Journal of Informatics and Mathematical Sciences
Vol. 11, No. 2, pp. 209–219, 2019
ISSN 0975-5748 (online); 0974-875X (print)
Published by RGN Publications
DOI: 10.26713/jims.v11i2.1214

http://www.rgnpublications.com

Research Article

Overcoming Adverse Effects of Correlations in Microarray Data Analysis

Linlin Chen[1,*] and Haiyan Su[2]

[1] *School of Mathematical Sciences, Rochester Institute of Technology, Rochester, USA*
[2] *Department of Mathematical Sciences, Montclair State University, Montclair, USA*
*__Corresponding author:__ lxcsma@rit.edu

Abstract. Due to the existence of the strong correlation between microarray gene expression levels, procedures which are commonly used to select the significant genes between two or more phenotypes cannot overcome the main problems: high instability of the number of false discoveries and low power. It may be impossible to completely understand these correlations due to the complexity of the biology nature. Gordon *et al.* [1] proposed a new multiple testing procedure to balance type I and II errors in an optimal way. However, the correlation structure of microarray data is still the main obstacle standing in the way of various gene selection procedures. To remove this obstacle, we improved the statistical methodology by exploiting the properties associated with the low dependency of the so-called delta-sequence proposed in Klebanov *et al.* [4]. Our study showed a similar behavior has been observed that both the mean and the standard deviation of the number of false positives are monotonically decreasing as a function of the threshold parameter. In addition, working with pairs, we have substantial reduction in both numbers, which means we gain power and stability in our new study.

Keywords. Correlation Structure; Microarray Gene Expresssion Data Analysis; Resampling

MSC. 62-07

Received: March 9, 2019 **Accepted:** April 2, 2019

1. Introduction

Due to the existence of the extremely strong and long-ranged correlation between gene expression levels, all procedures which are commonly used to select the significant genes between two or more phenotypes cannot overcome the main problems: high instability of the number of false discoveries and low power. It may be impossible to completely understand these correlations due to the complexity of the biology nature. Some believe that normalized data can be of some utility in correlation analysis. The adverse effects of normalization have been reported in conjunction with single-color microarrays in several publications ([3, 9, 5, 8]). Leaving aside the question of whether or not the currently used normalization procedures achieve their promulgated goal, it is a well-known fact that they distort to various degrees the correlation structure of microarray data ([9, 7]), the latter being the main concern in reference to the results reported in [6]. The popular view that the observed correlations between gene expression levels are solely attributable to an array-specific random effect caused by the technical noise is demonstrably false ([3, 4, 2]). The analysis of the MAQC data set by Klebanov *et al.* [3] shows that the level of random fluctuations of gene expression signals attributable to the technical noise in the contemporary Affymetrix platform is too low to cause a tangible bias in estimated correlation coefficients. There is also independent evidence discussed in paper [5] that normalization procedures distort the joint distribution of the true expression signals quite dramatically, even affecting their marginal distributions. Every known normalization procedure resorts to pooling (heavily dependent) observed signals across different probes (probe sets), thereby producing surrogate variables whose distributions differ from those of the true biological signals. In the context of testing for differentially expressed genes, this distortion of the true signal may induce an uncontrollable number of false discoveries, an effect especially pronounced in large sample studies where the control of type I errors may be entirely lost.

The idea of normalization was initially offered as an *ad hoc* expedience to improve significance testing for differentially expressed genes in two-sample comparisons. Even in this setting, the universal benefits of normalization are questionable. The situation is more obvious when the main focus is on correlation coefficients. Destroying correlations before studying them quantitatively is self-defeating and may lead to false biological conclusions. Therefore, it is best that normalization should not be used (or at least be used cautiously) when making inferences about the correlation structure of microarray data.

Klebnov and Yakovlev [2] introduced the δ-sequence of weak dependency. In this paper, we further study the application of this δ-sequence in the selection of the significantly changed genes. In Section 1, the δ-sequence is reviewed. The use of δ-sequence in conjunction with the Bonferroni adjustment and balancing type I and type II errors are presented in Section 2, with some real microarray data analysis results. The comparison with the univariate gene selection method is discussed in Section 3.

2. The δ-Sequence and Its Utility in Microarray Data Analysis

Klebnov and Yakovlev [2] proposed a new sequence of gene pairs (for each chip) that are weakly dependent random variables. This sequence, noted as the δ-sequence, provides a new method for identifying the most differentially expressed genes between phenotypes. In the present section, the δ-sequence is reviewed with its properties versus gene expression.

Suppose there are m genes and each gene has n independent replicates of gene expression measurements which are identically distributed. In the analysis, we use the logged expression levels, though the tendency of the correlation between the genes (or gene pairs) should be similar if we use the raw expression values.

The new sequence of gene pairs is formed in two steps. First, the genes are re-ordered according to their calculated variances in increasing order. Second, based on this reordered gene list, we form the sequence $\delta_i = y_{2i} - y_{2i-1}$, where y_j is the jth logged gene expression value for gene j, $j = 1,\dots,m$, where $i = 1,\dots,\frac{m}{2}$. When m is an odd number, the mth gene, which is the last one in the original list, can be paired with any other genes in the list. Under this situation, only two overlapping gene pairs are generated in the whole sequence and the impact on the results is not tangible.

The aforesaid is summarized in the form of the following algorithm:

1. Sort the gene lists with respect to their variance in increasing order.
2. Calculate $\delta_i = y_{2i} - y_{2i-1}$, where y_i is the ith gene expression value for gene i, $i = 1,\dots,\frac{\text{\# of genes}}{2}$. If the number of genes is odd, pair the last gene with any other randomly chosen gene.

The sequence produced is termed δ-sequence, and the properties of this new list of gene pairs have been examined thoroughly in the paper of Klebnov and Yakovlev [2]. They showed that the correlation within this new sequence is shifted toward left dramatically, compared with the correlation within their original gene values. Klebnov and Yakovlev [2] confirmed this property in various real data sets where the histogram of the correlation coefficients becomes symmetric and centered around zero. Figure 1 (Klebnov and Yakovlev [2]) is presented in their paper to illustrate the sample correlation between log-expression level of the single genes and the formed δ-sequence. To emphasis this nice property, we re-present their figure here. As seen, the correlation between the single genes tends to be very high while the elements in the δ-sequence are almost independent with the mean of the correlation coefficient of the all pairs of the elements around zero.

Gordon *et al.* [1] proposed a method of balancing type I and type II errors in multiple testing for differentially expressed genes. Because of the nature of the multiple testing procedure and the high dependency existing in the gene expression data, the testing has unusually high instability and low power. Having the new produced gene-paired sequence, which has the nice property of having a weak dependency, we improve the selection procedure by applying our novel method of balancing the type I error and type II error on this δ-sequence. To much preserve the natural information of the gene expression data, we use the re-sampling method, and the details of our study are presented in the following section.

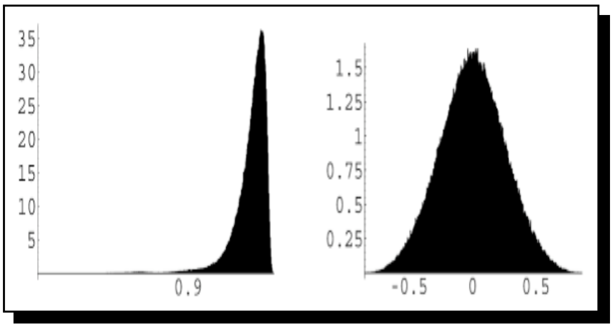

Figure 1. Histogram of correlation coefficients for all gene pairs from the first 2000 ordered genes (left panel) and for all pairs of the corresponding elements δ_i (right panel). The mean values are 0.97 for expression levels (left panel) and -1.5×10^{-4} for δ_i (right panel). HYPERDIP data set

3. The Use of δ-Sequence in Conjunction With the Bonferroni Adjustment and the Balancing Type I and Type II Errors

In the paper of Gordon *et al.* [1], a procedure was presented to balance type I and type II errors in significance testing for differential expression of individual genes. In the present section, we apply our new procedure on the δ-sequence in the process of selecting the significantly changed genes between two phenotypes.

3.1 Re-sampling

As one can expect from the paper of Gordon *et al.* [1], the study presented here is extensively computing expensive as there is double looping of the re-sampling for the real microarray data. To accomplish our study, we carried out our computing on a computer cluster. The study was designed similarly to the one described in [1]. We conduct our study by re-sampling from real data to preserve the actual correlation structure of gene expression levels as much as possible. The same data set, a set of microarray data of expression levels (Affymetrix GeneChip platform) of $m = 7084$ genes in $n = 88$ patients with hyperdiploid (HYPERDIP) acute lym-phoblasic leukemia identified through the St. Jude Children's Research Hospital Database (Yeoh *et al.* [10]), is used. 350 genes were randomly selected before the sub-sampling, and the standard deviation of their log-expression levels were estimated. These 350 pre-selected genes were fixed throughout all experiments.

For each step of the re-sampling procedures, two sub-samples of subjects (array) without overlapping, each of size $n = 30$, were generated from the collection of 88 available arrays. One sub-sample ($n = 30$) was modified manually by adding a constant shift (effect size) to the observed log-expression levels of the pre-defined 350 genes, and the second sub-sample stayed untouched. We use the calculated standard deviation described above as the effect size in our re-sampling study, and report the results obtained in the next section. A total 1500 pairs of sub-samples were generated and each of them was used to select differentially expressed genes by the proposed balancing method.

For each pair of sub-samples, the set S_i was formed by the permutation method. In doing so, for each looping, we permuted the gene list, formed a new sequence of gene pair by subtracting the gene expression levels of gene $2j-1$ from gene $2j$, where $j = 1,\dots,m/2$, and applied a two-sample test to the "new" log-expression measurements. A list of new gene pairs were selected at a significance level of 0.05. Each set S_i included genes for which the genes are included in the list obtained above. The procedure was repeated $k = 1500$ times and the frequencies of occurrence of each gene in the set S_i, $i = 1,\dots,k$, were estimated. Using the frequencies, a final set of differential expressed genes was identified from each sub-sample 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 details of the algorithm are summarized in the following:

1. Specify the penalties C_I and C_{II} for type I and type II errors, respectively, and compute $h = \frac{C_I}{C_I+C_{II}}$.
2. Randomly select $m = 350$ genes out of all genes.
3. Estimate the standard deviation σ of $m = 350$ genes, and use this σ to calculate the effect size: effect.size = σ.
4. Randomly draw two groups 30/30 (no-overlapping) out of 88 slides. For one group, add effect.size to each of the gene expression value of those pre-determined 350 genes.
5. Based on 30/30 group settings

   5.a. Calculate the standard deviation for each gene in one group, and sort the genes (two groups together) by the calculated standard deviation in increasing order. The genes mentioned below are based on the sorted genes.

   5.b. Obtain a list of modified genes by $y_{2n} - y_{2n-1}$, where y_i is the gene expression value for gene i, $i = 1,\dots,\frac{\text{\# of genes}}{2}$. We now have a list of 7084/2 paired-genes.

   5.c. Apply the Wilcoxon test to all the paired genes, and select the "significant" genes at the significance level α with the Bonferroni adjustment.

   5.d. 'Shift' method for δ-sequence to form the new gene list by subtracting the gene expression levels of gene $2i$ from gene $2i+1$, $i = 1,\dots,\frac{\text{\# of genes}}{2}$ e.g., 3-2, 4-3, ..., the last one-the first one.

   5.e. Apply the Wilcoxon test to all the paired genes formed above, and select the "significant genes" at the significance level α with the Bonferroni adjustment.

   5.f. Select the genes which are in the two selected gene lists, and compare with the pre-defined true 350 genes, and calculate the number of the true discoveries and the number of the false discoveries, and report the results.

6. Permutation 30/30 (K times), for each iteration,

   6.a. Permute all the genes, form a list of new gene list by subtracting the gene expression levels of gene $(2*n-1)$ from gene $(2*n)$.

   6.b. Apply Wilcoxon test on the data set from step 6.a, and obtain a list of significant genes (p-value < α $(= 0.05)$), denoted as S_i.

6.c. Repeat step 6.a and step 6.b for $K = 1500$ times, and then we will have a collection of subsets S_i of selected genes, where $i = 1, \ldots, K$.

6.d. For each gene j of all the genes (7084), calculate the proportion of set S_i $(i = 1, \ldots, K)$ that contain the gene j, denoted as b_j, if $b_j > h$, gene j is selected

6.e. Repeat step 6.d for all the genes, and get a list of genes selected, denoted it as S

6.f. Calculate the number of true discoveries and false discoveries.

7. Repeat step 4-6 M times.

The procedure described above is conducted on the HYPERDIP data set, and the results are shown in the next section.

4. Results and Discussion

The mean number of the discoveries and the corresponding standard deviation as functions of the threshold parameter h (for $h \geq 0.5$) are shown in Figure 2 together with the relationship between the true discoveries and false discoveries. The effect size of one σ is used to carry out the experiment in this figure. As shown, it is easy to see that the mean of the true discoveries decreases as the variation increases. The mean number of false discoveries decreases as the variation decreases. The number of the true discoveries increases to some point then becomes stable as the number of the false discoveries increases. The flatness of the number of the true discoveries after some point is related to the fact that the number of true positives is bounded from above by the value of 350 in our experiment, and the number of false discoveries is bounded by the total number of null hypothesis. In our study, there are 6783 null hypotheses, which is much larger than the number of the true alternative hypotheses. We have studied the different effect sizes, and with the smaller effect size (half σ), the mean and the variance of the number of false discoveries both increase a little bit but not much, as shown in Figure 3. The observation is the same as when the effect size is σ. It is a natural result, as one would expect.

To better demonstrate the different trend in the number of true discoveries, false discoveries and total discoveries and the variations, two more figures (Figure 4 and Figure 5) with the means (or standard variations) are also presented for easy and clear comparison.

Comparing with the results in paper ([1]), a similar behavior have been observed that both the mean and the standard deviation of the number of false positives are monotonically decreasing as a function of h. However, working with pairs, we have substantial reduction in both numbers, which means we gain power and stability in our new study. Figure 6 and Figure 7 confirmed our observation. This effect is attributable to the fact that the paired-genes have much weaker dependency than the single genes. The extremely strong and long-ranged correlation between gene expression levels is one of the reasons that causes the high standard deviation of the number of false discoveries.

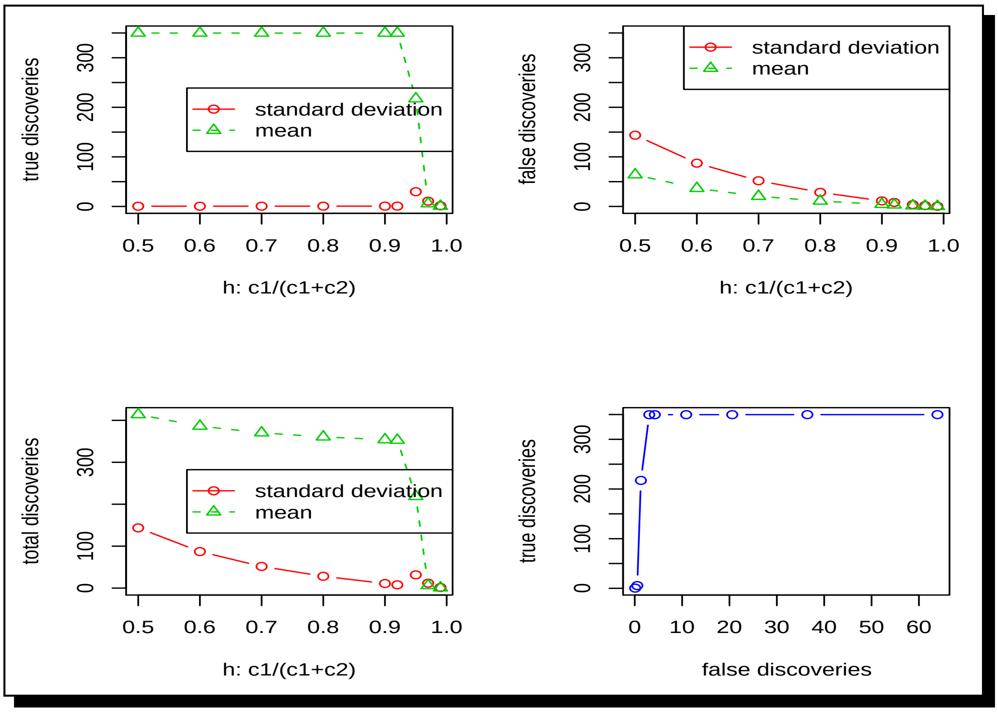

Figure 2. Mean and standard deviation of the number of true and false positives as functions of the parameter h. The total number of genes is 7084, the number of 'truly different' genes is 350, the effect size is equal to one σ. Other parameters are described in the text

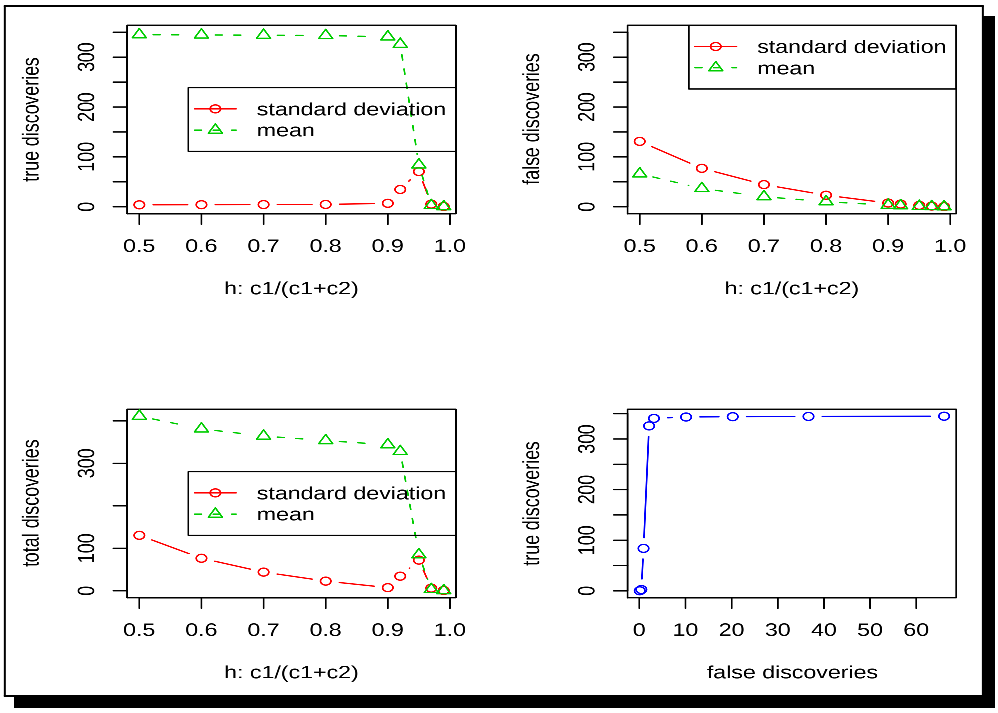

Figure 3. Mean and standard deviation of the number of true and false positives as functions of the parameter h. The total number of genes is 7084, the number of 'truly different' genes is 350, the effect size is equal to $\frac{\sigma}{2}$

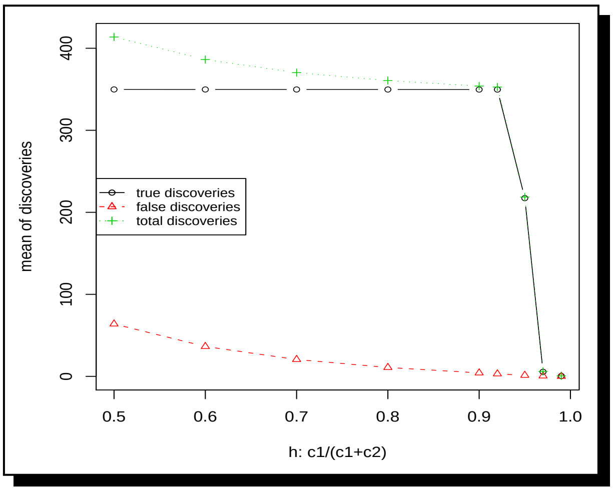

Figure 4. Mean of the number of the discoveries as functions of the parameter h

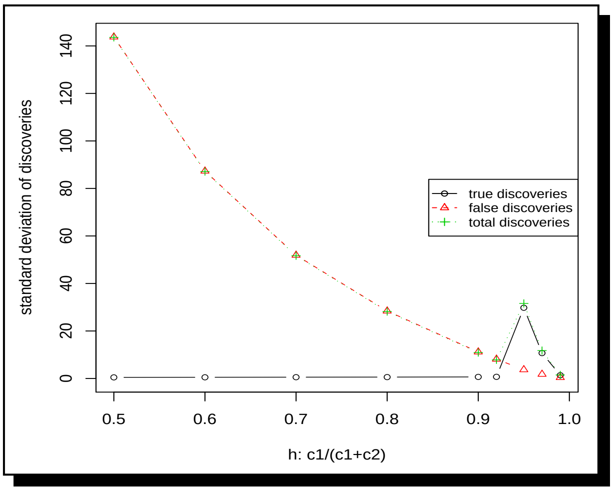

Figure 5. Standard deviation of the number of discoveries as functions of the parameter h

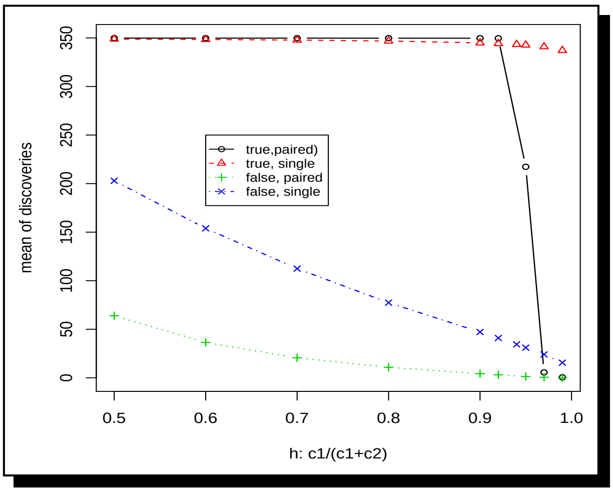

Figure 6. Mean of the number of the discoveries as functions of the parameter h

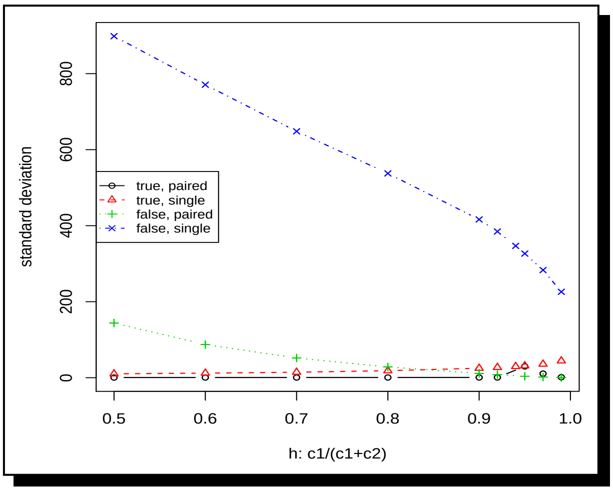

Figure 7. Standard deviation of the number of discoveries as functions of the parameter h

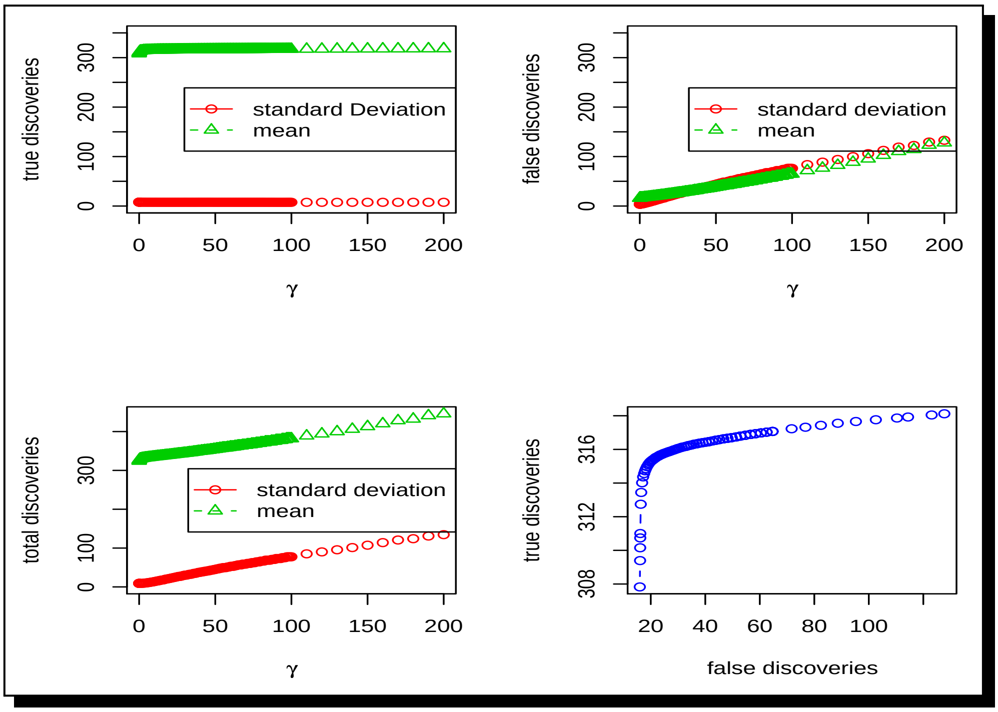

Figure 8. Mean and standard deviation of the number of true and false positives as functions of the parameter γ. The total number of genes is 7084, the number of "truly different" genes is 350, the effect size is equal to one σ. Other parameters are described in the text

Similar results are shown in Figure 8 with the Bonferroni procedure that controls the *Per-Family Error Rate* (PFER) at level γ. The tendency is the same to those in Figure 2. With the favorable property of the weak dependency between the paired genes, the test results confirm the large reduction of the number of mean false discoveries and the variation, compared with the analysis with single genes.

Similar to what we have talked about before, 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. 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. 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 re-sampling or simulations is computationally prohibitive.

In this paper, we improved the optimized procedure with the δ sequence in terms of improving power and stability by reducing the mean and standard deviation of the false positive rates. We need to point out that the way of pairing genes to form pair with the resultant weakly dependence sequences can be extended to more numbers of genes in a pair, which we will investigate further in the future.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] A. Gordon, L. Chen, G. Glazko and A. Yakovlev, Balancing type one and two errors in multiple testing for differential expression of genes, *Computational Statistics & Data Analysis* **53**(5) (2009), 1622 – 1629, DOI: 10.1016/j.csda.2008.04.010.

[2] L. Klebanov and A. Yakovlev, Diverse correlation structures in microarray gene expression data and their utility in improving statistical inference, *Annals of Applied Statistics* **1**(2) (2007), 538 – 559, DOI: 10.1214/07-AOAS120.

[3] L. Klebanov and A. Yakovlev, How high is the level of technical noise in micro array data?, *Biology Direct* **2** (2007), Article 0, DOI: 10.1186/1745-6150-2-9.

[4] L. Klebanov, C. Jordan and A. Yakovlev, A new type of stochastic dependence revealed in gene expression data, *Statistical Applications in Genetics and Molecular Biology* **5**(1) (2006), Article 7, DOI: 10.2202/1544-6115.1189.

[5] L. Klebanov, L. Chen and A. Yakovlev, Revisiting adverse effects of cross-hybridization in Affymetrix gene expression data: Do they matter for correlation analysis?, *Biology Direct* **2**(28) (2007), DOI: 10.1186/1745-6150-2-28.

[6] M. J. Okoniewski and C. J. Miller, Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations, *BMC Bioinformatics* **7** (2006), Article 276, DOI: 10.1186/1471-2105-7-276.

[7] A. Ploner, L. D. Miller, P. Hall, J. Bergh and Y. Pawitan, Correlation test to assess low-level processing of high-density oligonucleotide microarray data, *BMC Bioinformatics* **6** (2005), Article 80, DOI: 10.1186/1471-2105-6-80.

[8] X. Qiu and A. Yakovlev, Comments on probabilistic models behind the concept of false discovery rate, *J. Bioinformatics and Comput. Biol.* **4** (2007), 963 – 975, DOI: 10.1142/S0219720007002965.

[9] X. Qiu, L. Klebanov and A. Y. Yakovlev, Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes, *Statistical Applications in Genetics and Molecular Biology* **4** (2005), Article 34, DOI: 10.2202/1544-6115.1157.

[10] E. H. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. H. Pui, W. E. Evans, C. Naeve, L. Wong and J. R. Downing, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell* **1**(2) (2002), 133 – 143, DOI: 10.1016/S1535-6108(02)00032-6.