g and brain region r the between mice individual variability in expression levels:, $\hat{\sigma}^2_{gr} = \hat{\sigma}^2_{b,gr} \cdot n_{gr}/k_{gr}$ (where n_{gr} is the number of animals for which gene expression in gene g and brain region r was measured, and k_{gr} is the number of biological replicates taken to measure this expression level). We then assign the random values to the individual animals. Now the correlation and the significance from these values are more realistic. For instance, consider the correlation of Glo1 expression in the cerebellum and the proportion of time the mouse spends in the center of the arena (See Supplementary Material S2 for details). The estimated correlation with the behavioral trait is −0.62, with P-value <0.0001. Two different runs for the above combination yield correlations of −0.31 and −0.44, and the P-values are 0.029 and 0.001, respectively. Both are weaker than before and somewhat less significant, but closer to reality.

We simulated such data 1000 times and averaged the correlations and the logarithm of the P-values over the simulation. In the above example the simulation-averaged correlation is −0.36.

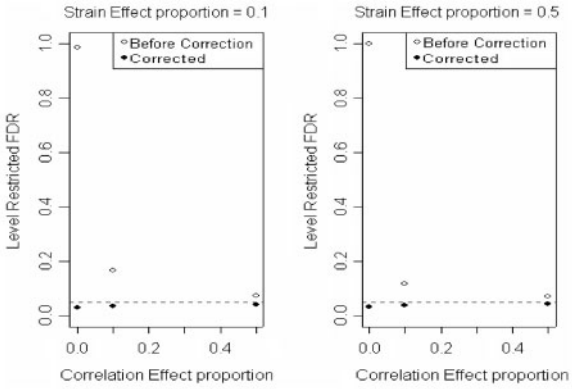

Fig. 3. Estimated FDR when restricted to correlation level. Uncorrected and corrected by $B(R_l,J)$.

2.5 A simulation study

A simulation study of the FDR level achieved by hierarchical testing was conducted on data simulated to have the same characteristics as our experiment data (except for number of genes). Expression data at the gene and brain-region level was set to include non-zero differences between strains for a proportion p_d of the cases. Of them, a proportion p_c was set to correlate with some of the behavioral traits. Correlation of each gene were introduced either to one of its brain regions (with probability p_a) or to all of them, and p_t of the traits were correlated with expression. Thus we incorporated expression correlation into the 'networks' of genes that are correlated with behavior. We did not incorporate the across brain region correlations, for at this stage of the simulation all we need is that the P-values will be valid. Effect and noise sizes were chosen so that the power under the alternative at the first stage will be high—from 0.5 to 1. Details are given in the Supplementary Material. The simple effect F-test was applied here at the first stage to select combinations of gene and brain region where strain differences in expressions were found. Next, the expression levels of the selected combinations were tested for association with the behavioral traits using Spearman's Correlation. The FDR controlling procedure in BH was applied in two ways: using the selected subset testing, and using the hierarchical testing scheme.

3 RESULTS

3.1 Results concerning methodology

For the subset selection method, the FDR is controlled at the desired level of *0.05*. For the hierarchical testing method, we first examine the value of δ^*. The values found are close to 1, around *1.04–1.06*. Running additional simulations on configuration 3 with q set to *0.01* and *0.1*, which yields δ^* of 1.03 and 1.06, respectively, we may reasonably conclude that δ^* keeps stable for any q. The $\delta \sim 1$ is consistent with results of previous simulations.

The full tree FDR should be controlled at the level of $2q\delta^*$. Consequently, the conservative bound on the FDR level is still $2q$, and as seen in Figure 2, it is achieved for all configurations. When the proportion of false null hypotheses is small, most of the rejected hypotheses are parent hypotheses. This was indeed the case where the full-tree FDR reaches its upper limit.

Next we examined the FDR of the end-node testing scheme, which under independence should be controlled at level $2Lq\delta^* = 4q\delta^*$. Again, $\delta \sim 1$, and the conservative bound of $4q$ still applies. This is evidently much too conservative a bound, the actual value that can be inferred from Figure 2 lies well below $2q$. The degree of this excess depends on the number of correlations. Where it is relatively small, this gap is relatively small. Where there are no correlations at all the two criteria are nearly equal. Here, the expected total number of discoveries remains the same as that of the full-tree scheme. In view of the above, for both full-tree and end-node FDR control we use the BH at level $q_1 = q_2 = q/2$, in order to control the FDR at level q.

Restricting our interest to the correlation-testing level, we focus on how much this FDR is increased due to the restriction. The multiplier $B(R_l,J)$, increases in the number of subfamilies visited relative to the number of rejections within them. Indeed, as shown by our simulation results in Figure 3, the multiplier, averaged over a specific configuration, can take very large values. Here, it reaches around 56 and 77, when there are no correlations (configurations 1 and 4), and thus very few rejections relative to the number of gene and brain-region combinations selected in the first stage. However, when there are correlations in the data, this multiplier is much smaller. For the cases with the smallest number of correlations the multiplier is around 5.3 and 3.1. For cases with a larger number of correlations it is already less then 2, here around 1.6 and 1.8. For the case with the largest number of correlation, the factor moves further towards 1, here 1.3.

The large values that the FDR multiplier can take are responsible for the large values that the Level-2 FDR reaches. Nevertheless, it seems that dividing Level-2 FDR by the estimated $B(R_l,J)$, will reduce it to the desired level. Thus, for the level-restricted testing scheme, a $q^* = q/\hat{B}(R_l,J)$ may achieve control at level q. While δ^* can in principle get as

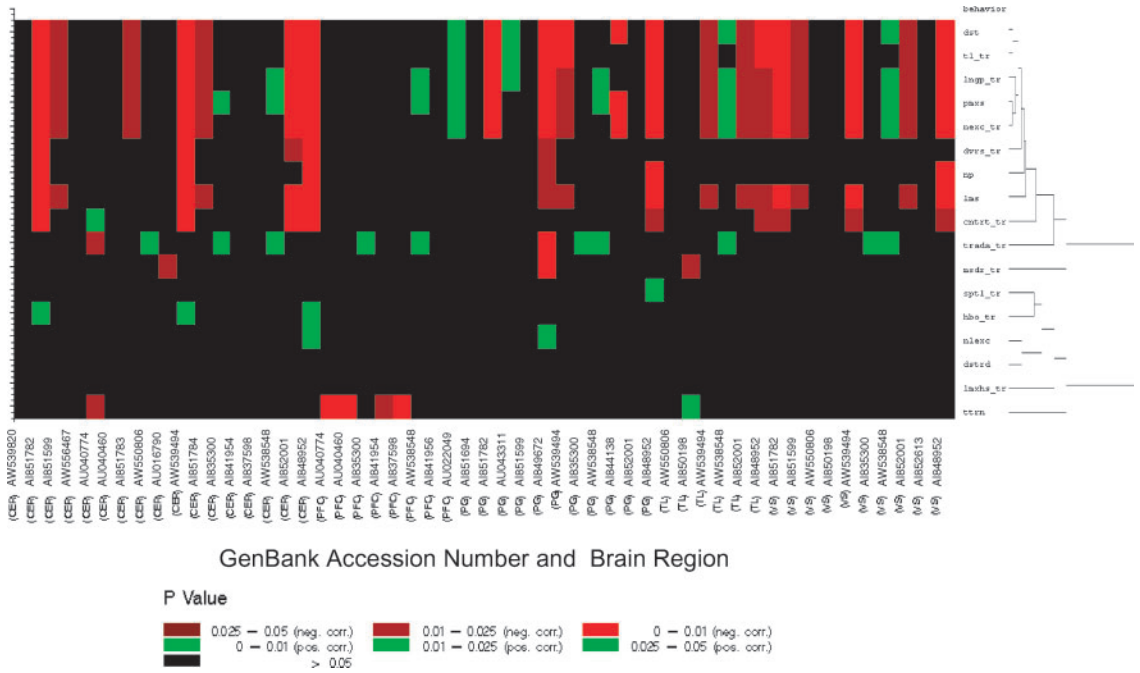

Fig. 4. Significance and Direction of Associations between Gene expression in Brain Regions and Behavioral Measures. Brain region information is given below gene identities: prefrontal cortex (PF), ventral striatum (VS), temporal lobe (TL), periaqueductal gray (PG) and cerebellum (CR).

high as 1.44, we see from Figure 3 that it is just around 1 over the configurations studied. Hence, using the minimal q^* such that $q^* \cdot \hat{B}(R_l,J) \leq q$, may offer Level-2 FDR $\leq q$.

Next we compare the power of both approaches with the BH procedure and two representative adaptive procedures: the adaptive BH (Benjamini and Hochberg, 2000) and the Average Estimate (Jiang, 2004). For estimating π_0 the Average Estimate considers an k regions $(i/k,1)$ for $i=k,k-1,\ldots1$, and estimate Π_0 for each as in Storey's λ (2002). The choice of the particular i to be used is done in a way similar the choice in the adaptive BH (Benjamini and Hochberg, 2000), that is starting from $i=k-1$ and continuing as long as the number of P-values in $((i/k,(i+1)/k]$ decreases. Even though it has not been shown to control the FDR analytically, it is claimed to be equivalent or slightly superior to many existing adaptive procedures and hence we included it in our study. Single-step methods were also found to have FDR ≤ 0.05.

It is well documented (Benjamini and Hochberg, 2000; Storey, 2002, and many others) that the adaptive methods offer better alternative to the BH when π_0 is medium or small. Our interest is in higher values of π_0 where discoveries are expected to be few and estimating higher values of π_0 offers little improvement, if any. Table 1 displays the power of the proposed methods to identify correlations compared to the single-step methods, for such higher values of π_0. (More on the configurations in Supplementary Material S3). When π_0 is near 1, that is when the potential findings are scarce, the hierarchical tree testing scheme achieves the highest power while controlling the FDR, in spite of requiring a lower q relative to other methods. When π_0 is much smaller than 1, the advantage of single-step adaptive methods take the lead.

Table 1. Power at three configurations (SE <.01)

Type of Method	Method	Power		
		$\pi_0=.820$ *(conf. 7)*	$\pi_0=.990$ *(conf. 3)*	$\pi_0=.998$ *(conf. 2)*
Single-step methods	BH	0.76	0.45	0.29
	Adaptive BH	0.78	0.45	0.29
	Average estimate	0.78	0.45	0.29
Proposed methods	Subset selection	0.59	0.44	0.34
	Hierarchical tree	0.59	0.49	0.50

3.2 Results on associations

Using the hierarchical procedure with an overall FDR of.05, 124 pairs of genes and brain regions were identified in the initial screening for strain differences; they involved 64 genes and all brain regions, though not equally represented (e.g. 17 genes in the CER versus 6 in TL). In the second stage 186 associations were found involving only 20 genes and all brain regions. At this level $B=(124+186)/187=1.66$, justifying the somewhat conservative value of 2 we chose. The results for the genes and brain regions are displayed in Figure 4. The familiar looking display does not carry the usual information: rather than expression levels it displays the direction and significance of the association between a behavioral trait (at the row) and the

expression level of the gene at the brain region identified in the column. The non-significant correlations are blackened, as we cannot be sure about the direction of the association even if it exists. The behavioral traits are clustered by the correlation over mice. The clustering process did not influence the expression data.

Using the subset selection procedure FDR screening at 0.025, (as in the screening stage of the hierarchical method) and testing for association at FDR level .05, 52 pairs of genes and brain regions were identified to be associated with at least one trait, i.e. 13 more genes than found by hierarchical procedure. However, only 161 associations were found with this method, 25 less than by the hierarchical one.

Genes identified in this study are now subject to biological experimentation in order to establish direct and causative associations. The full implication of the biological findings will be discussed when these results become available.

4 DISCUSSION

We suggested two approaches to control the FDR in multilayer screening strategies. In both methods only families corresponding to genes that passed the FDR screening are subsequently tested for their correlation with a behavioral endpoint. In the first method all these families are combined into a single family over which the FDR is controlled (i.e. subset selection procedure). In the hierarchical FDR testing each screened family is tested separately while controlling the FDR within it. In this case the overall FDR can still be controlled by lowering the level at each stage. The amount by which the level should be lowered depends on the goal of the researchers, whether in full-tree, fixed-layer or end-nodes.

If all screened families are approximately the same in terms of m_0/m and P-values distribution, the first method is superior as the testing is done at a higher q and the behavior of BH will be little affected by the amalgamation of the families. If a few families have high correlations while many others have none (or close to none) the second method has the advantage in spite of lower q at both stages. As expected, in our example the subset selection method was more powerful in the first stage, but the hierarchical was more powerful in identifying correlations for the families selected in the first stage. Even when theoretical bounds do not guarantee control of FDR within one stage l, it seems that a correction using $\hat{B}(R_l,J)$ will provide control at the desired level.

The single-step adaptive methods should be more powerful than both method when m_0/m is substantially smaller than 1. However, this is not the typical case in microarray data or in other high throughput studies, since researchers look for discoveries which are not abundant. In our case the proportion of associations discovered is less than 0.0001, so even if the number of true associations is larger by an order of magnitude we have $m_0/m > 0.999$. In contrast, the hierarchical method has the advantage of reducing the number of tested hypotheses, thus reducing the cost in power due to multiplicity. Thus, if the researcher identifies an initial screening criterion of the genes that would successfully eliminate cases with no chance of being selected in the second stage, the hierarchical testing scheme will yield the largest number of discoveries.

ACKNOWLEDGEMENTS

The research has been supported by NIH grant #DA015087 (GIE,NHL,YB) and Israel Science Foundation (DY). We thank the reviewers for helpful comments.

Conflict of Interest: none declared.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

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

Benjamini,Y. *et al.* (2001) Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.*, **125**, 279–284.

Benjamini,Y. *et al.* (2006) Two-Staged Linear Step-Up FDR Controlling Procedure. *Biometrika*, **93**, 491–507.

Benjamini,Y. and Yekutieli,D. (2001) The control of the False Discovery Rate under dependency. *Ann. Stat.*, **29**, 1165–1188.

Benjamini,Y. and Yekutieli,D. (2005) The False Discovery Rate Approach to Quantitative Trait Loci Analysis. *Genetics*, **171**, 783–789.

Chesler,E.J. *et al.* (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.*, **37**, 233–242.

Drai,D. *et al.* (2000) Statistical discrimination of natural modes of motion in rat exploratory behavior. *J. Neurosci. Methods*, **96**, 119–131.

Dudoit,S. *et al.* (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Efron,B. *et al.* (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

Greenland,S. and Robins,J. (1994) Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *Am. J. Epidemiol.*, **139**, 747–760.

Jiang,H. (2004) A two-step procedure for multiple pairwise comparisons in microarray experiments. PhD Thesis, Purdue University, West Lafayette, IN, USA.

Kafkafi,N. *et al.* (2005) Genotype-environment interactions in mouse behavior: A way out of the problem. *Proc. Natl Acad. Sci.*, **102**, 4619–4624.

Kerns,R.T. *et al.* (2005) Ethanol-responsive brain region expression networks: implications for behavioral responses to acute ethanol in DBA/2J versus C57BL/6J mice. *J. Neurosci.*, **25**, 2255–2266.

Kerr,M.K. *et al.* (2002) Statistical analysis of a gene expression microarray experiment with replication. *Statist Sinica*, **12**, 203–217.

Lee,N.H. and Saeed,A.I. (2006) Microarrays an overview. *Methods Mol. Biol.*, **353**, 265–300.

Letwin,N.E. *et al.* (2006) Combined application of behavior genetics and microarray analysis to identify regional expression themes and gene-behavior associations. *J. Neurosci.*, **26**, 5277–5287.

Pavlidis,P. (2003) Using ANOVA for gene selection from microarray studies of the nervous system. *Methods*, **31**, 282–289.

Pavlidis,P. and Noble,W.S. (2001) Analysis of strain and regional variation in gene expression in mouse brain. *Genome Biology*, **2**, 00421–0042.15.

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

Reiner,A. (2007) FDR control in two-sided tests with dependent test statistics. *Biom. J.*, **49**, 107–126.

Reiner,A. *et al.* (2003) Identifying differentially expressed genes using False Discovery Rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Robinson,W.S. (1950) Ecological correlations and the behavior of individuals. *Am. Sociol. Rev.*, **15**, 351–357.

Smyth,G.K. (2005) Limma: linear models for microarray data. In Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds.) *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.

Smyth,G.K. *et al.* (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21**, 2067–2075.

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

Storey,J.D. and Tibshirani,R. (2003) SAM thresholding and False Discovery Rates for detecting differential gene expression in DNA microarrays. In Parmigiani,G., Garrett,E.S., Irizarry,R.A. and Zeger,S.L. (eds.) *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York.

Winer,B.J. (1971) *Statistical Principles in Experimental Design*. Second ed. McGraw-Hill, Inc.

Yekutieli, D. (2007) Hierarchical False Discovery Rate Methodology. (A revision is under review in Journal of the American Statistical Association)

Yekutieli,D. *et al*. (2006) Multiplicity Issues Related to Complex Research Questions in Microarray Analysis. *Statist. Neerlandica*., **60**, 414–437.

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