BIOINFORMATICS

Vol. 19 no. 3 2003, pages 368–375
DOI: 10.1093/bioinformatics/btf877

Identifying differentially expressed genes using false discovery rate controlling procedures

Anat Reiner, Daniel Yekutieli and Yoav Benjamini*

Department of Statistics and Operations Research, The Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel

Received on August 14, 2001; revised on June 20, 2002; September 5, 2002; accepted on September 9, 2002

ABSTRACT

Motivation: DNA microarrays have recently been used for the purpose of monitoring expression levels of thousands of genes simultaneously and identifying those genes that are differentially expressed. The probability that a false identification (type I error) is committed can increase sharply when the number of tested genes gets large. Correlation between the test statistics attributed to gene co-regulation and dependency in the measurement errors of the gene expression levels further complicates the problem. In this paper we address this very large multiplicity problem by adopting the false discovery rate (FDR) controlling approach. In order to address the dependency problem, we present three resampling-based FDR controlling procedures, that account for the test statistics distribution, and compare their performance to that of the naïve application of the linear step-up procedure in Benjamini and Hochberg (1995). The procedures are studied using simulated microarray data, and their performance is examined relative to their ease of implementation.

Results: Comparative simulation analysis shows that all four FDR controlling procedures control the FDR at the desired level, and retain substantially more power then the family-wise error rate controlling procedures. In terms of power, using resampling of the marginal distribution of each test statistics substantially improves the performance over the naïve one. The highest power is achieved, at the expense of a more sophisticated algorithm, by the resampling-based procedures that resample the joint distribution of the test statistics and estimate the level of FDR control.

Availability: An R program that adjusts p-values using FDR controlling procedures is freely available over the Internet at www.math.tau.ac.il/∼ybenja.

Contact: anatr@post.tau.ac.il

*To whom correspondence should be addressed.

INTRODUCTION

Gene expression analysis across various biological conditions, cell cycle states, tissues and subjects may help identify differentially expressed genes. This type of information is a valuable pinpoint in the investigation of biological processes and functional disorders. cDNA microarrays have been recently used for monitoring expression levels of thousands of genes simultaneously. See Brown and Botstein (1999) for an overview.

The statistical significance of the differential expressions may be tested using replicated experiments. Considering a microarray data analyzed by testing each gene, 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.

In microarray experiments, mRNA is extracted out of specific animals and tissues under biological conditions that are functionally associated with the mechanism examined. Consequently, the genes tend to subgroup into highly correlated expression levels for reasons such as co-regulations. One of the sources of dependencies in microarray data, particularly in studies of cancer, is co-regulation based on genomic locations and gene expression biases based on the effects of aneuploidy. Furthermore, gene expression measurement errors may be dependent due to factors related to the RNA source, the normalization process and the pooled variability estimation. Multiple testing of such data will produce correlated test statistics. Thus, it is essential to account for the dependency structure between the test statistics.

The number of comparisons of gene expression level studied in a single article has been growing (literarily) at an exponential rate since the beginning of the 1990's, and by 1997, it has reached 4000 genes. While numerous

methods were available for controlling the family-wise type I error rate (FWE), which is the probability of committing even one error in the family of hypotheses, attention to the multiplicity problem in gene expression analysis has been virtually null until very recently.

Dudoit *et al.* (2002) is one of the first studies to recognize the importance of the multiplicity problem as one of the key statistical issues arising in microarray data analysis. The Westfall and Young step-down algorithm, herein WY (Westfall and Young, 1989), a permutation-based procedure, is used to adjust for multiplicity by controlling the FWE, without assuming *t* distribution of the test statistics of each gene's differential expression.

While in some cases FWE control is needed, the multiplicity problem in microarray data does not require a 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). Controlling this FDR criterion in the simultaneous testing of gene expression is the focus of this paper, since it admits more powerful procedures.

A couple of studies have already contemplated the use of the FDR controlling approach in microarray analysis. Tusher *et al.* (2001) applies a significance analysis algorithm to examine UV-damaged DNA. The transcriptional response of human cells to ionizing radiation was measured by microarrays. A two-stage *p*-value adjustment is applied. The estimated FDR is computed using permutations of the data, allowing the possibility of dependent tests. Therefore, as pointed out by the authors, it seems plausible that this estimated FDR approximates the strongly controlled FDR when any subset of null hypotheses is true. However, the authors noted that due to the limited number of possible distinct permutations, the number of distinct values that the *p*-value can take is limited. Consequently, the FDR estimate turns out to be too 'granular', so that either zero or 300 significant genes are identified, depending on how the *p*-value was defined. A similar result was obtained using the adaptation to dependent tests suggested by Benjamini and Yekutieli (2001b).

Dudoit *et al.* (2002) consider using the linear step-up FDR controlling procedure, herein BH (Benjamini and Hochberg, 1995), since it is less conservative than the WY procedure, but rules it out as it was known at the time of the study that the procedure required independence of the test statistics.

Additional work concerning the multiplicity issue in microarray data via FDR control is currently under progress by several groups. Sabatti *et al.* (2002) argues in favor of FDR control in microarray data analysis based on the general results obtained by Abramovich *et al.* (2000). These results suggest the optimality of FDR thresholding for estimating a sparse vector of means by its adaptation to the degree of sparsity. Efron *et al.* (2001) and Storey (2001) discuss a Bayesian interpretation of the FDR within the context of microarray data. *A-posteriori* probabilities of effect for individual genes are estimated, offering local FDR analysis.

Recent advances in FDR methodology offer improved ways of incorporating FDR control in gene expression analysis. The results of Benjamini and Yekutieli (2001b) extended the scope of applicability of the BH procedure to dependency situations, and Yekutieli and Benjamini (1999) introduced resampling-based procedures that control the FDR under dependency. We modify and adapt these advances to the setting of gene expression analysis, considering four procedures and studying their properties.

The first procedure considered is the BH as applied to the *p*-values corresponding to the *t* tests. The second uses the same BH, as applied to the marginal *p*-values estimated by resampling and then pooling the resampling distributions over genes. The two other procedures are based on estimating the joint distribution of the *p*-values and the FDR at a given potential threshold using resampling. They differ by the way this distribution is summarized, that is by the local FDR estimator used. We study these procedures using simulation, while adhering to the structure of the original data analyzed in Dudoit *et al.* (2002). We first find that all four procedures control the FDR at the desired level. All four are also more powerful than their FWE counterparts. We then compare their gain in power, since not all four are equally easy to implement.

METHODS

The false discovery rate criterion

The common approach in simultaneous testing is to construct a procedure that controls the FWE. Benjamini and Hochberg (1995) offer 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 then FWE control, and thus FDR controlling procedures are potentially more powerful. While some situations require FWE control, such as when the result of rejecting hypotheses yields an action (e.g. a drug is approved), in other cases FDR control is sufficient. The analysis of gene expression data is such a case, as its purpose is to extract genes that are potential candidates for further investigation. 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. Controlling the probability of at least one such rejection appears to be over-conservative and will result in reduced experimental efficiency due to unnecessary loss of power. Controlling the FDR instead allows control of the proportion of effort invested in vain, on the average, at the next stage of the investigation.

We define FDR as follows. Consider a family of m simultaneously tested null hypotheses of which m_0 are true. 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, V the number of true null hypotheses erroneously rejected, and S the number of false hypotheses rejected. Now let Q denote V/R when $R > 0$ and 0 otherwise. Then the FDR is defined as

$$FDR = E(Q).$$

As shown in Benjamini and Hochberg (1995), the FDR of a multiple comparison procedure is always smaller than or equal to the FWE, where equality holds if all null hypotheses are true. Thus control of the FDR implies control of the FWE when all hypotheses are true. In the context of gene expression analysis, this result means that if in reality no genes are differentially expressed and the FDR is controlled at some level q, then the probability of erroneously detecting any differentially expressed genes is less than or equal to q.

The linear step-up procedure (BH)

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

Benjamini and Hochberg (1995) show that when the test statistics are independent, this procedure controls the FDR at the level q. Actually, the FDR is controlled at level $FDR \leqslant q \cdot m_0/m \leqslant q$.

Benjamini and Yekutieli (2001b) further show that $FDR \leqslant q \cdot m_0/m$ for positively dependent test statistics as well. The technical condition under which the control holds is that of positive regression dependency on each test statistic corresponding the true null hypotheses (as defined there). In particular, the condition is satisfied by positively correlated normally distributed one-sided test statistics, and their studentized t-tests. The studentized form applies to the cDNA microarray data structure as a result of the tendency of the measurement errors of gene expressions to be positively correlated, due to common latent factors involved. When no real differential expression exists, these are the main sources of variability. Furthermore, since up-regulation and down-regulation are about equally likely to occur, the property of FDR control can be extended to two-sided tests (Yekutieli, 2002).

For more general cases, in which the positive dependency conditions do not apply, Benjamini and Yekutieli (2001b) prove that replacing q with $q/\sum_{i=1}^{m} \frac{1}{i}$ in the linear step-up procedure will provide control of the FDR. However, this modification may be too conservative for the microarray problem. In fact, the simulation study presented in this paper further supports our claim that working with q already controls the FDR.

The adaptive procedures

Since 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 $q^* = q\frac{m}{m_0}$ instead of q to gain more power.

Estimating m_0 from a set of p-values goes back to Schweder and Spjøvtoll (1982). Hochberg and Benjamini (1990) formulize their approach and synthesize a procedure that controls the FWE (see Turkheimer *et al.* (2001) for further progress). Benjamini and Hochberg (2000) suggest the adaptive procedure that combines the estimation of m_0 with the BH procedure. Storey (2001) suggests similar versions to estimate m_0, which are implemented in SAM (Storey and Tibshirani, 2003). Benjamini *et al.* (2001) suggest a similarly motivated two-stage procedure with proven FDR controlling properties.

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. 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.

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, each adjusted p-value is compared to the desired significance level, and if smaller, the hypothesis is rejected. Therefore, the way adjusted p-values are used and interpreted remains conveniently familiar, regardless of the adjustment procedure complexity.

For an FWE controlling procedure, the adjusted p-value of an individual hypothesis is the lowest level for which $FWE \leqslant \alpha$. For instance, for the Bonferroni procedure, the adjusted p-value is simply $P_i.m$. For Holm's procedure, where we set k to be the smallest i that satisfies $P_{(i)} > \frac{\alpha}{m+1-i}$, and reject all hypothesis $H_{(i)}, i = 1, \ldots, k-1$, the adjusted p-value can be calculated as $P_{(j)}^{Holm} = \max_{1 \leqslant i \leqslant j}\{P_{(i)} \cdot (m+1-i)\}$. For an FDR controlling procedure, the adjusted p-value

of an individual hypothesis is the lowest level of FDR for which the hypothesis is first included in the set of rejected hypotheses. Thus the adjusted p-value of $P_{(j)}$ using the BH procedure, is $P_{(j)}^{BH} = \min_{j \leqslant i}\{P_{(i)} \cdot \frac{m}{i}\}$.

Resampling FDR adjustments

For data containing high inter-correlations, generally designed multiple comparisons may be over-conservative in specific dependency structures. Resampling-based multiple testing procedures, introduced by Westfall and Young (1989), utilize the empirical dependency structure of the data to construct more powerful FWE controlling procedures.

In p-value resampling, the data is repeatedly resampled under the complete null hypotheses, and a vector of resample-based p-values is computed. The underlying assumption is that the joint distribution of p-values corresponding to the true null hypotheses, which is generated through the p-value resampling scheme, represents the real joint distribution under the null hypothesis. Thus, for each value of p, the number of resampling-based p-values less than p, denoted by $V^*(p)$, is an estimated upper bound to the expected number of p-values corresponding to true null hypotheses less than p.

The WY procedure estimates the FWE by

$$FWE^{\text{est}}(p) = \frac{\#(V^*(p) > 0)}{N},$$

where N is the number of resampling iterations. Then H_{0i} is rejected if $FWE^{est}(p_i) \leqslant \alpha$.

Yekutieli and Benjamini (1999) follow a similar path to achieve FDR p-value adjustments, but, unlike the FWE, the FDR is also a function of the number of false null hypotheses rejected. Therefore, for each value of p, they first conservatively estimate the number of false null hypotheses less than p, denoted by $\hat{s}(p)$, and then estimate the FDR adjustment by

$$FDR^{\text{est}}(p) = E_{V^*(p)} \frac{V^*(p)}{V^*(p) + \hat{s}(p)}.$$

Two estimation methods are suggested differing by their strictness level. The FDR local estimator is conservative on the mean, and the FDR upper limit bounds the FDR with probability 95%.

The two above methods use resampling to estimate the joint distribution of the p-values. A third alternative uses the BH procedure to control the FDR, but rather than using the raw p-values, it estimates the p-values by resampling from the marginal distribution and collapsing over all hypotheses in the following way, assuming exchangeability of the marginal distributions: For the kth gene, with an observed test statistics t_k, the estimated p-value is

$$P_k^{\text{est}} = \frac{1}{I}\sum_{i=1}^{I}\left[\frac{1}{N}\#\left(\left|t_i^{\cdot j}\right| \geqslant |t_k|\right)\right]$$

We next use the estimated p-values in the BH procedure to easily obtain the BH point estimate for the kth gene:

$$P_{(k)}^{BH} = \min_{k \leqslant i} \frac{P_{(i)}^{\text{est}} \cdot m}{i}$$

All above FDR adjustments can now be used to test the null hypotheses at some arbitrary value q. But rather than adhering to q, all p-value adjustments can be plotted simultaneously as a function of any monotone transformation of p (for example, the test statistic). Such a plot, suggested by Yekutieli and Benjamini (1999) and by Storey (2001), allows the researcher to decide on a meaningful rejection region while being warned of the overall type I error in terms of the FDR.

RESULTS

The data

We analyzed a cDNA microarray dataset used in Dudoit *et al.* (2002), that is publicly available on the web. The data consists of gene expression measurements of 6359 genes from a study of lipid metabolism in mice (Callow *et al.*, 2000). The goal of the experiment was to identify genes with altered expression in the livers of mice with very low HDL cholesterol levels compared to inbred control mice. The treatment group consisted of eight mice with the apolipoprotein AI knockout (this gene is known to play a pivotal rule in HDL metabolism) and the control group consisted of eight 'normal' C57B1/6 mice.

Results of multiple testing on the original data

We applied the normalization described in the paper through lowess smoothing of the log intensity ratio $\log_2$(Red/Green) versus the mean log intensity $\log_2 \sqrt{(\text{Red} \cdot \text{Green})}$. We first examined the p-values obtained directly from the 'raw' t-statistics with 14 degrees of freedom. Ignoring multiplicity, the actual number of raw p-values larger than 0.05 is 568 (out of 6359). On the other extreme, the Bonferroni adjustment points to eight rejections.

Applying the FDR controlling BH procedure on the raw p-values, we came up with the same eight genes identified as differentially expressed in the original analysis. This is not surprising. First, recall that a subgroup of genes was identified by the FWE controlling procedure in the original analysis through distinguishingly low p-values. The FWE adjusted p-values of that subgroup were all below 0.01 while the rest were all above 0.6. Second, we anticipate that the actual distribution of the test statistics is not quite

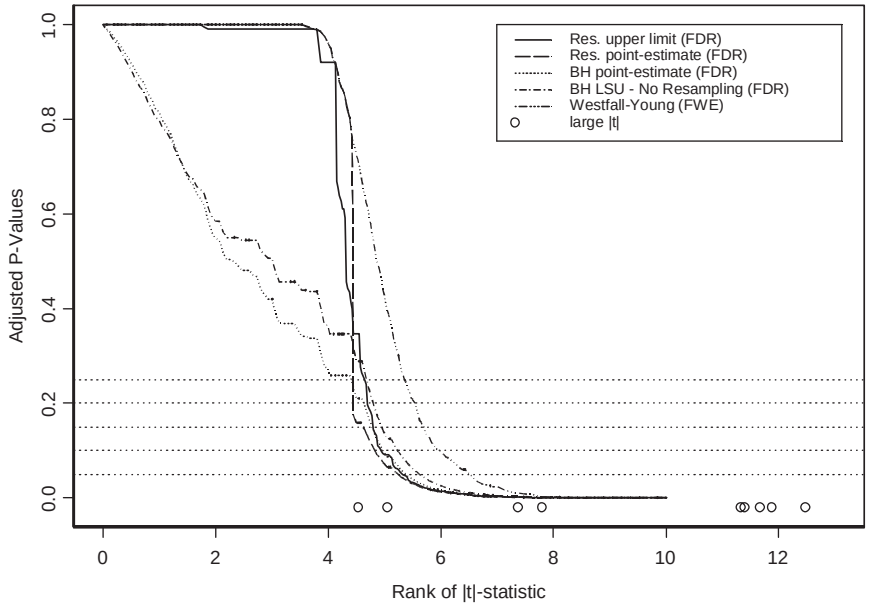

Fig. 1. FDR and FWE adjusted p-values — original data.

the same t-distribution underlying the derivation of the p-values.

We estimated the distribution of the t-statistics over 1000 resampling iterations. Adjusted p-values were calculated using the Westfall and Young algorithm, the two local FDR estimators of Yekutieli and Benjamini (1999), namely the resampling-based point estimator and the resampling-based 1–0.05 upper limit, and the BH point estimator. Figure 1 is a plot of the adjusted p-values versus the test statistics. The ten highest absolute t-values (except the largest one, 20.6, which is too far to the right) are marked on the plot. As seen, the FDR local estimators and the BH point estimator consistently produce much lower adjusted p-values then those produced by WY algorithm. The WY adjusted p-values decrease more slowly than the FDR adjusted p-values. As implied by the plot, at the 0.05 significance level, we may still reject the same eight hypotheses by all procedures. Increasing the FDR level to 0.1 allows rejection of only one more hypothesis. Using the WY algorithm leaves us the initial eight genes.

The FDR controlled and FWE controlled results for this experiment are very close, both being very different from the unadjusted results. This should come as no surprise since the most significant eight genes are separated from the others, as discussed earlier. In fact, it is reassuring that the reduced conservativeness of FDR controlling procedures does not trigger discovery of artifacts. In other cases typical of microarray data, where there is no clear distinction between differentially expressed genes and similarly expressed ones, we would expect to find that controlling the FDR allows the identification of more genes than controlling the FWE. We thus proceed to comparatively examine the performance of the multiple testing procedures under controlled occurrence of differential expression.

A comparative simulation study

Simulated data configuration We fixed the number of differentially expressed genes to 70, roughly 1% of the total number of genes in the experiment. Differential expression was generated using the weak l_p-ball model described in Abramovich *et al.* (2000), by which a sparse signal pattern was generated:

$$r \cdot n^{1/p} \cdot i^{-1/p}, \quad i = 1, \ldots, n$$

where p is the decay rate parameter, r is the decay function maximum and n is the number of values. We used $p = 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2$.

For each one of the 70 genes with the top differential expression measurement, the mean difference was subtracted from the group with the greater mean, thereby removing potential differences not attributed to noise. This modified data set served as the raw data for our simulation, where on each simulation repetition, the experiment and control groups were shuffled. No shuffling of the genes was performed, so that the original dependency structure was preserved. Next, we added the simulated sparse differential expression values to 70 randomly selected genes, thereby getting a single repetition in the simulation. We then applied the multiple testing procedures described earlier on

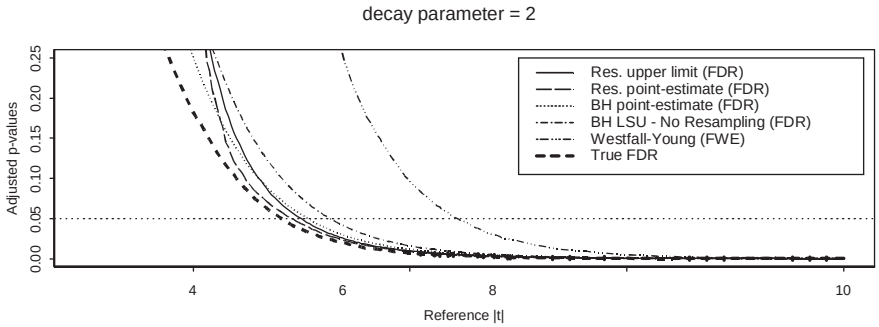

Fig. 2. FDR and FWE adjusted p-values — simulated data.

each repetition, this time with 100 resampling iterations. We repeated the simulation 400 times, calculating the average FDR and power over the repetitions.

Simulation study results Figure 2 presents the mean curves of the adjusted p-values versus the test statistics, for decay rate parameter 2 and FDR level below 0.25. The maximal standard error of the estimated FDR was below 0.003. The plot also includes the 'true FDR', which is the proportion of the absolute values of the t-statistics that exceed a reference point due to chance, out of the total number of absolute values of the t-statistics that exceed the same reference point. As seen in the plot, for all FDR controlling procedures, the adjusted p-values are larger then the corresponding true FDR, indicating guaranteed FDR control. This result holds for FDR level smaller then 0.5. As expected, all FDR controlling procedures produce FDR adjusted p-values much closer to the true FDR than the FWE adjusted p-values obtained by the WY algorithm.

Figure 3 plots the power of the various multiple testing procedures, for each configuration of effects. Here we also include Holm's non-resampling multiple testing procedure, that controls the FWE. As seen, all FDR controlling procedures obtain substantially more power than the FWE controlling procedures. The resampling point-estimator is the most powerful procedure, with the other two resampling estimators following very closely behind, with no consistent advantage of one over the other. Although the upper-limit resampling estimator estimates the joint distribution of the test statistics, its relative conservativeness does not allow increase of power over the BH resampling estimator, which estimates only the marginal distribution. The resampling upper-limit estimator does supply more protection, in that it further controls the empirical FDR with probability 0.95. The BH procedure performs relatively well, in spite of it assuming t distributed test statistics that are either independent or positively dependent, and not using resampling. Holm's procedure, performing the most poorly, reconfirms the advantage of resampling for FWE control under dependency.

DISCUSSION

Many researchers tend to dismiss the issue of multiplicity in microarray data analysis, as well as in similar very large parallel experiments that are becoming technologically feasible bioinformatical tools. They rely on the argument that these experiments merely serve for screening and their purpose is to supply the researcher with an initial pool of candidates. Therefore, statistical considerations that limit the power to generate candidate hypotheses should not be taken at this stage.

This argument is acceptable in the sense that protection is not needed against even a single type I error, so that the severe loss of power involved in such protection is not justified. However, the proportion of errors in the pool of candidates is of great economical significance since follow-up studies are costly, and thus avoiding multiplicity control is costly. Indeed, the FDR criterion is economically interpretable; when considering a potential threshold, the adjusted FDR gives the proportion of the investment that is about to be wasted on false leads. Thus the choice of the FDR level q is an economical one. It is for these reasons that multiplicity should be controlled when testing differently expressed genes in microarray analysis, and that it is best done using the FDR criterion.

Controlling the FDR at the screening stage of the research carries a benefit for the next research stages, as shown by Benjamini and Yekutieli (2001a). Consider a study with R_1 significant results, while controlling the FDR at level q_1. A follow-up study is conducted on the pool identified in the first stage, and a level α FWE controlling procedure is applied. It has been shown that the FWE of the combined two-stage study is αq_1. Alternatively, if a level q_2 FDR controlling procedure is used in the second stage, the combined two-stage FDR is shown to satisfy $E(V_2/R_2) \leqslant q_1 q_2$. Either way, the initial

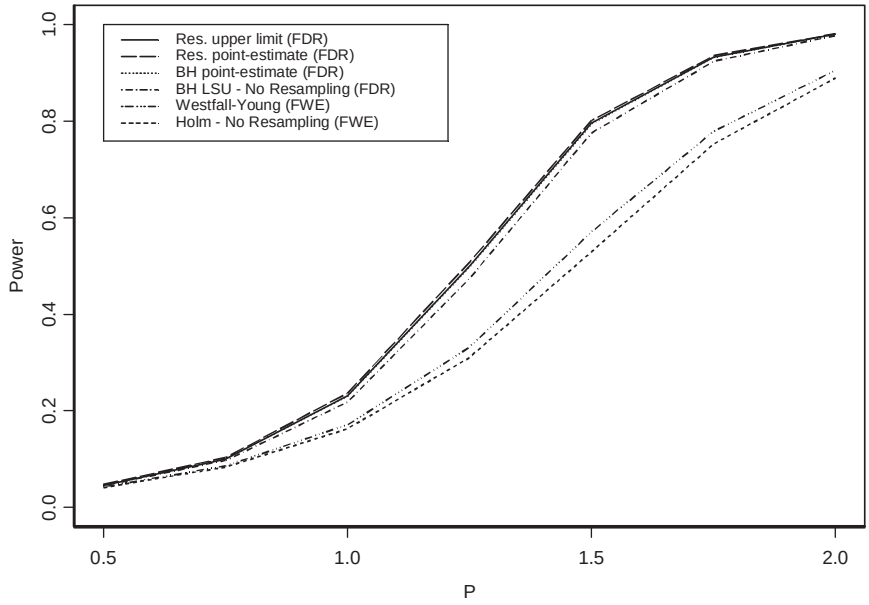

Fig. 3. Power by configuration.

chosen level, q_1, can be allowed to be quite high. For example, assume that in a microarray experiment 100 genes were identified using an FDR controlling procedure with $q = 0.2$. Next, using Bonferroni with α 0.25 at the second stage, that is, assessing individual significance by comparing to $0.25/100$, controls the FWE at the level of 0.05.

The FDR approach for identifying differentially expressed genes has been considered and discussed by Dudoit *et al.* (2002), Sabatti *et al.* (2002), Efron *et al.* (2001), Storey (2001), Storey and Tibshirani (2001, 2003), and Tusher *et al.* (2001). In some of these discussions FDR controlling procedures have been avoided by the researchers due to the rudimentary state of their development with regard to correlated test statistics. Recent theoretical developments have extended the scope of application of existing procedures and offered new ones. In this paper we described and illustrated four procedures that were shown to control the FDR at the desired level (less than 0.5). All FDR controlling procedures retain higher power than FWE controlling procedures, and are therefore highly useful for the discovery of differential genetic expression. The choice among the four is a matter of buying more power and better properties at the expense of more complicated computations.

It should be emphasized that a substantial increase in power is already gained when the p-values are estimated by resampling, and then used in the BH procedure. Shuffling the control and experiment groups created permutations adhering the original dependency structure among the genes. Collapsing the distributions of the test statistics for the genes to a single distribution, and using it to estimate the p-value at each gene, overcomes the discrete nature of the permutation distribution of a test statistics based on few observations, a problem described by Tusher *et al.* (2001). This procedure can be implemented in any statistical software that enables resampling. Still, the researcher may be better off using the more powerful resampling point estimate of Yekutieli and Benjamini (1999). A sample R program is available on http://www.math.tau.ac.il/∼ybenja. The resampling upper limit estimator offers both FDR control (which holds on the average) and a control on the empirical FDR level (up to probability $1 - q$). Thus one gives up (very little) in terms of power relative to the resampling point estimator, gaining further assurance on the actual proportion of false discoveries.

It is not quite clear how the dependency structure affects the performance of multiple testing procedures. Measurement error of microarray data tends to be positively dependent, and simple FDR controlling procedures such as the BH copes with such dependency. A co-regulation dependency may be a result of biological variability of the co-regulation and need not be positive. However, this is essentially the dependency between the 'typical' parameters when situations investigated vary. A possible model may separate this variability into a common co-regulation component and an individual one, in addition to the measurement error, and only the last two components affect the applicability of the simple procedures. An experiment inquiring the relationship between the different components and assessing their relative importance may therefore be highly informative.

Characterization of the dependency structure attributed to the above common co-regulation is one of the main research targets of microarray experiments. Its study is especially challenging due to the high complexity of biological functional pathways. Noise introduced by the non-differentially expressed genes may obscure this structure. Pre-selection of genes that pass an initial FDR testing at moderate q, may largely suppress this noise, thereby improving more specific analyses such as clustering and classification. Here FDR screening essentially serves as an initial dimension reduction technique.

ACKNOWLEDGEMENTS

This research has been partially supported by the F.I.R.S.T. grant from the Israeli Academy of Sciences and Humanities. Y.Benjamini has been partially supported by an NIH grant and a U.S.–Israel Binational Science Foundation grant.

REFERENCES

Abramovich,F., Benjamini,Y., Donoho,D. and Johnstone,I. (2000) Adapting to unknown sparsity by controlling the false discovery rate. *Technical Report No. 2000-19*. Department of Statistics, Stanford University.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. 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., Krieger,A. and Yekutieli,D. (2001) *Two-staged linear step-up FDR controlling procedure*, Technical Report, Department of Statistics and Operation Research, Tel-Aviv University, and Department of Statistics, Wharton School, University of Pennsylvania.

Benjamini,Y. and Yekutieli,D. (2001a) Quantitative traits loci analysis using the false discovery rate. Under revision.

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

Brown,P.O. and Botstein,D. (1999) Exploring the new world of genome with DNA microarrays. *Nature Genet.*, **21**, 33–37.

Callow,M.J., Dudoit,S., Gong,E.L., Speed,T.P. and Rubin,E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Res.*, **10**, 2022–2029.

Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sinica*, **12**, 111–139.

Efron,B., Tibshirani,R., Storey,J.D. and Tusher,V. (2001) Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, **96**, 1151–1160.

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

Holm,S. (1979) A simple sequentially rejective multiple test procedures. *Scand. J. Statist.*, **6**, 65–70.

Sabatti,C., Karsten,S.L. and Geschwind,D.H. (2002) Thresholding rules for recovering a sparse signal from microarray experiments. *Math Biosci.*, **176**, 17–34.

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

Storey,J.D. (2001) The positive false discovery rate: A Bayesian interpretation and the Q-Value. *Technical of Report 2001-12*. Department of Statistics, Stanford University.

Storey,J.D. and Tibshirani,R. (2001) Estimating false discovery rate under dependence, with applications to DNA microarrays. *Technical of Report 2001-28*. Department of Statistics, Stanford University.

Storey,JD 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), To appear in *The Analysis of Gene Expression Data: Methods and Software*. Springer, New York.

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

Tusher,V., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Westfall,P.H. and Young,S.S. (1989) p-value adjustment for multiple tests in multivariate binomial models. *J. Am. Stat. Assoc.*, **84**, 780–786.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan Infer.*, **82**, 171–196.

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