

Assessing Stability of Gene Selection in Microarray Data Analysis

Xing Qiu, Yuanhui Xiao, Alexander Gordon,
and Andrei Yakovlev

¹*Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Rochester, New York 14642, USA.*

Corresponding author: Andrei Yakovlev, Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, New York 14642, USA, ph: (585)275-6688, fax: (585)273-1031, e-mail: Andrei_Yakovlev@urmc.rochester.edu

E-mail addresses:

X. Qiu: xqiu@bst.rochester.edu

Y. Xiao: yxiao@bst.rochester.edu

A. Gordon: Alexander_Gordon@urmc.rochester.edu

A. Yakovlev: Andrei_Yakovlev@urmc.rochester.edu

Abstract

Background. The number of genes declared differentially expressed is a random variable and its variability can be assessed by resampling techniques. Another important stability indicator is the frequency with which a given gene is selected across subsamples. We have conducted studies to assess stability and some other properties of several gene selection procedures with biological and simulated data.

Results. Using cross-validation techniques we have found that some genes are selected much less frequently (across cross-validation samples) than other genes with the same adjusted p -values. The extent to which this type of instability manifests itself depends on a specific multiple testing procedure and the choice of a test statistic. The effect of correlation between gene expression levels on the performance of multiple testing procedures is studied by computer simulations.

Conclusions. Cross-validation represents a tool for reducing the set of initially selected genes to those with a sufficiently high selection frequency. Using cross-validation it is also possible to assess variability of different performance indicators. Stability properties of several multiple testing procedures are described at length in the present paper.

1 Background

The result of every analysis of microarray data is an outcome of a random experiment. For example, the number of genes declared differentially expressed and the estimated false discovery rate (FDR) should be treated as random variables and their variability has to be assessed in the same fashion that

the population variance is estimated in the usual statistical inference. The variance of the number of differentially expressed genes (as well as other outcomes of a given selection procedure) may depend on the chosen statistical test, method of multiple testing adjustment, effect sizes for different genes, and the correlation structure of the data. The latter factor deserves special attention. Although some normalization procedures may lead to a significant reduction in the correlation between gene expression levels, and thus between the associated test statistics, the remaining correlation may be strong enough to have a disastrous effect on the statistical inference from microarray data [1]. The effect of correlation between test statistics on the number of differentially expressed genes and estimates of the FDR was recently studied in detail by Qiu et al. [2] in conjunction with the empirical Bayes methodology for microarray data analysis. The effect of the other factors on variability of the most basic performance indicators of various testing procedures also invites a special investigation.

There is another facet of the problem to consider. Every specific analysis of microarray data results in a list of candidate genes that are deemed differentially expressed across the two conditions under study. The composition of this list is subject to random fluctuations and this effect also needs to be quantitatively assessed. Even if one concentrates on the selection of individual genes rather than gene combinations, the situation here is similar to that in the regression analysis aimed at selecting significant explanatory variables (covariates). When focusing on a specific variable, one can observe a certain degree of instability of this variable selection inherent in any pertinent statistical procedure. The term “stability” means “replication stability” for the selection of significant variables. This kind of stability is easy to as-

sess and interpret in simulation studies where the “true” set of differentially expressed genes is known. When analyzing biological data, one can resort to resampling techniques for this purpose [3]. In particular, one can apply a “leave-many-out” cross-validation procedure to the sample at hand and estimate the frequency with which a given gene has been selected across all cross-validation sub-samples. Then an additional selection criterion can be imposed by finally selecting only those genes that occur in the target set with a frequency greater than, say, 80%.

We have conducted a simulation study to evaluate the performance of several selection procedures in terms of the variability of such important indicators as the number of selected genes, FDR, and proportion of correctly rejected null hypotheses among all the hypotheses tested. All these indicators are directly accessible in computer simulations, thereby providing an explanatory insight into the performance of different procedures. From this perspective, the Bonferroni and Westfall-Young multiple testing procedures are explored in conjunction with the Student t , Kolmogorov-Smirnov, and Cramér-von Mises two-sample tests. The latter two tests are distribution free. The Bonferroni and Westfall-Young step-down procedures [4] are designed to control the familywise error rate (FWER). The FDR-based procedures are also explored; these are represented by the empirical Bayes method [5, 6, 7] and the Benjamini-Hochberg procedure [8]. It should be noted that our simulation studies do attempt to model the actual microarray data; their only purpose is to see which specific performance indicators may be sensitive to the presence of correlation in the data. Another study was concerned with actual biological data. We assessed probabilistic characteristics of the number of selected genes by resampling from a large set of data on two types of

childhood leukemia available from the St. Jude Children’s Research Hospital Database [9]. Using this data set, we also assessed the replication stability of gene selection and its dependence on adjusted p -values.

2 Results

2.1 Analysis of Biological Data

Table 1 presents the results of the leave-one-out (Del-1) and leave-seven-out (Del-7) cross-validation procedures applied to the selected set of biological data. In this study, the FWER is controlled at the level of 0.05. Shown in the parentheses is the percentage of “stable” genes relative to the mean (over the 200 cross-validation samples) number of selected genes computed under the additional requirement of at least 80% occurrence in the set of selected genes (target set). This percentage remains practically unchanged when changing the FWER control level for all the tests except the Kolmogorov-Smirnov test. The latter test displayed some irregularities when used in combination with leave-one-out cross-validation at different control levels of the FWER. These irregularities are attributable to granularity of p -values inherent in distribution-free tests in general and the Kolmogorov-Smirnov test in particular, and they do not merit detailed consideration here.

The results for the t -test in conjunction with the nonparametric empirical Bayes method, Benjamini and Hochberg procedure and its modification by Benjamini and Yekutieli [10] are displayed in Table 2. Note that the standard deviation of the number of selected genes is very high for these procedures. Comparing the two tables, it is clear that the proportion of highly stable (with at least 80% occurrence) genes appears to be virtually the same for all the tests and multiple testing procedures. However, the situation is not the

same when looking at less frequent genes as discussed below.

Shown in Figure 1 are the proportions of genes with different frequencies of occurrence in the target set among those genes that have been selected at least once in the course of leave-seven-out cross-validation. It is seen from this figure that the histograms are *U*-shaped so that one can distinguish two extreme groups of genes characterized by high and low stability, respectively. The proportions of genes in each “intermediate-frequency” category are relatively small. This phenomenon persists for all the statistical tests under study when the FWER is controlled either by the Bonferroni adjustment or by the Westfall-Young permutation algorithm.

It is clear from Figure 1 that the population of genes selected at least once across all cross-validations is heterogeneous with respect to their stability characterized by the frequency of occurrence in the target set. To gain a better insight into this heterogeneity, it makes sense to look at the relationship between the frequency of occurrence and the corresponding p -values. To this end, we produced scatter-plots for the frequency of occurrence in the set of selected genes across cross-validation sub-samples and the original adjusted p -values determined by the application of some testing procedures to the whole set of arrays. The results for the t -test with Bonferroni adjustment are given in Figure 2. For the leave-one-out cross-validation, the dependence appears to be almost linear but the scatter of points is wide, thereby suggesting that the leave-one-out cross-validation does not perturb the data sufficiently. The leave-seven-out cross-validation reveals a non-linear (but still monotonic) pattern showing that the relationship in question may be quite complex. In what follows, we will discuss only the observations resulted from the leave-seven-out cross-validation.

The results for the t -test and the Cramér-von Mises test with Bonferroni adjustment are compared in Figure 3. It is clear that the genes selected by the Cramér-von Mises test are uniformly more stable than those selected by the t -test. The difference is much less pronounced with the Westfall-Young algorithm as evidenced by Figure 4. Both multiple testing procedures yield similar scatter plots for the t -test showing its overall poor stability in comparison to the Cramér-von Mises test (Figure 5). In contrast, the stability of the Cramér-von Mises test can be increased substantially when using the more conservative Bonferroni adjustment in place of the Westfall-Young procedure (Figure 6). These results show that the stability of gene selection provides an important additional information on each gene in the target set and this information can be extracted from real data by resorting to resampling techniques.

The mean values and standard deviations of the number of genes selected by different multiple testing procedures are reported in Tables 1 and 2. It is also interesting to look at the shape of the corresponding distribution. Figure 7 shows that this shape varies widely for different procedures. The nearly symmetric form of this distribution in combination with a relatively small variance is an appealing feature of the Cramér-von Mises test.

2.2 Analysis of simulated data

To assess the effect of correlation between gene expression levels on the performance of gene selection procedures, we carried out simulation studies as described in Section 5.2.

Table 3 presents the most basic performance indicators for the sample size $n = m = 15$. Since the simulated data are normally distributed it comes as

no surprise that the t -test proves itself as the most powerful one among those under study. With this small sample size, however, even the t -test tends to be underpowered when used in combination with the Bonferroni adjustment or Westfall-Young adjustments. The power of the t -test is much higher with the Benjamini-Hochberg and nonparametric empirical Bayes procedures. The variance of the estimated power as well as the number of selected genes increases dramatically with increasing correlation between gene expression signals.

Table 4 shows the results for a larger sample size ($n = m = 43$). In this case, all the methods attain 100% power. For all the FWER controlling procedures, the mean number of selected genes is exactly 125 and the corresponding variance is quite small irrespective of the presence or absence of correlation between gene expression levels. The FDR estimates are also uniformly small for such procedures as indicated by Table 4. However, the effect of correlation on the standard deviation of the number of selected genes is still very strong (compare with Table 3) for the Benjamini-Hochberg and nonparametric empirical Bayes procedures, indicating the inherent instability of these procedures. It should be noted that there is also a dramatic effect of the correlation on the standard deviation of the FDR observed for the latter procedures (Table 4). The results for 1000 simulation runs were largely similar.

The histograms for the number of selected genes resulted from our simulation studies are included in the Additional Material Files [see the file “SIMULDIST”]. Note that the high variance observed for the BH/ t and EB/ t procedures (Figures 3 and 4 in the Additional Files) is mainly attributable to outliers.

3 Discussion

Numerous publications have considered the utility of multiple testing procedures in the context of microarray data analysis (see [11] for a review). However, little attention has been given to the replication stability of such procedures. Our results show that the variance of the number of differentially expressed genes can be very high for some multiple testing procedures (e.g., the nonparametric empirical Bayes and Benjamini-Hochberg procedures) even with reasonably large sample sizes. The variance of the true FDR is also quite high for such procedures. Our simulations show that the FWER-controlling procedures may also yield a high variance of the number of selected genes with smaller sample sizes. Whenever this is the case, the stability of membership in the list of candidate genes should be expected to be low. However, the reverse is not true. If the variance of the total number of selected genes is low, there still can be tangible variations in the stability of selection for individual genes, thereby affecting the composition of the resultant list of candidate genes. This obviously can have a strong effect on the ranking of candidate genes based on purely statistical criteria such as the magnitude of associated test-statistics (p -values) or estimated posterior probabilities.

The present study demonstrates that the proportion of highly stable (with frequencies of more than 80%) genes appears to be almost the same for all the selection procedures under study. However, the overall stability of gene selection varies among different methods. The Cramér-von Mises seems to be superior to other methods in this respect. It is difficult to control the stability of gene selection by an additional adjustment of p -values. Indeed, for the FWER-controlling procedures, the relationship between the original

(adjusted for multiple testing) p -values and the selection frequency appears to be non-linear. However, resampling techniques represent a universal tool for assessing the stability in question with the data at hand. As was mentioned in Section 1, our simulation studies play only a subsidiary role in our attempt to provide a quantitative insight into the issue of stability of gene selection procedures. However, such studies are important in that they show how the correlation between gene expression levels can affect the results of testing two-sample marginal hypotheses.

Tables 3 and 4 illustrate the importance of sample size. It is unfortunate that the issue of sample size is often disregarded in microarray studies which typically report findings based on a very small number of arrays. We find it already quite amazing that some traditional testing procedures controlling the FWER seem to do a good job (in terms of the stability of gene selection and overall power) with such moderate sample sizes as 40-50 per group. The situation is not the same for less conservative FDR-based procedures which appear to be much more sensitive to the strength of correlation in the data. Unfortunately, it has become standard practice to make statistical inferences from just 3-4 replicates when analyzing microarray data. We believe that the sample size requirements should substantially be elevated to make microarray technology produce reliable and biologically significant results. These requirements have little to do with the usual power calculations, because the analysis of microarrays is exploratory (not confirmatory) by its very nature. Therefore, recommendations on this issue should be general enough to cover a wide range of possible situations. It is only experience accumulated from numerous analyses of large data sets that can suggest a reasonable range of sample sizes. We believe that the proposed methodology will help the

scientific community gain such an experience.

4 Conclusions

Using cross-validation techniques we have found that some genes are selected much less frequently (across cross-validation samples) than other genes with the same adjusted p -values. The relationship between the stability of gene selection and the original (adjusted) p -values may be rather complex but cross-validation techniques can advantageously be used to select the most stable genes. Using cross-validation, it is also possible to assess variability of the number of selected genes. In reference to the latter indicator, all selection procedures are highly unstable when the sample size is small and correlations are present in the data. For the FWER-controlling procedures studied in the present paper by simulations, this property correlates well with the level of random fluctuations in the estimated power of a given procedure. With larger sample sizes, the more conservative FWER-controlling procedures appear to be more stable than the FDR-based procedures in the presence of correlations. The stability characteristics discussed in the present paper provide an additional information that should be utilized in gene selection procedures.

5 Methods

5.1 Biological Data

For the purposes of this study, use was made of the St. Jude Children's Research Hospital (SJCRH) Database on childhood leukemia which is publicly available on the following website: <http://www.stjuderesearch.org/data/ALL1/>. The whole SJCRH Database contains gene expression data on 335 subjects,

each represented by a separate array (Affymetrix, Santa Clara, CA) reporting measurements on the same set of $m = 12558$ genes. We selected two groups of patients with hyperdiploid (Hyperdip) and T-cell acute lymphoblastic leukemia (TALL), respectively. The groups were balanced to include 43 patients in each group. Since the nature of our study was purely methodological, the choice of the data set was quite arbitrary; it was dictated solely by sample size considerations. The microarray data thus chosen were background corrected and normalized using the Bioconductor RMA software. This software implements the quantile normalization procedure (Bolstad et al. 2003, Irizarry et al. 2003) carried out at the probe feature level. After this normalization, each gene is represented in the final data set by the logarithm (base 2) of its expression level.

5.2 Simulated Data

We simulated $2n$ independent multi-variate normal random vectors with exchangeable correlation structure, each representing log-intensities of 1255 genes of which the first 125 genes were designated to be differentially expressed. Two sets of simulations were conducted with the sample size chosen to be $n = 15$ and $n = 43$, respectively. In total, 200 independent data sets, each consisting of $2n$ simulated vectors, were generated for each sample size. The marginal distributions of the log-intensities of “Not Different” genes were standard normal, while the log-intensities of “Different” genes expressions followed the normal distribution with mean two and unit variance. The exchangeable pairwise correlation structure was superimposed on the normal vectors with independent components as discussed in [1]. In the present study, the correlated data were generated for a single value of the correlation

coefficient $\rho = 0.6$. We use the following self-explanatory notation for the four sets of simulated data: SIM15, SIM15CORR, SIM43, SIM43CORR.

5.3 Resampling Techniques

When analyzing biological data, we used a subsampling version of the delete- d jackknife method [14, 15], which is essentially equivalent to the leave- d -out cross-validation. For simplicity, this type of analysis will be referred to as the cross-validation analysis. We used two values of d , $d = 1$ (leave-one-out cross-validation) and $d = 7$ (leave-seven-out cross-validation), to perturb the data set. The total number of cross-validation samples was typically equal to 200. In a separate study, we ascertained that the results for 1000 cross-validation samples were largely similar. Let Z be the statistic (performance indicator) under study. The variance of Z is estimated by a resampling counterpart of the jackknife sample variance [15]:

$$V = \frac{n-d}{dB} \sum_{l=1}^B \left(Z_{n-d,l} - \frac{1}{B} \sum_{k=1}^B Z_{n-d,k} \right)^2,$$

where B is the total number of subsamples ($B = 200$), $Z_{n-d,j}$ is the statistic Z evaluated at the j th delete- d jackknife subsample.

5.4 Selection of Differentially Expressed Genes

When resorting to the Bonferroni adjustment, one needs to compute unadjusted p -values from the sampling distribution of the test statistic under consideration. For the t -test we used quantiles of the Student distribution. For the Kolmogorov-Smirnov and Cramér-von Mises tests, original numerical algorithms were developed and verified by comparison with the corresponding quantiles computed by other methods [16]. These algorithms will be presented at length in another paper. The Westfall-Young step-down algorithm

[4] bypasses the stage of computing unadjusted p -values and goes directly to the estimation of adjusted p -values at a given level of the FWER. We carried out 10,000 permutations to model a null distribution of each test statistic. We also used the multiple testing adjustment proposed by Benjamini and Hochberg [8] and its modification by Benjamini and Yekutieli [10]. The non-parametric empirical Bayes method by Efron et al. [5, 6, 7] was one more method of choice in the present paper. We used kernel smoothing (with the Gaussian kernel) for density estimation to implement the empirical Bayes method. The threshold level of the posterior probability was set at 0.95.

To distinguish between different statistical procedures, we use the following notation:

B/t – t -test with Bonferroni adjustment;

B/KS – Kolmogorov-Smirnov test with Bonferroni adjustment;

B/CVM – Cramér-von Mises test with Bonferroni adjustment;

WY/t – t -test with Westfall-Young algorithm;

WY/KS – Kolmogorov-Smirnov test with Westfall-Young algorithm;

WY/CVM – Cramér-von Mises with Westfall-Young algorithm;

BH/t – t -test with Benjamini-Hochberg adjustment;

BY/t – t -test with Benjamini-Yekutieli adjustment;

EB/t – t -test with gene selection by nonparametric empirical Bayes method.

5.5 False Discovery Rate and Power

We provide estimates of the FDR only for simulations. We do not report FDR estimates for biological data because only indirect methods [18, 19] are available in this case. Such methods introduce an additional variation in the estimates which is impossible to distinguish from that caused by a given

selection procedure. In our simulation studies, the true FDR was estimated directly as the proportion of false discoveries among all discoveries. Then the sample mean (across the 200 samples) of this nonparametric estimate is reported together with the corresponding standard deviation. It happened only once (when applying the Kolmogorov-Smirnov test with Bonferroni adjustment to a sample of size $n = 15$) that we set the estimated FDR at zero (see [17] for the definition of the positive FDR). Since the expression levels of the 125 differentially expressed genes are identically distributed, the power can be defined as the expected proportion of correct discoveries among the 125 true alternative hypotheses. We provide the usual nonparametric estimates of the power thus defined and its standard deviation.

6 Authors' Contributions

The basic idea behind this study emerged from discussions between AY and AG. The detailed study design was developed by all the members of the research team. YX and XQ carried out the needed computations and simulations.

7 Acknowledgements

This research is supported by NIH grant GM075299 (Yakovlev).

8 Additional files

The additional file "SIMULDIST" includes four figures representing histograms for the number of selected genes pertaining to the simulation studies reported in Section 2.2.

References

- [1] Qiu X, Klebanov L, Yakovlev AY: Correlation between gene expression levels and limitations of the empirical Bayes methodology in microarray data analysis. *Statistical Applications in Genetics and Molecular Biology* 2005, submitted.
- [2] Qiu X, Brooks AI, Klebanov L, Yakovlev A: The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* 2005, 6: # 120.
- [3] Sauerbrei W, Schumacher M: A bootstrapping resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1993, 11: 2093-2109.
- [4] Westfall PH, Young S: *Resampling-Based Multiple Testing*. Wiley, New York, 1993.
- [5] Efron B, Tibshirani R, Storey JD, Tusher V: Empirical Bayes analysis of a microarray experiment. *J Amer Statist Assoc* 2001, 96: 1151-1160.
- [6] Efron B: Robbins, empirical Bayes and microarrays. *Ann Statist* 2003, 31: 366-378.
- [7] Efron B: Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J Amer Statist Assoc* 2004, 99: 96-104.
- [8] Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995, 57: 289-300.

- [9] St. Jude Children's Research Hospital (SJCRH) Database on Childhood leukemia [<http://www.stjuderesearch.org/data/ALL1>].
- [10] Benjamini Y, Yekutieli D: The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 2001, 29: 1165-1188.
- [11] Dudoit S, Shaffer JP, Boldrick JC: Multiple hypothesis testing in microarray experiments. *Statistical Science* 2003, 18:71-103.
- [12] Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003, 19(2):185-193.
- [13] Irizarry RA, Gautier L, Cope LM: An R package for analyses of Affymetrix oligonucleotide arrays, In: *The Analysis of Gene Expression Data 2003*, Parmigiani G, Garrett ES, Irizarry RA, Zeger SL, eds, Springer, New York, pp. 102-119.
- [14] Politis DN, Romano JP: Large sample confidence regions based on subsamples under minimal assumptions. *Ann Statist* 1994, 22: 2031-2050.
- [15] Shao J, Tu D: *The Jackknife and Bootstrap*. Springer Series in Statistics, Springer, New York, 1995.
- [16] Conover WJ: *Practical Nonparametric Statistics*, 3d Edition. Wiley, New York, 1999.
- [17] Storey JD: The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann Statist* 2004, 31, 2013-2035.
- [18] Storey JD, Tibshirani R: Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003, 100, 9440-9445.

- [19] Reiner A, Yekutieli D, Benjamini Y: Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 2003, 19, 368-375.

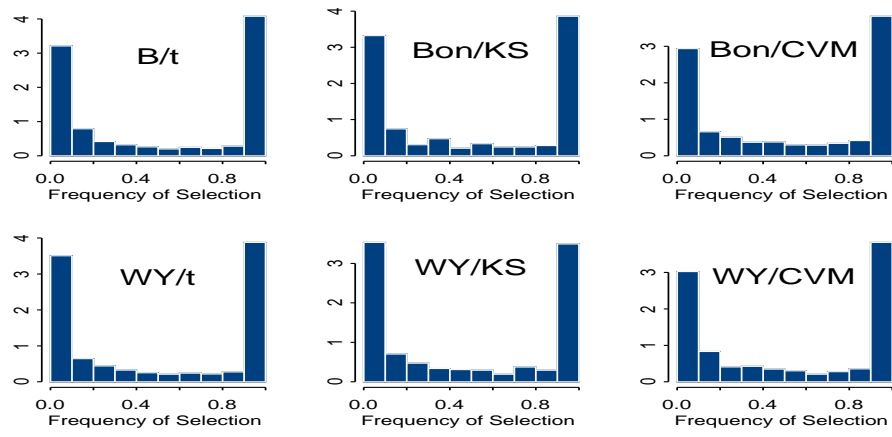


Figure 1: Histograms of the frequency of occurrence in the set of selected genes obtained by leave-seven-out cross-validation.

Table 1: Leave- d -out cross-validation analysis of biological data with $d = 1$ and $d = 7$

Method	Leave-one-out Cross Validation			Leave-seven-out Cross-validations		
	Mean number of selected genes	Standard deviation	Number of stable genes and its proportion to the mean number of selected genes	Mean number of selected genes	Standard deviation	Mean number of stable genes and its Proportion to the Mean of selected genes
B/KS	851	100	794(93.32%)	622	71	504(80.99%)
B/CVM	1043	152	954(91.49%)	1096	123	853(77.80%)
B/t	926	128	867(93.59%)	775	103	644(83.05%)
WY/KS	851	105	794(93.34%)	685	153	533(77.82%)
WY/CVM	1058	168	971(91.78%)	889	124	711(79.98%)
WY/t	1036	145	972(93.84%)	876	110	726(82.89%)

Table 2: Leave-7-out cross-validation analysis of biological data

Method	Mean number of selected genes	Standard deviation of the number of selected genes	Mean number of stable genes and its proportion to the mean number of selected genes
EB/t	1867	438	1481(79%)
BH/t	2726	445	2176(80%)
BY/t	1599	222	1282(80%)

Table 3: Simulating the basic characteristics of gene selection procedures, 125 differentially expressed genes, 200 simulation runs, $n = 15$. The table presents mean values over simulation runs. Standard deviations are given in parentheses.

Method	Number of Selected Genes		FDR		Power	
	S15	S15COR	S15	S15COR	S15	S15COR
B/KS	36(5.3)	34(18.8)	<0.0006	<0.0001	0.28(0.04)	0.27(0.15)
B/CVM	80(4.9)	80(25.0)	<0.0008	<0.0004	0.64(0.04)	0.64(0.20)
B/t	89(5.5)	88(24.7)	<0.0007	<0.0008	0.71(0.04)	0.70(0.20)
WY/KS	36(4.7)	53(26.3)	0	<0.0008	0.29(0.04)	0.43(0.21)
WY/CVM	81(5.6)	90(21.25)	<0.0003	<0.0008	0.65(0.04)	0.72(0.17)
WY/t	90(5.4)	98(19.66)	<0.0007	0.0009	0.72(0.04)	0.79(0.16)
BH/t	130(2.8)	139(73.5)	0.048(0.019)	0.051(0.135)	0.99(0.01)	0.99(0.03)
EB/t	116(3.0)	141(100.0)	0.012(0.006)	0.052(0.157)	0.92(0.02)	0.96(0.07)

Table 4: Simulating the basic characteristics of gene selection procedures, 125 differentially expressed genes, 200 simulation runs, $n = 43$. The table presents mean values over simulation runs. Standard deviations are given in parentheses.

Method	Number of Selected Genes		FDR		Power	
	SIM43	SIM43CORR	SIM43	SIM43CORR	SIM43	SIM43CORR
B/KS	125(0.3)	125(0.5)	<0.0003	<0.0001	1	1
B/CVM	125(0.3)	125(0.3)	<0.0005	<0.0005	1	1
B/t	125(0.2)	125(0.4)	<0.0003	<0.0006	1	1
WY/KS	125(0.4)	125(0.4)	<0.0002	<0.0003	1	1
WY/CVM	125(0.3)	125(0.4)	<0.0006	<0.0010	1	1
WY/t	125(0.2)	125(0.3)	<0.0003	<0.0008	1	1
BH/t	131(2.7)	140(90.0)	0.0427(0.0192)	0.0356(0.1219)	1	1
EB/t	125(0.2)	133(60.0)	0.0082(0.0012)	0.0246(0.1024)	1	1

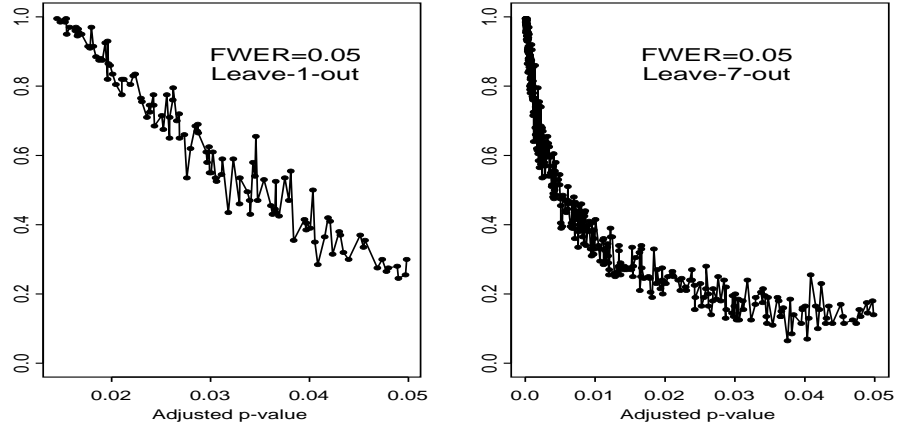


Figure 2: Frequency of occurrence in the set of selected genes versus adjusted p -values for the t -test with Bonferroni adjustment. Left panel: leave-one-out cross-validation, right panel: leave-seven-out cross-validation.

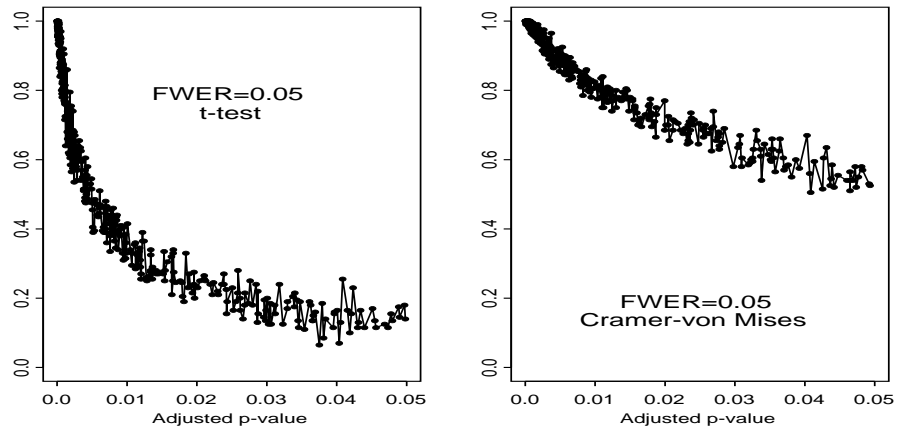


Figure 3: Frequency of occurrence in the set of selected genes versus adjusted p -values for the t - and Cramér-von Mises test with Bonferroni adjustment (leave-seven-out cross-validation).

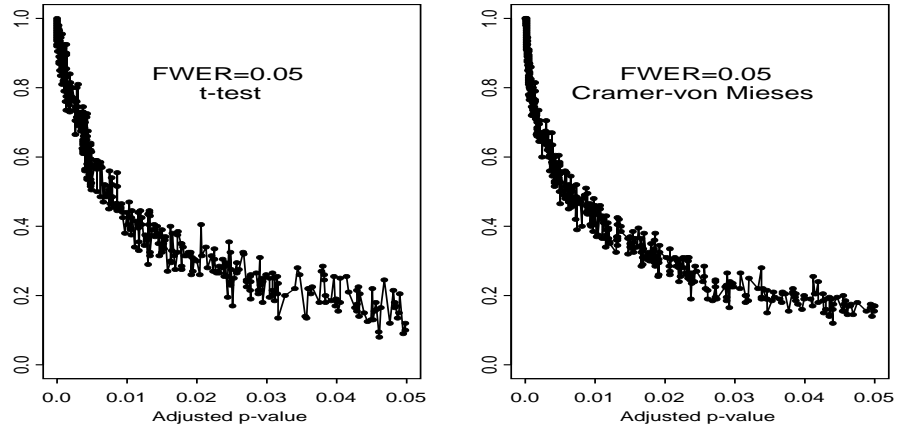


Figure 4: Frequency of occurrence in the set of selected genes versus adjusted p -values for the t - and Cramér-von Mises test with Westfall-Young algorithm (leave-seven-out cross-validation).

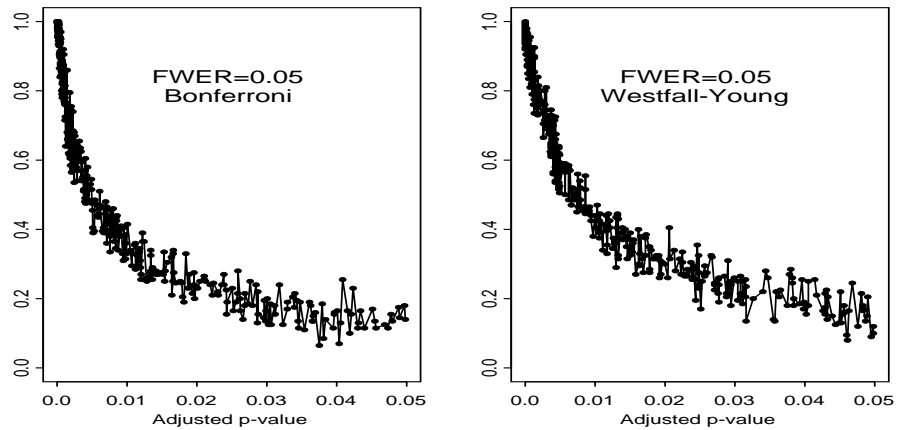


Figure 5: Frequency of occurrence in the set of selected genes versus adjusted p -value for the t -test with Bonferroni adjustment and Westfall-Young algorithm (leave-seven-out cross-validation).

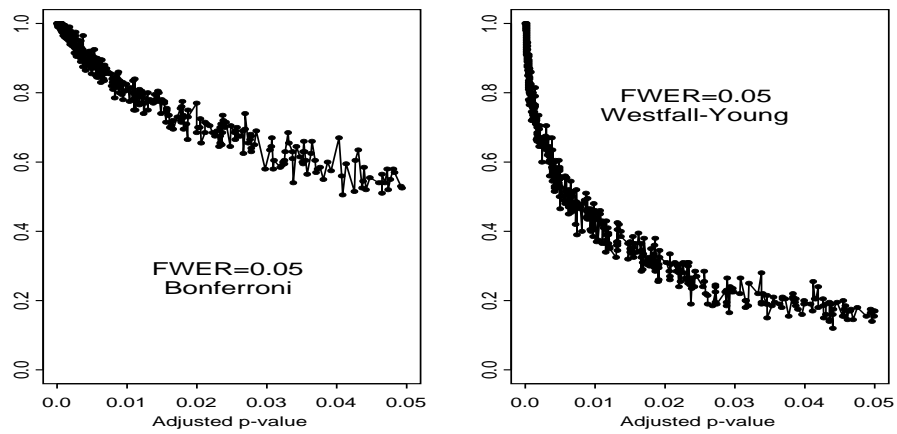


Figure 6: Frequency of occurrence in the set of selected genes versus adjusted p -values for the Cramér-von Mises test with Bonferroni adjustment and Westfall-Young algorithm (leave-seven-out cross-validation).

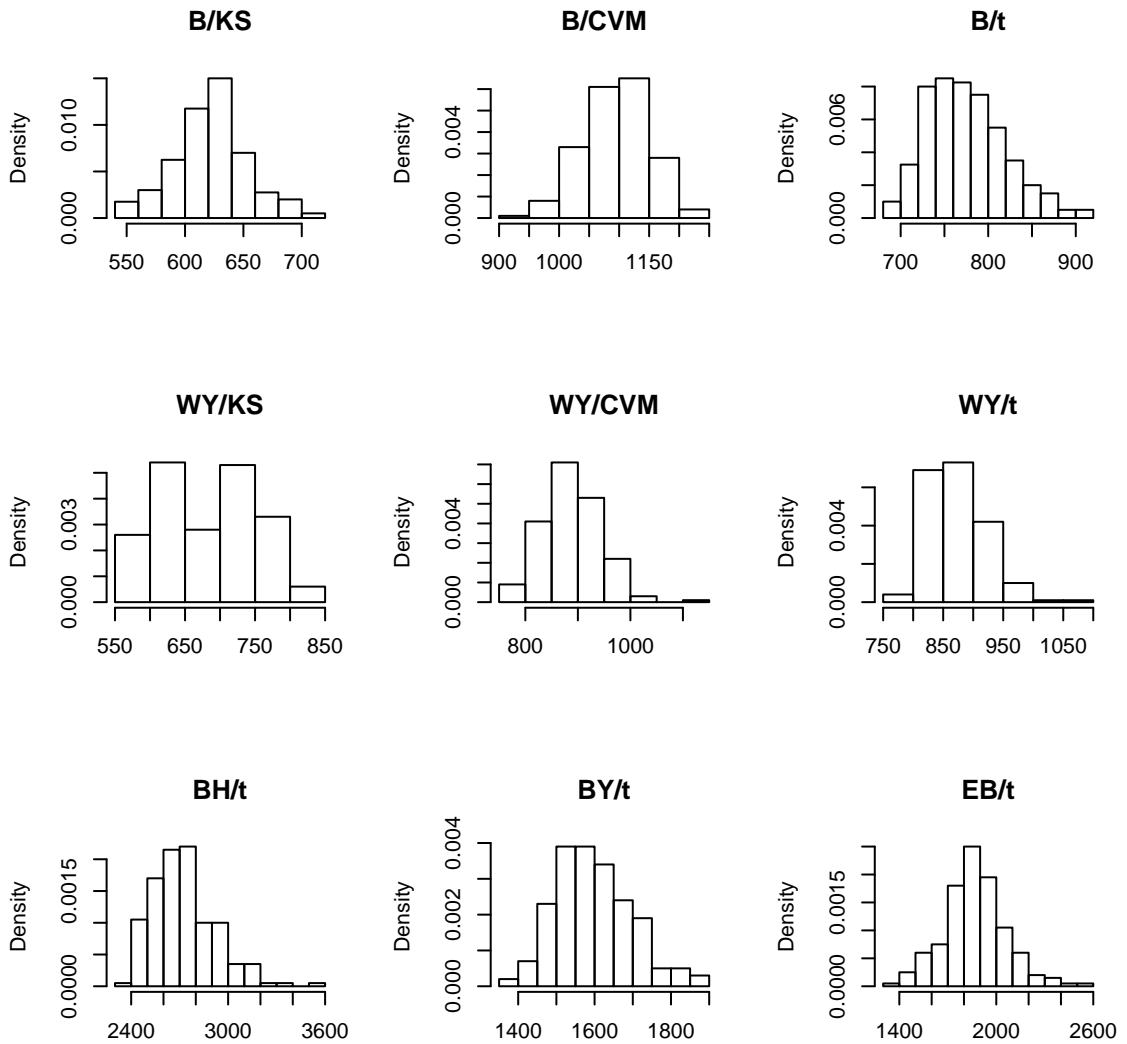


Figure 7: Histograms of the number of selected genes across 200 cross-validations for different methods applied to the biological data.