

# Some Comments on Instability of False Discovery Rate Estimation

Xing Qiu

Department of Biostatistics and Computational Biology, University of Rochester, 601  
Elmwood Avenue, Box 630, Rochester, New York 14642  
E-mail: [xqiu@bst.rochester.edu](mailto:xqiu@bst.rochester.edu)

Andrei Yakovlev

Department of Biostatistics and Computational Biology, University of Rochester, 601  
Elmwood Avenue, Box 630, Rochester, New York 14642  
E-mail: [Andrei.Yakovlev@urmc.rochester.edu](mailto:Andrei.Yakovlev@urmc.rochester.edu)

## Abstract

Some extended false discovery rate (FDR) controlling multiple testing procedures rely heavily on empirical estimates of the FDR constructed from gene expression data. Such estimates are also used as performance indicators when comparing different methods for microarray data analysis. The present communication shows that the variance of the proposed estimators may be intolerably high, the correlation structure of microarray data being the main cause of their instability.

## 1. Introduction

Since the seminal paper by Benjamini and Hochberg<sup>1</sup> was published in 1995, multiple testing procedures designed to control the false discovery rate (FDR) have been gaining in popularity. The original approach to designing an appropriate FDR controlling procedure was to find a data-driven thresholding rule for observed  $p$ -values such that the resultant FDR does not exceed a prechosen FDR level.<sup>1-4</sup> The second approach, taken by Storey,<sup>5</sup> is to fix the thresholding rule, and to find an estimate,  $\widehat{FDR}$ , for the FDR such that its expectation is greater than or equal to the true FDR over the prechosen rejection region. Storey et al.<sup>6</sup> developed a unifying framework by proposing a family of FDR controlling procedures arising from a family of conservatively biased point estimates of the FDR for a fixed rejection region. The authors argue that the goals of the two approaches are met with the proposed family. The main fact forming the basis of the

general paradigm proposed by Storey et al.<sup>6</sup> is that their estimator is biased up whenever the  $p$ -values corresponding to the true null hypotheses are independent, which guarantees control of the FDR “on average” under such conditions.

The FDR estimator introduced by Storey<sup>5</sup> and explored further by Storey et al.<sup>6</sup> and Storey and Tibshirani<sup>7</sup> includes an estimate,  $\hat{\pi}_0$ , of the expected proportion of true null hypotheses  $\pi_0$ . This element of FDR controlling procedures is believed to be beneficial as it leads to an increase in the overall “average power”. The same idea lies at the heart of adaptive counterparts of the original Benjamini-Hochberg step-up procedure.<sup>8,9</sup>

It is clear that estimation and control of the FDR are closely intertwined in modern approaches to the problem of large-scale multiple testing. However, it remains unknown to what extent the estimators for the FDR (and  $\pi_0$ ) underlying the proposed FDR controlling procedures are affected by the actual dependence structure of microarray data.

To the best of our knowledge, the majority of papers dealing with the concept of FDR have been focused solely on the expected values of the proposed estimators, leaving their higher order properties out of consideration. This view of the problem is far short of complete. Especially important is the variance of the estimated FDR because of its intimate connection to the accuracy with which the Type 1 errors are controlled. If this variance is high, one should also expect a high variance of the total number of rejected hypotheses, which is an important observable indicator of the performance of a given testing procedure. Qiu et al.<sup>10</sup> demonstrated that the empirical Bayes method for finding and ranking of differentially expressed genes is susceptible to this effect because of strong correlations between gene expression levels, and consequently between test-statistics (and  $p$ -values!) associated with different genes.<sup>11</sup> Their results also indicate that the variance of Storey’s estimator may be quite high in applications to microarray data analysis. While normalization procedures tend to destroy the correlation between test-statistics<sup>11</sup>, their application does not remedy the problem because of certain side effects. In particular, normalization procedures interfere in the true biological signal and reduce its variance, thereby inducing additional false discoveries. Quantitative insight into dependencies between genes represents one of the main challenges in microarray data analysis. Pavlidis et al.<sup>12</sup> and Qiu et al.<sup>13</sup> proposed resampling tools for assessing stability of gene selection procedures in the presence of correlations between gene expression levels.

This note is intended to demonstrate that the nature of microarray data is not conducive to the wide use of the multiple testing procedures based on the present-day estimators for the FDR because such estimators tend to have a high variance. The high variability of the FDR estimates tends to increase instability of the corresponding gene selection procedures.

## 2. Study Design

### 2.1. Computer simulations

We generated twelve sets of simulated data to study the performance of FDR estimators. Each set consists of 1000 independently generated pairs of samples of different sizes. Each sample includes  $n$  realizations of a random vector  $\mathbf{X}$  with log-normal marginal distributions. The components of  $\mathbf{X}$  represent expression levels of  $m = 1255$  genes, while each realization of  $\mathbf{X}$  represents a single array. To model the presence of differentially expressed genes in two-sample comparisons, the mean log-expression of the first 125 genes is set to be equal to 2 in one sample and to 0 in another. The variance of log-expressions is kept equal to 1 in both samples. The log-expressions of the remaining 1130 genes in both samples are generated from a standard normal distribution. All the samples under each comparison are of equal sizes. The matrix  $[x_{ij}]$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, n$ , represents the data for each group of  $n$  arrays.

Particular features of the first six sets of simulated data are given below.

1. **SIM43**: Each of the two samples under comparison includes  $n = 43$  arrays. For each array, the logarithms of expression levels  $x_{ij}$  of all genes are generated as independent normally distributed random variables. The mean values and variances of the marginal distributions for “different” and “not different” genes are specified as described above. A total of 1000 pairs of such samples comprise **SIM43**.
2. **SIM43CORR**: This data set is generated with the intention to demonstrate the role of correlation between expression levels on the estimated FDR. To this end, each sample consisting of  $n = 43$  arrays is generated from a joint log-normal distribution with exchangeable correlation structure as described in Qiu et al.<sup>10</sup> More specifically, we first generate a set of random variables  $y_{ij}$ ,  $i = 1, \dots, m$ ,  $j = 1, \dots, n$ , in exactly the same way as **SIM43**. We next generate an  $n$ -dimensional random vector with independent and identically distributed components, each component having a standard normal distribution. Denoting this vector by  $A = \{a_j\}$ ,  $j = 1, \dots, n$ , we define  $x_{ij} = \sqrt{\rho}a_j + \sqrt{1 - \rho}y_{ij}$ ,  $i = 1, \dots, m$ ;  $j = 1, \dots, n$ , so that for any  $i_1 \neq i_2$  and  $j$  we have all pairwise correlation coefficients  $\text{Corr}(x_{i_1j}, x_{i_2j}) = \rho$ . Once a value of  $\rho$  has been chosen, the process is repeated 1000 times with different random seeds. Our simulations were conducted with  $\rho = 0.4$ . Since our focus is only on proof of principle, the choice of  $\rho$  is of little consequence to the objectives of this study and exploring other values of  $\rho$  would serve no specific purpose. The marginal distributions of log-expressions are identical to those for the set **SIM43**.
3. **SIM15**: This data set is generated in just the same way as **SIM43** but the size of each of the two samples is equal to 15.

4. **SIM15CORR**: This data set is generated in just the same way as **SIM43CORR** but the size of each of the two samples is equal to 15.
5. Two additional data sets (**SIM5** and **SIM5CORR**) with  $n = 5$  were generated to represent an extreme case of a very small sample size.

To study the effect of the proportion of true alternative hypotheses we produced six counterparts of all the above described data sets with only 50 “different” genes. These data sets are denoted by **SIM43S**, **SIM15S**, **SIM5S**, **SIM43SCORR**, **SIM15SCORR**, **SIM5SCORR**, respectively.

## 2.2. Biological Data

For the purposes of this study, use was made of the St. Jude Children’s Research Hospital (SJCRH) Database on childhood leukemia.<sup>14</sup> This data set is publicly available on the following website: <http://www.stjude.com/research/data/ALL1>. The whole SJCRH Database contains gene expression data on 335 subjects, each represented by a separate array (Affymetrix, Santa Clara, CA) reporting measurements on the same set of  $m = 12558$  genes. We selected two groups of patients with hyperdiploid (**Hyperdip**) and T-cell acute lymphoblastic leukemia (**TALL**), respectively. The groups were balanced to include  $n = 43$  patients in each group. Since the nature of our study was purely methodological, the choice of the data set was quite arbitrary; it was dictated solely by sample size considerations.

We used a subsampling variant of the delete- $d$ -jackknife method<sup>15,16</sup> to estimate the variance of two specific estimators of the FDR (see Section 3). Let  $T(\mathbf{X})$  be the statistic of interest, where  $\mathbf{X}$  is the vector of observed expression measures. Using the delete- $d$ -jackknife method, one repeatedly computes  $T(\mathbf{X})$  after leaving out  $d$  vector-valued observations. Under this resampling scheme, the statistics are of the form:  $T_{r,\mathbf{s}}(X_j, j \in \mathbf{s}^c)$ , where  $\mathbf{s}$  is a subset of  $\{1, \dots, n\}$  of size  $d$ ,  $\mathbf{s}^c$  is the complement of  $\mathbf{s}$ ,  $d$  is an integer that depends on the sample size  $n$  ( $1 \leq d \leq n$ ), and  $r = n - d$ . Unlike the bootstrap, the delete- $d$ -jackknife is asymptotically valid without any smoothness requirements on the statistic  $T(\mathbf{X})$  whose sampling distribution we want to estimate.<sup>17</sup> In accordance with currently available recommendations,<sup>15,16</sup> we set  $d = 7$  which amounts to leaving out 7 arrays in each of the two groups under study at each resampling step. A total of  $B = 200$  subsamples were drawn without replacement to compute the following resampling counterpart of the delete- $d$ -jackknife variance:<sup>16</sup>

$$\text{Var}(Q^*) = \frac{n-d}{dB} \sum_{l=1}^B \left( Q_{n-d,l}^* - \frac{1}{B} \sum_{k=1}^B Q_{n-d,k}^* \right)^2,$$

where  $Q_{n-d,j}^*$  is an estimate of the FDR derived from the  $j$ -th delete- $d$  jackknife subsample. Two different forms of  $Q^*$  are considered in the next section.

### 3. Estimating the False Discovery Rate from Microarray Data

We tested two methods of FDR estimation. The first method was proposed Storey<sup>5</sup> and studied more thoroughly by Storey et al.<sup>6</sup> and Storey and Tibshirani.<sup>7</sup> In what follows this method will be referred to as the Storey-Tibshirani estimator (STE). The second estimator was proposed by Yekutieli and Benjamini<sup>4</sup> and discussed later by Reiner et al.<sup>9</sup> in the context of microarray data analysis. This estimator will be referred to as the Yekutieli-Benjamini estimator (YBE).

#### 3.1. The Storey-Tibshirani Estimator

The STE has the following form:

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 \cdot m \cdot t}{\#\{p_i \leq t\}}, \quad (1)$$

where  $m$  is the total number of hypotheses (genes),  $p_i$  is the  $p$ -value of the  $i$ -th test,  $t$  is a preset threshold for raw  $p$ -values and  $\hat{\pi}_0$  is an estimator of  $\pi_0$ , the latter being defined as the expected proportion of true null hypotheses  $m_0$  among the  $m$  hypotheses tested. As suggested by Storey and Tibshirani,<sup>7</sup> the estimator  $\hat{\pi}_0$  is constructed as follows:

1. For a range of  $\lambda$ , say  $0, 0.01, 0.02, \dots, 0.95$ , calculate

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1 - \lambda)}, \quad (2)$$

where  $\lambda$  is a tuning parameter.

2. Let  $\hat{f}$  be the natural cubic spline fit to  $\hat{\pi}_0(\lambda)$  as a function of  $\lambda$ .
3. Set the estimate of  $\pi_0$  to be  $\hat{\pi}_0 = \hat{f}(1)$ .

The estimator given by (1) is based on the assumption that the observed  $p$ -values for the true null hypotheses are uniformly distributed. This is definitely a valid assumption for each individual gene as long as the  $p$ -values computed for a given two-sample test under the null hypothesis are correct. A pitfall here is that the  $p$ -values associated with different genes are used to estimate  $\pi_0$  and  $FDR(t)$ . These  $p$ -values are heavily dependent random variables. Furthermore, our previous study of the correlation structure of  $t$ -statistics associated with different genes indicates the presence of a “long-range” correlation involving thousands of genes.<sup>11</sup> The same applies equally to the corresponding  $p$ -values. This may make the estimators  $\hat{\pi}_0$  and  $\#\{p_i \leq t\}$  highly sensitive to sample fluctuations, resulting in a high variance of  $\widehat{FDR}(t)$ .

### 3.2. The Yekutieli-Benjamini Estimator

Let  $t$  be a prechosen threshold for  $p$ -values and  $R(t)$  the corresponding number of rejected hypotheses. The total number of hypotheses to be tested is denoted by  $m$ . Let  $V(t)$  be the number of rejections (false discoveries) under the complete null hypothesis. The null hypothesis is modeled through permutations to generate values of  $V(t)$ . Let  $v_\beta$  denote the  $\beta$ -quantile of the permutation distribution of  $V(t)$ . The YBE is constructed as follows:

$$\widetilde{FDR}(t) = \begin{cases} \mathcal{E}_{V(t)} \left\{ \frac{V(t)}{V(t)+R(t)-tm} \right\} & \text{if } R(t) - v_\beta(t) \geq tm, \\ \mathcal{E}_{V(t)} \{ I(R(t) \geq 1) \} & \text{otherwise,} \end{cases} \quad (3)$$

where  $\mathcal{E}_{V(t)}$  is the empirical expectation over the permutations and  $I\{\cdot\}$  is the indicator function.

In formula (3), the term  $R(t) - tm$  mimics the number of false discoveries  $S(t)$ . Yekutieli and Benjamini<sup>4</sup> treat this term as a downward biased estimator for  $\mathbb{E}\{S(t)\}$  proceeding from the following argument:

$$\mathbb{E}\{R(t) - mt\} \leq \mathbb{E}\{S(t)\} + \mathbb{E}\{V(t)\} - m_0t = \mathbb{E}\{S(t)\},$$

where  $m_0$  is the number of true null hypotheses.

The algorithm given by formula (3) is quite time-and-space consuming and we used 100 permutations in its construction from each subsample. The level  $\beta = 0.95$  was used in our simulations reported in the next section.

### 3.3. Simulation Studies and Analysis of Biological Data

The above-described estimators were used in conjunction with Student's  $t$ -test which is the most popular choice in microarray studies. In our simulation studies, the FDR was also estimated nonparametrically as the ratio of the number of true rejections (discoveries) to the total number of rejected hypotheses in each sample. This estimator is referred to as the NPE. Since the NPE is unbiased, it can be used to make an approximate assessment of the biases inherent in the STE and YBE.

Tables 1 displays the results of FDR estimation from 1000 sets of the simulated data SIM15 (see Section 2.1) using all the three estimation methods at different threshold levels. Since the NPE is expected to have a high variance and is not our focus here, we present only the estimated mean values of this estimator. When compared with the NPE, the STE and YBE tend to be biased upwards, which is expected by their design. Their standard deviations are quite small for all the values of  $t$  in Table 1. The standard deviations of  $\widetilde{FDR}(t)$  increase dramatically when correlation is introduced as shown in Table 2. This appears to be even more so for the estimator  $\widetilde{FDR}(t)$ . The situation does not change much when the number of arrays

Table 1: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for independent expression levels and  $n = 15$  (SIM15).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0009	0.0011	0.0009	0.0004	0.0001
2.1	0.0156	0.0171	0.0154	0.0014	0.0014
4.1	0.0292	0.0323	0.0292	0.0021	0.0026
6.1	0.0420	0.0470	0.0426	0.0025	0.0039
8.1	0.0549	0.0613	0.0556	0.0030	0.0051
10.1	0.0675	0.0752	0.0683	0.0034	0.0063
12.1	0.0797	0.0887	0.0806	0.0038	0.0075
14.1	0.0916	0.1019	0.0926	0.0042	0.0087
16.1	0.1036	0.1147	0.1043	0.0048	0.0098
18.1	0.1151	0.1272	0.1157	0.0053	0.0109
20.1	0.1261	0.1394	0.1268	0.0056	0.0120

Table 2: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for correlated expression levels and  $n = 15$  (SIM15CORR).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0014	0.0011	0.0010	0.0008	0.0004
2.1	0.0150	0.0153	0.0155	0.0039	0.0052
4.1	0.0261	0.0280	0.0296	0.0059	0.0103
6.1	0.0367	0.0400	0.0436	0.0077	0.0155
8.1	0.0471	0.0515	0.0573	0.0093	0.0208
10.1	0.0569	0.0633	0.0709	0.0269	0.0261
12.1	0.0664	0.0750	0.0843	0.0383	0.0315
14.1	0.0754	0.0895	0.0976	0.0676	0.0369
16.1	0.0842	0.1063	0.1106	0.0999	0.0424
18.1	0.0930	0.1334	0.1234	0.1532	0.0479
20.1	0.1016	0.1751	0.1361	0.2199	0.0534

Table 3: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for independent expression levels and  $n = 43$  (SIM43).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0007	0.0010	0.0007	0.0003	0.0001
2.1	0.0143	0.0173	0.0149	0.0011	0.0013
4.1	0.0282	0.0323	0.0286	0.0015	0.0026
6.1	0.0416	0.0468	0.0420	0.0019	0.0038
8.1	0.0542	0.0609	0.0550	0.0022	0.0050
10.1	0.0669	0.0745	0.0677	0.0027	0.0061
12.1	0.0791	0.0879	0.0800	0.0031	0.0073
14.1	0.0911	0.1008	0.0921	0.0035	0.0085
16.1	0.1023	0.1136	0.1038	0.0040	0.0096
18.1	0.1140	0.1259	0.1152	0.0045	0.0107
20.1	0.1251	0.1380	0.1263	0.0049	0.0118

Table 4: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for correlated expression levels and  $n = 43$  (SIM43CORR).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0007	0.0009	0.0007	0.0007	0.0002
2.1	0.0120	0.0149	0.0153	0.0041	0.0050
4.1	0.0220	0.0272	0.0297	0.0061	0.0098
6.1	0.0318	0.0388	0.0439	0.0078	0.0148
8.1	0.0417	0.0505	0.0579	0.0220	0.0198
10.1	0.0510	0.0613	0.0717	0.0240	0.0249
12.1	0.0599	0.0718	0.0854	0.0261	0.0300
14.1	0.0686	0.0862	0.0988	0.0652	0.0352
16.1	0.0767	0.1072	0.1121	0.1167	0.0404
18.1	0.0848	0.1378	0.1252	0.1743	0.0456
20.1	0.0927	0.1747	0.1382	0.2271	0.0509

per group is increased from 15 to 43 for both the independent and correlated data (see Tables 3 and 4). The same tendencies were observed when the number of true alternative hypotheses was decreased from 125 to 50 (Tables 1A - 4A in the Supporting Material). The standard deviation of both estimates typically increases with decreasing the number of true alternatives.

The tendency towards higher instability of the FDR estimates in the presence of correlations can be seen even with a sample size as small as  $n = 5$  (Tables 5 and 6). The picture appears more blurred with fewer “different genes” as can be seen when comparing Tables 5 and 6 with their counterparts (Tables 5A and 6A) in the Supporting Material. This is attributable to much more variability in all the estimates presented in Tables 5A and 6A (Supporting Material) under

Table 5: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for independent expression levels and  $n = 5$  (SIM5).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0395	0.0442	0.0406	0.0383	0.0273
2.1	0.0745	0.0834	0.0767	0.0175	0.0169
4.1	0.0959	0.1079	0.0984	0.0171	0.0170
6.1	0.1141	0.1274	0.1160	0.0174	0.0179
8.1	0.1298	0.1443	0.1313	0.0175	0.0191
10.1	0.1443	0.1598	0.1452	0.0178	0.0200
12.1	0.1571	0.1742	0.1583	0.0181	0.0211
14.1	0.1691	0.1876	0.1705	0.0183	0.0222
16.1	0.1805	0.2001	0.1819	0.0182	0.0230
18.1	0.1916	0.2123	0.1930	0.0184	0.0241
20.1	0.2025	0.2240	0.2036	0.0187	0.0250

Table 6: Mean values and standard deviations for different FDR estimators and threshold values. Simulated data for correlated expression levels and  $n = 5$  (SIM5CORR).

$t \times 1255$	NPE/mean	YBE/mean	STE/mean	YBE/std	STE/std
0.1	0.0283	0.0471	0.0330	0.0375	0.0387
2.1	0.0820	0.1813	0.1608	0.1937	0.2553
4.1	0.0967	0.2469	0.1721	0.2624	0.2219
6.1	0.1097	0.3014	0.1806	0.3066	0.1994
8.1	0.1202	0.3554	0.1910	0.3419	0.1811
10.1	0.1289	0.4008	0.2068	0.3669	0.2381
12.1	0.1404	0.4409	0.2147	0.3826	0.1920
14.1	0.1492	0.4772	0.2255	0.3930	0.1850
16.1	0.1573	0.5224	0.2360	0.4016	0.1781
18.1	0.1648	0.5550	0.2455	0.4047	0.1734
20.1	0.1736	0.5922	0.2557	0.4053	0.1752

such conditions. Especially unreliable are the estimates for the smallest values of  $t$  because the corresponding numbers of rejections are too small.

Recall that the correlation coefficient of 0.4 was chosen to model the exchangeable correlation structure of simulated data. The effects of correlation in biological data may be even stronger. In this connection it is worth recalling some facts about the correlation structure of microarray gene expression data. When estimating pairwise correlation coefficients in all gene pairs from the **Hyperdip** data we found<sup>18</sup> that 26.3% of the pairs have their correlation coefficients in the range 0.5-0.75, while 71.4% of the pairs are characterized by correlation coefficients greater than 0.75. Furthermore, the characteristic property of the correlation structure of gene expression

Table 7: Mean values and standard deviations for the STE at different threshold values. The results are obtained by resampling from the TALL-Hyperdip data.

$t \times 12558$	STE/mean	STE/std
0.1	0.0001	0.0001
2.1	0.0021	0.0011
4.1	0.0037	0.0020
6.1	0.0052	0.0029
8.1	0.0066	0.0036
10.1	0.0079	0.0044
12.1	0.0092	0.0051
14.1	0.0104	0.0058
16.1	0.0116	0.0065
18.1	0.0127	0.0072
20.1	0.0139	0.0079

data is a “long-range” correlation manifesting itself in thousands of tightly dependent expression measures.<sup>11,19</sup> Since our simulations suggest that the YBE is less stable than the STE, we limited ourselves to the latter estimator in our analysis of the biological data described in Section 2.2. It comes as no surprise that the resampling procedure of Section 2.2 applied to the Hyperdip-TALL comparison results in high standard deviations as shown in Table 7. These coefficients tend to be even higher than those reported for the simulated data.

## 4. Discussion

The results of Section 3.3 indicate that the strong correlation between gene expression levels makes the estimates of the FDR highly unstable. In our opinion, the variance of such estimates is intolerably high. Quantitative assessment of the effect of this variability on the performance of the FDR controlling procedures that are built on the proposed estimators invites a special investigation. However, these estimators are also used for other purposes as estimators *per se*. For example, the estimated FDR served as a performance indicator in a comparative study of several methods for producing Affymetrix expression scores.<sup>20</sup> Great caution must be exercised in using the STE and YBE for this or similar purposes in view of their biasedness and variability.

From our simulations, it appears that the STE is more stable than the YBE in the presence of correlations. The severely limited number of permutations used in our study may provide part of the explanation. More computer intensive studies are needed to add clarity to the issue.

One source of variability of the STE is the plug-in estimator  $\hat{\pi}_0$ . This estimator is constructed by smoothing the underlying histogram of  $p$ -values. However, the  $p$ -values associated with different genes are heavily dependent random variables, resulting in high variations of the

histogram from sample to sample. Figure 1 shows two sample realizations of the histogram in the simulated data SIM43CORR. The shape of the histogram varies dramatically between the two samples. Two sample histograms obtained by resampling from the biological data, shown in Figure 2, are also very dissimilar, thereby indicating that the variance of  $\hat{\pi}_0$  may be quite high. It is rather common in individual samples that  $\hat{\pi}_0$  is either greater than 1 or less than 0. It seems reasonable to truncate the estimator at both extreme levels. However, while reducing the variance, this expedient introduces an additional bias and it is unclear how a reasonable compromise can be reached in this trade-off. Even the truncated version of  $\hat{\pi}_0$  appears to be highly unstable.

Both estimators  $\widehat{FDR}(t)$  and  $\widetilde{FDR}(t)$  studied in this paper perform quite well in the case of independent expression levels (Tables 1 and 3) but their performance in terms of their stability is deteriorated in the presence of high and “long-range” correlations (Tables 2, 4, and 7). This effect should always be borne in mind when designing statistical methods for microarray data analysis.

## Acknowledgements

The research of A. Yakovlev is supported in part by NIH/NIGMS Grant GM075299. We would like to express our gratitude to the anonymous reviewers for their thoughtful suggestions.

## References

1. Benjamini Y, Hochberg Y, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B* **57**: 289-300, 1995.
2. Benjamini Y, Liu W, A step-down multiple hypothesis procedure that controls the false discovery rate under independence, *J. Statist. Planning Inf.* **82**: 163-170, 1999.
3. Benjamini Y, Yekutieli D, The control of the false discovery rate in multiple testing under dependency, *Ann. Statist.* **29**: 1165-1188, 2001.
4. Yekutieli D, Benjamini Y, Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics, *J. Statist. Planning Inf.* **82**: 171-196, 1999.
5. Storey JD, A direct approach to false discovery rates, *J. R. Statist. Soc. B* **64**: 479-498, 2002.
6. Storey JD, Taylor JE, Siegmund D, Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B* **66**: 187-205, 2004.

7. Storey JD, Tibshirani R, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**: 9440-9445, 2003.
8. Benjamini Y, Hochberg Y, On the adaptive control of the false discovery rate in multiple testing with independent statistics, *J. Educ. Behav. Statist.* **25**: 60-83, 2000.
9. Reiner A, Yekutieli D, Benjamini Y, Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* **19**: 368-375, 2003.
10. Qiu X, Klebanov L, Yakovlev AY, Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes, *Statistical Applications in Genetics and Molecular Biology* **4**: Article 34, 2005.
11. Qiu X, Brooks AI, Klebanov L, Yakovlev A, The effects of normalization on the correlation structure of microarray data, *BMC Bioinformatics* **6**: Article # 120, 2005.
12. Pavlidis P, Li Q, Noble WS, The effect of replication on gene expression microarray experiments, *Bioinformatics* **19**: 1620-1627, 2003.
13. Qiu X, Xiao Y, Gordon A, Yakovlev A, Assessing stability of gene selection in microarray data analysis, *BMC Bioinformatics* **7**: #50, 2006.
14. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell* **1**(2): 133-143, 2002.
15. Efron B, Tibshirani R, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, New York, 1993.
16. Shao J, Tu D, *The Jackknife and Bootstrap*. Springer Series in Statistics, Springer, NY, 1995.
17. Politis DN, Romano JP, Large sample confidence regions based on subsamples under minimal assumptions, *Ann Statist* **22**: 2031-2050, 1994.
18. Almudevar A, Klebanov LB, Qiu X, Salzman P, Yakovlev AY, Utility of correlation measures in analysis of gene expression, *NeuroRx*, in press.
19. Klebanov L, Jordan C, Yakovlev A, A new type of stochastic dependence revealed in gene expression data, *Statistical Applications in Genetics and Molecular Biology* **5**: Article 7, 2006.

20. Shedden K, Chen W, Kuick R, Ghosh D, MacDonald J, Cho KR, Giordano TJ, Gruber SB, Fearon ER, Taylor JMG, Hanash S, Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data, *BMC Bioinformatics* **6**: Article # 26, 2005.

## Authors' Profiles

Xing Qiu received his M.A. in Mathematics from the University of Rochester, U.S.A, in 2000 and a Ph.D. in Mathematics from the University of Rochester, U.S.A., in 2004. He has been a Post-doctoral Fellow in the Department of Biostatistics and Computational Biology, University of Rochester, U.S.A., since 2004. He is coauthor of three peer reviewed papers in the field of microarray data analysis.

Andrei Yakovlev received his Ph.D. in Biology from the Institute of Physiology, Academy of Sciences, U.S.S.R., in 1973 and a Ph.D. in Mathematics from Moscow State University in 1981. He served as Head of the Department of Biomathematics, Central Institute of Radiology (1978-1988) and Chair of the Department of Applied Mathematics, St. Petersburg Technical University (1988-1992), St. Petersburg, Russia, and Director of Biostatistics, Huntsman Cancer Institute, University of Utah (1996-2002). He is currently Professor and Chair, Department of Biostatistics and Computational Biology, University of Rochester, U.S.A. He is author or co-author of 4 books and over 180 peer reviewed papers in biomathematics and biostatistics. He is an Elected Fellow of the Institute of Mathematical Statistics and American Statistical Association, and an Elected Member of the Russian Academy of Natural Sciences and International Statistical Institute. He is a recipient of the Alexander von Humboldt Award, John Simon Guggenheim Fellowship, and Distinguished Scholarly and Creative Research Award of the University of Utah.

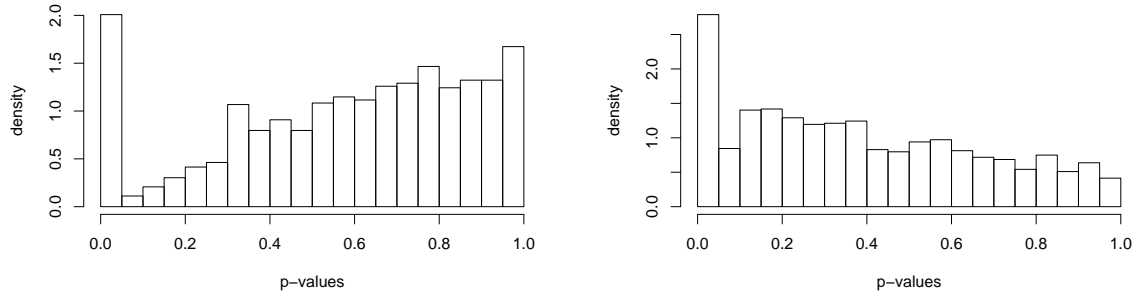


Figure 1: Two sample realizations of the histogram of  $p$ -values in the simulated data SIM43CORR

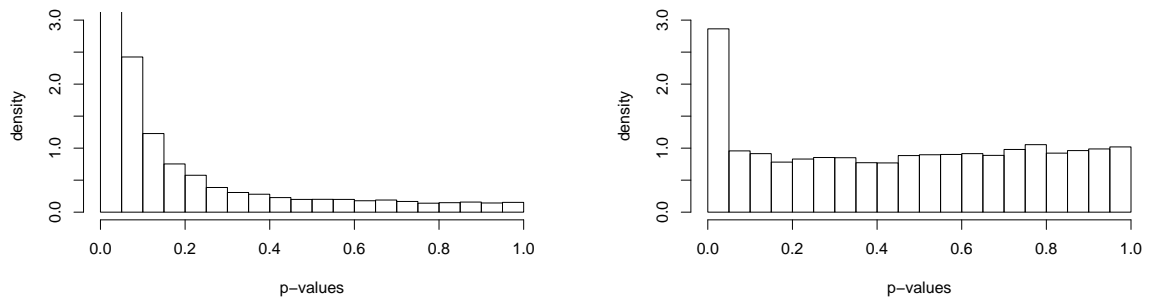


Figure 2: Two sample realizations of the histogram of  $p$ -values in the biological data TALL-Hyperdip