Genet. Sel. Evol. 36 (2004) 191–205
© INRA, EDP Sciences, 2004
DOI: 10.1051/gse:2003058

Original article

Bootstrapping of gene-expression data improves and controls the false discovery rate of differentially expressed genes

Theo H.E. Meuwissen[a*], Mike E. Goddard[b]

[a] Institute for Animal Science, Agricultural University of Norway, 1432 Ås, Norway
[b] Institute of Land and Food Resources, University of Melbourne, Parkville, 3052 Australia, and Victorian Institute of Animal Science, Attwood, Victoria, 3049 Australia

(Received 17 December 2002; accepted 23 October 2003)

Abstract – The ordinary-, penalized-, and bootstrap *t*-test, least squares and best linear unbiased prediction were compared for their false discovery rates (FDR), *i.e.* the fraction of falsely discovered genes, which was empirically estimated in a duplicate of the data set. The bootstrap-*t*-test yielded up to 80% lower FDRs than the alternative statistics, and its FDR was always as good as or better than any of the alternatives. Generally, the predicted FDR from the bootstrapped *P*-values agreed well with their empirical estimates, except when the number of mRNA samples is smaller than 16. In a cancer data set, the bootstrap-*t*-test discovered 200 differentially regulated genes at a FDR of 2.6%, and in a knock-out gene expression experiment 10 genes were discovered at a FDR of 3.2%. It is argued that, in the case of microarray data, control of the FDR takes sufficient account of the multiple testing, whilst being less stringent than Bonferoni-type multiple testing corrections. Extensions of the bootstrap simulations to more complicated test-statistics are discussed.

microarray data / gene expression / non-parametric bootstrapping / *t*-test / false discovery rates

1. INTRODUCTION

DNA microarrays can measure the expressions of tens of thousands of genes simultaneously, which provides us with a new, very powerful tool for the study of gene regulatory and metabolic networks [2, 13]. Typically two treatments are compared for the level of expression of many genes. Such data might traditionally be analyzed using one *t*-test for each gene. However, the large number of tests makes the type I error rates large and hard to control [15]. Statistical testing is further complicated by the small number of replicates within each

[*] Corresponding author: theo.meuwissen@iha.nlh.no

treatment (1-20); and by gene-expression data not following a (Log-) Normal distribution [14].

Because the number of genes on a microarray is often very large, and every gene is tested for its treatment effect, statistical significance testing should account for the large number of tests performed. Traditionally, multiple statistical tests are conducted while controlling the probability of making one or more type I errors (*e.g.*, the Bonferoni multiple testing correction). Benjamini and Hochberg [1] and Tusher *et al.* [15] suggest the control of the false discovery rate (FDR), *i.e.* the proportion of rejected null-hypotheses that are actually true. When some of the null-hypotheses are false, FDR control is less strict than controlling the type I error probability. We will assume here that the gene-expression experiment is conducted to point us to the genes affected by the treatment, and that further research will sort out the details of this effect. In this case, some erroneously rejected null-hypothesis can be accepted, as long as there are not too many relative to the total number of detected genes. Hence, controlling the FDR may be a reasonable strategy.

The non-Normality of gene-expression data hampers the ranking of the gene-expression effects for their statistical significance, *i.e.* we do not know which genes are most likely to have a real effect. Although, the t-test statistic is optimal for Normally distributed data, Tusher *et al.* [15] found better rankings using the penalized t-statistic:

$$t_p = \frac{m_1 - m_2}{a + \sqrt{s_1^2/n_1 + s_2^2/n_2}} \tag{1}$$

where m_i (s_i^2) is the mean (variance) of the gene-expressions under the ith treatment, and a is a constant which is added to avoid small s_i^2 values resulting in large and thus apparently significant t-statistics. Lönnstedt and Speed [10] and Efron *et al.* [5] used the 90th percentile of the standard errors of all the genes, *i.e.*, for 90% of the genes, a is larger than the usual denominator of the t-statistic ($\sqrt{s_1^2/n_1 + s_2^2/n_2}$). Although, adding a to the denominator avoids large t-statistics due to small (underestimated) standard errors, the statistical justification for this addition is lacking, and hence the value of a is based on heuristics and empirical evidence.

Kerr and Churchill [9] propose the use of linear models for the analysis of appropriately transformed gene-expression data, either using Least Squares (LS, [12]) with homogeneous or heterogeneous error variance (heterogeneous error variance implies that a separate error variance is estimated for every gene). Alternatively, they suggest the use of random effects models which use BLUP (best linear unbiased prediction) for the estimation of the gene-expression effects. Another alternative is to use non-parametric bootstrapping in order to account for any non-normality of the transformed gene-expression data [8]. However, which of these methods is most appropriate for the analysis of gene-expression data is not clear. The aim of this study is to compare ordinary -, penalized -, and bootstrap *t*-tests, and LS and BLUP models with homogeneous and heterogeneous error variance for their false discovery rates of differentially expressed genes. The false discovery rates were empirically assessed by finding the differentially expressed genes in a first data set, and confirming their expression in a second data set. The methods were compared in two publicly available data sets: the leukemia data of Golub *et al.* [6], and the apoAI knockout mice data of Callow *et al.* [4].

2. METHODS

2.1. Leukemia data

The advantage of the leukemia data of Golub *et al.* [6] is that it actually contains data from two replicated experiments contrasting gene expressions in acute myeloid leukemias (AML) to those of acute lymphoblastic leukemia (ALL). The two data sets are called TRAIN and INDEPEND. TRAIN is used here to estimate and rank the gene-expression effects for their significance, and INDEPEND is used for verifying the effect. The leukemia data are described in detail by [6] and available at www.genome.wi.mit.edu.

The TRAIN data consisted of 38 bone marrow samples: 27 ALL and 11 AML samples. The INDEPEND data consisted of 17 AML and 17 ALL samples. Light-intensities (foreground minus background), that were smaller than 50, were considered not clearly above background, and were treated as missing records. The deletion of low intensity records may have biased the average expression of genes with extremely low expressions upwards, but this bias is conservative in the sense that it reduces the difference in expressions between AML and ALL, *i.e.* it reduces the false discovery rate. The records were log-transformed before being analyzed.

2.2. ApoAI knockout mice data

The apoAI data are described in detail by Callow *et al.* [4] and are available at stat-www.berkeley.edu/users/terry/zarray/html/apodata.html. This data consisted of 8 samples from knockout mice and 8 samples from control mice.

In order to obtain again a test and a control data set, the data were arbitrarily split into two sub sets called: DATA1 and DATA2. Each sub set consisted of 4 of the knockout mouse arrays and 4 of the control mouse arrays. The apoAI-knockout mouse data contained some light-intensities (foreground minus background) of 0 which were treated as missing records. Records were log-transformed before being analyzed.

2.3. False discovery rates

The FDR of the, say, 200 most significant genes is predicted using [1]:

$$\mathrm{FDR(TNS)} = \min_{i \geq \mathrm{TNS}} \left[\frac{N * P(i)}{i} \right] \tag{2}$$

where N = the total number of genes in the analysis (N = 5284 and 6384 for the leukemia data and apoAI data, respectively); TNS = total number of significant genes (*e.g.* 200); $P(i)$ = P-value of the ith most significant gene as estimated from normal distribution theory or the bootstrap-t-test. In equation (2), $N * P(i)$ equals the expected number of false positives out of i-significant genes, and the minimization over i ensures that FDR increases monotonically as TNS increases.

In the case of the leukemia data, we used TRAIN to predict the FDR and INDEPEND to verify this prediction. For this verification, an empirical estimate of the FDR was obtained by counting how many of the significant effects fail to be in the same direction when estimated in the INDEPEND set. Under the null-hypothesis of no treatment effect, 50% of the INDEPEND estimates will be in the opposite direction to those of the TRAIN data. Thus, an empirical estimate of the false discovery rate is FDRe = 2 $*$ NOD/TNS, where NOD is the number of significant effects in TRAIN that are in the Opposite Direction in INDEPEND, and TNS is the total number of significant effects in TRAIN. A more formal justification for the FDRe estimated is given in the Appendix, together with a simulation study to test this empirical estimate of the FDR. A second estimate of FDRe is obtained by swapping the two data sets, *i.e.* determining significance in INDEPEND and checking the direction of the effect in TRAIN. However, this second estimate of the FDRe is not independent of the first, *i.e.* its information content is lower than that of an independent second estimate. In the case of the apoAI data, TRAIN is replaced by DATA1 and INDEPEND by DATA2.

2.4. Methods of analyses

The t-test statistic is obtained by applying equation (1) and setting $a = 0$. The penalized t_p-test is also obtained from equation (1), with a equal to the 90th percentile of the standard errors of all the genes [5, 10]. For the Least Squares (LS; [12]) analysis the model fitted was:

$$\mathrm{y}_{ijk} = \mu + \mathrm{m}_i + \mathrm{g}_j + \mathrm{t}_k + (\mathrm{g} * \mathrm{t})_{jk} + \mathrm{e}_{ijk},$$

where y_{ijk} = log-transformed light-intensity; μ = overall mean; m_i = effect of ith mRNA-sample; g_j = effect of jth gene; t_k = effect of kth treatment; and $(\mathrm{g} * \mathrm{t})_{jk}$ = the gene-by-treatment effect. In the LS model, the error variance was either estimated across all genes (homogeneous error variance) or estimated within each gene (heterogeneous error variance). The test-statistic used in the LS model was $\mathrm{z}_i = ((\mathrm{g} * \mathrm{t})_{i1} - (\mathrm{g} * \mathrm{t})_{i2})/se_i$, where se_i = standard error of the estimate of $((\mathrm{g} * \mathrm{t})_{i1} - (\mathrm{g} * \mathrm{t})_{i2})$.

The BLUP model [7] equals the LS model except that the gene*treatment effects are assumed random, *i.e.* they are assumed to be sampled from a distribution of effects with mean 0 and variance $\sigma^2_{gti} = \sigma^2_{ei}/\lambda$ where λ is the usual BLUP variance ratio (error variance over gene-by-treatment variance), σ^2_{ei} = error variance, which was either assumed homogeneous (same for all genes) or heterogeneous (different for each gene). The error variances, σ^2_{ei}, and the variance ratio, λ, which was assumed constant across all genes, were estimated by residual maximum likelihood (REML; [11]). The test-statistic for the BLUP model was also $z_i = ((\mathrm{g} * \mathrm{t})_{i1} - (\mathrm{g} * \mathrm{t})_{i2})/se_i$, but estimates of $((\mathrm{g} * \mathrm{t})_{i1} - (\mathrm{g} * \mathrm{t})_{i2})$ and se_i differ from those of the LS model. Compared with the LS model, BLUP will regress the $(\mathrm{g} * \mathrm{t})_{ik}$ effects back to zero when the information on the $(\mathrm{g} * \mathrm{t})_{ik}$ effects is small.

The following steps were followed to calculate non-parametric bootstrap-t-test P-values. For each gene with n_i records:

1. Calculate t-statistic, t_real, from the real data using the ordinary t-test;
2. sample, with replacement and without respecting the treatments, records from the real data to form a bootstrap-simulated data set under the null-hypothesis;
3. calculate the t-statistic from the bootstrap simulated data, *i.e.* under the null-hypothesis;
4. repeat steps 2 and 3 N_boot times and calculate the bootstrap-P-value as: $P_\mathrm{boot} = (\mathrm{count}(|t_\mathrm{k}| > |t_\mathrm{real}|) + 1)/(N_\mathrm{boot} + 1)$ where t_k = t-statistic of the kth bootstrap simulation out of a total of N_boot simulations, and $\mathrm{count}(|t_\mathrm{k}| > |t_\mathrm{real}|)$ denotes the number of simulations in which t_k is more extreme than t_real. N_boot was 100 000 simulations.

Table I. The empirical false discovery rate (FDRe) when the 200 most significant AML- *versus* ALL-effects in the TRAIN data were tested in the INDEPEND data, and *vice versa*.

Method	FDRe (%)
t-test	22
t_p-test[1]	20
Bootstrap-t-test	4
LShom[2]	20
LShet[2]	15
BLUPhom[3]	15
BLUPhet[3]	9

[1] The 90th-percentile was $a = 0.346$ in equation (1). [2] LShom (LShet) = Least squares analysis with homogeneous (heterogeneous) error variance. [3] BLUPhom (BLUPhet) = Best linear unbiased prediction of leukemia effects, *i.e.* leukemia effects are random effects, and error variance was homogeneous (heterogeneous).

When the gene-by-treatment effects were ranked for their statistical significance, they were ranked for this bootstrap-P-value. In the case of an ordinary t-test the ranking on t_{real} or on their corresponding P-value are the same, but this is not the case for the bootstrap-t-test, where every gene has its own "table of P-values" which comes from the bootstrap simulations.

3. RESULTS

3.1. The leukemia data

Empirical false discovery rates for the seven methods for finding differentially expressed genes are shown in Table I. The bootstrap-t-test yielded the lowest FDRe of 4%, *i.e.* only 8 out of the 200 genes are expected to be false discoveries, which shows that the power of the experiment of Golub *et al.* [6] was high. However, for the ordinary t-test and the t_p-test, 44 and 40 out of the 200 genes are expected to be false discoveries, respectively. On average the linear model based methods achieved somewhat better FDR than the t- and t_p-test. Correction for heterogeneity of error variances further improved their FDR. The latter is expected since the distribution of the error variances covers a wide range of values and is skewed (Fig. 1), *i.e.* the variances are highly heterogeneous.

The bootstrap-t-test also gave tables of P-values that account for the non-normality in the data; 2545 genes had a nominal P-value below 0.01, and all the 200 most significant genes had a P-value of $1{*}10^{-5}$ indicating that, if we want

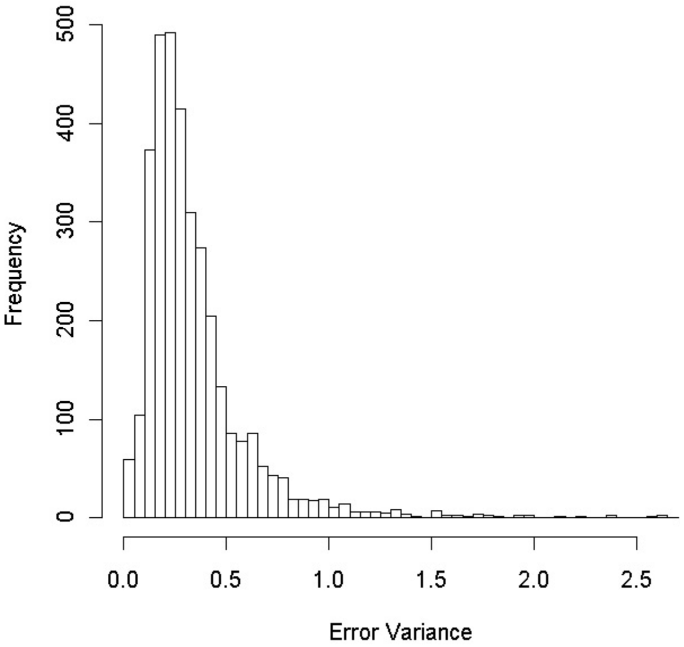

Figure 1. Histogram of the error variances (in squared log light-intensity units) as estimated by the least squares model with heterogeneity of error variances (only estimates based on more than 30 records are included).

to distinguish between these genes, we needed more bootstrap samples. These P-values resulted in a false discovery rate of FDR(200) = 2.6×10^{-4} (Eq. (2)). The empirical estimate of FDR (FDRe = 4%) is substantially higher than this predicted estimate (FDR(200) = 2.6×10^{-4}), which is probably because the INDEPEND data is not a true replication of the TRAIN data. This was described by Golub *et al.* [6] as: "INDEPEND contains a much broader range of samples, including samples from peripheral blood rather than bone marrow, from childhood AML patients, and from different reference laboratories that used different sampling protocols".

3.2. The apoAI knockout mice

Table II shows the empirical FDR for the apoAI knockout mice data. When the 200 most significant effects were considered, FDRe approached 100% for most methods, except for the Bootstrap-t-test for which about 1 out of every 3 significant genes was a false positive. This shows that all methods had little power to detect the true effects in DATA1, which contains half of the data of the

Table II. The empirical false discovery rate (FDRe, %) when the 200 and 8 most significant apoAI-knockout effects in DATA1 were tested in DATA2, and *vice versa*.

	No. of most significant effects	
Method	200	8
t-test	91	38
t_p-test[1]	83	38
Bootstrap-*t*-test	37	13
LShom[2]	76	25
LShet[2]	91	38
BLUPhom[3]	83	13
BLUPhet[3]	71	63

[1] The 90th-percentile was $a = 0.267$ in equation (1). [2] LShom (LShet) = Least squares analysis with homogeneous (heterogeneous) error variance. [3] BLUPhom (BLUPhet) = Best linear unbiased prediction of apoAI-effects, *i.e.* apo-AI effects are random effects, and error variance was homogeneous (heterogeneous).

experiment of Callow *et al.* [4], and in DATA2, which contains the other half. However, these authors reported only eight significant genes. The Bootstrap-*t*-test achieved 1 false discovery out of eight significant genes, while the other methods had substantially higher FDRe, except for BLUP with homogeneous error variance which achieved the same FDRe as Bootstrap-*t* in this case.

When the Bootstrap-*t*-test was used in the complete data of apoAI knockout mice of Callow *et al.* [4], the eight most significant genes were the same genes as those found significant by a normal quantile – quantile plot of the *t*-test statistics [4]. These eight genes are also described in detail by [4]. Using the bootstrap-*t*-test, 629 had a nominal *P*-value below 0.01, and the *P*-values and FDR of the 11 most significant genes are given in Table III. The predicted false discovery rate was low at FDR(8) = 2.4%. Also the 10 most significant genes had a FDR well below 5% (FDR(10) = 3.2%). At a FDR below 5%, in addition to the eight differentially regulated that were reported by [4], two more genes were found differentially regulated: Incenp (accession no. W13505) was up-regulated, and Serpinf1 (accession no. AA691483) was down-regulated.

The two subsets of the apoAI data, DATA1 and DATA2, contained only 4 control and 4 knockout expressions per gene. In these small data sets, the bootstrap predictions of *P*-values and thus of FDR for the, say 8, most highly significant genes should be overestimated because the tails of the distribution of the data are imprecisely estimated due to the small sample size (see Discussion section). When more genes are considered, *e.g.* the 200 most significant genes, FDR(200) is 26 and 28% for DATA1 and DATA2, respectively, which is

Table III. Predicted false discovery rates and nominal P-values using bootstrapping for different numbers of significant genes in the complete apoAI-knockout data of [3].

No. of most significant genes	FDR(%)	Nominal-P (%)
1	1.6	0.001
2	1.6	0.001
4	1.6	0.001
8	2.4	0.003
10	3.2	0.005
11	5.4	0.010

somewhat lower than the empirical estimate of 37% (Tab. II). As in the TRAIN data, the predicted FDR is somewhat optimistic, but in the same order of magnitude as the empirical estimate of the FDR.

4. DISCUSSION

Seven alternative methods for the analysis of gene expression data were compared for their empirical false discovery rates of differentially expressed genes. The bootstrap-*t*-test yielded an up to 80% lower FDR than the alternative tests, and was in all comparisons as good as or superior to the best of the alternatives. The improved ranking for significance by the bootstrap-*t*-test is due to the non-normality of the log-transformed light intensities, whose distribution is better approximated by the bootstrap simulations than by a log-normal distribution. It is expected that as the number of micro-arrays per experiment increases due to improved and cheaper micro-array technology, the increased number of data points will further improve the approximation of the distribution of the data by bootstrap simulations, and thus will improve the rankings based on the bootstrap-*t*-test.

The linear models (LS and BLUP) generally gave lower FDRe than the ordinary and penalized *t*-test. In the case of Table I, correction for heterogeneity of variance was superior to assuming homogeneous variances. In the case of Table II, correction for heterogeneity of variance did not seem to improve FDRe, possibly because of the small sizes of DATA1 and DATA2 which results in poor estimates of the variances of the individual genes. In general, the linear models are more flexible than the *t*-tests for analyzing microarray data in that they can analyze data where many factors are affecting the records [9]. In situations where many factors are affecting the records, it is therefore worthwhile to device bootstrapping methods that use linear models and their corresponding

F-tests instead of t-tests. However, the general idea of the bootstrap simulations will remain the same namely:

1. Randomly sample with replacement treatment identifiers to the records;
2. calculate the F-statistic or other statistic;
3. repeat steps 1 and 2 to make a bootstrap-P-table of the F-statistic from which the P-value of the real data F-statistic can be found.

From the bootstrap simulations, the FDR could be predicted using equation (2) in any one experiment. If the size of the experiment is sufficiently large to approximate the distribution of the data even within its tails, these predicted FDR were too low in the publicly available data that were used here, *i.e.* they were too optimistic about the true FDR. In small data sets, such as DATA1 and DATA2, the predicted FDR of the most significant gene-by-treatment effects are expected to be substantially overestimated as shown in the next paragraph.

As an example, suppose a gene has 4 log-light intensity records of {1, 2, 3, 4} for treatment 1, and 4 for treatment 2: {11, 12, 13, 14}. These data show an extremely significant treatment effect: P-value of the two-sided t-test is 3.4×10^{-5}. However, the P-value of the bootstrap-t-test is only 6×10^{-3}. When sampling with replacement from the forementioned data, it is relatively likely to sample an even more significant data set by: (1) sampling with replacement the first 4 records out of the set {1, 2, 3, 4} and the second 4 records out of the set {11, 12, 13, 14} (probability is $1/2^8 = 0.004$); (2) sampling duplicated records in such a way that the t-statistic becomes even more extreme than that of the original data (probability is approximately $1/2$). Hence, the expected P-value of the bootstrap-t-test is approximately $0.004{*}1/2{*}2$, where the factor 2 is due to the two-sided testing. Note that this P-value is rather insensitive to the actual treatment effect, *i.e.* the data set {1, 2, 3, 4, 21, 22, 23, 24} gives the same P-value. This upward bias of P-values of highly significant gene*treatment effects disappears quickly when we have more records, *e.g.* with eight records per treatment the minimum bootstrap-t-test P-value is approximately: $1/2^{16} = 1.5 \times 10^{-5}$. Hence, the bootstrap-$t$-test needs at least 16 records per gene in order to avoid a severely upwards bias in the P-values of the most significant genes, and thus a too high predicted FDR for the most significant effects.

Hence, a small number of records makes the approximation of the distribution of the data too poor in the tails to make predictions about the P-values of highly significant genes. If the smallest bootstrap-P-values are close to the minimum possible P-value, $P_{\min}$, an upward bias of these P-values is expected, where:

$$P_{\min} = \left(\frac{n_1}{N}\right)^{n_1} \left(\frac{n_2}{N}\right)^{n_2}$$

with n_1 (n_2) = number of records in treatment 1 (2), and N is total number of records per gene. Also, the number of bootstrap simulations should be large enough to be able to estimate low P-values, *i.e.* the number of bootstrap simulations should preferably exceed $100/P$.

An alternative non-parametric approach to assess the FDR is the permutation of the data, as for instance used in significance analysis of microarrays (SAM; [15]). Random permutations of the data are obtained by sampling from the data without replacement (instead of with replacement as in bootstrapping). Obviously, if there are sufficiently many records per gene, sampling with or without replacement gives very similar results. However, if there are few records per gene, there are only few possible permutations of the data, *e.g.* in DATA1 there are $\binom{8}{4} = 70$ possible permutations which makes that the smallest possible P-value is 0.014 (*i.e.* significance at the 1% level is not possible). SAM overcomes this problem of few possible permutations by determining the P-values (and thus FDR) across all genes, but this has the drawback that all genes have to be ranked on the same test-statistic, in the case of SAM: the t_p-statistic (1). Thus although SAM gives an improved estimate of the true FDR, its FDR is the same as that of the t_p-test.

Benjamini and Hochberg [1] show that equation (2) controls the FDR when the $P(i)$-values are independent. Benjamini and Yekutieli [2] show that this is also the case for positively dependent $P(i)$-values. Measurement errors of light-intensities tend to be positively correlated, *e.g.* spatial effects that are not corrected for can increase all light-intensities in a region. Biological effects may lead to negative correlations between gene expression levels, however this will also result in positive correlations between the $P(i)$-values because double-sided test-statistics are used here (*i.e.* both down- and up-regulated genes have low $P(i)$-values). In situations where strong negative correlations between the $P(i)$-values are expected, resampling techniques can be used to adjust the FDRs [16]. The empirical estimate of FDR (FDRe) is not expected to be biased by correlations between the genes, but its standard error will be increased. However, if the two data sets that are used to calculate FDRe are not independent, FDRe can be severely biased.

With a FDR of 2.6%, 200 genes were found up/down regulated when comparing AML to ALL samples. Golub *et al.* [6] used 50 genes to distinguish the two cancer types, which seems a conservative number given the results in Table I. A list of the 200 genes most affected by the cancer types is available from the authors. In the apoAI data, the bootstrap-t-test indicated that at least two more cDNA's, Incenp and Serpinf1, were affected by the apoAI knockout (at an FDR < 5%), where Incenp was up-regulated while all other significant genes were down regulated. Incenp and Serpinf1 were not discovered by the t-test at an adjusted P-value < 20% [4], which shows that the bootstrap-t-test increases the power of detecting differentially regulated genes substantially.

In conclusion, non-parametric bootstrapping of gene expression data substantially improved the detection of differentially expressed genes compared to the alternative methods in two publicly available data sets. Since the results are only based on two data sets, more investigations are needed to confirm these results, but the differences in FDR were so large that we expect these results to hold also in other data sets. Furthermore, an empirical estimate of the FDR, FDRe, was developed and tested in Monte Carlo simulations. This FDRe can be used to compare current and future micro-array analysis methods for their FDR.

ACKNOWLEDGEMENTS

We are grateful to T.R. Golub *et al.* and M.J. Callow *et al.* for making their data publicly available.

REFERENCES

[1] Benjamini Y., Hochberg Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. Royal Stat. Soc. B 57 (1995) 289–300.

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

[3] Brown P.O., Botstein D., Exploring the new world of the genome with DNA microarrays, Nat. Genet. Suppl. 21 (1999) 33–37.

[4] Callow M.J., Dudoit S., Gong E.L., Speed T.P., Rubin E.M., Microarray expression profiling identifies genes with altered expression in HDL-deficient mice, Genome Res. 10 (2001) 2022–2029.

[5] Efron B., Tibshirani R., Storey J.D., Tusher V., Empirical Bayes analysis of a micro-array experiment, J. Amer. Stat. Assoc. 96 (2001) 1151–1160.

[6] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiury M.A., Bloomfield C.D., Lander E.S., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[7] Henderson C.R., Applications of linear models in animal breeding, University of Guelph, Canada, 1984.

[8] Kerr M.K., Churchill G.A., Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 8961–8965.

[9] Kerr M.K., Churchill G.A., Statistical design and the analysis of gene expression microarray data, Genet. Res. Camb. 77 (2001) 123–128.

[10] Lönnstedt I., Speed T.P., Replicated microarray data, Stat. Sinica 12 (2002) 31–46.

[11] Patterson H.D., Thompson R., Recovery of inter-block information when block sizes are equal, Biometrika 58 (1971) 545–554.

[12] Searle S.R., Linear models, Wiley, New York, 1971.

[13] Slonim D.K., From patterns to pathways: gene expression data analysis comes of age, Nat. Genet. Suppl. 32 (2002) 502–508.

[14] Smyth G.K., Yang Y.H., Speed T.P., Statistical issues in cDNA microarray data analysis, in: Brownstein M.J., Khodursky A.B. (Eds.), Functional Genomics: Methods and Protocols, Methods in Molecular Biology, Humana Press, Totowa, 2002.

[15] Tusher V., Tibshirani R., Chu G., Significance analysis of microarrays applied to the ionizing radiation response, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 5116–5121.

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

APPENDIX: AN EMPIRICAL ESTIMATE OF THE FALSE DISCOVERY RATE (FDR)

We assume that we have two data sets in which the effect of the same two treatments on gene-expression levels is investigated. Typically two such data sets are obtained by splitting the microarrays of a larger experiment, at random, in two approximately equally sized sets of microarrays. The data sets are called DATA1 and DATA2. We analyse both data sets with the data analysis method, of which we want to estimate the FDR. This results in an estimate of the difference between the two treatments, b_1 and b_2, for DATA1 and DATA2, respectively. Let b_{true} denote the true difference between the treatments. Our method of analysis also results in a statistic that is used to determine the statistical significance of the treatment. This statistic is called t_1 and t_2 for DATA1 and DATA2, respectively, and the treatment is called significant for DATA1 if t_1 exceeds some critical value t_α.

Now let us consider the genes where the analysis of DATA1 results in a significant treatment effect. Some of these genes may be false positives, *i.e.* H0 is true: the treatment has no effect on the expression level. The alternative hypothesis is Ha: treatment does affect the expression level. If a gene was found significant in DATA1 and Ha is true we expect that the gene will show a similar effect in DATA2. Moreover, if the effect of a significant gene has the opposite effect in DATA2 compared with DATA1, *i.e.* $\text{Sign}(b_1) \neq \text{Sign}(b_2)$, we start to doubt that Ha is true for this gene. Hence, we start to believe that this gene is a false positive. The probability that $\text{Sign}(b_1) \neq \text{Sign}(b_2)$ given that DATA1 gave a significant treatment effect is:

$$\begin{aligned}
&\text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|t_1 > t_\alpha) \\
&\qquad = \text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|t_1 > t_\alpha, \text{H0}) * \text{P}(\text{H0}|t_1 > t_\alpha) \\
&\qquad\quad + \text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|t_1 > t_\alpha, \text{Ha}) * \text{P}(\text{Ha}|t_1 > t_\alpha) \\
&\qquad = \text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{H0}) * \text{P}(\text{H0}|t_1 > t_\alpha) \\
&\qquad\quad + \text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{Ha}) * \text{P}(\text{Ha}|t_1 > t_\alpha) \\
&\qquad = {}^1\!/_2 * \text{P}(\text{H0}|t_1 > t_\alpha) + \text{P}(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{Ha}) * \text{P}(\text{Ha}|t_1 > t_\alpha) \\
&\qquad \approx {}^1\!/_2 * \text{P}(\text{H0}|t_1 > t_\alpha)
\end{aligned} \tag{A.1}$$

The one-but-last equality in (A.1) assumes that $P(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{H0}) = {}^1\!/_2$, which is the case if the estimates of the difference between the treatments, b_1 and b_2, are independent and have a symmetric distribution around zero under H0 (and that $b_1 = 0$ (or $b_2 = 0$) has an infinitesimal probability which is the case for any variable with a continuous distribution). The last approximation in (A.1) is based on the assumption that $P(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{Ha})$ is small. Note that $P(\text{Sign}(b_1) = \text{Sign}(b_{\text{true}})|\text{Ha})$ is very high, since the probability that the estimate b_1 was in the wrong direction and significantly different from zero is small (smaller than the type-I-error rate). Hence, $P(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{Ha}) \approx P(\text{Sign}(b_{\text{true}}) \neq \text{Sign}(b_2)|\text{Ha})$ is assumed small, and the latter will generally be the case when the experiment has reasonable power.

Note that, in (A.1), $P(\text{H0}|t_1 > t_\alpha)$ equals the false discovery rate, and thus an empirical estimate of $P(\text{H0}|t_1 > t_\alpha)$ can be obtained from:

$$\text{FDRe} = 2 * P(\text{Sign}(b_1) \neq \text{Sign}(b_2)|t_1 > t_\alpha).$$

From the above derivation it is also clear that FDRe overestimates the FDR when $P(\text{Sign}(b_1) \neq \text{Sign}(b_2)|\text{Ha})$ is not negligible, *i.e.* when the power of the experiment is small. The performance of FDRe as an estimate of the FDR was tested in a simulation study. Even when the power of the experiment was as low as 9%, FDRe overestimated the true FDR by only 5.8%-points (Tab. A.I). When the size of the experiment was doubled or quadrupled, the overestimation reduced to 2.5 or 0.2%-points, respectively. Hence, FDRe yields a very good estimate of FDR in reasonably large experiments, and yields a conservative estimate (*i.e.* an over-estimate) of the true FDR in small experiments.

Table A.I. True and estimated false discovery rates (FDR) and the power of the experiment when the size of the experiment, denoted by the number of microarray-slides (*n*), is varied. Each slide shows the difference between a treatment and a control.

n	FDR_{true}	FDRe	Power
5	0.479	0.537	0.094
10	0.162	0.187	0.166
15	0.044	0.046	0.439

There are two data sets (DATA1 and DATA2) each with *n* slides. Each slide contains 100 000 genes, such that reliable estimates of FDR_{true} and FDRe are obtained. The log light-intensity of each gene in each treatment at each slide is simulated as the sum of a treatment effect and a measurement error which is sampled from $N(0, 1)$. The treatment effect is 0 for the control treatment and for 90% of the genes that show no treatment effect. The treatment effect is 1 for 10% of the genes. The statistical significance of the treatment was tested by a *t*-test at a type-I-error rate of 0.01.