UW Biostatistics Working Paper Series

1-29-2004

Calibrating Observed Differential Gene Expression for the Multiplicity of Genes on the Array

Yingye Zheng
Fred Hutchinson Cancer Research Center, yzheng@fhcrc.org

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

1. Introduction

The advance of high throughput technologies has considerable implications for research in the areas of cancer detection and prevention. In a gene expression array experiment, the expression levels of thousands of genes are monitored simultaneously. Such exploratory studies promise to identify transcripts that show high expression levels in cancer tissues as compared to normal tissues, to pinpoint the biological processes for cancer at the most basic level, and to discover cDNAs encoding proteins that could be potentially useful markers for cancer screening and diagnosis.

Typically, microarray experiments involve exploring enormous numbers of genes on a relatively small set of subjects. For example, in a study concerning gene expression profiling and clinical outcome of breast cancer (Van 't Veer et al., 2002), tumor tissue from 34 patients who developed distant metastases within 5 years and 44 patients who were free of disease for at least 5 years were analyzed to compare the hybridizations on an array of 25,000 cDNAs. Statistical analysis of data from such studies is challenging for several reasons. First, when thousands or tens of thousands of genes are under consideration from a single experiment, performing separate significant tests for each gene greatly increases the type I error. Second, the expression levels of genes tend to cluster as they may function on the same biological pathways and thus co-regulate under the experimental conditions examined. As a consequence, the test statistics can be far from independent. Third, because of concerns about cost or rarity of the target population, a microarray study is usually carried out on a small number of subjects. In this situation the underlying distributional assumptions for the test statistics, which are based on large sample theory, may not be valid or precise enough. It is essential to take into account these problems in the analysis of data.

It is also important to recognize that an appropriate statistical approach depends on the scientific objectives of the study. In this article, we consider microarray studies that are aimed to explore a large pool of genes and select for more careful investigation a subset of genes that are differentially expressed in two tissue types (e.g., cancer versus healthy tissue). In practice, the gene selection process entails several steps. As an initial step, one needs to characterize the capacity of each gene in discriminating between the different tissue types. The choice of statistic is crucial to the entire process. The classic measure of discrimination, such as the two-sample t-statistic or the Mann-

Whitney U-statistic are often considered at this stage. Two additional measures that are related to the Receiver Operating Characteristic (ROC) curve are suggested in Pepe et al. (2003) when there is emphasis on the discriminating capacity over a particular range of the distribution. Once the statistic for discrimination is calculated for each gene, the next step towards selecting genes is to rank the genes based on their evidence for differential expression. At the final stage, one chooses to further investigate the genes that rank well, for example, one might narrow down future research to the top k ranking genes. One important question at this stage is what subset of genes should be selected, i.e., at what k should one draw the line so that the selection process is statistically more rigorous than just choosing some arbitrary k?

In this article we focus on statistical methods for this question. Note that we are not concerned with combining information across genes, a consideration that may or may not follow the analysis that simply ranks the genes. Our focus is on determining a set of genes that each appear to be differentially expressed. In section 2, we first review existing statistical methods that can be adopted for gene selection and then describe a new approach. We compare the performance of our proposed approach with the existing methods using simulation studies in section 3. We further illustrate our new approach with an application to the breast cancer data and close in section 5 with some remarks about the methodology.

2. Selecting Genes

2.1 Existing Methods

Statistical methods for microarray analysis has been a burgeoning area of statistical research in recent years (for review, see Dudoit, Shaffer, Boldrick (2002)). The problem of identifying differentially expressed genes can be translated into the framework of multiple hypothesis testing, where each gene corresponds to a single hypothesis test, and rejecting one hypothesis is equivalent to claiming that the gene is differentially expressed. Table 1 describe the situation when m genes (or hypotheses) are tested. We suppose m_0 of the m genes are not differentially expressed, or are true null hypotheses. We denote by R the number of rejected hypothesis, V the number of false positives, and T the number of false negatives. Only m and R are observable quantities.

An appropriate test procedure should aim to keep both V (the type I error) and T (the type II

error) small. In the univariate setting, the usual strategy is to first prespecify an acceptable type I error α, then seek a test with the most power (smallest type II error) among the class of tests with the same α. To generalize to the multivariate setting, the approach is to define a multiple testing procedure in terms of the adjusted p-value $\tilde{p}_j$ for hypothesis j, which takes all other tests that are involved into consideration, rather than the individually unadjusted p-value p_j. One then rejects H_j if $\tilde{p}_j \leq \alpha$. The adjusted p-values are usually derived in such a way that some type I error rate is controlled at level α.

One type I error rate, the family-wise error rate (FWER), is defined as

$$FWER = P[V \geq 1]. \tag{1}$$

It is the probability of reporting at least one false positive in the family of hypotheses. The step-down algorithm of Westfall and Young (1993) is an example of a multiple testing procedure that controls FWER. The procedure defines the jth adjusted p-value as $\tilde{p}_j = P[min_{1 \leq l \leq m} P_l \leq p_j | H_0^c]$. Here H_0^c denotes the complete null hypothesis, where all the null hypotheses are true (i.e., $m = m_0$) and P_l is the unadjusted p-value for the l^{th} hypothesis denoted with capital letter here because it is a random variable. The joint distribution of $(P_1, ..., P_m)$ can be estimated by permuting the columns of the gene by array data matrix. This algorithm thus takes into account the potential dependence structure amongst genes. Compared with the popular Bonferroni procedure, the approach is less conservative.

Benjamini and Hochberg (1995) suggested a multiple testing procedure that aimed to control a different type I error rate, namely, the false discovery rate (FDR). In their definition,

$$FDR = E(V/R|R > 0)P(R > 0) \tag{2}$$

The concept of FDR is appealing in the context of gene discovery for several reasons. First, FDR has a straightforward interpretation. It is the expected proportion of false positives among genes for which H_0 is rejected, and approximately, it is $P[H_0|rejected]$ since $P[R = 0]$ is typically small. In many applications it can be less stringent than controlling FWER. More importantly, when the goal of a microarray study is to narrow down to a small subset of genes as potential candidates for scrutinization in the next stage of research, one can usually tolerate a small number of false

positives in exchange for higher power. Controlling FDR directly translates into controlling the amount of unnecessary effort invested in a few false positives in the next stage of gene discovery. several procedures have been proposed to control FDR. For example, Benjamini and Hochberg (1995) described a linear step-up procedure. Suppose we order the unadjusted p-values as $p_{(1)} \leq p_{(2)} \ldots \leq p_{(m)}$, with corresponding ordered null hypotheses $H_{(1)}$, $H_{(2)}$, ..., $H_{(m)}$. The adjusted p-value for $H_{(j)}$ is $\tilde{p}_{(j)}^{BH} = min_{k=j,\ldots m}\left\{min(\frac{m}{k}p_{(k)}, 1)\right\}$. We reject $H_{(1)}, \ldots, H_{(k)}$ for $k = max\{j : \tilde{p}_{(j)}^{BH} \leq \alpha\}$ for a desired FDR level α. It can be shown that for independent and continuous test statistics, the procedure yields $FDR = \alpha * m_0/m$, which is $\leq \alpha$. Furthermore, the same level of FDR control holds for positively dependent test statistics as well in the sense defined by Benjamini and Yekutieli (2001). One example of positive dependency structure is positively correlated normally distributed test statistics. When m_0/m is substantially smaller than 1, it is tempting to consider an adaptive procedure so FDR is controlled exactly at level α. For example, Storey (2002) suggested to first estimate m_0, and reject $H_{(1)}, \ldots, H_{(k)}$ for $k = max\{j : \tilde{p}_{(j)}^{BH} * \hat{m}_0/m \leq \alpha\}$. To estimate m_0, Storey suggests the following procedure

$$\widehat{m}_0(\lambda) = \frac{\Sigma_{i=1}^{m} I\{p_i \geq \lambda\}}{1-\lambda}, \tag{3}$$

where λ is in the interval $(0, 1)$ and can be chosen using cross-validation, for example. The adaptive procedure is usually more powerful because it is less conservative, being based on the bound $\frac{m}{\hat{m}_0}\alpha$ rather than α for $\tilde{p}_{(j)}^{BH}$.

Different from the multiple testing procedures described above, the SAM (significance analysis of microarrays) procedure (Efron et al., 2000, Tusher et al., 2001) chooses rejection regions from the distributional properties of the test statistics. The original SAM procedure proposed by Efron et al. (2000) makes use of the ordered test statistics $t_{(1)} \geq t_{(2)} \ldots \geq t_{(m)}$ and a resampling technique. Under the assumption that none of the genes is truly differentially expressed, the labeling of the two groups, cases and controls, can be interchanged. One performs B permutations of the labels and obtains $t_{(j),b}$ for $b = 1, \ldots, B$. The expected value $\bar{t}_{(j)}$ for the jth order statistic under H_0^c can then be estimated based on the permuted samples. For a fixed threshold Δ, genes with $|t_{(j)} - \bar{t}_{(j)}| \geq \Delta$ are claimed significant by the SAM procedure. The SAM procedure can be tailored to control FDR^*, a quantity similar to the FDR that is defined above. This requires estimating the FDR^* for each

Δ from the permutation samples under H_0^c (Storey, 2002) and then choosing the Δ that yields the desired FDR^* level. The strength of the approach is that it offers great flexibility in choosing rejection regions while controlling for FDR^* at a desired level. However, the procedure is based on the implicit assumption that the distributions of $t_{(j)} - \bar{t}_{(j)}$ are homogenous. Furthermore, the procedure controls $FDR^* = E(V|H_0^c)/R$, which is not the same as the FDR as originally defined by Benjamini and Hochberg (1995).

2.2 The AP Method

We propose a new multiple hypothesis testing procedure here. Consider calculating the following adjusted p-value for the kth ordered gene: $p^*_{(k)} = P[|T^0|_l \geq |t_{(k)}||H_0^c, l \leq k]$. This is the probability under H_0^c of observing a statistic as extreme or more extreme than the observed kth order statistic $t_{(k)}$ among the top k order statistics $T^0_{(l)}, l \leq k$. In calculating $p^*_{(k)}$, $T^0_{(l)}$ is a random variable for $t_{(l)}$ under H_0^c. The proposal is to declare genes whose adjusted p-values are $\leq \alpha$ as significant (details below). The idea has some intuitive appeal in our opinion. Given the observed order statistic $t_{(k)}$, it asks how likely it is that under the complete null hypothesis H_0^c the test statistic for genes $(1), (2), ..., (k)$ would exceed $t_{(k)}$. The quantity $p^*_{(k)}$ calibrates $t_{(k)}$ to the distributions of the order statistics under H_0^c. This seems like a natural step. It is similar to SAM in this regard. However, SAM rejects on the basis of $|t_{(j)} - \bar{t}_{(j)}|$ with a cut-off Δ, that is the same for all genes. Our procedure on the other hand acknowledges that the distribution of $T^0_{(j)}$ may not be symmetric about its mean and that its variance may depend on the order (j). Moreover, we will show in section 3 that the operating characteristics of our procedures are comparable (and sometimes better than) existing procedures. We suggest estimating $p^*_{(k)}$ using a resampling procedure to avoid assumptions about the joint distribution of test statistics and to take into account the potential dependence structure amongst genes. In summary, implementation of our procedure consists of the following steps:

1. Compute the order statistics $t_{(1)} \geq t_{(2)} \cdots \geq t_{(m)}$.

2. Perform B permutations of the group labels and obtain $t_{(j),b}$ for each permutation sample b.

3. Compute

$$p^*_{(k)} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{k}\sum_{l=1}^{k} I(|t_{(l),b}| \geq |t_{(k)}|) \tag{4}$$

4. Monotonize the p-values: $\tilde{p}_{(j)} = min_{k=j,\ldots m}\left\{min(p^*_{(k)}, 1)\right\}$

5. Reject $H_{(j)}$ if $\tilde{p}_{(j)} \leq$ some chosen α.

3. Simulation Study

We present the results from numerical studies in this section.

3.1 Compare the proposed AP procedure with SAM

We first evaluate the performance of our resampling-based p-values procedure (hereafter referred to as AP) with SAM, as both procedures make use of the distributions of the order statistics under H_0^c. We generated $m = 500$ gene expression values $\mathbf{X}_1, \cdots, \mathbf{X}_m$ for n_0 control and n_1 case subjects. For the small sample study, we chose $n_0 = n_1 = 20$ and for the moderate sample study we choose $n_0 = n_1 = 50$. We set equal numbers of true null and alternative hypotheses, i.e., $m_0 = 50\%m$. Two scenarios for generating expression levels of different genes were used. In the first scenario, $\mathbf{X}_k \sim N(0, \sigma_0)$ with $\sigma_0 \sim N(1, 0.5)$, for $k = 1, \ldots, m$ for controls and for $k = 1, \ldots, m_0$ for cases. For cases the expression levels of regulated genes $\mathbf{X}_k$, $k = m_0 + 1, \ldots, m$ are generated as $\mathbf{X}_k \sim N(2.0, \sigma_1)$, where $\sigma_1 \sim N(1.5, 0.5)$. Although we allow some variation in terms of dispersion of the distributions of different genes, these distributions still come from the same location-scale family and are symmetric. In the second scenario, we used gamma distributions for generating the expression data. The gene expression values for all m genes of controls and m_0 null genes for cases are specified as $\mathbf{X}_k \sim \gamma(1, 1)$, and the expression levels of regulated genes for cases are generated as $\mathbf{X}_k \sim \gamma(2, 0.8)$, $k = m_0 + 1, \ldots, m$. Thus in this scenario we assume genes expression levels are not symmetrically distributed.

For each simulation configuration, we generated $S = 500$ datasets and performed both our proposed procedure and the SAM procedure. The two-sample test statistic we used to gauge differential expression is the Mann-Whitney U-statistic (denoted by AUC) or equivalently the Wilcoxon ranksum statistic. The SAM procedure is therefore slightly different from the original algorithm of Tusher

et al.(2001). For each dataset, we first calculate AUC for each gene and order them as $AUC_{(1)}$, $AUC_{(2)}$,..., $AUC_{(m)}$. We then take $B = 1000$ permutations of the group labels. For each permutation b we obtain the corresponding ordered AUC statistics: $AUC^b_{(1)}$, $AUC^b_{(2)}$,..., $AUC^b_{(m)}$. For the SAM procedure, we calculate $d_j = |AUC_{(j)} - \overline{AUC}^B_{(j)}|$ for the jth ordered AUC, where $\overline{AUC}^B_{(j)}$ is the average of $AUC^b_{(j)}$ across the B permutation samples. We reject the corresponding jth gene if the value of d_j exceeds some prespecified quantity, Δ. For AP, we calculate the adjusted p values for the ordered genes based on the same permutation samples using the procedure as described in the previous section, and reject a gene if the adjusted p value is less than a prespecified value, α. For each dataset s and each procedure, we record R_s, the number of genes that are claimed to be differentially expressed, and V_s, the number of genes rejected among all the genes that are in truth not differentially expressed. Let $Q_s = V_s/R_s$ if $R_s \neq 0$, and 0 if $R_s = 0$, we then calculate FDR as

$$FDR = \frac{1}{S}\sum_{s=1}^{S} Q_s, \tag{5}$$

and average power as

$$Power = \frac{1}{S}\sum_{s=1}^{S} \frac{R_s - V_s}{m - m_0}, \tag{6}$$

where $(R_s - V_s)/(m - m_0)$ is the proportion of differentially expressed genes that are claimed to be significant. Note that the average power is equivalent to the true positive rate (TPR). In addition, we record the false positive rates (FPR) as

$$FPR = \frac{1}{S}\sum_{s=1}^{S} \frac{V_s}{m_0}, \tag{7}$$

Decision criteria for the AP and the SAM procedures are defined by thresholds, α and Δ respectively, that are on completely different scales. To compare the performances of the two procedures we therefore use ROC curves. That is, for each procedure, we plot TPR versus FPR as the threshold varies across its entire possible range. The ROC curve is a one-one monotone function from $(0, 1)$ to $(0, 1)$ that is a well accepted measure for comparing decision procedures. Better decision procedures are characterized by higher ROC curves. For $n_0 = n_1 = 20$, if the gene expression levels are of gamma variates, the ROC curve based on the adjusted p-values from the AP procedure dominates the curve based on the SAM procedure (Figure 2, top left panel), especially over the region where FP is less

than .2. This indicates that for this particular simulation configuration, the AP procedure has higher accuracy at distinguishing between cases and controls than the SAM procedure. Furthermore, plots of average power versus FDR (Figure 2, top right panel) again show that AP is a more powerful procedure than SAM for this simulation configuration. For example, with FDR of 0.05, the average power is 0.71 for AP, compared with 0.45 for SAM; for FDR at 0.1, the average power is 0.86 for AP but 0.81 for SAM. However, these differences are not observed in the simulation situation where samples are generated solely from normal distributions (Figure 1). In addition, the superior performance of AP over SAM diminishes at moderate sample size such as $n_0 = n_1 = 50$ (bottom panels, Figure 1 and Figure 2) . These results may not be too surprising. A possible explanation is that SAM assumes homogenous and symmetric distributions of test statistics for all genes and that this an assumption is more likely violated with smaller sample sizes. On the other hand, the AP procedure naturally incorporates the variation in the distribution of the test statistic from gene to gene, and does not require any specific distributional assumption. It is thus a more robust and more powerful procedure, particularly in small samples.

3.2 **Controlling FDR and power**

We next compare the performance of the AP procedure with multiple testing procedures that proposed to control the FDR. We investigate to what extent factors such as the dependency structure and number of genes impact on the performance as measured by the FDR and average power.

We generate $m = 40, 200, 1000$ gene expression values. For the small sample study, we choose $n_0 = n_1 = 20$ and for the moderate sample we choose $n_0 = n_1 = 50$. We consider different numbers of true null hypotheses m_0 with $m_0 = 50\%m$, $75\%m$, or $90\%m$. The gene expression values are specified as $X_k \sim N(0,1)$, $k = 1, \ldots, m$ for controls and $k = 1, \ldots, m_0$ for cases, while X_k has a mixture distribution with $p = .7$, and $X_k \sim (1-p) * N(0,1) + p * N(1,2)$, $k = m_0 + 1, \ldots, m$ for cases. The gene expression values are correlated in groups of 10. Specifically, within each cluster of 10 consecutive genes, we let the correlations among the first five genes and the correlations among the second five genes (denote by r_1) be positive, however the correlation between the first 5 genes and the second 5 genes (denote by r_2) can be either positive or negative. In summary we consider the following

correlation structures: $(r_1 = 0, r_2 = 0)$, $(r_1 = 0.3, r_2 = 0.3)$, $(r_1 = 0.3, r_2 = -0.3)$, $(r_1 = 0.6, r_2 = 0.6)$, $(r_1 = 0.6, r_2 = -0.6)$, When the correlations are negative, the 'positive dependence' condition of Benjamini and Yekutieli (2001) does not hold. For each gene k, we calculate the test statistics AUC_k from the $\mathbf{X}_k$ for the n_0 cases and n_1 controls. For each simulation configuration, $S = 500$ datasets were generated. We implemented the linear step-up procedure (hereafter referred to as BH), the adaptive procedure (hereafter referred to as Adapt) with $\lambda = 0.5$ and our resampling-based AP procedure. We did not consider the SAM procedure in this set of simulation studies since SAM controls a different FDR than do the other procedures.

First consider a small sample study with 20 cases, 20 controls and 40 genes. Figure 3 displays FDR versus threshold (α) for studies with different correlation structures and different numbers of true null hypotheses m_0. For the BH procedure, the FDRs are less than α in all cases, and they get closer to α as the percentage of the true null hypotheses increases. In fact, they are very close to the value $\frac{m_0}{m}\alpha$. This is consistent with the theoretical result which states that the FDR is controlled at level $\frac{m_0}{m}\alpha$ for continuous and positively dependent test statistics. Interestingly, even in situations with both positive and negative correlations, where the positive dependence requirement is not satisfied, it appears the BH procedure still controls FDR at a level that is comparable with level of control achieved for positive dependence situations. For our AP procedure, the FDRs are higher than those from the BH procedure, but less than the threshold α for $m_0/m \leq 75\%$. Furthermore, the FDRs from the AP procedure increase as the number of the null hypotheses increase, as was seen for the BH procedure. The FDR exceeds α in the setting where $m_0/m = 90\%$. Thus the AP procedure does not necessarily control the FDR (nor was it intended to). For the adaptive procedure, the FDRs are almost always higher than for the other two procedures, particularly when m_0/m is small. In contrast to the AP and BH procedures, the adaptive procedure is sensitive to the underlying correlation structure. When the genes are statistically independent, the FDRs from the adaptive procedure are close to the corresponding α. However, when genes are positively correlated, we see in many cases that the attained FDR is often greater than the corresponding α at which it wishes to control the FDR. When genes are both positively and negatively correlated, the adaptive procedure seems to underestimate m_0 and thus the FDRs tend to be lower than α. A possible explanation for

this is that the weak dependence assumption required for the adaptive procedure probably does not hold when the number of genes is small.

Corresponding to the configurations in Figure 3, Figure 4 shows average power versus threshold. In general, the adaptive procedure is more powerful than the other two procedures, particularly in settings where the number of true null genes are small. Moreover, the AP procedure is always more powerful than the BH procedure. However the advantage of the adaptive procedure diminishes when the majority of the genes are not differentially expressed in truth. For example, when $m_0 = 75\%m$, the AP procedure is at least as powerful as the adaptive procedure.

Similar patterns are found in a study with the same number of genes but bigger sample sizes, $n_0 = n_1 = 50$, and in studies with larger numbers of genes: m=200 (see Table 2, Table 3, Table 4), and m = 1000 (data not shown). As the number of genes increases, the adaptive procedure better estimates m_0/m, but still seems to be problematic when genes are both negatively and positively correlated.

One phenomenon we observed from our simulation studies as well as from the literature is that a procedure with higher FDR is usually more powerful. This may simply result from its using a less stringent criterion for declaring a gene to be significant. Ideally a procedure should be compared against the class of procedures that controls FDR at the same level. To compare the powers of the three multiple testing procedures when operating at the same FDR levels (and here at different thresholds α), we plot average power as a function of FDR. In Figure 5, we use data from the simulation studies with correlation structure $r_1 = 0.6$, $r_2 = -0.6$ for $m = 200$ and $m = 1000$. It appears that for each value of m_0/m the operating characteristics of the three procedures lie on a single curve. That is for the three testing procedures we considered in the simulation studies, the same power can be achieved if we are willing to tailor the threshold so the same level of FDR is achieved. Furthermore, the figure suggests that there may exist a fundamental relationship between FDR and average power for a given data structure, regardless of the multiple testing procedure that one chooses, or the total number of genes in the pool.

In summary, the simulation studies demonstrate that the three procedures differ in regards to their attained FDR when the same thresholds α is applied to each. However, they have the same trade-off

between increasing FDR and increasing average power, so in this sense they are not fundamentally different in their operating characteristics. Nevertheless in practice one needs to specify α and proceed with selecting genes accordingly. The adaptive procedure is attractive because it is least conservative (and hence more powerful) since it seeks to control FDR at a level close to α. However, the adaptive procedure assumes the same conditions as those of BH, and requires m_0/m be well estimated. Thus for some correlation structures we found that the actual FDR of the procedure can exceed the nominal level. This may be worrisome in applications. The new AP procedure does not make any distributional assumptions and is strikingly more powerful than the BH procedure. However, like the SAM procedure it is not designed to control the FDR and was observed to not control FDR when m_0/m is large. These results are encouraging.

4. Analysis of the Breast Cancer Data

We analyze a publicly available cDNA microarray dataset from a study of breast cancer reported by Van't Veer et al. (2002). The data consist of approximately 25,000 gene expression measurements from 44 cases found to have good prognosis cases and 34 who had a poor prognosis. The goal of the study is to identify a subset of genes that are predictive of the prognostic status of breast cancer patients. Although Van't Veer et al proceeded to combine data across genes for prediction, we are concerned here only with the first step to select a set of genes which are each associated with prognosis.

The gene expression measurement is the logarithm of the ratio of the intensities of the red to green fluorescent dyes, where green dye is used for the reference pool and red is used for the experimental tissue. In the study of Van't Veer et al. (2002), as a first step the authors selected some 5000 genes by applying gene filtering techniques that are described in the paper. To investigate properties of our new multiple testing procedure, we follow the same gene filtering procedure and obtain a sample of 4866 genes. We use the AUC test statistic to describe how well a gene discriminates those subjects that develop distant metastases within 5 years (poor prognosis status) from those who are disease free beyond 5 years (good prognosis status). Figure 6 displays the distribution of the AUCs for the 4866 genes. The AUC statistics for most of the genes are between 0.5 to 0.6, indicating that the majority of the genes are not differentially expressed.

We first calculate the unadjusted p-value for each gene using the Wilcoxon rank-sum test, the test

that uses the AUC value as its test statistic. 839 out of 4866 genes (17.24%) have a p-value less than 0.05, suggesting that the problem with multiple comparisons may be quite substantial here. If we perform the Bonferroni adjusted procedure, only two genes have an adjusted p-value less than 0.05. We performed the BH linear step-up procedure, the adaptive procedure and our new AP procedure. For the adaptive procedure, we obtain an estimated $m_0/m = 0.70$ with a smoothing method suggested by Storey(2003). For the AP method, the adjusted p-values are calculated based on 1000 random permutations. Figure 7 displays the p-values for the top 200 genes using the above multiple testing procedures along with the unadjusted p-values. When we choose to reject genes at the 0.05 level, four genes are rejected by the BH procedure, 7 genes are claimed as significant by both the AP and the adaptive procedures. When we choose to reject on the basis of $\alpha < 0.1$, we find that 133, 197 and 317 genes are selected by the BH, AP and the adaptive procedures respectively. These results are consistent with our numerical finding that the adaptive procedure is usually the most powerful procedure. Assuming that our simulation study results apply to this dataset, with $m_0/m = 0.70$, the FDR is controlled at level less than or equal to α for the BH or AP procedures, however for the adaptive procedure FDR could be potentially higher than α depending on the correlation structure of the data. This means that for the new procedure, among the 4866 genes we considered, on average at most 20 genes out of the 197 genes could be false positive. If we can afford the time and costs that are spent in vain on the 20 genes, we may benefit from studying a relatively bigger pool of potentially informative genes. In this particular dataset, our new algorithm appears to be effective.

We next compare the AP procedure with SAM. The SAM software allows one to interactively change Δ to control FDR if desired. Storey (2001) argues that a positive FDR, $pFDR = E\left[\frac{V(\Delta)}{R(\Delta)}|R(\Delta) > 0\right]$, may be a better quantity than Δ since it provides more meaningful interpretation. Corresponding to the pFDR, he suggests estimating q value, the probability that a null gene is true conditioning on observing a statistic as extreme or more extreme. We note that although our AP value and q-value have different interpretations, both are individual measures that take into account the problem with multiplicity, and both can be used to calibrate differential gene expression. We thus compare our AP values with q-values reported by the SAM software. To facilitate the comparison for a gene j we consider a statistic d_j, that is based on the two sample t-statistic, but with a small constant s_0

added to the denominator, following the SAM procedure. Among the top 200 genes ranked on $d_{(j)}$, $j = 1, \ldots, 200$, more significant genes will be identified if we choose to draw the line based on the q values of the SAM procedure, compared with the selection procedure using the same AP value (top panel Figure 8). On the other hand, the AP values appear to be more fine tuned with the test statistics. As can be seen in the bottom panel of Figure 8, the AP value (without the final monotonization) decreases gradually as the value of the test statistic increases. Unlike the q-value which assign equal values to SAM scores range 3.5 to 5, AP values acknowledge such differences and gives distinct p values to the scores in that spectrum. In summary, the SAM procedure appears to be more powerful as it discovers more significant genes given the same significant level. However, the AP procedure is also attractive as it corresponds more closely with the values of the statistics under consideration. Furthermore, the estimation procedure for AP values are simpler than that for the q-values which also rely on more assumptions about the dependence structures and require m_0/m and pFDR be well estimated.

5. Discussion

This manuscript concerns the issue of selecting a subset of genes that are differentially expressed. We propose a new approach that to deciding which genes to select for further study. Genes are ranked according to some statistic and the procedure dictates at which k to draw the line.Genes above k are pursued further. statistically more rigorously select top k genes. Similar to many of the existing multiple hypothesis testing procedures, we take into account the problem of multiplicity by calculating adjusted p-values for all genes simultaneously and reject genes if their adjusted p-values do not exceed a predetermined value α. Similar to SAM, these adjusted p-values are computed based on the distributions of order statistics under H_0^c.

A strength of the approach we have presented is that the methodology can accommodate many complications such as dependence amongst genes. Although the proposed method does not directly control FDR, our simulation studies show that rejection based on the new adjusted p-value method is as powerful as those methods that aim to control FDR, given the same FDR level. However, for many of the existing FDR controlling procedures to perform well, certain assumptions about correlation structures are needed. These assumptions may or may not hold in practice. Furthermore, although

the proposed method is in spirit similar to the SAM procedure, our procedure has intuitive appeal and it acknowledges that the distribution of jth order statistic may not be symmetric about its mean and that its variance may depend on the order (j). Indeed, our simulation study shows that for small sample sizes where SAM is expected to perform well, our proposed method is more powerful than SAM if the gene expression levels are not symmetrically distributed.

Our study also leads to some interesting findings on the operating characteristics of our new method and some existing multiple hypothesis testing procedures. In the diagnostic testing setting, it is well known that increasing TPR usually is accompanied by decreasing 1-FPR. A similar trade-off can be observed in the relationship between AP and FDR at least based on our simulation studies. With the emergence of many new methods for choosing rejection regions in a microarray study, it is important in our opinion to take into consideration this aspect of the operating characteristics when the performance of a new method is evaluated.

REFERENCES

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

Benjamini, Y., and Hochberg, Y. (2000), "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics [Formerly: @J(JEdStat)]*, 25(1), 60–83.

Benjamini, Y., and Yekutieli, D. (2001), "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, 29(4), 1165–1188.

Dudoit, S., Shaffer, J., and Boldrick, J. (2002), "Multiple hypothesis testing in microarray experiments," *U.C. Berkeley Division of Biostatistics Working Paper Series*, 110.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000), *Microarrays and their use in a comparative experiment* Technical report, Department of Statistics, Stanford University.

Pepe, M. S., Longton, G., Anderson, G. L., and Schummer, M. (2003), "Selecting differentially expressed genes from microarray experiments," *Biometrics*, 59, 133–142.

Storey, J. D. (2001), *The positive false discovery rate: A Bayesian interpretation and the q-value* Available at **http://www.stat.berkeley.edu/∼ storey/**.

Storey, J. D. (2002), "A direct approach to false discovery rates," *Journal of the Royal Statistical Society, Series B, Methodological*, 64(3), 479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2002), *A unified Estimation approach to false discovery rates* Available at **http://www.stat.berkeley.edu/∼ storey/**.

Storey, J. D., and Tibshirani, R. (2002), *SAM thresholding and False discovery rates for detecting differential gene expression in DNA microarrays* Available at **http://www.stat.berkeley.edu/∼ storey/**.

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

Van't Veer, L. J., Dai, H., Van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., Van de Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G., Kerkhoven, R. M., Roberts, C., Linsley, P., and Friend, S. H. (2002), "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, 415, 530–536.

Westfall, P. H., and Young, S. S. (1993), *Resampling-based multiple testing: examples and methods for P-value adjustment* John Wiley & Sons.

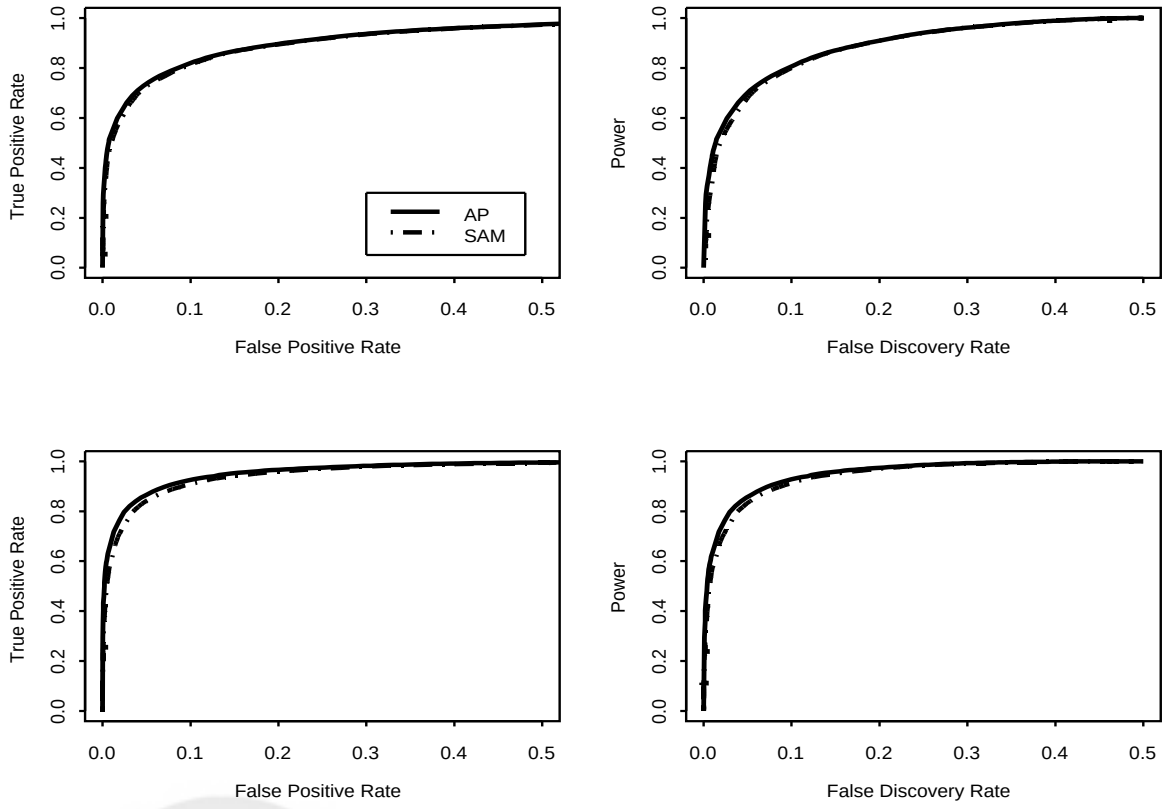

Figure 1: Simulation results for the AP and SAM procedures with expression levels from normal distributions. Top panels show data with $n_0 = n_1 = 20$, bottom panels show data with $n_0 = n_1 = 50$. Left panel shows data with ROC curves. Right panel shows Average power versus FDR.

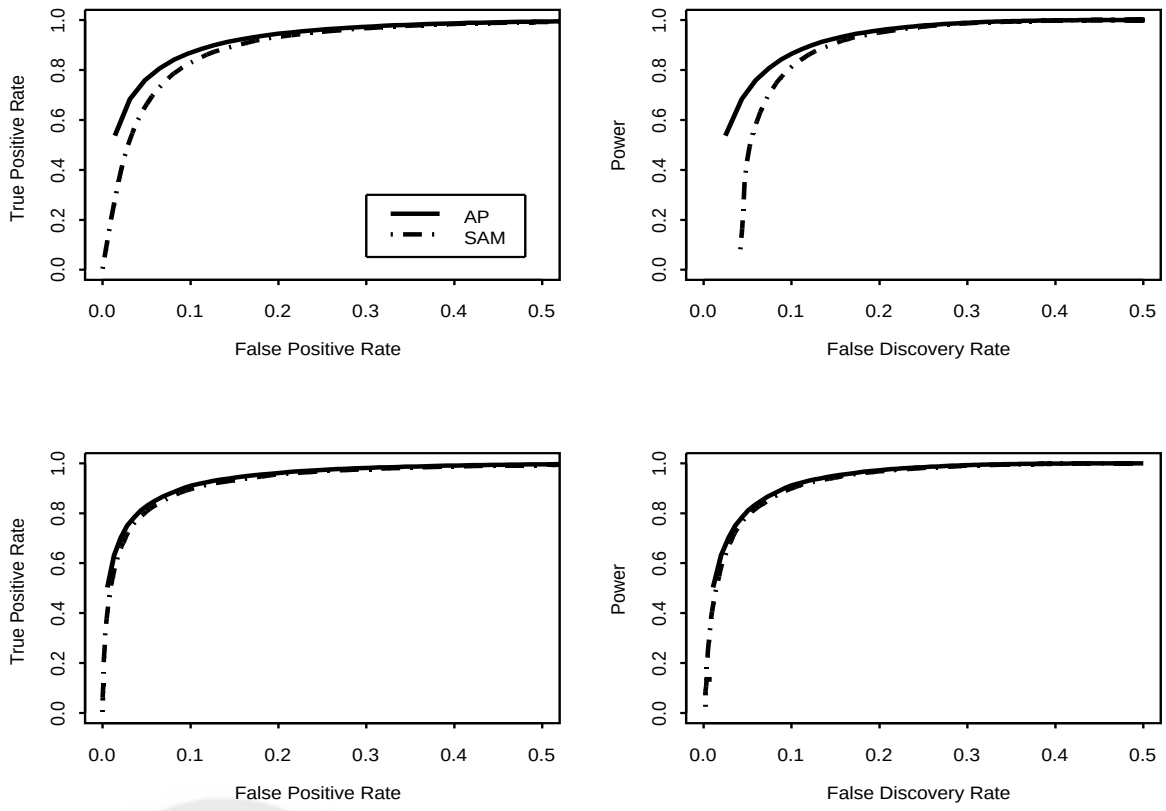

Figure 2: Simulation results for the AP and SAM procedures with expression levels from gamma distributions. Top panels show data with $n_0 = n_1 = 20$, bottom panels show data with $n_0 = n_1 = 50$. Left panel shows data with ROC curves. Right panel shows Average power versus FDR.

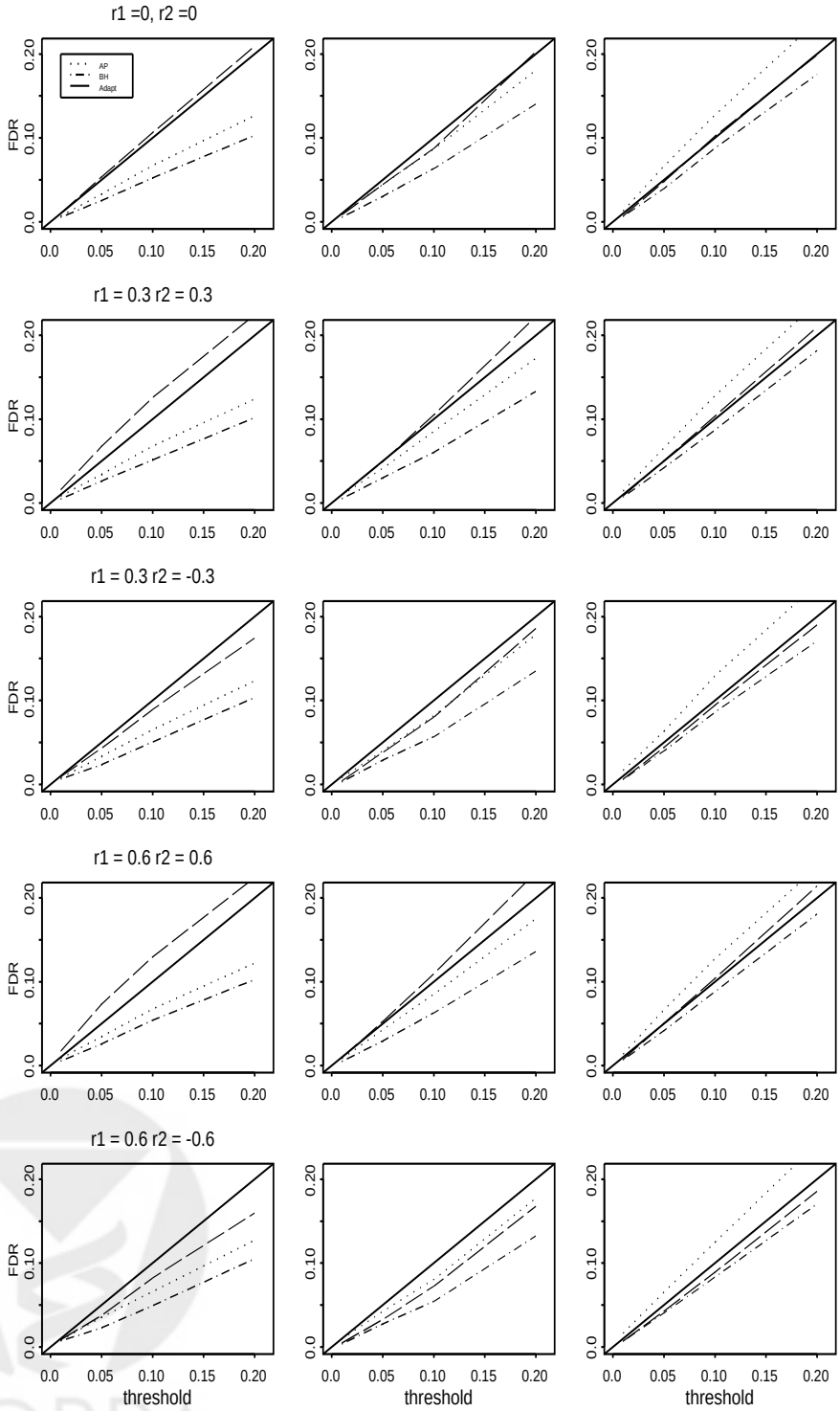

Figure 3: FDR versus α for m=40, $n_D = n_{\bar{D}} = 20$. Each row corresponds to a different correlation structure. The first column presents results from simulations with $m_0/m = 50\%$, the second column with $m_0/m = 75\%$, and the third column with $m_0/m = 90\%$. The solid diagonal line represents the benchmark where FDR is equal to the trhreshold α for declaring genes as differentially expressed using the multiple testing procedure.

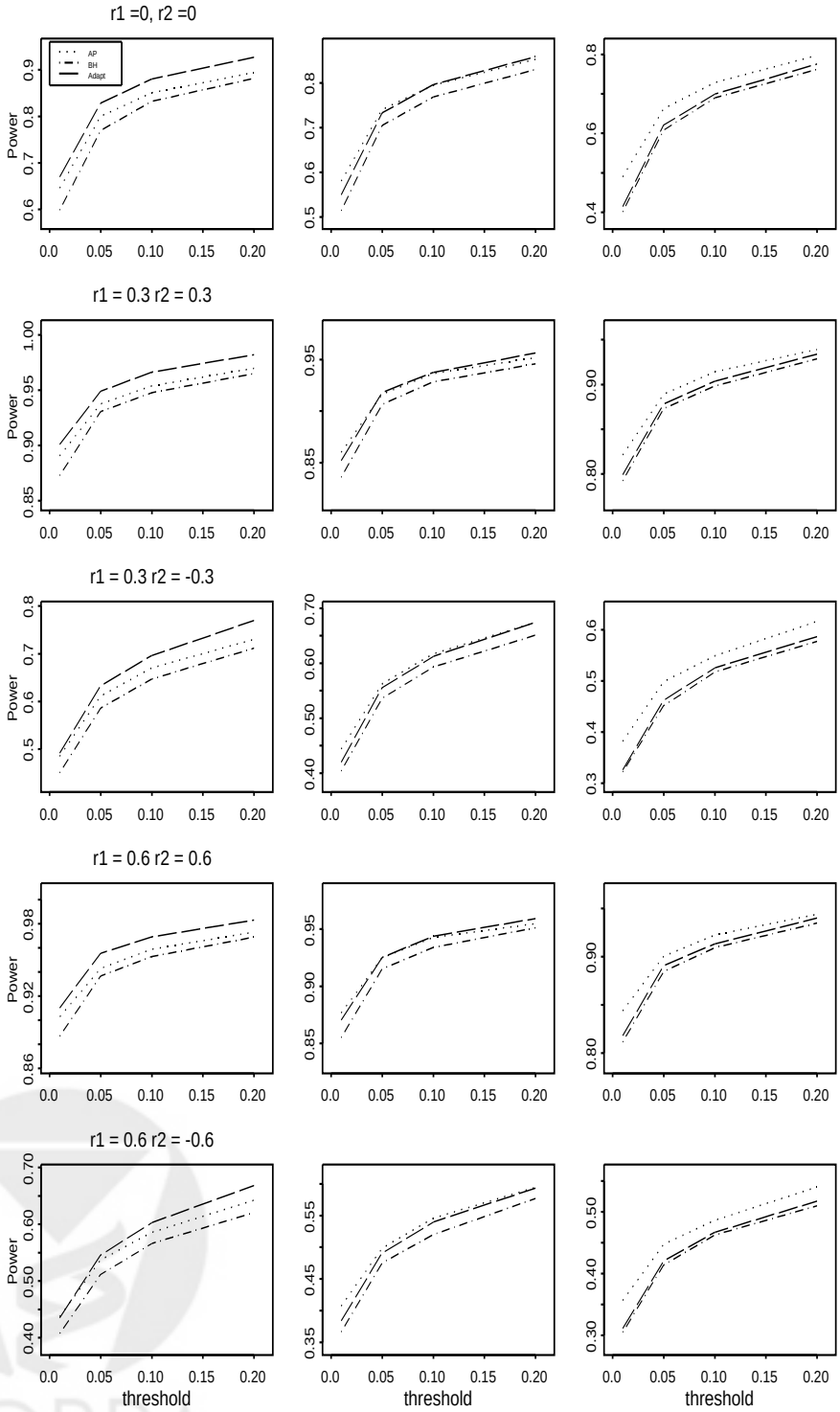

Figure 4: FDR versus α for m=40, $n_D = n_{\bar{D}} = 20$. Each row corresponds to a different correlation structure. The first column presents results from simulations with $m_0/m = 50\%$, the second column with $m_0/m = 75\%$, and the third column with $m_0/m = 90\%$.

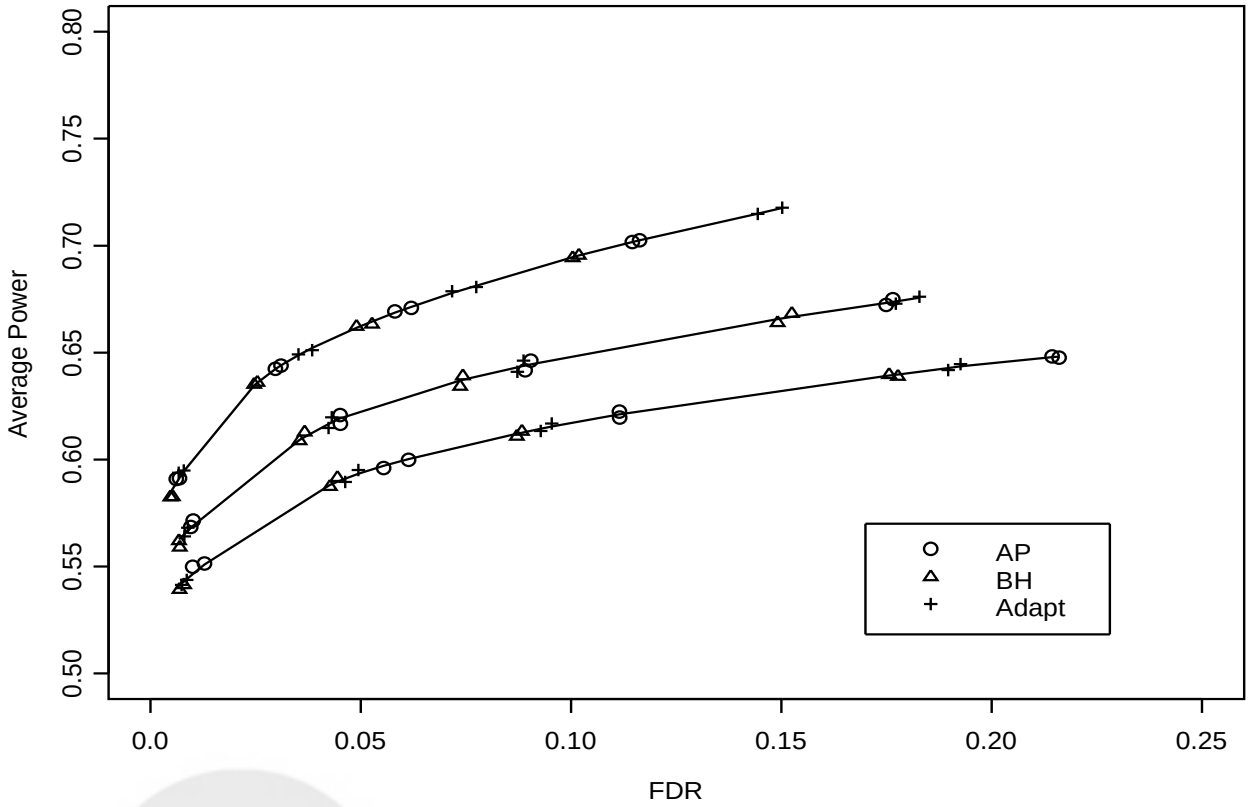

Figure 5: Average Power versus FDR from simulation studies with m=200 and m=1000. Separate lines are fitted with smoothing splines for $m_0 = 50\%m$(the line on top), $m_0 = 75\%m$(the line in the middle), and $m_0 = 90\%m$(the line at the bottom).

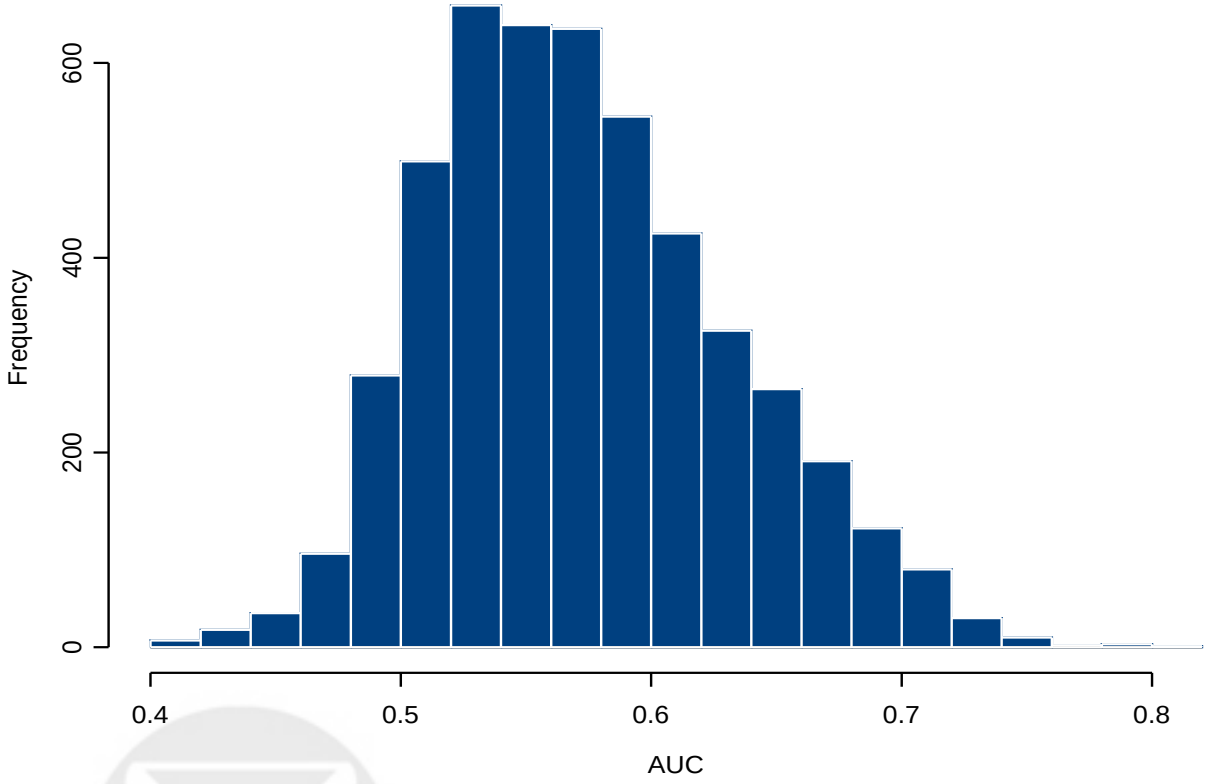

Figure 6: Histogram of AUC for 4866 genes from the breast cancer study

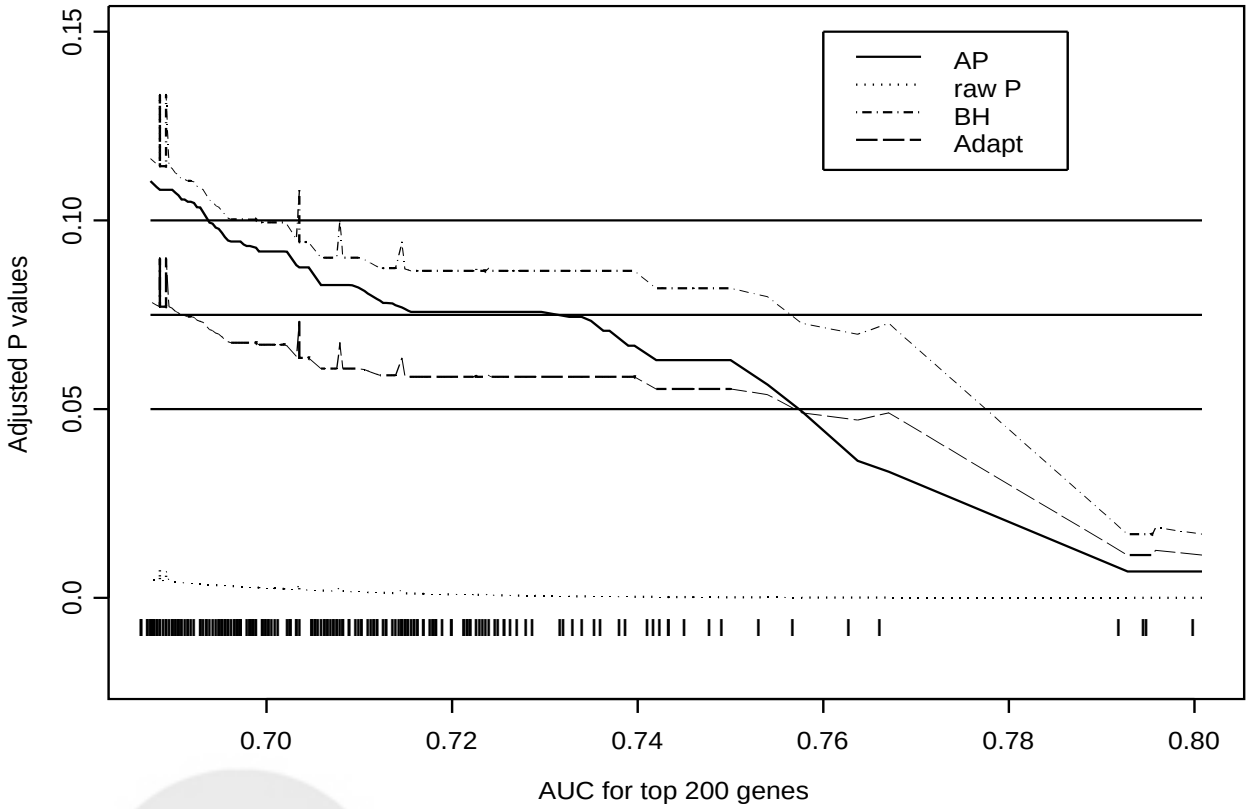

Figure 7: P values from different multiple testing procedures for the top 200 genes from the breast cancer study. The first horizontal solid line indicates reject region for $\alpha = 0.1$, the second line is for $\alpha = 0.075$, the third line is for $\alpha = 0.05$. |s represents the distribution of AUC.

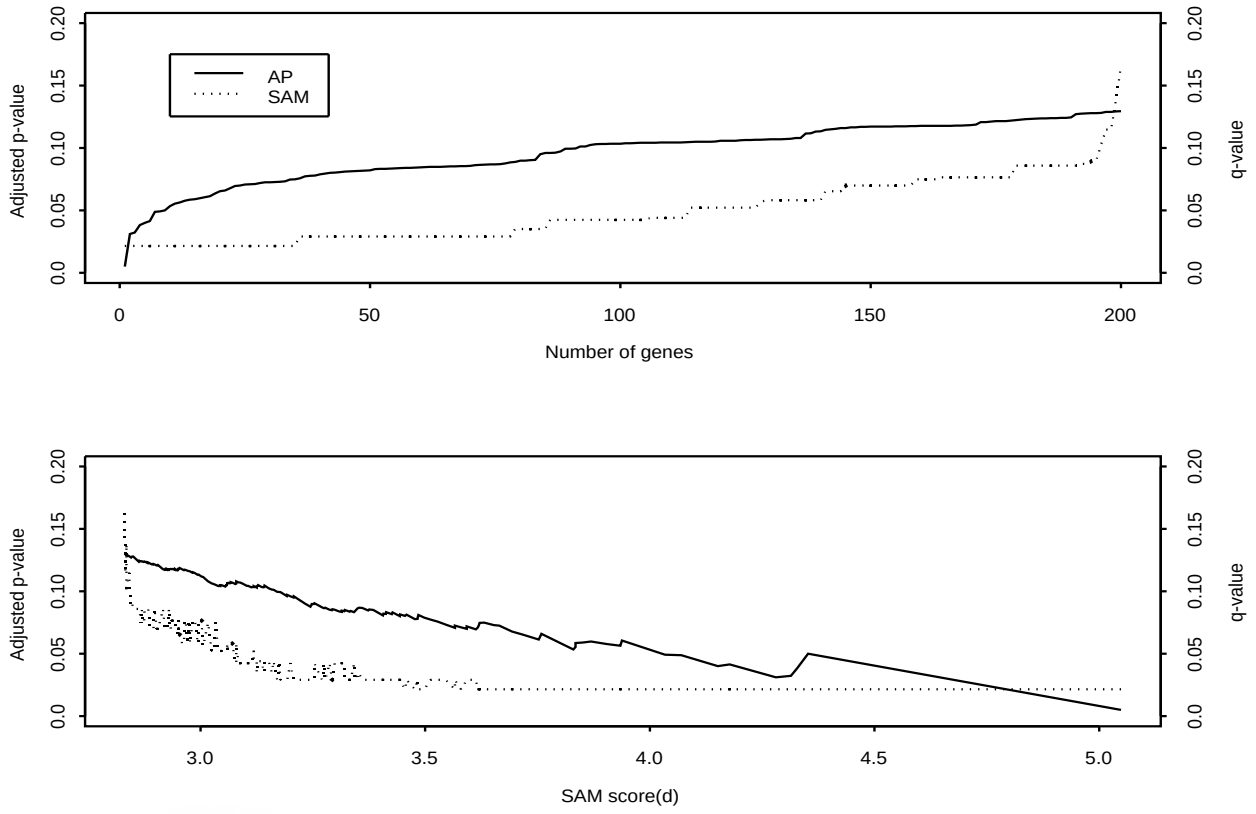

Figure 8: Comparison of the AP procedure with SAM using the breast cancer data. Top panel shows the p-values from the AP procedure (solid line) or the q-value from the SAM procedure (dotted line) for the top 200 genes. Bottom panel shows adjusted p-values/q values versus the test statistics (d-scores) from SAM for the top 200 genes.

Table 1: Outcomes from multiple tests with m genes

	# not rejected	#rejected	
# H_0	U	V	m_0
# H_a	T	S	m_1
total	m-R	R	m

Table 2: Simulation studies for m=200, r_1=0, r_2=0.

α	FDR			Average Power		
	AP	BH	Adapt	AP	BH	Adapt
	$m0 = 50\%m$					
0.01	0.007	0.005	0.010	0.866	0.859	0.877
0.05	0.032	0.027	0.051	0.907	0.902	0.919
0.1	0.061	0.052	0.100	0.923	0.920	0.936
0.2	0.116	0.103	0.199	0.940	0.937	0.955
	$m0 = 75\%m$					
0.01	0.010	0.008	0.010	0.846	0.836	0.845
0.05	0.045	0.037	0.049	0.889	0.883	0.891
0.1	0.090	0.074	0.100	0.908	0.903	0.910
0.2	0.172	0.149	0.198	0.927	0.923	0.931
	$m0 = 90\%m$					
0.01	0.013	0.006	0.007	0.489	0.401	0.414
0.05	0.066	0.039	0.048	0.662	0.608	0.621
0.1	0.127	0.089	0.102	0.728	0.689	0.699
0.2	0.239	0.177	0.199	0.797	0.762	0.775

Table 3: Simulation studies for m=200, r_1=0.6, r_2=0.6.

α	FDR			Average Power		
	AP	BH	Adapt	AP	BH	Adapt
	$m0 = 50\%m$					
0.01	0.007	0.006	0.012	0.964	0.963	0.968
0.05	0.031	0.026	0.057	0.976	0.975	0.980
0.1	0.061	0.0527	0.111	0.981	0.980	0.985
0.2	0.117	0.104	0.214	0.986	0.985	0.990
	$m0 = 75\%m$					
0.01	0.009	0.007	0.010	0.959	0.957	0.959
0.05	0.044	0.036	0.051	0.973	0.971	0.974
0.1	0.087	0.071	0.103	0.978	0.976	0.979
0.2	0.172	0.148	0.206	0.983	0.982	0.984
	$m0 = 90\%m$					
0.01	0.014	0.007	0.008	0.844	0.811	0.818
0.05	0.068	0.042	0.051	0.900	0.885	0.891
0.1	0.128	0.088	0.104	0.923	0.909	0.913
0.2	0.237	0.181	0.214	0.944	0.935	0.940

Table 4: Simulation studies for m=200, r_1=0.6, r_2=-0.6.

α	FDR			Average Power		
	AP	BH	Adapt	AP	BH	Adapt
	$m0 = 50\%m$					
0.01	0.007	0.005	0.008	0.591	0.583	0.595
0.05	0.031	0.025	0.038	0.644	0.636	0.651
0.1	0.062	0.053	0.078	0.671	0.664	0.681
0.2	0.116	0.102	0.150	0.703	0.696	0.718
	$m0 = 75\%m$					
0.01	0.010	0.007	0.009	0.572	0.562	0.568
0.05	0.045	0.037	0.043	0.621	0.613	0.620
0.1	0.090	0.074	0.089	0.646	0.639	0.646
0.2	0.177	0.153	0.183	0.675	0.669	0.676
	$m0 = 90\%m$					
0.01	0.016	0.007	0.007	0.356	0.305	0.311
0.05	0.065	0.041	0.043	0.447	0.414	0.420
0.1	0.125	0.083	0.089	0.486	0.462	0.467
0.2	0.242	0.170	0.185	0.540	0.510	0.517