Methodology article

A constrained polynomial regression procedure for estimating the local False Discovery Rate

Cyril Dalmasso*[1], Avner Bar-Hen[2] and Philippe Broët[1]

Address: [1]JE 2492 – Univ. Paris-Sud, 16 avenue Paul Vaillant Couturier, F94807 Villejuif, France and [2]UMR AgroParisTech/INRA 558, 16 rue Claude Bernard, 75231 Paris, France

Email: Cyril Dalmasso* - dalmasso@vjf.inserm.fr; Avner Bar-Hen - avner@inapg.fr; Philippe Broët - broet@vjf.inserm.fr

* Corresponding author

This article is available from: http://www.biomedcentral.com/1471-2105/8/229

Abstract

Background: In the context of genomic association studies, for which a large number of statistical tests are performed simultaneously, the local False Discovery Rate (*lFDR*), which quantifies the evidence of a specific gene association with a clinical or biological variable of interest, is a relevant criterion for taking into account the multiple testing problem. The *lFDR* not only allows an inference to be made for each gene through its specific value, but also an estimate of Benjamini-Hochberg's False Discovery Rate (*FDR*) for subsets of genes.

Results: In the framework of estimating procedures without any distributional assumption under the alternative hypothesis, a new and efficient procedure for estimating the *lFDR* is described. The results of a simulation study indicated good performances for the proposed estimator in comparison to four published ones. The five different procedures were applied to real datasets.

Conclusion: A novel and efficient procedure for estimating *lFDR* was developed and evaluated.

Background

The use of current high-density microarrays for genomic association studies leads to the simultaneous evaluation of a huge number of statistical hypotheses. Thus, one of the main problems faced by the investigator is the selection of genes (or gene products) worthy of further analysis taking multiple testing into account.

Although the oldest extension of the classical type I error rate is the family-wise error rate (*FWER*), which is defined as the probability of falsely rejecting at least one null hypothesis (e.g., the lack of relationship between gene-expression changes and a phenotype), *FWER*-based procedures are often too conservative, particularly when numerous hypotheses are tested [1]. As an alternative and less stringent error criterion, Benjamini and Hochberg

introduced, in their seminal paper [2], the False Discovery Rate (*FDR*), which is defined as the expected proportion of false discoveries among all discoveries. Here, a discovery refers to a rejected null hypothesis.

Assuming that the test statistics are independent and identically distributed under the null hypothesis, Storey [3] demonstrated that, for a fixed rejection region Γ, which is considered to be the same for every test, the *FDR* is asymptotically equal to the following posterior probability:

$$FDR(\Gamma) = \Pr(H = 0 | T \in \Gamma) \qquad (1)$$

where H is the random variable such that $H = 0$ if the null hypothesis, noted H_0, is true; $H = 1$ if the alternative hypothesis, noted H_1, is true; and T is the test statistic consid-

ered for all tested hypotheses. However, one drawback is that the *FDR* criterion associated with a particular rejection region Γ refers to all the test statistics within the region without distinguishing between those that are close to the boundary and those that are not [4].

For this purpose, Efron [5] introduced a new error criterion called the local False Discovery Rate (*lFDR*) which can be interpreted as a variant of Benjamini-Hochberg's *FDR*, that gives each tested null hypothesis its own measure of significance. While the *FDR* is defined for a whole rejection region, the *lFDR* is defined for a particular value of the test statistic. More formally:

$$lFDR(t) = \Pr(H = 0 | T = t). \tag{2}$$

As discussed by Efron [6], the local nature of the *lFDR* is an advantage for interpreting results from individual test statistics. Moreover, the *FDR* is the conditional expectation of the *lFDR* given $T \in \Gamma$:

$$FDR(\Gamma) = E(lFDR(T) | T \in \Gamma). \tag{3}$$

In this context, most of the published procedures for estimating *lFDR* proceed from a two-component mixture model approach, in which the marginal distribution of the test statistic can be written:

$$f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t). \tag{4}$$

Here, f_0 and f_1 are the conditional density functions corresponding to null and alternative hypotheses, respectively, and $\pi_0 = \Pr(H = 0)$. Using these notations, *lFDR* can be expressed as:

$$lFDR(t) = \pi_0 \frac{f_0(t)}{f(t)}. \tag{5}$$

A variety of estimators have been proposed that either consider a full model-based approach (for a few [7-10]) or estimate an upper bound of *lFDR* without any assumption for f_1. It is worth noting that, in this latter framework, the probability π_0 is not identifiable [11]. Thus, from equation (5), only an upper bound estimate can be obtained for *lFDR*.

Four procedures that do not require a distributional hypothesis for f_1 were introduced by Efron [6,12], Aubert *et al.* [13], Scheid and Spang [14] and Broberg [15]. These methods are based on the separate estimations of π_0, f_0 and f from the calculated *p*-values. For the last three procedures [13-15], the *p*-values are supposed to be uniformly distributed under the null hypothesis, while Efron's approach estimates f_0 from the observed data.

Herein, we describe a novel and efficient procedure for estimating *lFDR*. While classical approaches are based on the estimation of the marginal density *f*, we propose directly estimating π_0 and $1/f$ (equation 5) within the same framework.

To situate our procedure among the four published, we briefly recall below their individual principles.

Efron (2004) [12]
For this procedure, the *p*-values are transformed into *z*-values for which the theoretical distribution (under the null hypothesis) is a standard normal distribution. To take into account that f_0 may be different from the theoretical null distribution, the parameters are estimated from the observed distribution of the *z*-values as summarized below.

The density *f* is non-parametrically estimated using a general Poisson linear model, in which $\log(f(z))$ is modeled as a natural spline function with seven degrees of freedom. Then, the null distribution parameters are estimated as follows. The expectation is taken as arg max($\hat{f}(z)$) and the variance is deduced by quadratically approximating $\log(\hat{f}(z))$ for central *z*-values (for which $f_1(z)$ is supposed to be null). The proportion π_0 is then estimated by the ratio of the means $\overline{f(z)} / \overline{f_0(z)}$ calculated from these central *z*-values. The *lFDR* is finally estimated by $lFDR(z) = \widehat{\pi_0} \widehat{f_0(z)} / \widehat{f(z)}$. It should be noted that in addition to the normality assumption for the *z*-values under the null hypothesis, the procedure is also based on the assumptions that central *z*-values mainly consist of true null hypotheses and that the proportion $(1 - \pi_0)$ of modified genes is small. In particular, Efron recommends using this procedure for $\pi_0 > 90\%$.

Aubert et al. (2004) [13]
Assuming that the *p*-values are uniformly distributed under the null hypothesis ($f_0 = 1$), the procedure is based on the separate estimations of π_0 and *f* .

Ordering the *p*-values ($p_{(1)} \leq ... \leq p_{(m)}$), as Aubert *et al.* [13] did, a natural estimator of *f* is:

$$\hat{f}(p_{(i)}) = \frac{\hat{F}(p_{(i+1)}) - \hat{F}(p_{(i-1)})}{p_{(i+1)} - p_{(i-1)}} = \frac{2m}{p_{(i+1)} - p_{(i-1)}} \tag{6}$$

where $\hat{F}$ is the empirical cumulative distribution function of the *p*-values. The resulting estimator for this *lFDR*

is then $\widehat{lFDR}(p_{(i)}) = \dfrac{m\pi_0(p_{(i+1)} - p_{(i-1)})}{2}$. However, as noted by Aubert *et al.* [13], the variance of this estimator is large. A more stable estimator, related to the moving average methodology and corresponding to a generalization of the estimator 6, was given by the authors [13]. To estimate the probability π_0, Aubert *et al.* [13] proposed using an existing procedure, like those proposed by Storey and Tibshirani [16] or Hochberg and Benjamini [17].

Scheid and Spang (2004) [14]

As for the procedure proposed by Aubert *et al.*, the *p*-values are supposed to be uniformly distributed under the null hypothesis. Thus, this procedure is based on the separate estimations of π_0 and f. The marginal distribution f is estimated by dividing the interval [0, 1] into 100 equidistant bins from which a corresponding histogram is derived. A smoothing spline with seven degrees of freedom is then used to estimate f.

The probability π_0 is estimated by a stochastic downhill algorithm (summarized below) with the intention of finding the largest subset of genes that could follow a uniform distribution. A penalized Kolmogoroff-Smirnoff score related to the uniform distribution is calculated for the whole gene set:

$$S(J) = \max_{i \in J} | F_J(u_i) - u_i | + \lambda \frac{m - |J|}{m} \log(m - |J|)$$

(7)

where m is the total number of genes, J is the set of genes under consideration (first, the whole set of genes), F_J is the empirical cumulative distribution for the set J, and λ is a tuning parameter adaptively chosen (for details on the choice of, λ see [14]). Then, iteratively, genes are excluded so that the Kolmogoroff-Smirnoff score decreases. In practice, the procedure stops when the score is not reduced in $2m$ iterations. The score penalty takes into account the sample size m and avoids overfitting. At the end of the procedure, π_0 is estimated by the proportion of the remaining genes. Then, *lFDR* is estimated by $\widehat{lFDR} = \widehat{\pi_0}/f$.

Broberg (2005) [15]

The procedure proposed by Broberg to estimate *lFDR* is also based on the assumption that the *p*-values are uniformly distributed under the null hypothesis. Then, as for the two previous methods, the procedure is based on the separate estimations of π_0 and f. The marginal density f of the *p*-values is estimated by a Poisson regression, similar

to the procedure proposed by Efron. To enforce monotony, Broberg proposed using the Pooling Adjacent Violators algorithm (see [15] for details).

The probability π_0 is then estimated by $\min_{p \in [0,1]} \hat{f}(p)$. Then, *lFDR* is estimated by $\widehat{lFDR} = \dfrac{\widehat{\pi_0}}{\hat{f}}$.

Limitations of these estimators

Through different estimations of π_0, f_0 and f, these four procedures attempt to estimate an upper bound of *lFDR*. However, each of these methods has its own drawback. Efron's procedure [6,12] is restricted to situations in which $\pi_0 > 90\%$. The method of Aubert *et al.* [13] yields an estimator with a large variance. Sheid and Spang's procedure [14] is based on an iterative algorithm and requires extensive computational time (for large datasets). Finally, Broberg's approach [15] sometimes substantially underestimates *lFDR*. Our procedure, developed in details under Methods, is based on a polynomial regression under monotony and convexity constraints of the inverse function of the empirical cumulative distribution. Thus, an estimated upper bound of *lFDR* with small variability can be expected, regardless of the true value of π_0.

Results

Here, we compared, through simulations, our method to the four procedures described above. The five procedures are then applied to real datasets.

Simulated data

To compare our new estimator to the four previously published procedures, we performed a simulation study. Data were generated to mimic a two-class comparison study with normalized log-ratio measurements for m genes ($i = 1,...,m$) obtained from 20 experiments corresponding to two conditions ($j = 1, 2$), each with 10 replicated samples ($k = 1,...,10$), which corresponds to classical sample sizes for differential gene-expression studies. Two total numbers of genes were considered: one small ($m = 500$) and one larger ($m = 5,000$). In each case, all values were independently sampled from a normal distribution, $X_{i,j,k} \sim N(\mu_{ij}, 1)$. For the first condition ($j = 1$), all data were simulated with $\mu_{i1} = 0$. For the second condition ($j = 2$), a proportion π_0 of genes was simulated with $\mu_{i2} = 0$ (unmodified genes), while modified genes were simulated using three different configurations: (a) $\mu_{i2} = 1$ for the first half, $\mu_{i2} = 2$ for the second half; (b) $\mu_{i2} = 0.5$ for the first half, $\mu_{i2} = 1$ for the second half; and (c) $\mu_{i2} = 0.5$ for the first third, $\mu_{i2} = 1$ for the second third and $\mu_{i2} = 2$ for the last third.

In this way, we tried to mimic realistic situations with different patterns. Here, the distribution of modified genes is

a simple mixture of two components with small expression differences (configuration (a)) and large expression differences (configuration (b)), or a more complex mixture with three components (configuration (c)).

Four different π_0 values were considered. Because Efron's procedure was developed for situations with π_0 values greater than 0.90, we used $\pi_0 = 0.9$ and $\pi_0 = 0.98$. We also considered two lower values of π_0 that correspond to realistic situations not considered by Efron ($\pi_0 = 0.8$ and $\pi_0 = 0.6$). In total, $2 \times 3 \times 4 = 24$ different cases were considered.

To evaluate the behavior of the five procedures in the context of dependent data, we also generated datasets with so-called clumpy dependence (that is, datasets for which the measurements on the genes are dependent in small groups, with each group being independent of the others).

We applied the protocol described in [18] and [19] as follows: First, an independent dataset matrix (x_{ijk}) was generated, as described above. Then, for each group of 100 genes, a random vector $\mathbf{A} = \{a_{jk}\}$, where $j = 1, 2$ and $k = 1,..., 10$ was generated from a standard normal distribution. The data matrix (y_{ijk}) was then built so that: $y_{ijk} = \sqrt{\rho}\, a_{jk} + \sqrt{1 - \rho}\, x_{ijk}$ with $\rho = 0.5$. Thus, in each group, the genes have the same correlation, that is to say for $i_1 \neq i_2$, $Corr(y_{i_1 j}, y_{i_2 j}) = 0.5$. To render the results comparable with those obtained in the independent setting, the expectations μ_{ij} used for generating the matrix (x_{ijk}) were divided by $\sqrt{1 - \rho}$ so that the expectations of the random variables Y_{ijk} correspond to those described in configurations (a), (b) and (c) for independent data. We also considered other ρ values that gave similar results (data not shown).

In each case, the *p*-values, calculated under the null hypothesis $H_0 : \mu_{i1} = \mu_{i2}$, were obtained from the Student's statistic. Then, we estimated *lFDR* from our procedure, referred to as *polfdr*, and the four procedures presented in the background section, referred to as *locfdr* (Efron), *LocalFDR* (Aubert *et al.*), *twilight* (Scheid and Spang), *pava.fdr* (Broberg). Although these procedures were not designed to estimate the probability π_0 independently of *lFDR*, we also compared the estimators of π_0 obtained from the five procedures.

For each case, 1,000 datasets were simulated. To compare the different estimators, we considered three different criteria that are described below.

Criterion 1

Since the main contribution of *lFDR* is that it gives each tested hypothesis its own measure of significance, a small bias for any value within the whole interval [0, 1] can be preferable to a smaller bias limited to a subset of values within the interval. For this purpose and to assess the amplitude of the bias for the five procedures, we considered the infinity norm of the integrated error over the interval [0, 1] defined as follows:

$$b_1 = \max_{p \in [0,1]} \left| E\{\widehat{lFDR}(p) - lFDR(p)\} \right| \qquad (8)$$

and estimated by:

$$\hat{b}_1 = \max_{i=1,...,m} \left| \frac{1}{1,000} \sum_{k=1}^{1,000} \{\widehat{lFDR}(p_i^{(k)}) - lFDR(p_i^{(k)})\} \right|$$

$$(9)$$

where $p_i^{(k)}$ $i = 1,...,m$ are the m *p*-values corresponding to the kth dataset (among the 1,000 simulated datasets for each case). Here, the theoretical values $lFDR(p_i^{(k)})$ are calculated from a numerical approximation of the non-centered Student's distribution [20].

The estimated values of b_1 for independent data are reported in the Table 1. Although these values were always less than or equal to 0.17 for the *polfdr* procedure, the highest b_1 values for the *LocalFDR*, *pava.fdr*, *twilight* and *locfdr* procedures were 0.20, 0.21, 0.43 and 0.87, respectively. These results also showed that the *locfdr* method tended to substantially overestimate *lDFR*. For example, Figure 1 shows the expected *lFDR* as a function of *p* for each estimator with $m = 500$, $\pi_0 = 0.8$ and configuration (c) (the figures corresponding to all the other cases are provided in additional files). For these figures, the horizontal scale was log-transformed to better demonstrate the differences between the methods for small *p*-values. For dependent datasets, the bias of the five estimators increased. While the bias of our estimator was always less than or equal to 0.17, the highest bias values for the methods *pava.fdr*, *LocalFDR*, *twilight*, *locfdr* were 0.20, 0.23, 0.41 and 0.87, respectively (see additional files, Table 10).

Criterion 2

As noted under Background, the five methods were designed to estimate an *lFDR* upper bound. However, a negative bias can occur in some cases, leading to more false positive results than expected. In this context, we propose investigating with the five procedures the minimal negative bias, denoted b_2, over the interval [0, 1]:
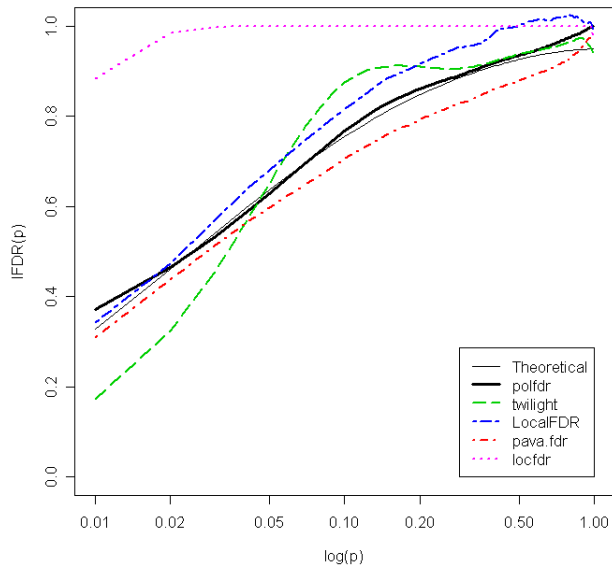
Table 1: Estimated values of b_1 for the five estimators in each independent simulated case.

Case	m	π_0	Configuration	polfdr	twilight	LocalFDR	pava.fdr	Locfdr
1	500	0.6	(a)	0.032	0.047	0.067	0.133	0.869
2			(b)	0.170	0.149	0.195	0.160	0.836
3			(c)	0.118	0.123	0.155	0.096	0.843
4		0.8	(a)	0.062	0.131	0.041	0.116	0.695
5			(b)	0.071	0.097	0.105	0.061	0.599
6			(c)	0.051	0.156	0.079	0.057	0.555
7		0.9	(a)	0.071	0.268	0.041	0.115	0.312
8			(b)	0.054	0.116	0.052	0.047	0.376
9			(c)	0.050	0.315	0.049	0.095	0.265
10		0.98	(a)	0.073	0.387	0.163	0.139	0.113
11			(b)	0.051	0.105	0.029	0.135	0.098
12			(c)	0.061	0.260	0.120	0.157	0.109
13	5,000	0.6	(a)	0.035	0.038	0.026	0.212	0.869
14			(b)	0.171	0.167	0.165	0.167	0.839
15			(c)	0.118	0.129	0.117	0.065	0.843
16		0.8	(a)	0.056	0.129	0.013	0.092	0.441
17			(b)	0.071	0.110	0.073	0.068	0.502
18			(c)	0.051	0.156	0.053	0.039	0.406
19		0.9	(a)	0.083	0.268	0.039	0.056	0.183
20			(b)	0.033	0.123	0.036	0.032	0.297
21			(c)	0.057	0.316	0.043	0.029	0.184
22		0.98	(a)	0.035	0.427	0.183	0.035	0.052
23			(b)	0.046	0.071	0.035	0.027	0.081
24			(c)	0.034	0.293	0.141	0.035	0.047

$$b_2 = \left| \min_{p \in [0,1]} \left(E[\widehat{lFDR}(p) - lFDR(p)] \times 1_{\{E[\widehat{lFDR}(p) - lFDR(p)] < 0\}} \right) \right| \tag{10}$$

and estimated by:

$$\hat{b}_2 = \left| \min_{i=1,\dots,m} \left(\frac{1}{1000}\sum_{k=1}^{1000}\{\widehat{lFDR}(p_i^{(k)}) - lFDR(p_i^{(k)})\} \times 1_{\left\{\frac{1}{1000}\sum_{k=1}^{1000}\{\widehat{lFDR}(p_i^{(k)}) - lFDR(p_i^{(k)})\} < 0\right\}} \right) \right| \tag{11}$$

Results for independent datasets (Table 2) indicated that all the estimators have non-negligible minimal negative biases. However, while b_2 was always less than or equal to 0.08 for our method, the maximal b_2 values were 0.11, 0.18, 0.21 and 0.43 for the estimators *locfdr*, *LocalFDR*, *pava.fdr* and *twilight*, respectively. More precisely, while our estimator slightly underestimated *lFDR* in some cases, when π_0 was close to 1, the *twilight* method tended to underestimate *lFDR* for small *p*-values (see Figure 1) and the *pava.fdr* method tended to substantially underestimate *lFDR* for all *p*-values (for example, see Figure 2). The *pava.fdr* method underestimation can be attributed to the upper bound of π_0, which is estimated by $\min[\hat{f}(p_{(i)})]$, because $E\{\min[\hat{f}(p_{(i)})]\} \leq \min[E\hat{f}(p_{(i)})]\}$. Thus, even

Figure 1
Expected lFDR as a function of log(p) for each estimator with m = 500, π_0 = 0.8 and configuration (c).

Table 2: Estimated values of b_2 for the five estimators in each independent simulated case.

Case	m	π_0	Configuration	polfdr	twilight	LocalFDR	pava.fdr	locfdr
1	500	0.6	(a)	0.015	0.047	0.000	0.133	0.000
2			(b)	0.000	0.016	0.000	0.000	0.000
3			(c)	0.000	0.039	0.000	0.010	0.000
4		0.8	(a)	0.057	0.131	0.000	0.116	0.000
5			(b)	0.000	0.071	0.000	0.024	0.000
6			(c)	0.011	0.156	0.000	0.057	0.000
7		0.9	(a)	0.071	0.268	0.041	0.115	0.046
8			(b)	0.005	0.116	0.013	0.047	0.031
9			(c)	0.040	0.315	0.049	0.095	0.050
10		0.98	(a)	0.073	0.387	0.163	0.139	0.113
11			(b)	0.051	0.105	0.029	0.135	0.098
12			(c)	0.061	0.260	0.120	0.157	0.109
13	5,000	0.6	(a)	0.011	0.019	0.000	0.212	0.000
14			(b)	0.000	0.018	0.000	0.000	0.000
15			(c)	0.000	0.041	0.000	0.000	0.000
16		0.8	(a)	0.056	0.129	0.005	0.092	0.000
17			(b)	0.000	0.079	0.000	0.000	0.000
18			(c)	0.016	0.156	0.000	0.003	0.000
19		0.9	(a)	0.083	0.268	0.039	0.056	0.001
20			(b)	0.000	0.123	0.021	0.000	0.000
21			(c)	0.057	0.316	0.043	0.029	0.000
22		0.98	(a)	0.027	0.427	0.183	0.035	0.023
23			(b)	0.010	0.071	0.035	0.027	0.017
24			(c)	0.018	0.293	0.141	0.035	0.021
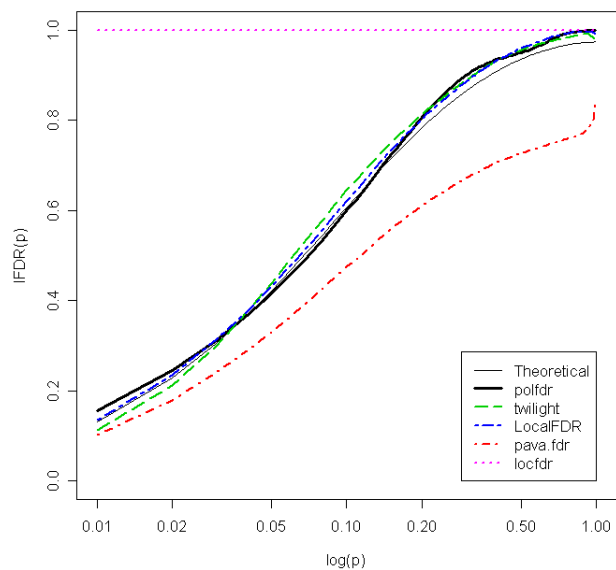
though this method can sometimes lead to a low bias (because its negative bias compensates for the gap between the upper bound and the true value), this estimator can generate high negative bias (see Figure 2). These results also indicated that even though the *locfdr* method tended to overestimate *lFDR* for the majority of *p*-values, it also tended to underestimate *lFDR* for *p*-values close to 1.

Criterion 3
To evaluate the accuracy of the five procedures at all points simultaneously, we estimated the root mean integrated square error (*RMISE*) of the five estimators which is defined by:

$$RMISE = \sqrt{E\left[\int_0^1 \left(\widehat{lFDR}(p) - lFDR(p)\right)^2 dp\right]} \quad (12)$$

and estimated by:

$$\widehat{RMISE} = \sqrt{\frac{1}{1,000}\sum_{k=1}^{1,000}\sum_{i=1}^{m}\left[\left(\widehat{lFDR}(p_i^{(k)}) - lFDR(p_i^{(k)})\right)^2 \times \left(p_{(i+1)}^{(k)} - p_{(i)}^{(k)}\right)\right]}$$

$$(13)$$

As shown in Table 3, these results indicated that, except for the *pava.fdr* method (which can substantially underestimate *lFDR*, as shown above), our method gave the lowest *RMISE* in 15/24 cases. For the 6 cases with π_0 close to

Figure 2
Expected lFDR as a function of log(p) for each estimator with m = 5000, π_0 = 0.6 and configuration (a).

Table 3: Estimated RMISE for the five estimators in each independent simulated case.

Case	m	π_0	Configuration	polfdr	twilight	LocalFDR	pava.fdr	locfdr
1	500	0.6	(a)	0.071	0.093	0.194	0.136	0.208
2			(b)	0.157	0.155	0.235	0.121	0.340
3			(c)	0.118	0.122	0.221	0.090	0.279
4		0.8	(a)	0.067	0.085	0.187	0.122	0.144
5			(b)	0.095	0.094	0.201	0.087	0.193
6			(c)	0.083	0.089	0.194	0.091	0.157
7		0.9	(a)	0.089	0.085	0.180	0.112	0.076
8			(b)	0.080	0.081	0.178	0.090	0.110
9			(c)	0.075	0.088	0.183	0.106	0.078
10		0.98	(a)	0.093	0.106	0.172	0.089	0.043
11			(b)	0.078	0.100	0.170	0.077	0.045
12			(c)	0.081	0.098	0.170	0.079	0.044
13	5,000	0.6	(a)	0.036	0.040	0.061	0.191	0.234
14			(b)	0.149	0.153	0.152	0.133	0.343
15			(c)	0.101	0.113	0.117	0.037	0.278
16		0.8	(a)	0.029	0.047	0.060	0.088	0.119
17			(b)	0.069	0.077	0.087	0.056	0.185
18			(c)	0.052	0.071	0.074	0.032	0.143
19		0.9	(a)	0.048	0.056	0.060	0.054	0.056
20			(b)	0.041	0.050	0.065	0.037	0.099
21			(c)	0.039	0.063	0.063	0.035	0.064
22		0.98	(a)	0.042	0.069	0.062	0.027	0.021
23			(b)	0.035	0.031	0.056	0.023	0.029
24			(c)	0.039	0.052	0.060	0.025	0.023

one (π_0 = 0.98), the *locfdr* method yielded the lowest *RMISE*. For the last 3 cases, the difference between our method's RMISE and the lowest value (obtained with the twilight estimator) did not exceed 0.4% (case 7). Moreover, these results also indicated that the *LocalFDR* estimator, despite a small bias in all cases had a higher *RMISE* than our estimator due to its wide variance.

For dependent data, the *RMISE* of the five estimators increased and the differences were smaller. Our method yielded the lowest *RMISE* for 7/24 cases (see the Table 12 in additional files).

However, because in practice, some investigators might want to select only genes with low *lFDR*, we also reported the results obtained with the 3 criteria over the interval [0, 0.2] (See additional files). They showed that our method maintained good performances compared to the four others. Other thresholds for the *p*-values were considered (10% and 40%) and gave similar results (data not shown).

To compare the performance of the different estimators of the parameter π_0 obtained with the different methods, we evaluated their expectations and their root mean square errors.

Table 4 gives the means of the five estimators of the parameter π_0 over the 1,000 simulated independent datasets (results for dependent datasets are provided in additional files, Tables 13–14). The average bias over the 24 simulated datasets was the smallest for our new method (0.1%) with a maximal positive bias of 12% (for m = 5, 000, π_0 = 60% and configuration (b)) and a maximal negative bias of 4% (for m = 500, π_0 = 98% and configuration (c)). It is worth noting that the method with the highest positive bias was *locfdr* (29%), while the one with the highest negative bias was *pava.fdr* (13%).

The estimated root mean square errors for each estimator of the parameter π_0 are given in Table 5. Note that the root mean square errors of our estimator were less than or equal to 0.126 for the 24 simulated datasets, while it could reach 0.130, 0.132, 0.145 and 0.292 for *locfdr*, *LocalFDR*, *twilight* and *pava.fdr* methods, respectively.

Concerning computing time, our procedure was rapid, while the *twilight* method was cumbersome and impracticably long for large numbers of tested hypotheses. For example, the means of computing times on a personal computer (over 20 simulated datasets) for m = 5, 000, π_0 = 0.6 and configuration (c) were 50s, 2s, 1s, 1s and 1s for the methods *twilight*, *LocalFDR*, *polfdr*, *pava.fdr* and *locfdr*,

Table 4: Mean of all estimates of π_0 for the five estimators in each independent simulated case.

Case	m	π_0	Configuration	polfdr	Twilight	LocalFDR	pava.fdr	locfdr
1	500	0.6	(a)	0.604	0.613	0.523	0.852	0.604
2			(b)	0.707	0.718	0.665	0.890	0.716
3			(c)	0.656	0.677	0.604	0.839	0.669
4		0.8	(a)	0.787	0.806	0.721	0.849	0.791
5			(b)	0.841	0.860	0.792	0.915	0.849
6			(c)	0.812	0.839	0.767	0.890	0.828
7		0.9	(a)	0.863	0.897	0.824	0.918	0.886
8			(b)	0.903	0.915	0.876	0.954	0.912
9			(c)	0.888	0.907	0.842	0.934	0.899
10		0.98	(a)	0.940	0.947	0.938	0.983	0.943
11			(b)	0.953	0.949	0.949	0.989	0.937
12			(c)	0.951	0.954	0.948	0.988	0.947
13	5,000	0.6	(a)	0.614	0.613	0.469	0.851	0.616
14			(b)	0.720	0.718	0.707	0.888	0.725
15			(c)	0.670	0.676	0.604	0.838	0.680
16		0.8	(a)	0.801	0.806	0.729	0.848	0.805
17			(b)	0.853	0.859	0.842	0.916	0.861
18			(c)	0.833	0.841	0.803	0.888	0.841
19		0.9	(a)	0.877	0.903	0.857	0.918	0.900
20			(b)	0.920	0.929	0.914	0.954	0.929
21			(c)	0.901	0.918	0.883	0.934	0.915
22		0.98	(a)	0.968	0.974	0.971	0.982	0.975
23			(b)	0.974	0.980	0.979	0.989	0.980
24			(c)	0.972	0.978	0.975	0.986	0.978

respectively. For a larger number tested hypotheses $m = 50,000$ (not considered in the simulation study), the means of computing times were 7,261s, 162s, 108s, 2s and 1s, respectively.

Real data
Our method, together with *twilight*, *LocalFDR*, *locfdr* and *pava.fdr*, was applied to two datasets from genomic breast-cancer studies (Hedenfalk *et al*. [21] and Wang *et al*. [22]).

Data from Hedenfalk et al. [21]
Hedenfalk *et al*. [21] investigated the gene-expression changes between hereditary (*BRCA1*, *BRCA2*) and non-hereditary breast cancers. The initial dataset consists of 3,226 genes with expression log-ratios corresponding to the fluorescent intensities from a tumor sample divided by those from a common reference sample. Like Aubert *et al*. [13], we focused on the comparison of *BRCA1* and *BRCA2*, and used the same *p*-values which were calculated for each gene from a two-sample *t*-test.

Figure 3 shows the estimated *lFDR* as a function of the *p*-values for the five estimators. The five procedures yielded different results. For example, the estimated *lFDR* for 3 different genes are reported in Table 6. These results show clear differences between the five methods. In particular, the *locfdr* method gave 1 for the three genes, which can be explained by a π_0 value smaller than 0.9. Indeed, the estimated π_0 values were, respectively, 0.67, 0.67, 0.66, 0.66

and 1 for the *polfdr*, *twilight*, *LocalFDR*, *pava.fdr* and *locfdr* methods. Concerning the four remaining procedures, the highest differences for the three genes were respectively 3%, 7% and 5%.

Data from Wang et al. [22]
Wang *et al*. [22] wanted to provide quantitative gene-expression combinations to predict disease outcomes for patients with lymph-node negative breast cancers. Over 22,000 expression measurements were obtained from Affymetrix oligonucleotide microarray U133A GeneChips for 286 samples. The expression values calculated by the Affymetrix GeneChip analysis software MAS5 are available on the GEO website [23] with clinical data. For normalisation, the quantile method [24] was applied on log-transformed data.
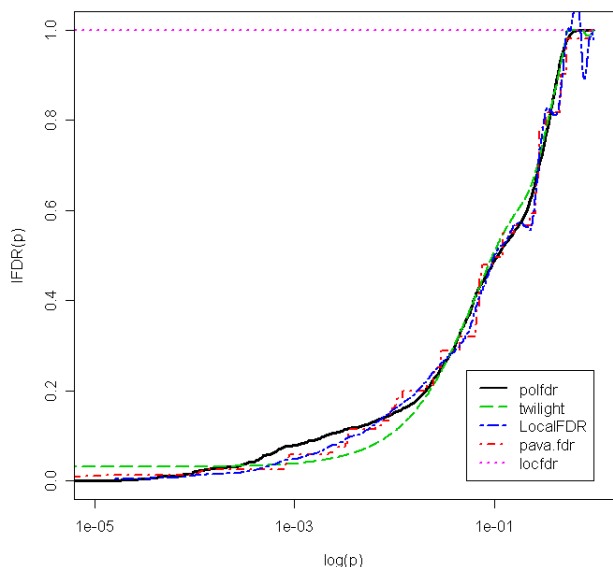
Here, we focused on identifying gene-expression changes that distinguish patients who experienced a tumour relapse within 5 years, from patients who continued to be disease-free after a period of at least 5 years. The *p*-values were calculated for each gene from a two-sample *t*-test and the five methods were applied.

Figure 4 shows the estimated *lFDR* as a function of the *p*-values for the 5 estimators. As noted above, *FDR* can be estimated from *lFDR* using equation (3) via the mean of the estimated *lFDR* over the rejection region Γ. When selecting all genes so that the estimated *FDR* is less than

Table 5: Mean square error of all estimates of π_0 for the five estimators in each independentsimulated case.

Case	M	π_0	Configuration	polfdr	twilight	LocalFDR	pava.fdr	locfdr
1	500	0.6	(a)	0.048	0.084	0.089	0.255	0.052
2			(b)	0.126	0.145	0.088	0.292	0.130
3			(c)	0.086	0.116	0.054	0.241	0.089
4		0.8	(a)	0.052	0.090	0.096	0.057	0.056
5			(b)	0.078	0.109	0.064	0.120	0.080
6			(c)	0.065	0.099	0.067	0.096	0.065
7		0.9	(a)	0.074	0.080	0.093	0.039	0.053
8			(b)	0.063	0.080	0.075	0.065	0.062
9			(c)	0.060	0.084	0.088	0.050	0.056
10		0.98	(a)	0.077	0.076	0.069	0.040	0.064
11			(b)	0.067	0.072	0.053	0.041	0.071
12			(c)	0.064	0.066	0.056	0.041	0.060
13	5,000	0.6	(a)	0.023	0.029	0.132	0.251	0.024
14			(b)	0.124	0.121	0.109	0.288	0.127
15			(c)	0.075	0.081	0.015	0.238	0.083
16		0.8	(a)	0.017	0.032	0.073	0.049	0.021
17			(b)	0.061	0.066	0.046	0.116	0.065
18			(c)	0.043	0.050	0.014	0.089	0.047
19		0.9	(a)	0.039	0.031	0.045	0.021	0.019
20			(b)	0.034	0.042	0.027	0.055	0.035
21			(c)	0.029	0.036	0.023	0.036	0.024
22		0.98	(a)	0.025	0.025	0.013	0.012	0.018
23			(b)	0.024	0.023	0.009	0.015	0.018
24			(c)	0.023	0.024	0.011	0.014	0.018

5%, our method selected 325 genes while the *pava.fdr* and *LocalFDR* methods selected 367 and 229 genes, respectively, and the *twilight locfdr* methods did not select any gene. It is worth noting that these strong differences have substantial consequences on the following analyses. The estimated π_0 values were, respectively, 0.711, 0.720, 0.714, 0.723 and 0.914 for the *polfdr*, *pava.fdr*, *LocalFDR*, *twilight* and *locfdr* methods.

Discussion

In the simulations, for independent datasets, the results indicated good performances for our procedure compared to the four previously published methods. Indeed, while the infinity norm b_1 was small in every simulated case with our procedure, it could be large for *twilight* and *locfdr* procedures. Moreover, despite the fact that the five estimators were designed with conservative biases, the *twilight* procedure could generate substantial negative bias for small *p*-values, the *locfdr* procedure underestimated the *lFDR* for *p*-values close to 1, and *pava.fdr* tended to underestimate *lFDR* for all *p*-values. In addition, and compared to *LocalFDR*, our method gave smaller *RMISE* in all cases. When considering only the lowest *p*-values, the simulation results showed the same trend. In summary, our new estimator exhibited more stable behavior than the four others.

For dependent datasets, simulation results led to similar conclusions. Indeed, correlations between genes do not

Figure 3
Estimated lFDR as a function of log(p) for each estimator for the Hedenfalk et al. dataset.

Table 6: *lFDR* estimations for three genes in Hedenfalk et al. data.

p-value	Rank	*polfdr*	*twilight*	*LocalFDR*	*pava.fdr*	*locfdr*
0.00041	36	0.05	0.03	0.02	0.03	1
0.01294	297	0.16	0.13	0.18	0.20	1
0.30534	1604	0.73	0.75	0.77	0.78	1

affect the marginal distribution of the *p*-values but increase the variability of the different methods and the bias of the estimators of π_0.

It is worth noting that a major assumption underlying our procedure, like *twilight*, *LocalFDR* and *pava.fdr*, relies on the distribution of the *p*-values under the null hypothesis. Because the uniformity assumption is sometimes not tenable [12], Efron's procedure estimates the null distribution parameters from the observed marginal distribution. However, a limitation of that approach is the need for additional assumptions concerning the proportion of true null hypotheses. Another way to address the problem of the null distribution is how the *p*-values are calculated, notably using sampling methods (for a few [25-27]).

Conclusion

Herein, we proposed a novel, simple and efficient procedure for estimating the *lFDR*. Estimating its value is essential for genomic studies, as it quantifies gene-specific evidence for being associated with the clinical or biological variable of interest. Moreover, it enables calculation of the *FDR*.

As seen from the simulation results, our new estimator performed well in comparison to *locfdr*, *twilight*, *LocalFDR* and *pava.fdr*. As discussed above, our method yielded a positive bias for *lFDR* that reflects the conservative estimation of the probability π_0. However, this limitation is compensated for by the fact that no assumption is required for f_1.

Finally, we think that extending our approach to multidimensional settings could be useful, as recently discussed by Ploner *et al.* [28], but will require additional investigations.

The R function *polfdr* that implements the procedure is available on the polfdr website [30].

Methods

As for the procedures proposed by Aubert *et al.*, Scheid and Spang and Broberg, we make the assumption that, under the null hypothesis, the *p*-values are uniformly distributed. However, instead of estimating the density *f* (and then taking the reciprocal of the estimate), we directly estimate the reciprocal of *f*.

1/f estimation

Let's consider $\varphi = F^{-1}(p)$, the inverse cumulative distribution function of the *p*-values. Then, $\forall p \in [0, 1]$, $\varphi(F(p)) = p$ and $1/f$ is the first derivative of the function φ. Indeed, since $\varphi \circ F$ is the identity function:
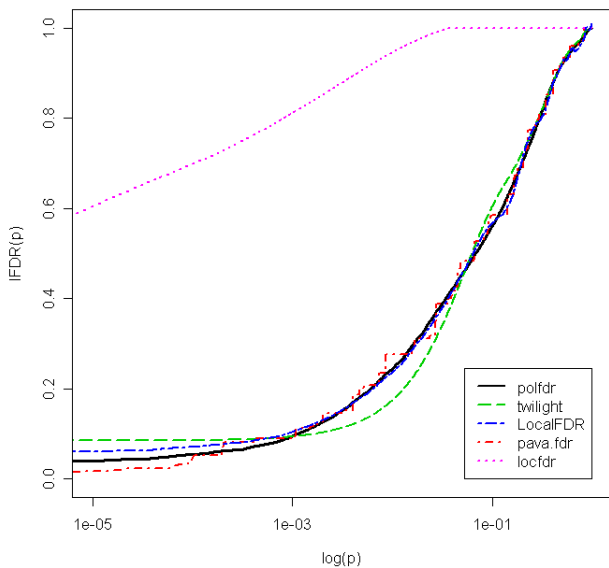
$$\frac{d\varphi(F(p))}{dp} = 1. \quad (14)$$

Moreover:

$$\frac{d\varphi(F(p))}{dp} = \frac{dF(p)}{dp} \times \frac{d\varphi(F(p))}{dF(p)} = f(p) \times \frac{d\varphi(F(p))}{dF(p)}. \quad (15)$$

Thus:

$$\frac{1}{f(p)} = \frac{d\varphi(F(p))}{dF(p)} \quad (16)$$

Figure 4
Estimated lFDR as a function of log(p) for each estimator for the Wang et al. dataset.

Equation 16, illustrated in the Figure 5, is linked to the geometrical relationship between the *FDR* and *lFDR*, as noted by Efron [6].

Because the *lFDR* (and thus $1/f$) is non-negative, the function φ is non-decreasing. Moreover, assuming that *lFDR* is non-decreasing with p (that is to say that, the closer a *p-value* is to one, the greater the probability that the null hypothesis is true), the function φ is convex. Then, we propose using a convex 10-degree polynomial for φ.

Therefore, we consider the following linear formulation to represent the relationship between the *p*-values and the empirical cumulative distribution function:

$$\mathbf{p} = \widetilde{\mathbf{F(p)}}\mathbf{A} + \mathbf{E} \qquad (17)$$

where $\mathbf{p} = t(p_{(1)},...,p_{(m)})$ is the column vector of observed *p*-values, $\widetilde{\mathbf{F(p)}} = \left(\widetilde{\mathbf{F(p)}}^{0},...,\widetilde{\mathbf{F(p)}}^{d} \right)$, $\widetilde{\mathbf{F(p)}}$ is the vector of the empirical cumulative distribution function of the *p*-values, $\mathbf{A} = t(a_0,...,a_d)$ is the column vector of the polynomial's coefficients, d is the degree of the polynomial, and $\mathbf{E}$, the error term, is a random vector for which the expectation is 0.

The estimator of the polynomial regression coeffcients' vector $\mathbf{A}$ can be obtained by solving the following least-square minimization problem with constraints:

$$\min_{\mathbf{CA} \geq 0} \left\| \widetilde{\mathbf{F(p)}}\mathbf{A} - \mathbf{p} \right\|^2 \qquad (18)$$
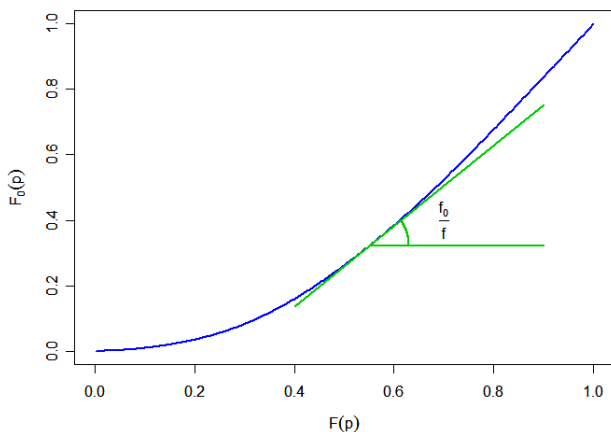
Figure 5
Graph of the null cumulative distribution versus the marginal cumulative distribution.

where

$$\mathbf{C} = \begin{pmatrix} 0 & \cdots & d(d-1)\left(\dfrac{1}{m}\right)^{d-2} \\ \cdots & k(k-1)\left(\dfrac{i}{m}\right)^{k-2} & \cdots \\ 0 & \cdots & d(d-1)\left(\dfrac{m}{m}\right)^{d-2} \\ 0 & \cdots & d\left(\dfrac{1}{m}\right)^{d-1} \\ \cdots & k\left(\dfrac{i}{m}\right)^{k-1} & \cdots \\ 0 & \cdots & d\left(\dfrac{m}{m}\right)^{d-1} \end{pmatrix}$$

$$(19)$$

We impose the constraints $\mathbf{CA} \geq 0$ on our minimization problem due to the convexity and monotony of φ, which can be written: $\forall i \in \{1,...,m\}$,

$$\varphi''(i/m) = \sum_{k=2}^{d} \left\{ k(k-1)\left(\frac{i}{m}\right)^{k-2} \times a_k \right\} \geq 0 \qquad \text{and}$$

$$\varphi'(i/m) = \sum_{k=1}^{d} \left\{ k\left(\frac{i}{m}\right)^{k-1} \times a_k \right\} \geq 0.$$ Quadratic programming is used to calculate the solution ([29]). Finally, an estimate of $1/f(p) = \varphi'(p)$ is deduced from the estimated regression coefficients.

π_0 estimation

Classical approaches attempted to estimate π_0 from $f(1)$, which is the lowest upper bound of π_0 based on the mixture model (4). Indeed, if no assumption is made for f_1, π_0 is not identifiable and $f(1)$ is the lowest upper bound based on the equation (4). Here, we propose using the same model to estimate π_0 that is used to estimate $1/f$. Therefore, we consider the reciprocal of the function φ. However, due to higher bias and variance at the boundaries of the domain, estimating π_0 from a value close (but not equal) to 1 is more appropriate. In order to obtain a less sensitive estimator with respect to φ', it is reasonable to estimate π_0 at the point where φ'' is at its minimum:

$$\hat{\pi}_0 = \frac{1}{\varphi'(\arg\min_{x>a}(\varphi''(x)))}. \qquad (20)$$

In practice, we propose setting $a = 0.5$. Note that the estimation of π_0 is not sensitive to the choice of a and other values can be considered.

Authors' contributions

CD, ABH and PB have equally contributed to this work. All authors read and approved the final manuscript.

Additional material

Additional file 1

Figures_independent

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-229-S1.pdf]

Additional file 2

Figures_dependent

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-229-S2.pdf]

Additional file 3

Tables

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-229-S3.pdf]

References

1. Hochberg Y, Tamhane A: *Multiple Comparison Procedures* Wiley; 1987.
2. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57**:289-300.
3. Storey JD: **A direct approach to false discovery rates.** *J R Stat Soc Ser B* 2001, **64**:479-498.
4. Glonek G, Salomon P: **Comment on 'Resampling-based multiple testing for microarray data analysis' by Ge Y, Dudoit S, Speed T.** *TEST* 2003, **12**:1-44.
5. Efron B, Tibshirani R, Storey J, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J Am Stat Assoc* 2001, **96**:1151-1160.
6. Efron B: **Local false discovery rates.** *Technical Report* 2005 [http://www-stat.stanford.edu/~brad/papers/False.pdf].
7. Liao JG, Lin Y, Selvanayagam ZE, Shih WJ: **A mixture model for estimating the local false discovery rate in DNA microarray analysis.** *Bioinformatics* 2004, **20**:2694-701.
8. Pan W, Lin J, Le C: **A mixture model approach to detecting differentially expressed genes with microarray data.** *Funct Integr Genomics* 2003, **3**:117-24.
9. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **5**:155-76.
10. Broët P, Lewin A, Richardson S, Dalmasso C, Magdelenat H: **A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments.** *Bioinformatics* 2004, **20**:2562-2571.
11. Langaas M, Lindqvist B, Ferkingstad E: **Estimating the proportion of true null hypotheses, with application to DNA microarray data.** *J R Stat Soc Ser B* 2005, **67**:555-572.
12. Efron B: **Large-scale simultaneous hypothesis testing: the choice of a null hypothesis.** *J Am Stat Assoc* 2004, **99**:96-104.
13. Aubert J, Bar-Hen A, Daudin JJ, Robin S: **Determination of the differentially expressed genes in microarray experiments using localFDR.** *BMC Bioinformatics* 2004, **5**:125.
14. Scheid S, Spang R: **A stochastic downhill search algorithm for estimating the local false discovery rate.** *IEEE Transactions on Computational Biology and Bioinformatics* 2004, **1**:98-108.
15. Broberg P: **A comparative review of estimates of the proportion unchanged genes and the false discovery rate.** *BMC Bioinformatics* 2005, **6**:199.
16. Storey JD, Tibshirani R: **Statistical significance for genome-wide studies.** *Proc Natl Acad Sci* 2003, **100**:9440-9445.
17. Hochberg Y, Benjamini Y: **More powerful procedures for multiple significance testing.** *Stat Med* 1990, **9**:811-818.
18. Storey JD, Tibshirani R: **Estimating false discovery rates under dependence, with applications to DNA microarrays.** In *Technical Report 2001–28* Department of Statistics, Stanford University; 2001.
19. Qiu X, Klebanov L, Yakovlev A: **Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes.** *Stat Appl Genet Mol Biol* 2005, **4**:Article34.
20. Johnson NL, Kotz S, Balakrishnan N: *Continuous Univariate Distributions, chapters 28 and 31 Volume 2*. Wiley, New York; 1995.
21. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg A, Trent J, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty E, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Sauter G: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344**:539-548.
22. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EM, Atkins D, Foekens JA: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.** *Lancet* 2005, **365**:671-679.
23. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo]
24. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance.** *Bioinformatics* 2003, **19**:185-193.
25. Pan W: **On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression.** *Bioinformatics* 2003, **19**:1333-1340.
26. Guo X, Pan W: **Using weighted permutation scores to detect differential gene expression with microarray data.** *J Bioinform Comput Biol* 2005, **3(4)**:989-1006.
27. Xie Y, Pan W, Khodursky AB: **A note on using permutation based false discovery rate estimate to compare different analysis methods for microarray data.** *Bioinformatics* 2005, **21**:4280-4288.
28. Ploner A, Calza S, Gusnanto A, Pawitan Y: **Multidimensional local false discovery rate for microarray studies.** *Bioinformatics* 2006, **22**:556-565.
29. Gill PE, Murray W, Wright MH: *Practical Optimization* London: Academic Press; 1981.
30. **polfdr** [http://ifr69.vjf.inserm.fr/polfdr]