Published in final edited form as:

Comput Stat Data Anal. 2009 March 15; 53(5): 1622–1629. doi:10.1016/j.csda.2008.04.010.

Balancing Type One and Two Errors in Multiple Testing for Differential Expression of Genes

Alexander Gordon^a, Linlin Chen^{b,*}, Galina Glazko^b, and Andrei Yakovlev^b

^a Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Boulevard, Charlotte, North Carolina, U.S.A

^b Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, U.S.A

Abstract

A new procedure is proposed to balance type I and II errors in significance testing for differential expression of individual genes. Suppose that a collection, $\mathcal{A}_k$, of k lists of selected genes is available, each of them approximating by their content the true set of differentially expressed genes. For example, such sets can be generated by a subsampling counterpart of the delete- d -jackknife method controlling the per-comparison error rate for each subsample. A final list of candidate genes, denoted by S^* , is composed in such a way that its contents be closest in some sense to all the sets thus generated. To measure “closeness” of gene lists, we introduce an asymmetric distance between sets with its asymmetry arising from a generally unequal assignment of the relative costs of type I and type II errors committed in the course of gene selection. The optimal set S^* is defined as a minimizer of the average asymmetric distance from an arbitrary set S to all sets in the collection $\mathcal{A}_k$. The minimization problem can be solved explicitly, leading to a frequency criterion for the inclusion of each gene in the final set. The proposed method is tested by resampling from real microarray gene expression data with artificially introduced shifts in expression levels of pre-defined genes, thereby mimicking their differential expression.

Keywords

Multiple testing; error rates; resampling; microarray data

1. Introduction

A sample of expression levels of m genes measured by single-color array technologies is represented by n independent and identically distributed copies (across arrays) of a random vector $\mathbf{Z} = Z_1, \dots, Z_m$ with joint distribution $W(z_1, \dots, z_m)$. The components of $\mathbf{Z}$ are stochastically dependent and this dependence is extremely strong and long-ranged [13]. Since the dimension of the vector $\mathbf{Z}$ is typically high relative to the number of observations (replicates of experiments), the univariate testing is the dominant method of dimension reduction in microarray studies. The most standard practice is to test the hypothesis of no differential

*Corresponding author. Tel.: +1 5852756696; Fax: +1 5852731031, Email address: linlin.chen@urmc.rochester.edu (Linlin Chen). On February 27, 2008, Dr. Andrei Yakovlev tragically passed away. We deeply grieve the loss of our colleague, advisor, and friend who was a source of inspiration for all around him.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

expression for each gene [19]. Formulated in terms of the marginal distributions of all components of $\mathbf{Z}$, this hypothesis amounts to stating that the expression levels of a particular gene are identically distributed in two (or more) phenotypes. The most basic issue to be addressed in this setting is that of multiple hypothesis testing [5].

There are several ways to guard against type I errors (false discoveries) when testing multiple hypotheses. One approach is to provide control of the family-wise error rate (FWER), defined as the probability of making at least one type I error among all hypotheses tested. A step-down multivariate resampling algorithm originally proposed by Westfall and Young [30] falls into this category. Dudoit and co-workers [6] implemented the Westfall-Young algorithm based on the t -statistic. The utility of this algorithm was later considered in conjunction with the N -statistic [15] and distribution-free tests [31]. Another popular method used to control the FWER is the Bonferroni procedure. In applications to microarray data analysis, the FWER-controlling version of the Bonferroni procedure appears to be overly conservative, raising concerns about its usefulness where the magnitude of multiple testing is high. However, the Bonferroni procedure is also known to control another meaningful characteristic of type I errors, namely, the per family error rate (PFER) defined as the expected number of false discoveries. By controlling the PFER rather than the FWER, one can gain much more power in detecting alternative hypotheses [11,18]. Proceeding from this simple fact, Gordon et al. [11] suggest that the Bonferroni procedure be extended by setting its parameter γ (see Section 3) at any desirable nominal level (subject to the constraint: $\gamma < m$, where m is the total number of hypotheses) that can be even greater than one. The Bonferroni procedure thus extended controls the PFER at level γ . Yet another approach that has become especially popular in recent years is focused on controlling the false discovery rate (FDR), defined as the expected proportion of falsely rejected null hypotheses among all rejections [1,2,26,29].

Whichever error rate (FWER, PFER, FDR) a given multiple testing procedure (MTP) is designed to control, it is typically organized as follows:

1. A two-sample test is chosen to produce p -values associated with each gene, the t -test being the most popular choice in microarray studies.
2. The computed p -values are arranged in the ascending order.
3. A specific MTP is applied to determine a p -value cut-off so that all the genes with p -values smaller than the cut-off value are declared differentially expressed.

This approach places the emphasis on type I errors and their control with the choice of a two-sample statistic remaining to be the sole determinant of the resultant power. The relationship between type I and II errors escapes attention, being perceived solely as a property of the selected statistical test and the data to be analyzed. It should be noted that the test multiplicity significantly affects the overall power of a given MTP, which may become extremely low when increasing the number of tests but keeping control of type I errors at a constant level.

Unlike the above-mentioned approaches, a version of the empirical Bayes method introduced in microarray analysis by Efron [7] and explored further by Efron [8–10] is not intended to automatically control any type I error rate at a pre-specified level. What this method does is more in the spirit of the Bayesian approach to hypothesis testing, which is known to be optimal in the sense that it minimizes a linear combination of the probabilities of type I and type II errors (see [4] for the case of simple hypotheses). If the mixture model underlying the nonparametric empirical Bayes method (NEBM) by Efron can be estimated well (which is the case whenever the numbers of null and alternative hypotheses are large and their associated test-statistics are independent and identically distributed), this method is expected to have the same optimality property. In this very special sense, the NEBM attempts to account for both types of errors in simultaneous testing of multiple hypotheses.

In the present paper, we consider a more direct approach to the usual trade-off between the two types of error and its theoretical underpinning from the frequentist perspective. A procedure developed in Section 2 is designed to optimize the selection of differentially expressed genes with this trade-off taken explicitly into account. The proposed procedure exploits the advantages both of a properly chosen distance between gene sets and of resampling techniques. If the practitioner is more concerned about false discoveries rather than the resultant power, the balance provided by this procedure can be shifted towards type I relative to type II errors. However, the procedure provides a more general framework allowing for any relative importance of the type I among both types of errors.

The paper is organized as follows. Section 2 presents a theoretical justification of the proposed method. In Section 3, the method is tested by resampling from real microarray gene expression data with artificially introduced shifts in marginal distributions of expression levels for a set of pre-defined genes. The results of testing and their statistical implications are discussed in Section 4.

2. Method

The decision to include a gene in the list of differentially expressed genes is subject to type I errors (false inclusion of a gene in the list) and type II errors (failure to include a truly “different” gene in the list). Let us look at this process of decision making from the following perspective. Suppose that we have to pay a penalty for committing an error with the cost depending on the error type. Let c_I or c_{II} be the costs of type I and II errors, respectively. Denote the set of all true differentially expressed genes by X and that of the genes declared to be in X by S . Then the total penalty is

$$\delta(S, X) = c_I |S \setminus X| + c_{II} |X \setminus S|, \quad (1)$$

where $A \setminus B$ denotes the difference between the sets A and B , that is, the set of all elements of A that do not belong to B , and the symbol $|A|$ is used for the cardinality of the set A . The function $\delta(S, X)$ given by (1) has all the properties of a metric except for symmetry. In the special case of $c_I = c_{II}$, it reduces to the usual symmetric distance between two sets. It is desirable to find a set S^* that minimizes the total penalty. Since $\delta(X, X) = 0$ the best possible choice is $S^* = X$, of course. However, the set X is unknown and we need to find an approximate solution to this optimization problem.

Suppose that a collection $\mathcal{S}_k$ of sets S_i , $i = 1, 2, \dots, k$, of selected genes is available to approximate the content of the true set X . The way these sets are generated is irrelevant, as is the quality (accuracy) of approximation. We replace the problem of minimization of $\delta(S, X)$ (i.e., the asymmetric distance from S to X) as a function of S by the problem of minimization of the function $\delta(S, S_i)$ (i.e., the asymmetric distance from S to S_i) averaged over $i = 1, 2, \dots, k$. This suggests the following optimization problem:

$$f(S) = \frac{1}{k} \sum_{i=1}^k \delta(S, S_i) \equiv \frac{1}{k} \sum_{i=1}^k (c_I |S \setminus S_i| + c_{II} |S_i \setminus S|) \rightarrow \min. \quad (2)$$

If there are m genes in total, we have 2^m possible choices of S , so the problem looks computationally prohibitive. However, it does have a simple explicit solution. Namely, it can be shown that the optimization problem formulated in (2) is equivalent to

$$g(S) = \sum_{i \in S} \left(\frac{c_I}{c_I + c_{II}} - b(j) \right) \rightarrow \min, \quad (3)$$

where $b(j)$ is the proportion of sets S_i ($i = 1, \dots, k$) that contain the j th gene ($j = 1, \dots, m$). The latter problem has the following obvious solution: the optimal set S^* should include all genes j for which $b(j) > h \equiv c_I/(c_I + c_{II})$ and include no genes with $b(j) < h$; genes with $b(j) = h$ (if any) may be included in S^* or not – their inclusion does not affect the value of $g(S^*)$. Now we need to prove the equivalence of problems (2) and (3).

Proposition

Problems (2) and (3) are equivalent.

Proof—To solve problem (2), we have to minimize the function

$$f(S) = c_I \frac{1}{k} \sum_{i=1}^k |S \setminus S_i| + c_{II} \frac{1}{k} \sum_{i=1}^k |S_i \setminus S|.$$

Denote the indicator function of a set $A \subset \{1, 2, \dots, m\}$ by 1_A , so that for any pair of such sets A, B we have

$$|A \setminus B| = \sum_{j \in A} 1_{B^c}(j),$$

where B^c is the complement of B , and

$$|B \setminus A| = \sum_{j \notin A} 1_B(j).$$

Therefore, for each i ($1 \leq i \leq k$)

$$c_I |S \setminus S_i| + c_{II} |S_i \setminus S| = c_I \sum_{j \in S} 1_{S_i^c}(j) + c_{II} \sum_{j \notin S} 1_{S_i}(j),$$

so that

$$\begin{aligned}
f(S) &= \frac{1}{k} \sum_i [c_I |S \setminus S_i| + c_{II} |S_i \setminus S|] \\
&= c_I \frac{1}{k} \sum_i \sum_{j \in S} 1_{S_i^c}(j) + c_{II} \frac{1}{k} \sum_i \sum_{j \notin S} 1_{S_i}(j) \\
&= c_I \sum_{j \in S} \frac{1}{k} \sum_i 1_{S_i^c}(j) + c_{II} \sum_{j \notin S} \frac{1}{k} \sum_i 1_{S_i}(j) \\
&= c_I \sum_{j \in S} \frac{\#\{i: j \notin S_i\}}{k} + c_{II} \sum_{j \notin S} \frac{\#\{i: j \in S_i\}}{k} \\
&= c_I \sum_{j \in S} a(j) + c_{II} \sum_{j \notin S} b(j),
\end{aligned}$$

where $b(j)$ and $a(j)$ are the proportions of those $i \in \{1, 2, \dots, k\}$ for which j belongs or does not belong to S_i , respectively. The last expression can be re-written as

$$\begin{aligned}
&c_I \sum_{j \in S} a(j) + c_{II} \sum_{j=1}^m b(j) - c_{II} \sum_{j \in S} b(j) \\
&= \sum_{j \in S} (c_I a(j) - c_{II} b(j)) + c_{II} \sum_{j=1}^m b(j) \\
&= \sum_{j \in S} (c_I (1 - b(j)) - c_{II} b(j)) + \text{const} \\
&= \sum_{j \in S} (c_I - (c_I + c_{II}) b(j)) + \text{const} \\
&= (c_I + c_{II}) \sum_{j \in S} \left[\frac{c_I}{c_I + c_{II}} - b(j) \right] + \text{const}.
\end{aligned}$$

The equality

$$f(S) = (c_I + c_{II})g(S) + \text{const},$$

shows that optimization problems (2) and (3) are equivalent and this completes the proof.

The proposed procedure does not address the question of how to construct the collection of approximating sets S_i , but rather how to summarize the information provided by them in an optimal way. For example, the sets S_i can be generated by the subsampling version of the delete- d -jackknife method combined with the selection of all hypotheses rejected at a pre-defined significance level. A set S^* composed of only those genes with the frequency of occurrence in the sets S_i exceeding a threshold level of $c_I/(c_I + c_{II})$ will satisfy the optimality criterion given by (1). Therefore, the set S^* is to be reported as the list of genes finally selected by this particular method. The MTP thus designed allows one to find an optimal set of genes, the size and composition of which depend on the perceived relative importance of type I and II errors and the basic statistical method used for hypothesis testing.

3. Resampling Study

We designed a special study to test the proposed method. This study is extremely computer intensive as it requires two loops of resampling from real microarray data. There is, however, significant data parallelism available, which can be exploited. For this reason, we tested the proposed procedure with the aid of a cluster computer. Unlike the study presented below, the method itself is not excessively time-consuming - it compares well with any permutation-based gene selection procedure.

The study was designed as follows. In an effort to preserve the actual correlation structure of gene expression levels as much as possible, we carried out our study by resampling from real

data. For this purpose, use was made of a set of microarray data reporting expression levels (A3ymetrix GeneChip platform) of $m=7084$ genes in $n=88$ patients with hyperdiploid acute lymphoblastic leukemia identified through the St. Jude Children's Research Hospital Database [32]. This set of genes was identified after removing all probe sets with dubious definitions as recommended in [3]. Prior to subsampling, 350 genes were randomly selected and the standard deviations of their log-expression levels were estimated from the whole group of 88 arrays. The composition of this subset of 350 pre-defined genes was fixed throughout all experiments. At every step of the resampling procedure, two subsamples of subjects (arrays), each of size $n = 30$, were generated from the collection of available arrays. To preclude possible ties from occurring, these subsamples were obtained by randomly splitting into two equal parts a larger subsample of 60 subjects drawn without replacement from the group of 88 patients. One subsample ($n = 30$) was modified by adding a constant shift (effect size) to observed log-expression levels of the pre-defined 350 genes. No changes were made to the second subsample of size $n = 30$. We report the results obtained for the effect size equaling one standard deviation (σ) of the log-expression signals for each individual gene, but smaller and larger effect sizes were also studied. A total of 1500 pairs of subsamples were generated and each of them was used to select differentially expressed genes by the proposed method.

For each pair of subsamples, the sets S_i were formed by the delete- d -jackknife method [21]. In doing so, we left out $d = 5$ subjects from each subsample and applied a two-sample test to log-expression measurements for the remaining subjects. The chosen value of the parameter d is suggested by the asymptotic jackknife theory [27] as the only available guide for making this choice. To avoid rigid parametric assumptions, we used the Mann-Whitney test with exact p -values computed by a standard function in R. Before deciding on this software, it was tested by independent computations using our original code. Each set S_i included genes for which the two-sample hypothesis was rejected at a significance level of 0.05. The procedure was repeated $k = 1500$ times and the frequencies of occurrence of each gene in the sets S_i , $i = 1, \dots, k$, were estimated. Using these frequencies, a final set of differentially expressed genes was identified from each subsample of size $n = 30$ and the numbers of false and true discoveries were recorded, with their mean and standard deviation serving as the main performance indicators.

The aforesaid can be summarized in the form of the following algorithm:

- Step 1. Specify the penalties C_I and C_{II} for type I and II errors, respectively, and compute $h = C_I / (C_I + C_{II})$.
- Step 2. a) Randomly draw 60 subjects without replacement and divide them randomly into two groups, each containing $n = 30$ subjects. b) In one of the two subsamples, modify 350 pre-defined genes by adding a constant shift (effect size) to their log-expression levels. The effect size is a multiple of the standard deviation (σ) of log-expression levels estimated from the group of 88 subjects.
- Step 3. Leave out $d = 5$ arrays from each group and apply the exact Mann-Whitney test to the log-expression measurements provided by the remaining arrays in order to select all differentially expressed genes at a significance level of 0.05. These genes make up the current set S_i .
- Step 4. Repeat Step 3 k times ($k = 1500$) to generate a collection of subsets S_i , $i = 1, \dots, k$, of selected genes.
- Step 5. For each gene j , compute the proportion b_j of sets S_i containing gene j , and check the condition $b_j \geq h$. Select all genes for which this condition is met and determine the numbers of false and true positives.

- Step 6. Repeat Steps 2–5 N times ($N = 1500$) and estimate the mean and standard deviation of the numbers of false and true discoveries from the N subsamples of size $n = 30$.

The mean number of true positives and the corresponding standard deviation as functions of the threshold parameter h (for $h \geq 0.5$) are shown in Figure 1A. The experiment presented in this figure was carried out with the effect size equaling one σ . It is clear that the variability of true discoveries increases with decreasing the mean power, a regularity observed for all multiple testing procedures. This effect is attributable to the fact that the number of true positives has a relatively low upper bound. Therefore, it is natural that the variance of the number of true positives gets smaller when its mean value is approaching this bound. The number of false discoveries is bounded by the total number of null hypotheses; the latter is believed to be much larger than the number of true alternative hypotheses in real microarray data. In our resampling study, the total number of null hypotheses is equal to 6734.

As one would expect, both the mean and the standard deviation of the number of false positives are monotonically decreasing functions of h (Figure 1B). The slope of the standard deviation of false discoveries increases slightly in the same range of h where its counterpart for the true discoveries increases and the mean power begins to decline. Since the proportions b_j are estimated by the corresponding relative frequencies, it becomes difficult to estimate the performance indicators (from a limited number of subsamples!) in the neighborhood of $h = 1$, which is why their behavior in this region cannot be shown in Figure 1. The high standard deviation of the number of false discoveries is attributable to the multiplicity of tests, as well as to the extremely strong and long-ranged correlations between gene expression levels alluded to in Section 1 (see Section 4 for further discussion). The mean and variance of the number of false discoveries both increase with smaller effect sizes, which tendency is deemed natural.

For comparison, Figure 2 presents the same indicators for the Bonferroni procedure with parameter γ . (This procedure rejects the null hypotheses whose observed p -values do not exceed γ/m ; as was said, it controls the PFER at level γ .) The dynamics of these indicators are similar to those depicted in Figure 1 (with the argument changing in the opposite direction). For the Bonferroni procedure, however, these indicators can be observed over a wide range of the parameter γ , even at its relatively small values.

The ROC curve for the proposed procedure is given in Figure 3 for two effect sizes: one σ (solid line) and 0.5σ (dashed line). As expected from Figures 1 and 2, this estimated ROC is virtually identical (in the common range of the mean number of false discoveries) to that yielded by the Bonferroni procedure, which observation will be discussed in the next section.

The behavior of the mean and variance of the total number of rejections produced by the Bonferroni procedure in the neighborhood of $\gamma = 0$ deserves a closer look. In particular, one can see from Figure 4 that the standard deviation of the total number of rejected hypotheses attains a minimum in the region of small values of γ almost concurrently with a sharp increase in the mean power (Figure 2). A similar feature was described in the paper by Gordon et al. [11] in relation to the Bonferroni and Benjamini-Hochberg procedures in a somewhat differently designed study. Since the total number of rejections is the only observable performance indicator in applications to biological data, this feature may be of practical utility in the analysis of at least some large data sets where the noted phenomenon is pronounced well. The same behavior of the standard deviation of the total number of rejections is expected from the proposed procedure in the neighborhood of $h = 1$, but testing this conjecture by resampling or simulations is computationally prohibitive.

4. Discussion and Conclusion

There may be many ways to generate a collection $\mathcal{F}_k = \{S_i\}_{i=1}^{i=k}$ of approximating gene sets for the purpose of gene expression profiling, but the most fundamental question still remains: how to form a final set that summarizes the information contained in S_i in an optimal way? The suggestion by Stolovitzky [28] to take their intersection does not work because the intersection of $\{S_i\}_{i=1}^{i=k}$ depends on k , tending to an empty set as k grows. In the present paper, we propose a frequency-based solution to the problem that satisfies a certain criterion of optimality. Furthermore, the proposed approach allows for balancing type I and II errors in the construction of the ultimate set of differentially expressed genes.

The idea of ranking genes by the frequency of their occurrence in a target set was first introduced by Qiu et al. [24] in conjunction with currently practiced multiple testing procedures. More specifically, the authors proposed to generate k subsamples by the subsampling version of the delete- d -jackknife method, apply a given MTP to each of them, and finally select only those genes that have been declared differentially expressed in more than hk subsamples, with the choice of h being arbitrary (e.g., $h = 0.8$). They perceived this procedure as a method of stability assessment rather than a selection procedure in its own right. While the procedure developed in the present paper looks similar, the parameter h is now endowed with a statistical meaning, being derived directly from the relative costs of type I and type II errors.

The ROC of the proposed procedure was constructed in terms of the mean numbers (or proportions) of false and true discoveries. When estimated from the same data, this ROC is virtually identical to that produced by the Bonferroni procedure. A similar observation was made by the authors of [11] in regard to the Bonferroni and Benjamini-Hochberg procedures. These facts deserve attention as they suggest the existence of a wide class of p -value-only-based MTPs with practically identical ROC curves. A claim that a certain procedure is less conservative than another is usually based on comparing two different sections of their possibly indistinguishable ROC curves. The assertion (supported by empirical data) that different MTPs, under certain mild conditions, should have practically equal ROC curves is, in fact, natural; however, rigorous theoretical results supporting it are not readily available. A similar insight into higher-order characteristics, such as the variance, of the numbers of false and true discoveries represents an even more challenging problem (see [11] for some simulation results of this nature). These facts also suggest that caution should be exercised when claiming one procedure to be uniformly better than another in terms of their operational characteristics. The assessment and comparison of MTPs lie in a different plane. Different procedures answer different questions or meet different optimality criteria. The practitioner chooses a procedure that has the most intuitive appeal. From this perspective, the method introduced in the present paper enriches the arsenal of available MTPs.

The approach we employed to model the effects of differential expression in Section 3 disregards their complex multivariate nature. At the same time, it represents the most rigorous way known to us of testing various selection procedures. We suggest that every newly proposed method for finding differentially expressed genes be tested by this kind of experimentation with real data. Exploring possible ways of modeling multivariate changes in the joint distribution of gene expression signals that better preserve the correlation structure of microarray data is another interesting problem for future research.

It is clear from Figures 1 and 2 that the high variability of the number of false discoveries from subsample to subsample is a detrimental property one should be particularly concerned about. Unfortunately, this instability of type I errors manifests itself in all MTPs whenever they are applied directly to heavily dependent expression signals of multiple genes, an issue considered

in several publications from different angles [11–13, 17, 20, 23–25]. Especially prone to this instability are methods that explicitly resort to pooling expression measurements across genes, the NEBM and adaptive FDR-based procedures representing relevant examples. The recourse to normalization procedures does not provide a satisfactory solution to the problem because of their distorting effects on the true expression signals. Such effects are especially pronounced in large sample studies where control of type I errors may be entirely lost. The adverse effects of currently used normalization procedures will be discussed at length in a forthcoming paper. An effective cure for this difficulty can be furnished by exploiting the property of weak correlation between elements of the so-called δ -sequence recently discovered in several sets of microarray data [13]. The new paradigm arising from the existence of the δ -sequence in biological data leads to a new methodology for selecting differentially expressed genes in non-overlapping gene pairs [13,17]. The potential of the above-proposed gene selection procedure in conjunction with the δ -sequence has yet to be explored.

Acknowledgments

This research is supported in part by NIH/NIGMS grants GM075299 and GM079259 (A. Yakovlev) and by Alfred P. Sloan Research Fellowship (G. Glazko). We would like to express our gratitude to the leadership of the Laboratory for Laser Energetics, University of Rochester for providing us with access to their SGI Altix XE computing cluster.

References

1. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995;57:289–300.
2. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 2001;29:1165–1188.
3. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WR, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Research* 2005;33(20):e175. [PubMed: 16284200]
4. DeGroot, MH. Probability and Statistics. 2. Addison-Wesley Publishing Company; 1986.
5. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments, *Statistical Science*. 2003;18:71–103.
6. Dudoit S, Yang YH, Speed TP, Callow MJ. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002;12:111–139.
7. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001;96:1151–1160.
8. Efron B. Robbins, empirical bayes and microarrays. *Annals of Statistics* 2003;31:366–378.
9. Efron B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* 2004;99:96–104.
10. Efron B. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association* 2007;102:93–103.
11. Gordon A, Glazko G, Qiu X, Yakovlev AY. Control of the mean number of false discoveries, Bonferroni, and stability of multiple testing. *Annals of Applied Statistics* 2007;1(1):179–190.
12. Klebanov L, Yakovlev AY. Treating expression levels of different genes as a sample in microarray data analysis: Is it Worth a Risk? *Statistical Applications in Genetics and Molecular Biology* 2006;5 Article 9.
13. Klebanov L, Yakovlev AY. Diverse correlation structures in microarray gene expression data and their utility in improving statistical inference. *Annals of Applied Statistics* 2007;1(2):538–559.
14. Klebanov L, Jordan C, Yakovlev AY. A new type of stochastic dependence revealed in gene expression data. *Statistical Applications in Genetics and Molecular Biology* 2006a;5(1) Article 7.
15. Klebanov L, Gordon A, Xiao Y, Yakovlev AY. A new permutation test motivated by microarray data analysis. *Computational Statistics and Data Analysis* 2006b;50(12):3619–3628.

16. Klebanov L, Chen L, Yakovlev AY. Revisiting adverse effects of cross-hybridization in A3ymetrix gene expression data: Do they matter for correlation analysis? *Biology Direct* 2007;2 Article 28.
17. Klebanov L, Qiu X, Yakovlev AY. Testing differential expression in non-overlapping gene pairs: A new perspective for the empirical Bayes method. *Journal of Bioinformatics and Computational Biology*. 2007 in press.
18. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004;124:379–398.
19. Lee, M-L. *Analysis of Microarray Gene Expression Data*. Kluwer; Boston: 2004.
20. Owen A. Variance of the number of false discoveries. *Journal of the Royal Statistical Society Series B* 2005;67:411–426.
21. Politis DN, Romano JP. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* 1994;22:2031–2050.
22. Qiu X, Brooks AI, Klebanov L, Yakovlev AY. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* 2005a;6 Article 120.
23. Qiu X, Klebanov L, Yakovlev AY. Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology* 2005b;4 Article 34.
24. Qiu X, Xiao Y, Gordon A, Yakovlev AY. Assessing stability of gene selection in microarray data analysis. *BMC Bioinformatics* 2006;7 Article 50.
25. Qiu X, Yakovlev AY. Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology* 2006;4(5):1057–1068. [PubMed: 17099941]
26. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003;19:368–375. [PubMed: 12584122]
27. Shao, J.; Tu, D. *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer; New York: 1995.
28. Stolovitzky G. Gene selection in microarray data: the elephant, the blind men and our algorithms. *Current Opinion in Structural Biology* 2003;13:370–376. [PubMed: 12831889]
29. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B* 2004;66:187–205.
30. Westfall, PH.; Young, SS. *Resampling-based multiple testing: examples and methods for p-value adjustment*. John Wiley and Sons; New York: 1993.
31. Xiao Y, Gordon A, Yakovlev AY. The L_1 -version of the Cramer-von-Mises test for two-sample comparisons in microarray data analysis. *EURASIP Journal of Bioinformatics and Computational Biology* 2006;1–9. Article ID 85769.
32. Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, Behm FG, Raimondi SC, Relling MV, Patel A, Cheng C, Campana D, Wilkins D, Zhou X, Li J, Liu H, Pui CH, Evans WE, Naeve C, Wong L, Downing JR. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002;1(2):133–143. [PubMed: 12086872]

Fig. 1. Mean (solid lines) and standard deviation (dashed lines) of the number of true (A) and false (B) positives as functions of the parameter h . The total number of genes is 7084, the number of “truly different” genes is 350, the effect size is equal to one σ . Other parameters are described in the text.

Fig. 2. Mean (solid lines) and standard deviation (dashed lines) of the number of true (A) and false (B) positives resulted from the Bonferroni procedure at different values of the parameter γ .

Fig. 3. ROC for the proposed procedure. x-axis: mean number of false discoveries divided by the total number of null hypotheses; y-axis: mean number of true discoveries divided by the total number of true alternatives.

Fig. 4. Mean and standard deviation of the total number of rejections resulted from the Bonferroni procedure in 1500 subsamples.