Gene expression

Estimation of false discovery proportion under general dependence

Yudi Pawitan1,*; Stefano Calza1,2 and Alexander Ploner1

1Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden and
2Department of Biomedical Sciences and Biotechnology, Brescia, Italy

Received on June 29, 2006; revised on October 9, 2006; accepted on October 10, 2006
Advance Access publication October 17, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Wide-scale correlations between genes are commonly observed in gene expression data, due to both biological and technical reasons. These correlations increase the variability of the standard estimate of the false discovery rate (FDR). We highlight the false discovery proportion (FDP, instead of the FDR) as the suitable quantity for assessing differential expression in microarray data, demonstrate the deleterious effects of correlation on FDP estimation and propose an improved estimation method that accounts for the correlations.

Methods: We analyse the variation pattern of the distribution of test statistics under permutation using the singular value decomposition. The results suggest a latent FDR model that accounts for the effects of correlation, and is statistically closer to the FDP. We develop a procedure for estimating the latent FDR (ELF) based on a Poisson regression model.

Results: For simulated data based on the correlation structure of real datasets, we find that ELF performs substantially better than the standard FDR approach in estimating the FDP. We illustrate the use of ELF in the analysis of breast cancer and lymphoma data.

Availability: R code to perform ELF is available in http://www.meb.ki.se/~yudpaw.

Contact: yudi.pawitan@ki.se

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Wide-scale correlations between genes in expression data can occur for a number of reasons, e.g. because genes are organized in active biochemical pathways, or due to unmeasured underlying variables, including imperfect normalization [Ploner et al., 2005]. Standard estimation of false discovery rate (FDR) tends to ignore both biological and technical correlations [e.g. Storey and Tibshirani, (2003); Pound and Cheng (2004); Pawitan et al. (2005a, b)], with the argument that the methods work under weak dependence. A number of recent papers, however, have emphasized the serious effects of correlation in microarray data analysis (e.g. Klebanov and Yakovlev, 2006; Qiu et al., 2006; Efron, 2006, (http://www.stat. stanford.edu/~brad/)). The purpose of this paper is (1) to analyse the variation of standard FDR estimates for correlated genes, (2) to introduce the concept of a latent FDR conditional on the correlation effects and (3) to describe a novel procedure for estimating the latent FDR (ELF) that improves on the standard FDR estimate.

1.1 Terminology

For simplicity, all our examples are based on the comparison of two independent groups using a standard t-statistic z. Because the methods we use are computational, requiring no theoretical derivation, they can be extended to other statistics and more general comparison problems with minor and obvious modifications.

Suppose we test m genes, with corresponding statistics z1, . . . , zm. For a fixed critical value c, we define the number of non-differentially expressed (nonDE) genes and the number of genes declared DE as

\[ V(c) = \text{Number of null } |z_i| \geq c \] (1)

\[ R(c) = \text{Number of } |z_i| \geq c, \]

and the false discovery proportion (FDP) as

\[ \text{FDP}(c) = \frac{V(c)}{R(c)}, \]

except in the case of \( R(c) = 0 \), in which case we just set \( \text{FDP}(c) = 0 \). For convenience we will drop the argument c and refer to the FDP.

In short, the FDP is the unknown and random proportion of false discoveries among the genes declared to be DE. The standard FDR is simply the marginal average of the FDP:

\[ \text{FDR} = E(\text{FDP}). \]

In the following, we define the latent FDR as the conditional average of the FDP given a random effect b that captures the correlation effects:

\[ \text{LFDR} = E(\text{FDP} | b). \]

Consequently, the standard FDR is just the average of the latent FDR over the random effects b:

\[ \text{FDR} = E_b(\text{LFDR}). \]

The basic idea of this paper is familiar from regression analysis: the conditional variance \( \text{var}(\text{FDP} | b) \) can be much smaller than the marginal variance \( \text{var}(\text{FDP}) \), so the LFDR estimate can have significantly lower variance than the FDR estimate.
Section 2.1 Notation

Functions appear in regular type with brackets for the argument, e.g. as \( f(z) \) for the density function of \( z \). Vectors appear in bold, e.g. \( \mathbf{x} \). If some extra information \( \mathbf{x} \) is available, then the optimal predictor is the conditional mean \( \mathbb{E}[\mathbf{y} | \mathbf{x}] \) of \( \mathbf{y} \) given \( \mathbf{x} \). In practice, we will estimate \( \mathbb{E}[\mathbf{y}] \) and \( \mathbb{E}[\mathbf{y} | \mathbf{x}] \), but for the problem of prediction, they will both be interpreted as predictors of the unknown \( \mathbf{Y} \).

To highlight the novel contributions of this paper: (1) We focus on the FDP as the target of estimation. (2) We use the singular value decomposition (SVD) to explain the variability in the FDP. This is based on the variation of the null distribution of the test statistics generated by permutation. (3) Based on the SVD results, we propose a random-effect model for a latent null distribution function, with a corresponding latent FDR that captures the effects of correlation. (4) We describe an ELF procedure, which is based on a Poisson model with identity link, and demonstrate that it improves on the standard FDR estimate as an estimate of the FDP.

2 METHODS

2.1 Notation

Suppose we have a hypothetical experiment that has generated a list of 100 candidate genes, and we know that the true FDP is 0.25 and FDR = 0.10. This means that the specific list of genes found in the experiment contains 25 non-DE genes, however, if we replicate the same experiment repeatedly, using the same threshold \( c \) in each case, then the candidate lists produced by these replications will contain on average only 10 non-DE genes. Obviously, the FDP is the quantity a researcher would really want to know in order to assess the quality of a gene list found in his or her specific experiment.

Suppose, in validating the specific gene list, we were able to perform precise determination of DE status, e.g. by using more accurate methods, such as the PCR on larger sample size, then we will come closer to the true FDP (of the list in the original experiment), not the FDR. The standard estimation procedure tries to estimate the FDR and a large discrepancy between FDR and FDP occurs when genes are correlated.

We demonstrate this problem for simulated data in Figure 1. All plots in the figure are based on comparing two groups; see Section 2.2 for details. In the first row of Figure 1, we show results for independent genes, in the second and third row results for realistic correlation structures between genes, simulated from the breast-cancer and lymphoma datasets described in Section 2.2.

From each of these three scenarios, we have simulated multiple realizations with exactly the same settings. For each the dataset generated in this manner, we can compute the true FDP, the true FDR and the standard FDR estimate (Section 2.6) as a function of the t-statistic. Note that the true FDP varies from realization to realization, but that the true FDR is fixed.

When the genes are independent (Fig. 1a), the true FDPs (grey lines) vary tightly around their mean (solid black line), which is theoretically purposes identical to the true FDR (dotted black line). Figure 1b shows also that, for independent genes, the standard FDR estimate (on the vertical axis) works reasonably well as an estimate of the true FDP (on the horizontal axis). This situation changes dramatically in Figure 1c–f, where the genes are correlated: while the average FDPs in Figure 1c and e are still close to the true FDR, the individual true FDP curves vary greatly around that mean between the different realizations. More importantly, because of the severe variability, the standard FDR estimates in Figure 1d and f are no longer useful estimates of the underlying true FDPs. Note however that Figure 1b, d and f show that the FDR estimates are nearly unbiased, regardless of the amount of correlation between genes, so the negative effects of the correlations seem to be limited to the variability.

1.3 Estimation of FDP

Conceptually, there might be a confusion between the estimation of FDP and FDR, because in practice an FDR estimate can be interpreted as an FDP estimate. Let us put this in the context of predicting the outcome of a random variable \( Y \) without extra information, the optimal predictor of \( Y \) in terms of mean-squared error is its marginal mean \( \mathbb{E}[Y] \). If some extra information \( x \) is available, then the optimal predictor is the conditional mean \( \mathbb{E}[Y | x] \). In practice, we will estimate \( \mathbb{E}[Y] \) and \( \mathbb{E}[Y | x] \), but for the problem of prediction, they will both be interpreted as predictors of the unknown \( Y \).

(Although FDP is a random quantity, the term ‘predictor of FDP’ sounds awkward, so we stick to ‘estimate of FDP’.) Thus, the best theoretical estimate of the FDP for uncorrelated data is its marginal mean, i.e. precisely the FDR. In practice, the FDR of course needs to be estimated, and the resulting FDR estimate can be interpreted as an empirical estimate of the FDP. Now if we have correlated genes, we can construct a conditional mean of the FDP, given extra information about the underlying correlation structure. This conditional average is the LFDR, which is a theoretically superior estimate of the FDP compared to the marginal average (the FDR).

To highlight the novel contributions of this paper: (1) We focus on the FDP as the target of estimation. (2) We use the singular value decomposition (SVD) to explain the variability in the FDP. This is based on the variation of the null distribution of the test statistics generated by permutation. (3) Based on the SVD results, we propose a random-effect model for a latent null distribution function, with a corresponding latent FDR that captures the effects of correlation. (4) We describe an ELF procedure, which is based on a Poisson model with identity link, and demonstrate that it improves on the standard FDR estimate as an estimate of the FDP.
2.2 Real and simulated datasets

We demonstrate our method using simulated data as well as two published datasets that have been used extensively as examples for FDR methodology: the BRCA breast cancer dataset was collected from patients with hereditary breast cancer, carrying mutations of either the BRCA1 gene (n1 = 7) or the BRCA2 gene (n2 = 8), as described in Hedenfalk et al. (2001). Expression was measured originally for 3226 genes, but following Storey and Tibshirani, (2003), we removed 56 extremely volatile genes and analysed only the remaining m = 3170 genes.

The lymphoma dataset contains 240 cases of diffuse large B-cell lymphoma described in Rosenwald et al. (2002), which was collected on a custom-made DNA microarray chip yielding measurements for m = 7399 probes. The average clinical followup for patients was 4.4 years, with 138 deaths occurring during this period. For our purposes, we ignore the censoring information and only compare expression between the 102 survivors and the 138 non-survivors.

All expression data were analysed on the log scale. Global normalization was used to set the array mean to zero and standard deviation to one.

To mimic realistic correlation structures between genes, we sample the error distribution from the residuals of real datasets as follows:

- subtract the group-wise means for each gene;
- randomly sample arrays without replacement from the whole collection, allocating n1 arrays to group 1 and n2 to group 2;
- for DE genes, add an effect size ± Dσ1, in group 2, with equal proportions of under- and over-expression and σ, the sample standard deviation of gene i, and D the effect size, or log fold-change.

For the BRCA data, we simulate two groups with n1 = 7 and n2 = 8 arrays, for the lymphoma data, we use n1 = n2 = 10 arrays. In both cases, we set the proportion of nonDE genes to π0 = 0.9, and the effect size to D = 1.

For comparison, we also run simulations with independent genes. In this case, we assume m = 3000 genes per array (comparable to the BRCA data), and n1 = n2 = 10 arrays. For arrays in group 1, the log-expression values are normally distributed with mean zero and variance σ2 = 1. In group 2, the nonDE genes are also log-normal with mean zero and variance σ2 = 1, while the DE genes are log-normal with mean ± D and variance σ2 = 1. We set the proportion of nonDE genes to π0 = 0.9, and the effect size D = 1.

In the simulations for Figure 1, each realization of the data generates m statistics with known DE status, so we can compute both the true FDR and the standard FDR estimate as functions of the critical value for each realization. The true theoretical FDR is computed according to the theory in Pawitan et al. (2005a).

To avoid a potential confusion, we note that we use two different types of replication of the data throughout: ‘permutations’ and simulated ‘realizations’. The former refers to artificial datasets generated by permuting the group labels from a single observed dataset. The latter refers to independent replication of datasets from a simulation experiment as outlined in this section.

2.3 Singular value decomposition

Because the great majority of genes are generally nonDE, the variation of the FDP will be dominated by the distribution of the test statistics for nonDE genes. As we do not know the DE status, the null distribution is generated by permuting the group labels. To simplify the analysis, we partition the range of the observed statistics into B equipropped bins with width Δ. Let the histogram-count vector y = (y1, . . . , yB) be the number of statistics z that fall into each bin. In practice, the bin width should be selected so that the histogram is quite smooth; since there is typically a large number of genes, this is easy to achieve. We use Δ = 0.1 in all our examples.

Let X = (x1, . . . , xn) be the array data, with N = n1 + n2, and g = (g1, . . . , gn) the vector of group labels. Each permuted dataset (X, g*) is generated by a random re-arrangement of group labels g* = g1, . . . , gn. We compute test statistics for the permuted data, so each permutation contributes a single count vector y*. Let y be the mean vector of y’s over K permutations and the B × K matrix Y* the mean-corrected matrix of count-vectors y* (K = 50 in our examples). The singular value decomposition of Y* is

\[ Y' = UV^* = \sum_{k=1}^{K} \lambda_k u_k v_k. \]

For each vector we have therefore

\[ y_i' - y^* = \sum_{k=1}^{K} \lambda_k v_k u_k, \]

and the rank-1 approximation of y_i using a single singular vector is

\[ y_i' = \hat{y} + \lambda_1 v_1 u_1. \]

This rank-1 approximation will be the basis of our model, where the u_1 term captures the dependence between genes.

The SVD is closely connected to principal component analysis (PCA); indeed, the scaled singular vector \( \lambda_1 u_1 \) in Equation (3) is just the first principal component of Y*. The main difference is that PCA is performed via the correlation or covariance matrices, so it does not immediately suggest linear models (2) and (3).

2.4 Mixture model and latent FDR

We assume that the observed statistics follow a mixture model for DE and nonDE genes, as outlined in Efron et al. (2001):

\[ f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z), \]

where \( f_0(z) \) and \( f_1(z) \) are the densities of the statistics from nonDE and DE genes, and \( \pi_0 \) is the proportion of nonDE genes.

Since the permutation distribution reflects the null distribution of statistics from nonDE genes, Equation (3) immediately suggests a Poisson model for the contribution of nonDE genes:

\[ y \sim \text{Poisson}(\mu = m\Delta \pi_0 f_0(z)). \]

for \( f_0(z) \) follows a linear model

\[ f_0 = \phi_0 + b \phi_1, \]

and, from Equation (3), \( \phi_0 \) and \( \phi_1 \) correspond to \( y^* \) and \( u_1 \), respectively. \( (\phi_0 \) and \( \phi_1 \) are scaled so that they integrate to one.) This correspondence is important, as it shows how the method can be applied to arbitrary test statistics. The parameter b in Equation (4) varies between realizations; it is a random effect parameter that captures the variation of the null distribution due to correlation. From Equation (3), the average of b is equal to zero.

Let \( f_0(z) \) be the continuous density function corresponding to the discrete version \( f_0(z) \) in Equation (4), and similarly for \( f_0(z) \) and \( f_1(z) \). The density of the observed statistics is now

\[ f(z) = \pi_0 f_0(z) + \pi_0 f_1(z) + (1 - \pi_0) f_0(z). \]

Assuming that conditional on b, the statistics are independent, the count statistic \( V(z) \) in Equation (1) is a cell of a multinomial distribution on the 2-by-2 table. Conditioning on \( R(c) = r \), the contribution from the null genes \( V(z) \) is binomial with size r and probability

\[ p \equiv \int_{f_0} f_0(z) \text{d}z / \int_{f_0} f(z) \text{d}z. \]

Conditional on \( R = r \), the FDP is a sample proportion, so its mean is equal to the true proportion:

\[ \text{E}(\text{FDP} | b, R = r) = p, \]

and we obtain the latent FDR as

\[ \text{LFDR} \equiv \text{E}(\text{FDP} | b) = \text{E} \left\{ \text{E}(\text{FDP} | b, R = r) \right\} = \frac{\pi_0 \int_{f_0} f_0(z) \text{d}z}{\int_{f_0} f(z) \text{d}z}. \]
Note that the LFDR is the FDR based on \( f_d(z) \) as null density. In contrast, the standard FDR is based on \( f_0(z) \), which is the average of \( f_0 \) over a collection of different \( b \)’s.

### 2.5 Estimated latent FDR: ELF

The practical procedure for estimating the latent FDR, which we refer to as ELF, is based on the notion that the random effect \( b \) is realized in the observed data. For the count vector \( y \) observed for the data, we fit a Poisson model with identity link (Pawitan, 2001, Section 6.3):

\[
y \sim \text{Poisson}(\mu = m \lambda f),
\]

where \( f \) follows a linear model

\[
f = \pi_0 f_0 + \pi_1 b f_1 + (1 - \pi_0) f_1.
\]

In the estimation we found an improvement of the performance if we apply the standard FDR is based on \( f_0 \), \( f_1 \) and \( b \) as weights, in effect emphasizing the center of the distribution. We can then show that the procedure is equivalent to the unweighted least squares method.

Given estimates of the parameters, we can estimate the LFDR in view of Equation (6) as

\[
\text{LFDR} = \frac{m \pi_0 \int_{|z| > c} f(z) dz}{\text{number of } |z| > c},
\]

where the denominator is an estimate of \( m \int_{|z| > c} f(z) dz \).

The predictors and parameters in model (7) are computed by the following procedure:

1. Perform \( K \) permutations of group labels. Each permuted dataset generates a histogram-count vector \( y^* \).
2. Compute the predictor \( \phi_0 \) from the average vector \( y^* \) by scaling so that it integrates to one.
3. Construct a matrix \( Y^* \) from the \( y^* \)’s. Compute the predictor \( \phi_1 \) from the smoothed first singular vector \( u_1 \) of \( Y^* \).
4. Since \( f_1 \) is unknown, the regression is performed in two steps. First, fit the reduced model \( y \sim \text{Poisson}(\mu = m \lambda f_0) \), where

\[
f = \beta_0 \phi_0 + \beta_1 \phi_1 + \beta_2 f_1,
\]

and compute the residual vector \( r = y - \hat{y} \). Estimate \( f_1 \) by smoothing the original vector \( r(n \Delta) \) as a function of \( z \).
5. Fit the full model (7)

\[
f = \beta_0 \phi_0 + \beta_1 \phi_1 + \beta_2 f_1,
\]

and re-estimate the full set of coefficients (\( \beta_0, \beta_1, \beta_2 \)).

Note that \( \pi_0 \) is automatically estimated in the procedure as the coefficient of \( \phi_0 \).

Because of the relationships between the parameters, it may be possible to improve the estimate of \( \pi_0 \), but we will not attempt this here. We have also found that attempts at iterating the two-step procedure generally increase variability. This may be a problem of trading bias for variability, where reducing the one increases the other.

We can show that for small values of \( b \) in Equation (4), say \( |b| < 0.2 \), the null density \( f_d(z) \) is well approximated by the density of \( N(0, \sigma^2 = (1 - b^2 \sqrt{2})^2) \), but not otherwise. Our simulations indicate that this approximation is suitable for reducing the variability in the estimation of \( f_d(z) \). For our real data examples, the results are practically identical with and without the approximation.

### 2.6 Comparison procedures

The first and most natural comparison is of course with the standard FDR estimate (e.g. Storey and Tibshirani, 2003; Pawitan et al., 2005b). Let \( p_1, \ldots, p_m \) be the ordered \( P \)-values from \( m \) test statistics, which may be obtained using a permutation null distribution. We can then give the standard estimate of FDR as a function of the \( P \)-values,

\[
\text{FDR}(p_k) = m \frac{\pi_0 \int_{|z| > k} f_d(z) dz}{m \lambda f_0(z) dz},
\]

Monotonicity is imposed by taking the cumulative minimum over all \( P \)-values to the right. In order to express the FDR as a function of the critical value of the test statistic, the \( P \)-values are transformed back to the original test-statistic scale.

We also compare ELF to the procedure proposed by Efron (B. Efron, 2006, unpublished data), which works for any normalized statistic \( z \). For the same histogram-count vector \( y \) as above, we fit a quadratic Poisson model with log-link,

\[
y \sim \text{Poisson}(\mu) \quad \log \mu(z) = \beta_0 + \beta_1 z + \beta_2 z^2, \quad -1 < z < 1,
\]

where as usual, \( \mu(z) \) is the continuous function version of \( \mu \). Note the important range restriction \(-1 < z < 1 \) for this model.

The null density \( f_d(z) \) is then estimated by the density of \( N(0, \sigma^2 = -0.5 \beta_2) \). The estimation of \( \pi_0 \) is based on an independent argument (B. Efron, 2006, unpublished data). Given \( f_d(z) \) and \( \pi_0 \), the latent FDR is computed as in Equation (8).

### 3 RESULTS

#### 3.1 SVD of permutation distributions

Figures 2 and 3 contrast the variation in the distribution of the test statistics between permutations for data with independent and for real microarray data, respectively. Figure 2a shows the histogram-count vectors of 50 permutations of a simulated dataset without gene correlations as solid grey curves; the grey area created by the overlapping curves shows the variation in histogram-counts between permutations. The mean pattern (solid line) is indistinguishable from the normal model, (which is drawn in...
dotted line, but is completely overlayed by the solid line). Figure 2b shows the singular values sorted by magnitude; it is evident there is no dominant component. This is further underlined in Figure 2c, which plots the actual components of the first singular vector over the corresponding histogram bins, together with a robust smoothing curve: the values appear to vary randomly around zero, with no clearly recognizable pattern.

Figure 2d finally relates the variation in histogram-count vectors as seen in Figure 2a to the variation in the first singular component: the histogram counts of the two permutations with the respective largest and smallest value of \( v_{11} \) in the rank-1 approximation [Equation (3)] are shown as dots, and their corresponding rank-1 approximations are shown as dark solid lines. These extreme permutations do not look special among the other 48 permutations, and the variation between the histograms is not associated with the difference in \( v_{11} \).

In contrast, the same plots for the lymphoma data as seen in Figure 3 show variation between different permutations that is both greater and much more strongly structured. In Figure 3a, there is larger variation of the individual histogram counts around their common mean, though the mean itself (solid) is again well enough approximated by the normal distribution to hide its dotted line. Figure 3b shows that the variation is dominated by one large singular value, associated with the pattern seen in Figure 3c. A consequence of this pattern is that for negative coefficients \( b \) in Equation (4), \( f_0 \) tends to have a heavier tail than the average, and vice versa for positive coefficients. Subsequent singular vectors do not have large contributions to the variation, as shown in the Supplementary material, though for the lymphoma data, the second vector seems to capture some asymmetry in the null distribution. (The third vector appears to be noise.) In practice, the inclusion of the second singular vector may be worthwhile if there is some asymmetry in the center of the observed distribution.

Figure 3d highlights again the two most extreme permutations, i.e. those associated with the maximum and minimum \( v_{11} \) in the rank-1 approximation Equation (3). Evidently, these two permutations span the range of observed variation in histogram counts adequately. The panel provides also graphical evidence that the rank-1 approximations (solid lines) describe the actual counts (dotted) sufficiently well. This suggests that we will be able to model the effects of correlation between genes on the distribution of the test statistics by a single factor that is easily computed using the SVD analysis.

It is worth emphasizing that the correlation between genes does not affect the average shape of the distribution, which is indistinguishable from the normal model in both figures. This explains also why the standard FDR estimate, which is largely based on the average permutation distribution, is very nearly unbiased, see Figure 1b, d and f. The correlation effect is visible only in the deviation from the mean, and the SVD analysis shows that this deviation is simple enough to be mostly determined by a single factor. (The phenomenon that correlation affects variability but not bias is well known in statistics; e.g. the ordinary least-squares method is still unbiased when the errors are correlated, but it is not efficient.)

We obtain similar results for the BRCA data, as well as simulated data based on the BRCA and lymphoma data, which are shown in the Supplementary material. There we also demonstrate that the observed SVD patterns are stable over different independent realizations of the data.

### 3.2 Performance of ELF

Figure 4 compares the performance of ELF, standard FDR and Efron’s procedure in estimating the FDP for 25 simulated lymphoma datasets. (a–c): estimated FDP on the vertical axis plotted against true FDP on the horizontal axis, each grey line corresponding to one realization of the data, and the mean shown as solid black line. (d): variance of the three procedures as seen in (a–c) as function of the true FDP.

Estimation of FDP under general dependence
for Efron’s procedure in Figure 4b has visibly larger variance than ELF. Figure 4c is similar to Figure 1f for the standard FDR estimation, and suffers also from large variability compared to ELF. Figure 4d explicitly compares the variance of the three procedures, indicating that ELF performs better and more consistently than either Efron’s procedure or the standard FDR.

The simulations also indicate that ELF’s estimation of \( \pi_0 \) is positively biased, although less so than Efron’s method. The positive bias is conservative, i.e. ELF may overestimate the FDP, but it is less of a concern than the large variability of the standard estimate, which may be extremely liberal. To understand the upward bias, we note that there is some confounding between \( \phi_1(z) \) and \( f_1(z) \) in model (5). We expect specifically the contribution to \( f(z) \) from genes with small effects to be confounded by \( \phi_1(z) \).

3.3 Analysis of real datasets

Figure 5 shows the FDP estimates using the standard FDR, ELF and Efron’s methods for the breast-cancer and lymphoma datasets. Figure 5a and c show the observed histogram counts of the test statistics (scattered points), together with the null distribution as estimated by ELF (solid line), Efron’s method (dashed) and the standard FDR approach (dotted line).

For the breast-cancer data, Efron’s procedure seems too conservative, as the null distribution is too wide, so the corresponding FDP estimate is high, with a minimum of >60% for up-regulated genes, and >80% for down-regulated genes. The tendency to be conservative is already obvious in the simulation studies. On the other hand, the standard FDR estimate appears optimistic: there are 575 genes with estimated FDR <20%. From the theory and the simulation, the low FDR is confounded by correlations between genes. ELF produces an estimate intermediate between Efron’s and the standard FDR estimate, declaring 120 genes to be DE with estimated LFDR <20%.

The estimated \( \pi_0 \) from the ELF, Efron’s and the standard FDR procedure are 0.87, 0.94 and 0.72, respectively. The standard estimate is likely optimistic, since the effect of correlation partially shows up as gene effects. We do not attempt to use more sophisticated methods for estimating \( \pi_0 \), since they generally do not take into account the correlation effects.

The obvious question now is which of these results is the most credible. Ideally, we should have a validation dataset from a larger series of BRCA1 and BRCA2 tumors. This is in fact available from the same group of researchers in Lund, Sweden, who conducted the original study. In his PhD thesis, Vallon-Christersson, (2005) analysed in detail functional and molecular differences between BRCA1 and BRCA2 breast cancers. One important finding is that the former tend to be estrogen-receptor negative, while the latter tend to be estrogen-receptor positive. Considering the importance of receptor status for breast tumor biology, we should expect significant differences in the gene expression of BRCA1 and BRCA2.

For the lymphoma data, Efron’s procedure again looks too conservative, allowing no discovery with FDP <20%. The success of cross-validated prediction of the survival status based on gene-expression data (Rosenwald et al., 2002) suggests that there must be DE genes among the top-ranking genes. Using the standard method, there are 540 genes with estimated FDR <20%, while ELF found 35 genes with estimated FDP <20%. The \( \pi_0 \) estimates from ELF, Efron’s and standard procedures are 0.92, 0.95 and 0.83.

4 DISCUSSION

In this paper, we have addressed the issue of gene–gene correlations in microarray data analysis. One important conceptual step is to move away from trying to estimate the FDR, but instead to target the FDP directly. This idea unifies the variety of FDR estimation methods as providing an estimate of the FDP. In any case, even when an FDR estimate is used in practice, the scientist will likely treat it as an FDP estimate, so the emphasis on the FDP seems natural.

The correlation induces large variability in the FDP, so the standard FDR estimate is often far from the true FDP. The standard FDR methodology is based on model (5), but without the second term which captures the correlation effects. ELF improves on the standard estimation by allowing for this term. In effect, the standard FDR approach is based on a comparison of the observed distribution of t-statistics to the average distribution across different permutations. In comparison, ELF tries to find a better null distribution than the average permutation distribution.

Most FDR estimation methods treat the collection of test statistics from different genes as a sample from a distribution. This allows, for example, mixture modelling and estimation of parameters. This approach might be questioned, as the statistics have the same logical footing as variables measured on a single experimental unit. However, there are other areas of statistics, such as time series analysis, stochastic processes or image analysis, where statistical inference is performed on a single realization of time series or image data. For such analyses, what matters is whether we can capture the dependence structure in the data adequately. Here we try to do the same in the context of microarray data analysis, where we use the SVD as a tool for discovering the dependence structure that allows us to perform proper statistical inference.

Under general dependence, estimation of the FDP turns out to be much harder than estimation of the FDR. In particular, it is difficult
Thus of Efron, and we show that it is possible to improve the standard procedure must allow an empirical null distribution that accounts for the presence of substantial gene–gene correlation, but that the procedure must allow an empirical null distribution that accounts for the effects of correlation. Our paper is very much in the spirit of Efron (B. Efron, (2006), unpublished data) is more only solution is to increase the sample size rather than the number can estimate the FDR in small datasets at all, and suggest that the corresponding lower bound can be defined in a similar manner, yielding a confidence envelope for the FDP. In contrast to our approach, however, these envelope procedures do not try to estimate the FDP itself. Specifically when the genes are correlated, the envelope functions will be very wide and can be a source of pessimism.

Figure 6 shows the 90% confidence envelope for the breast-cancer and lymphoma data, computed using the method described in Meinshausen, (2006). The width of the envelope is not surprising in view of the variability seen in Figure 1d and f. In practice, the wide envelope means that the standard FDR estimate is untrustworthy.

Klebanov and Yakovlev, (2006) appear pessimistic whether we can estimate the FDR in small datasets at all, and suggest that the only solution is to increase the sample size rather than the number of genes. Efron (B. Efron, (2006), unpublished data) is more optimistic, showing that it is still possible to estimate FDR in the presence of substantial gene–gene correlation, but that the procedure must allow an empirical null distribution that accounts for the effects of correlation. Our paper is very much in the spirit of Efron, and we show that it is possible to improve the standard FDR estimate as a predictor of the FDP.

There are, however, several key differences between our approach and Efron’s:

(1) Instead of theoretically-derived normal-based formulas, we use a fully-computational and data-driven approach via the SVD. This has the advantage that ELF can be easily generalized for arbitrary test statistics, as long as the permutation procedure can be used to compute the key components of the model.

(2) In Efron’s procedure, the crucial model fit is performed in a single step for a restricted range of test statistics, as the model is supposed to capture only nonDE genes. In ELF, model fitting is a two-step procedure for the full range of statistics, capturing the mixture of DE and nonDE genes.

Although ELF improves on the performance of Efron’s and the standard procedure, we observed some positive bias in its estimation of π0, so further improvement is possible. In this context, we also note that standard estimation of π0 is essentially based on comparing the observed distribution of test statistics with the average permutation distribution, and therefore requires a re-formulation that takes correlation effects into account. This issue needs further investigation.

In some areas of application, studies with small sample sizes, say 3-versus-3 arrays, are still common. In these cases the permutation approach is not feasible, so one cannot obtain the empirical shape of the singular vector. Correlation effects on FDR are likely to be severe in these small studies, and a realistic method for dealing with this problem will require further research.

Conflicts of Interest: none declared.

REFERENCES


Ploner,A. et al. (2005) Using correlations to evaluate low-level analysis procedures for high-density oligonucleotide microarray data. BMC Bioinformatics, 6, 80.


