Power and stability properties of resampling-based multiple testing procedures with applications to gene oncology studies

Dongmei Li[1] and Xuemei Li[2]

[1]Department of Public Health Science, The University of Hawaii at Manoa,
1960 East-West Road, Honolulu, HI 96822
[2]Shandong Huayu Technology College, Dezhou, Shandong 253034, P. R. China

Abstract

Resampling-based multiple testing procedures are widely used in genomic studies to identify differentially expressed genes and to conduct genome-wised association studies (GWAS). However, the power and stability properties of these popular resampling-based multiple testing procedures have not been extensively evaluated yet. Our study focus on investigating the power and stability of several resampling-based multiple testing procedures frequently used in microarray studies through simulations and gene oncology examples and provide some guidance and cautions for researchers using those methods in microarray data analysis.

Key Words: Resampling, multiple testing procedures, power, stability

1. Introduction

With the rapid development of biotechnology, microarrays became widely used in biomedical and biological fields to identify differentially expressed genes, detect transcription factor binding sites, and map complex traits using single nucleotide polymorphisms (SNPs) (Kulesh et al., 1987; Schena et al., 1995; Lashkari et al., 1997; Pollack et al., 1999, Buck and Lieb, 2004; Mei et al., 2000; Hehir-Kwa et al., 2007). The multiple testing error rate associated with thousands even millions hypotheses testing need to be taken into account. A popular multiple testing error rate being controlled in many multiple hypotheses testing procedures is the family-wise error rate (FWER) (Hochberg and Tamhane, 1987; Shaffer, 1995), which is defined as the probability of at least one false rejection. Another less stringent multiple testing error rate commonly used is the false discovery rate (FDR) (Benjamini and Hochberg, 1995) defined as the expected proportion of falsely rejected null hypotheses.

Resampling-based multiple testing procedures are widely used in microarray studies especially when the sample size is small or the distribution of test statistic is non-normally distributed or unknown. Meanwhile, resampling-based multiple testing procedures can also take the dependence structure among p-values or test statistics into account. The commonly used resampling techniques include permutation tests and bootstrap methods. A permutation test is a type of non-parametric statistical significance test in which a reference distribution is constructed by calculating all possible values of test statistics from permuted observations under a null hypothesis. The theory of permutation tests is based on the works of Fisher and Pitman in the 1930s (Good, 2005).

Permutation test is distribution-free and can provide exact p-values when the sample size is small. The bootstrap method was first introduced by Efron (1979) and further discussed by Efron and Tibshirani (1994) which is a way of approximating the sampling distribution from just one sample. Instead of taking many simple random samples from the population to find the sampling distribution of a sample statistic, the bootstrap method repeatedly samples with replacement from one random sample. Efron (1979) showed that the bootstrap method can (asymptotically) correctly estimate the variance of a sample median and the error rates in a linear discrimination problem (outperforming cross-validation). Freedman (1981) showed that the bootstrap approximation to the distribution of least square estimates is valid. Hall (1986) further showed the reduction of error coverage probability by the bootstrap method, from O ($n^{-1/2}$) to O (n^{-1}), which makes the bootstrap method one order more accurate than the delta method.

In this article, we focus on investigating the power and stability properties of several commonly used resampling-based multiple testing procedures: 1) the permutation test (Westfall and Young, 1993); 2) the permutation-based significant analysis of microarray (SAM) procedure (Tusher et al., 2001); 3) the bootstrap multiple testing procedure (Pollard and van der Laan, 2005). The permutation test and SAM procedure were claimed to control either the FWER or FDR at 5%. The bootstrap multiple testing procedure proposed by Pollard and van der Laan was claimed to control either FWER or FDR asymptotically at 5%.

1.1 Permutation test (Westfall and Young, 1993)

To carry out a permutation test based on a test statistic that measures the size of an effect of interest, we proceed as follows:

1. Compute the test statistics for the observed data set such as two sample t-tests.
2. Permute the original data in a way that matches the null hypothesis to get permuted resamples, and construct the reference distribution using the test statistics calculated from permuted resamples.
3. Calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw *p*-value by computing the proportion of permutation test statistics that are as extreme as or more extreme than the observed test statistic.

Westfall and Young (1993) proposed two methods to adjust the raw *p*-values to control the multiple testing error rates: one is single-step minP procedure and the other is single-step maxT procedure.

The single-step minP adjusted *p*-values are defined by (Ge et al., 2003)

$$\tilde{p}_i = \Pr(\min_{1\le l\le m} P_l \le p_i | H_M),$$

The single-step maxT adjusted *p*-values are defined in terms of the test statistics T_i themselves, namely (Ge et al., 2003)

$$\tilde{p}_i = \Pr\left(\max_{1\le l\le m} |T_l| \ge |t_i| \middle| H_M\right).$$

where H_M denotes the complete null hypothesis and P_l denotes the random variable for the raw *p*-value of the *l*th hypothesis.

1.2 Significant Analysis of Microarray (SAM) procedure (Tusher et al., 2001)

The Significance Analysis of Microarrays (SAM) was first introduced by Tusher et al. (2001) for identifying genes with statistically significant changes in expression by assimilating a set of gene-specific *t* tests. In SAM, each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Then, a scatter plot of the observed relative difference versus the expected relative difference estimated by permutation is used to select statistically significant genes based on a fixed threshold.

The SAM procedure can be summarized as follows based on the description of SAM in Tusher et al. (2001).

1. Compute a test statistic t_i for each gene i $(i = 1, \cdots, g)$.
2. Compute order statistics $t_{(i)}$ such that $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(g)}$.
3. Perform B permutations of the responses/covariates $y_1, \ldots, y_n$. For each permutation b, compute the permuted test statistics $t_{i,b}$ and the corresponding order statistics $t_{(1),b} \leq t_{(2),b} \leq \cdots \leq t_{(g),b}$.
4. From the B permutations, estimate the expected values of order statistics by $\bar{t}_{(i)} = (\frac{1}{B}) \sum_{b=1}^{B} t_{(i),b}$.
5. Form a quantile-quantile (Q-Q) plot (SAM plot) of the observed $t_{(i)}$ versus the expected $\bar{t}_{(i)}$.
6. For a given threshold Δ, starting at the origin, and moving up to find the first $i = i_1$ such that $t_{(i)} - \bar{t}_{(i)} > \Delta$. All genes past i1 are called significant positive. Similarly, starting at the origin, moving down to the left and find the first $i = i_2$ such that $\bar{t}_{(i)} - t_{(i)} > \Delta$. All genes past i_2 are called significant negative. Define the upper cut point $Cut_{up}(\Delta) = \min\{t_{(i)}: i \leq i_1\} = t_{(i1)}$, and the lower cut point $Cut_{low}(\Delta) = \max\{t_{(i)}: i \geq i_2\} = t_{(i1)}$.
7. For a given threshold, the expected number of false rejections E(V) is estimated by computing the number of genes with ti,b above $Cut_{up}(\Delta)$ or below $Cut_{low}(\Delta)$ for each of the B permutations, and averaging the numbers over B permutations.
8. A threshold Δ is chosen to control the Fdr $(Fdr = E(V)/r$ under the complete null hypothesis, at an acceptable nominal level.

1.3 Bootstrap method (Pollard and van der Laan, 2005)

The bootstrap method based on estimated null distribution of test statistics introduced by Pollard and van der Laan (2005) proceeds as follows:

1. Compute the test statistic for the observed data set.

2. Resample the data with replacement within each group to obtain bootstrap resamples, compute the test statistic for each resampled data set, and construct the reference distribution using the centered and/or scaled resampled test statistics.

3. Calculate the critical value of a level α test based on the upper α percentile of the reference distribution, or obtain the raw P-value by computing the proportion of bootstrapped test statistics that are as extreme as or more extreme than the observed test statistic.

The MTP function based on the bootstrap method includes both single-step minP and maxT adjusted P-values as well as step-down minP and step-down maxT adjusted P-values. The single-step maxT and minP adjusted P-values are defined as before.

The step-down minP adjusted P-values are defined by

$$\tilde{\mathrm{p}}_{\mathrm{r_i}} = \max_{\mathrm{k}=1,\dots,\mathrm{i}}\{\,\Pr\left(\min_{\mathrm{l}=\mathrm{k},\dots,\mathrm{m}} \mathrm{P}_{\mathrm{r_l}} \le \mathrm{p}_{\mathrm{r_k}}\middle|\mathrm{H_M}\right)\},$$

The step-down maxT adjusted P-values are defined by

$$\tilde{\mathrm{p}}_{\mathrm{S_i}} = \max_{\mathrm{k}=1,\dots,\mathrm{i}}\{\,\Pr\left(\min_{\mathrm{l}=\mathrm{k},\dots,\mathrm{m}} |\mathrm{T}_{\mathrm{s_l}}| \le |\mathrm{t}_{\mathrm{s_k}}|\middle|\mathrm{H_M}\right)\},$$

where $|t_{s_1}| \ge |t_{s_2}| \ge \cdots \ge |t_{s_m}|$ denote the ordered test statistics (Ge et al., 2003).

2. Simulation Studies

2.1 Simulation setup

Simulation studies were conducted to compare the power and stability of the resampling-based multiple testing procedures. According to Rubin et al. (2006), the power is defined as the expected number of true positives. The stability is measured as the variance of true discoveries and variance of total discoveries.

In our simulation studies, each set includes 100 independently generated two groups of samples with equal sample size of 3 and 12 in each group. The total number of genes (m) is set to be 2000 with the fraction of true null hypotheses (m_0/m) at 50%. In the two-group comparisons, the standardized logarithms of gene expression levels are generated from multivariate normal distribution. One group has 50% of genes with means at μ and the remaining with means at 0. All genes in the other group have means at 0. The mean expression level μ on log2 scale is set to be from 1 to 6 with step 0.50 for one set of simulation study. The variances of the standardized logarithm of gene expression levels are equal to 1 in both groups. Thus the mean differences of μ in gene expression between the two groups are also the Cohen’s d effect sizes. The pairwise correlation coefficients of test statistics are assumed to be 0 in the simulation study. The test statistics used are equal variance *t*-test throughout the simulation study. The FWER/FDR level is set to 5% (α = 0.05).

The mt.maxT and mt.minP function in R were used to evaluate the Westfall and Young permutation test. The sam function in R was used for the SAM procedure. The Bootstrap method proposed by Pollard and van der Laan (2005) was executed by the MTP function in R. The MTP function includes both maxT and minP methods and both single step and step-down procedures, which results in four different functions: single-step maxT, single-step minP, step-down maxT, and step-down minP.

2.2 Simulation Results

Based on the simulation results shown in Figure 1, the bootstrap minP procedures outperformed all other procedures in our study. When the FWER is controlled at 5% and sample size is as small as 3 in each group, bootstrap single-step and step-down minP procedures were the least conservative and most powerful resampling-based multiple testing procedures among all the procedures compared in this study. The two permutation-based maxT procedure and minP procedure had no power to detect any significant differences with powers around 0 although their FWER are also controlled at very low level of around 0. Both the conservativeness and power of the two boostrap single-step and step-down maxT procedures were between the permutation procedures and the bootstrap minP procedures. All the resampling-based multiple testing procedures showed high stability with estimated variance of true discoveries and total number of discoveries around zero. Meanwhile, the estimated FWER, power, and stability were constant across different effect sizes.

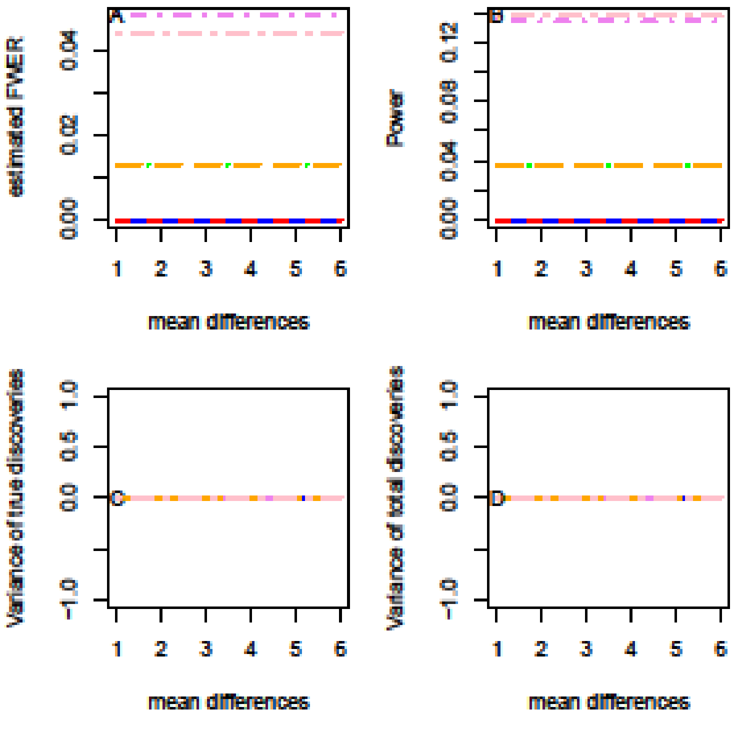

Figure 1: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FWER controlled at 5% and small sample size of 3 in each group ($m_0 =$ 50%). Solid blue line: Permutation single-step maxT procedure (mt.maxT function); Red dashed line: Permutation single-step minP (mt.minP function); Green dotted line: Bootstrap single-step maxT (MTP ss.maxT function); Violet dashed

line: Bootstrap single-step minP (MTP ss.minP function); Orange dashed line: Bootstrap step-down maxT (MTP sd.maxT function); Pink dashed line: Bootstrap step-down minP (MTP sd.minP function).

Figure 2 indicated the properties of those resampling procedures when the FDR is controlled at 5% and sample size is as small as 3 in each group. Bootstrap single-step and step-down minP procedures were the least conservative and most powerful resampling-based multiple testing procedures as that in the FWER controlling situation. The SAM procedure and the two permutation-based maxT and minP procedures had no power to detect any significant differences with powers around 0 although their FDR are also controlled at very low level of around 0. Both the conservativeness and power of the two boostrap single-step and step-down maxT procedures were in the middle of the SAM, permutation procedures and the bootstrap minP procedures. All the resampling-based multiple testing procedures showed high stability with 0 estimated variance of true discoveries and total number of discoveries. Meanwhile, the estimated FDR, power, and stability were constant across different effect sizes.

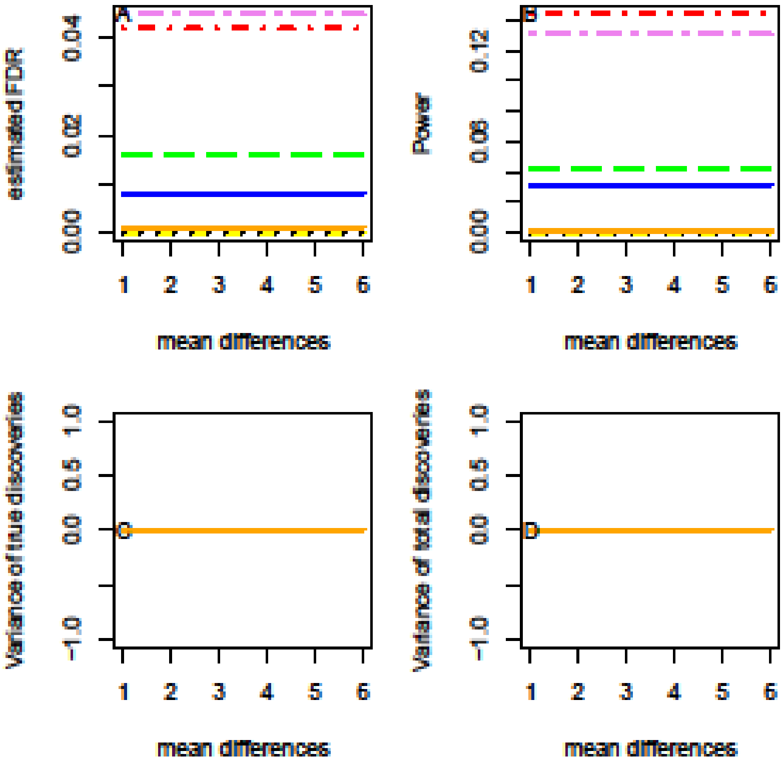

Figure 2: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FDR controlled at 5% and small sample size of 3 in each group (m_0= 50%). Yellow dashed line: Permutation single-step maxT procedure (mt.maxT function); Black dashed line: Permutation single-step minP (mt.minP function); Solid blue line: Bootstrap single-step maxT (MTP ss.maxT function); Red dashed line: Bootstrap single-step minP (MTP ss.minP function); Green dashed line: Bootstrap step-down maxT (MTP sd.maxT function); Violet dashed line: Bootstrap step-down minP (MTP sd.minP function); Solid orange line: SAM procedure (sam function).

When the sample size is moderate such as 12 in each group, both permutation and boostrap maxT and minP procedures can control the FWER at levels around 0 shown in Figure 3. Bootstrap step-down minP procedures had the highest power followed by bootstrap single-step minP procedure. Permutation single-step maxT has higher power than bootstrap maxT and permutation single-step minP procedures. All the resampling-based multiple testing procedures showed high stability with very small estimated variance of true discoveries and total number of discoveries. Meanwhile, the estimated FWER and power were constant across different effect sizes.

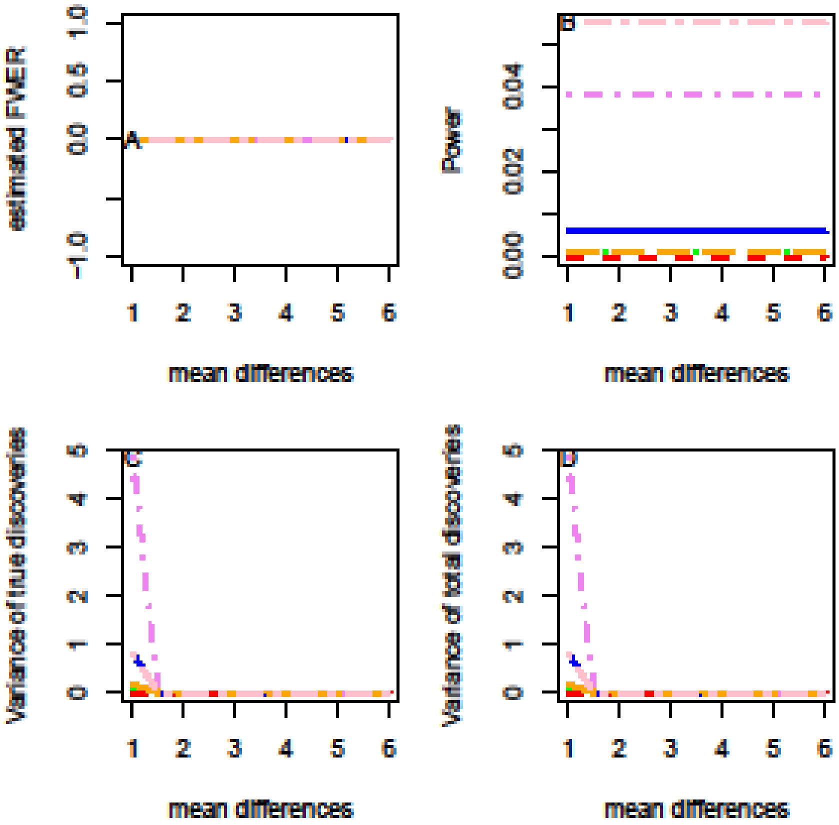

Figure 3: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FWER controlled at 5% and moderate sample size of 12 in each group (m_0 = 50%). Solid blue line: Permutation single-step maxT procedure (mt.maxT function); Red dashed line: Permutation single-step minP (mt.minP function); Green dotted line: Bootstrap single-step maxT (MTP ss.maxT function); Violet dashed line: Bootstrap single-step minP (MTP ss.minP function); Orange dashed line: Bootstrap step-down maxT (MTP sd.maxT function); Pink dashed line: Bootstrap step-down minP (MTP sd.minP function).

When sample size increases to 12 in each group and FDR is controlled at 5%, it was shown in Figure 4 that the estimated FDR from the SAM procedure was higher than 5% although it also had the highest power than the other procedures.

Meanwhile, the estimated variance of total discoveries from the SAM procedure was much larger than all the other procedures when the effect size is around 1. Permutation single-step maxT and minP had the second largest power with small estimated FDR and variance of true discoveries and total discoveries. The four bootstrap MTP procedures had low power and similar stability as the permutation maxT and minP procedures. Figure 4 also showed the bootstrap single-step and step-down minP has slightly higher power than the bootstrap single-step and step-down maxT method.

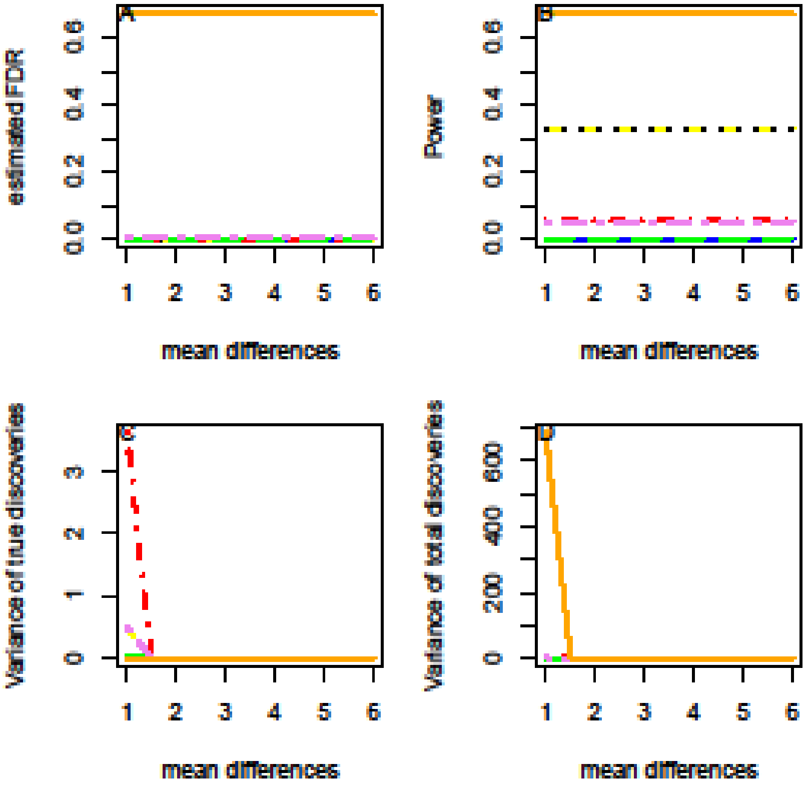

Figure 4: Power and stability properties of resampling-based multiple testing procedures for independent test statistics with FDR controlled at 5% and small sample size of 12 in each group (m_0 = 50%). Yellow dashed line: Permutation single-step maxT procedure (mt.maxT function); Black dashed line: Permutation single-step minP (mt.minP function); Solid blue line: Bootstrap single-step maxT (MTP ss.maxT function); Red dashed line: Bootstrap single-step minP (MTP ss.minP function); Green dashed line: Bootstrap step-down maxT (MTP sd.maxT function); Violet dashed line: Bootstrap step-down minP (MTP sd.minP function); Solid orange line: SAM procedure (sam function).

2.3 Cancer microarray example

Ovarian cancer is the fifth leading cause of all cancer death of women in the United States (Jemal et al., 2010). Microarray experiments were conducted to identify differentially expressed genes between chemotherapy favorable patients and chemotherapy unfavorable patients (Moreno et al., 2007). Those differentially expressed genes could be used to predict the possible responses of new ovarian cancer patients to chemotherapy to develop optimal treatment for each patient. The gene expression data from 12625 genes and 43 patients' mRNA samples obtained from Moreno et al.’s ovarian cancer microarray study were used to show the differences in the number of total discoveries) among those resampling-based multiple testing procedures with FWER or FDR controlled at 5%. The pre-processing of the ovarian cancer data set was done using the RMA background correction, quantile normalization, and robust linear model summarization. The raw *p*-values and adjusted *p*-values of comparisons between the chemotherapy favorable group (20 subjects) and chemotherapy unfavorable group (23 subjects) were calculated using the resampling-based multiple testing functions in the multtest package and siggenes package in R.

The gene expression levels of 12625 genes from 43 subjects on log2 scale were used to compare total number of discoveries of the resampling-based multiple testing procedures (Table 1). When the FWER is controlled at 5%, the two bootstrap minP procedures had more rejections than the permutation maxT procedure and the bootstrap maxT procedure. The permutation minP procedure rejected none of the null hypothesis. The single-step procedures had slightly more rejections than the step-down procedures. When the FDR is controlled at 5%, SAM rejected the largest number of null hypotheses, which is slightly higher than the two permutation-based resampling procedures. Bootstrap multiple testing procedures rejected much less null hypotheses compared to the permutation test procedures. The minP procedures rejected more compared to the maxT procedures within the bootstrap methods and the single-step procedures had slightly higher rejections than the step-down procedures.

3. Discussion

This paper investigates the power and stability properties of several popular resampling-based multiple testing procedures under independent case for small and moderate sample size data using available functions in R. Our simulation and real data example show that the bootstrap single-step and step-down minP procedures perform the best for both small sample size data (3 in each group) and moderate sample size data (12 in each group) when FWER control is required; the bootstrap single-step and step-down minP procedures are also the best when FDR control is desired for small sample size data (3 in each group); the permutation maxT and minP procedures performs the best for moderate sample size data when FDR control is required. The SAM procedure overestimates FDR although it has the higher power than the permutation and bootstrap maxT and minP procedures.

Simulation results also show the power of permutation maxT and minP procedures is 0 when the sample size is as small as 3 in each group. The results are predictable through the way that the permutation test procedures were conducted. The estimated null distribution of test statistics from the permutation maxT or minP procedure only has limited values when the sample size is very small. For bootstrap MTP procedures, it is interesting to notice the minP procedures always perform slightly better than the maxT procedure.

Our current investigation only focuses on the independent cases. To examine the effect of different correlation structures on the power and stability properties for those resampling-based multiple testing procedures will be an interesting topic to investigate for future studies. Meanwhile, further investigations are needed for extending our simulation distribution from multivariate normal to other distributions such as lognormal and binomial distribution and examining the power and stability properties of those resampling-based multiple testing procedures under the non-normal distribution situations.

Table 1: Comparisons of number of total discoveries for the resampling-based multiple testing procedures for the ovarian cancer example

Resampling Methods	**Rejected number of hypothesis**	
	FWER controlled at 5%	**FDR controlled at 5%**
Permutation mt.maxT	**1739**	**5336**
Permutation mt.minP	**0**	**5336**
Boostrap MTP (ss.maxT)	**1050**	**641**
Bootstrap MTP (sd.maxT)	**1029**	**673**
Bootstrap MTP (ss.minP)	**2782**	**2933**
Bootstrap MTP (sd.minP)	**2734**	**2826**
SAM (sam)		**5503**

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*. Series B (Methodological), 57(1):289 – 300.

Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83:349 – 360.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218 – 1228.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer, 3rd edition.

Hehir-Kwa, J., Egmont-Petersen, M., Janssen, I., Smeets, D., Geurts van Kessel, A., and Veltman, J. (2007). Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: A platform comparison based on statistical power analysis. *DNA Research*, 14:1 – 11.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, 14(4):1431 – 1452.

Jemal, A., Siegel, R., Xu, J., and Ward, E. (2010). Cancer statistics. *CA Cancer J Clin*, 56:106–130.

Kulesh, D. A., Clive, D. R., Zarlenga, D. S., and Greene, J. J. (1987). Identification of interferon-modulated proliferation-related cDNA sequences. Proceedings of the National Academy of Sciences, 84:8453 – 8457.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences*, 94:13057 – 13062.

Mei, R., Galipeau, P. C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R. K., Chee, M. S., Reid, B. J., and Lockhart, D. J. (2000). Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Research*, 10:1126 – 1137.

Moreno, C. S., Matyunina, L., Dickerson, E. B., Schubert, N., Bowen, N. J., Logani, S., Benigno, B. B., and McDonald, J. F. (2007). Evidence that p53-mediated cell-cyclearrest inhibits chemotherapeutic treatment of ovarian carcinomas. *PLoS One*, 2(5):e441.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., S., J. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23:41 – 46.

Pollard, K. S. and van der Laan, M. J. (2005). Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data. *Journal of Statistical Planning and Inference*, 125:85 – 100.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significant analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116 – 5121.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: examples and methods for P-Value adjustment*. New York: Wiley.