

Conservative adjustment of permutation p -values when the number of permutations is limited

Yinglei Lai

ylai@gwu.edu

Department of Statistics and Biostatistics Center,
The George Washington University,
2140 Pennsylvania Avenue, N.W., Washington D.C. 20052, U.S.A.

Abstract

The permutation procedure is widely used to assess the significance level (p -value) of a test statistic. This approach is asymptotically consistent. In genomics and proteomics studies, p -values are required to be evaluated at a “tiny” level. However, due to small sample sizes or limited computer resources, only a limited number of permutations can be obtained. Therefore, it is necessary to understand the accuracy of these permutation p -values.

In this study, we show through the theory of order statistics that a considerable proportion of p -values will be under-evaluated by the permutation procedure. To solve this problem, we propose to conservatively adjust permutation p -values. The adjustment requires no parametric assumption on the distribution of test statistic. The solution can be expressed by a normalized incomplete beta function. The related normal distribution approximation is also discussed. Simulations are conducted to illustrate the proposed method and two microarray gene expression data sets are considered for applications.

Keywords: p -value; permutation; microarray.

1 Introduction

Microarrays (Golub et al., 1999) and mass spectra (Adam et al., 2002) are widely used for biological and medical studies. These technologies can simultaneously measure a large number of variables (genes, m/z ratios). However, the sample sizes of these data are generally small. One important application of these high-throughput technologies is to identify variables that can significantly distinguish different sample groups, such as normal against disease groups. After calculating the relevant test statistics, a crucial issue is to evaluate their significance levels (p -values). Because the distributions of test statistics are usually

unknown (especially when the sample sizes are small), the permutation procedure (permuting sample labels and re-calculating test statistics) is widely used to assess the p -values of calculated test statistics (Dudoit et al., 2003). This approach is asymptotically consistent when the number of permutations goes to infinity. Since a large number of variables are screened simultaneously, their p -values must be evaluated at a “tiny” level (e.g. 10^{-6}) so that the issue of multiple hypothesis testing can be addressed (Benjamini and Hochberg, 1995). However, in general, only a limited number of permutations can be obtained because of small sample sizes or limited computer resources. For example, to study diabetes, Herman et al. collected expression profiles for 12488 genes but the group sample size is only 3 or 5 (GEO accession number GSE1419). For this data set, there are only 56 different permutations for each gene. Even when sample sizes are relatively large (e.g. > 10), it is generally difficult for most computers to handle a huge number (e.g. $> 10^8$) of permuted test values. Therefore, it is necessary to understand the accuracy of these p -values when the number of permutations is limited.

Berger (2000) discussed some advantages and disadvantages of permutation tests. Recently, Klebanov et al. (2006) briefly discussed the number of permutations required when the confidence level and interval of a p -value were specified. However, when the number of permutations is limited, it is necessary to study how to avoid the under-evaluation of p -values, which may lead to false positives.

The theory of permutation p -value is closely related to the theory of order statistics. In this study, we first discuss this relationship and then study the conservative property of permutation p -values. To reduce the likelihood of under-evaluation of p -values, we propose to conservatively adjust permutation p -values. The adjustment requires no parametric assumption

on the distribution of test statistic. The solution can be expressed by a normalized incomplete beta function. The related normal distribution approximation is also discussed. Simulations are conducted to illustrate the proposed method and two microarray gene expression data sets are considered for applications.

2 Methods

2.1 Permutation Test

Without loss of generality, we briefly describe the permutation procedure for two-sample comparison. Suppose measurements are collected for a variable from two populations. A sample group label is assigned to each measurement. The absolute value of Student's t -test is used to test the hypotheses H_0 : two population means are equal. *v.s.* H_1 : two population means are different. If the variable is normally distributed and two population variances are equal, then the theoretical t -distribution can be used to give the p -value. If the distribution of the variable is not certain, then the permutation procedure can be used to evaluate the p -value. For each permutation, the group labels are randomly reassigned and the t -test is recalculated. After r permutations, we obtain r permuted t -test values $\{T_k : k = 1, 2, \dots, r\}$. These r values are sorted in increasing order: $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(r)}$. Conventionally, we assign $(r - k + 1)/r$ as the permutation p -value of $T_{(k)}$, [or equivalently, $T_{(k)}$ is the $(k - 1)/r$ permutation quantile,] $k = 1, 2, \dots, r$. For example, let t be a calculated test statistic. We run 19 random permutations. Since the original test value t is also a result of a certain permutation, there are 20 permuted test values. After ordering, if t is on the 19th position, then its permutation p -value will be 0.1.

Remark: We may have different definitions for the permutation p -value of $T_{(k)}$. It can also be defined as $(r - k + 1)/(r + 1)$ [or equivalently, $T_{(k)}$ is the $k/(r + 1)$ permutation quantile]. Similar results will be obtained for this definition if the theoretical framework in this study is followed.

If there are multiple variables in the data set, then the permuted test values from different variables can be pooled and the p -values can be evaluated based on this pool. This approach is typically used in microarray and mass spectra data analyses.

The number of permutations is limited when sample sizes are relatively small. For two-sample data, this number can be generally calculated as:

$$r = \binom{n_1 + n_2}{n_1} = \frac{(n_1 + n_2)!}{n_1!n_2!},$$

where n_1 and n_2 are the sample sizes in groups 1 and 2, respectively. Notice that some test statistics are symmetric about group labels when two group sample sizes are equal, i.e. the same test value will be obtained if group labels are exchanged. One example of these test statistics is the absolute value of Student's t -test. For these test statistics, the number of different permutations will be

$$r = \begin{cases} \frac{(n_1 + n_2)!}{2n_1!n_2!} & n_1 = n_2; \\ \frac{(n_1 + n_2)!}{n_1!n_2!} & n_1 \neq n_2. \end{cases}$$

2.2 Order Statistics

Let $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(r)}$ be the order statistics of r independently identically distributed (*i.i.d.*) random variables T_1, T_2, \dots, T_r and $F(t)$ be the cumulative probability distribution (*c.d.f.*) of these random variables. Without loss of generality, let $F(t)$ be continuous. Based on the theory of ordered statistics (Balakrishnan and Cohen, 1991, Page 13), the *c.d.f.* of $T_{(k)}$ is

$$\begin{aligned} F_{(k)}(t) &= \mathbf{P}\{T_{(k)} \leq t\} \\ &= \sum_{i=k}^r \binom{r}{i} [F(t)]^i [1 - F(t)]^{r-i} \\ &= \mathbf{P}\{B[r, F(t)] \geq k\}, \end{aligned} \tag{1}$$

where $B[r, F(t)]$ is a binomial random variable with number of trials r and probability of success $F(t)$. It is well known that the above probability is the cumulative probability at $F(t)$ of the beta distribution with parameters k and $r - k + 1$ (also called the normalized incomplete beta function): $I_{F(t)}(k, r - k + 1)$.

2.3 Conservative Level of a Permutation Quantile

According to the above discussion, $T_{(k)}$ is the $(k - 1)/r$ permutation quantile. Here, we define the conservative level of a quantile estimator \hat{q} as

$$\gamma = \mathbf{P}\{\hat{q} \in [q, \infty)\},$$

where q is the quantile to be estimated by \hat{q} . Let q be the $p = (k - 1)/r$ quantile of the test statistic [$F(q) = p$]. Based on Equation (1), the conservative level of $T_{(k)}$ is:

$$\begin{aligned} \gamma &= \mathbf{P}\{T_{(k)} \in [q, \infty)\} \\ &= \mathbf{P}\{T_{(k)} \geq q\} \end{aligned}$$

$$\begin{aligned}
&= 1 - \mathbf{P}[T_{(k)} < q] \\
&= 1 - \lim_{a \nearrow p} \mathbf{P}[T_{(k)} \leq F^{-1}(a)] \\
&= 1 - \lim_{a \nearrow p} I_a(k, r - k + 1).
\end{aligned}$$

Based on the normal distribution approximation, we will show later that this number is never close to one even when r is large but finite.

2.4 Conservatively Adjusted Permutation p -value

The above discussion shows that the conservative level of a permutation quantile is not satisfactory. Therefore, a considerable proportion of p -values will be under-evaluated by the permutation procedure. To obtain more conservatively evaluated p -values, we suggest the following adjustment. Instead of using $T_{(k)}$ to estimate the $p = (k - 1)/r$ quantile q , we propose to use it to estimate a smaller quantile b . A $100(1 - \alpha)\%$ conservatively adjusted permutation quantile of $T_{(k)}$ is a number b such that

$$\mathbf{P}\{T_{(k)} \in [b, \infty)\} = 1 - \alpha, \quad (2)$$

with $\alpha \in (0, 1)$. The number $1 - a = 1 - F(b)$ is called the $100(1 - \alpha)\%$ conservatively adjusted permutation p -value of $T_{(k)}$. Based on Equation (1), we have

$$I_a(k, r - k + 1) = \alpha.$$

Therefore, a is the α quantile of the beta distribution with parameters k and $r - k + 1$. The solution of a can be easily implemented by current statistical software, such as R.

2.5 Normal Approximation

When $k = r$, it is straightforward to have

$$\gamma = 1 - p^r = 1 - (1 - 1/r)^r \approx 1 - e^{-1} = 0.632.$$

Notice that $p = (k - 1)/r$ and $r \gg 1$. We can also obtain a similar result for $k = r - 1$

$$\begin{aligned}
\gamma &= 1 - p^r - rp^{r-1}(1 - p) \\
&= 1 - (1 - 2/r)^r(3r - 2)/(r - 2) \\
&\approx 1 - 3e^{-2} = 0.594.
\end{aligned}$$

For other k , it will be difficult to simplify γ . It is well known that the standard normal distribution can be used to approximate the *c.d.f.* of a binomial random variable $B(n, p)$:

$$\mathbf{P}[B(n, p) \leq x] \approx \Phi\left[\frac{(x + 0.5) - np}{\sqrt{np(1 - p)}}\right], \quad (3)$$

for an integer x , $0 \leq x \leq n$. $\Phi(\cdot)$ is the *c.d.f.* of the standard normal distribution. A rule of thumb for this approximation is that $0 < np \pm 3\sqrt{np(1 - p)} < n$ [or equivalently, $9/(n + 9) < p < n/(n + 9)$].

Based on Equation (1), the conservative level of $T_{(k)}$ can be approximated as:

$$\begin{aligned}
\gamma &= 1 - \lim_{a \nearrow p} \mathbf{P}[T_{(k)} \leq F^{-1}(a)] \\
&= \lim_{a \nearrow p} \{1 - \mathbf{P}[B(r, a) \geq k]\} \\
&= \lim_{a \nearrow p} \mathbf{P}[B(r, a) \leq k - 1] \\
&= \lim_{a \nearrow p} \mathbf{P}[B(r, a) \leq rp] \\
&\approx \lim_{a \nearrow p} \Phi\left[\frac{(rp + 0.5) - ra}{\sqrt{ra(1 - a)}}\right].
\end{aligned}$$

Notice that $p = (k - 1)/r$. When r is finite, we have

$$\gamma \approx \Phi\left[\frac{0.5}{\sqrt{rp(1 - p)}}\right]$$

$$\begin{cases} \leq \Phi\left[\frac{0.5}{\sqrt{1 - 1/r}}\right] \approx \Phi(0.5) = 0.691 & (p = 1/r, 1 - 1/r); \\ \geq \Phi\left[\frac{0.5}{\sqrt{r/4}}\right] \approx \Phi(0) = 0.5 & (p = 1/2); \end{cases}$$

for $p = (k - 1)/r$, $k = 2, 3, \dots, r$. Notice that the corresponding permutation p -value is $1 - p$ and $p = 0$ ($k = 1$) is usually not considered in practice. However, asymptotically, we have

$$\gamma = \lim_{a \nearrow p} \lim_{r \rightarrow \infty} \Phi\left[\frac{p - a + 0.5/r}{\sqrt{a(1 - a)}/r}\right] = \lim_{a \nearrow p} \Phi(+\infty) = 1.$$

Therefore, permutation quantiles are only asymptotically fully conservative. Even when the number of permutations is large but finite, the conservative levels of permutation quantiles are not satisfactory.

Based on Equations (1) and (3), we can obtain an approximated solution for a , the $100(1 - \alpha)\%$ conservatively adjusted permutation p -value of $T_{(k)}$:

$$\frac{rp + 0.5 - ra}{\sqrt{ra(1 - a)}} \approx z_\alpha,$$

where z_α is the $1 - \alpha$ quantile of the standard normal distribution. Solving this quadratic equation with restriction $a < p$ (Notice that $b < q$), we have:

$$\begin{aligned}
a & \approx \{r[2rp + z_\alpha^2 + 1] \\
& \quad - \sqrt{r^2[2rp + z_\alpha^2 + 1]^2 - 4(r^2 + rz_\alpha^2)(rp + 0.5)^2}\} \\
& \quad / [2(r^2 + rz_\alpha^2)]
\end{aligned} \quad (4)$$

When $r \rightarrow \infty$, we have $a \rightarrow p$ for any $0 < \alpha < 1$.

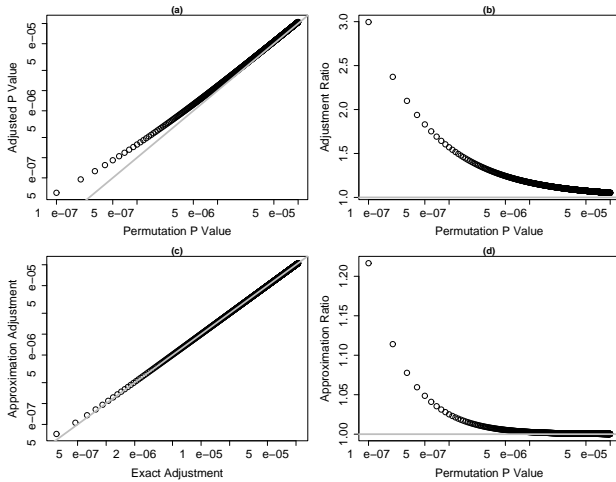


Figure 1: (a): The comparison of the 95% conservatively adjusted and the original permutation p -values. (b): The ratios between the 95% conservatively adjusted and the original permutation p -values. (c): The comparison of the 95% conservatively adjusted p -values from the normal approximation and the exact calculation. (d) The ratios between the 95% conservatively adjusted p -values from the normal approximation and the exact calculation.

3 Results

3.1 Adjustment Effects

We present two plots to illustrate the adjustment effects. Assume that there are $r = 10^7$ permuted test values and top 10^3 of them are of interest. The range of permutation p -values is $[10^{-4}, 10^{-7}]$. Figure 1(a) compares the 95% conservatively adjusted and the original permutation p -values and Figure 1(b) shows their ratios. The adjustment effects are different for different levels of p -values. The adjustment ratio increases as p -value decreases. In the figure, the ratio is about 3 when the permutation p -value is 10^{-7} and is close to one when the permutation p -value is 10^{-4} .

Two additional plots are presented to illustrate the normal approximation. Figure 1(c) compares the 95% conservatively adjusted p -values from the normal approximation and the exact calculation and Figure 1(d) shows their ratios. Overall, the approximation is close. The ratio is about 1.2 when the permutation p -value is 10^{-7} . The ratio is less than 1.05 when the permutation p -value is greater than 5×10^{-7} .

3.2 A Simulation Study

Because of randomness, with probability one, $T_{(k)}$ is different from the $(k-1)/r$ quantile, $k = 1, 2, \dots, r$,

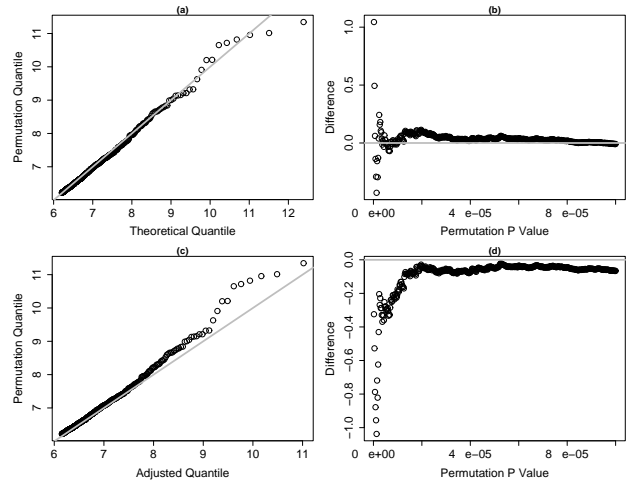


Figure 2: (a): The comparison of the permutation and the theoretical quantiles. (b): The differences between the theoretical and the permutation quantiles. (c): The comparison of the original and the 95% conservatively adjusted permutation quantiles. (d) The differences between the 95% conservatively adjusted and the original permutation quantiles.

when r is finite. To understand these deviations, we simulated expression measurements for 10000 genes and two groups with 6 samples in each group. Without loss of generality, the gene expressions in the first group were simulated from the standard normal distribution $N(0, 1)$. For the second group, expressions of 7000 genes were simulated from $N(0, 1)$ (70% non-differentially expressed) and expressions of the rest 3000 genes were simulated from $N(1, 1)$ (30% differentially expressed). We performed a complete permutation procedure (462 permutations) for each gene and pooled $r = 4620000$ permuted absolute t -test values to evaluate quantiles. Figure 2(a) compares the top 462 test values (permutation quantiles) $\{T_{(4620000-462+k)}, k = 1, 2, \dots, 462\}$ against the corresponding quantiles from the theoretical t -distribution and Figure 2(b) shows the differences between the theoretical and the permutation quantiles. A considerable proportion of quantiles are underestimated (positive differences in Figure 2.b). Therefore, their corresponding p -values are undervalued. Figure 2(c) compares the original against the 95% conservatively adjusted permutation quantiles and Figure 2(d) shows the differences between the 95% conservatively adjusted and the original permutation quantiles. These quantiles are conservatively adjusted (negative differences in Figure 2.d).

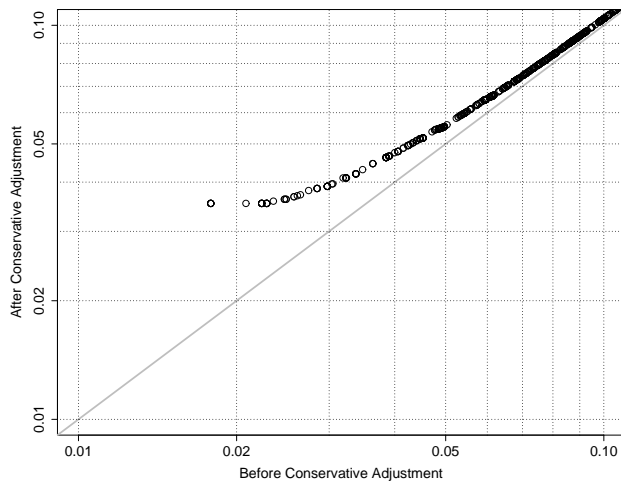


Figure 3: The comparison of the false discovery rates before and after the 99% conservative adjustments for the permutation p -values based on the T-cell microarray gene expression data set.

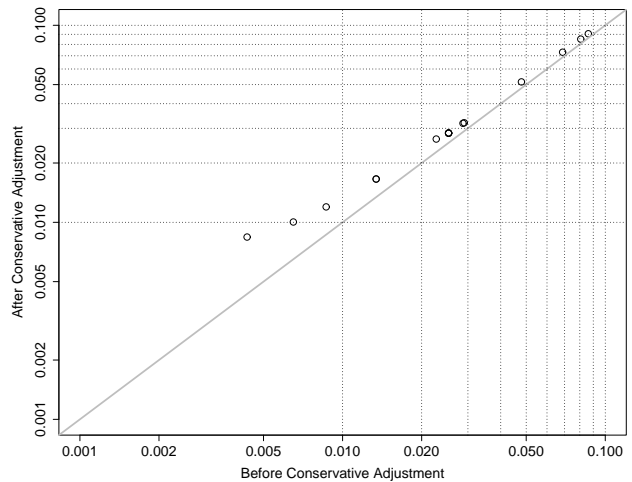


Figure 4: The comparison of the false discovery rates before and after the 90% conservative adjustments for the permutation p -values based on the microarray gene expression data set for a smoke study.

3.3 Applications

We apply the proposed method to two microarray gene expression data sets, which are publicly available in the database of Gene Expression Omnibus (GEO) with accession numbers GSE1419 and GSE3320. The first data set was collected to compare the gene expression profiles between T regulatory and T effector cells. It contains 12488 gene expression profiles for 8 samples (3 regulatory and 5 effector). The second data set was collected to compare the gene expression profiles between normal non-smokers and normal smokers. It contains 22283 gene expression profiles for 11 samples (5 non-smokers and 6 smokers).

Many different test statistics have been proposed for identifying differentially expressed genes for microarray data. The purpose of this study is to conservatively adjust permutation p -values. The proposed adjustment requires no parametric assumption on the distribution of test statistic. Without loss of generality, we used the absolute value of Student's t -test to identify differentially expressed genes. For the both data sets, all possible permutations were considered. The permutation p -values were evaluated based on the pool of permuted test values from all different genes. The false discovery rate (FDR) control procedure proposed by Benjamini and Hochberg (1995) was used for multiple comparison adjustments (R function `p.adjust` with option `method="BH"`). Other multiple comparison adjustment methods, such as an FDR estimation procedure (R package `qvalue`) proposed by Storey and Tibshirani (2003), can also be considered. Since per-

mutation p -values are conservatively adjusted (to be larger), similar results will be obtained if we use other adjustment methods.

For the first data set, 56 permutations were performed for each of 12488 genes to generate 56×12488 permuted test values. Since there were 162 genes with FDRs (based on the original permutation p -values) less than 0.05, we chose a more stringent 99% conservative adjustment for permutation p -values. Figure 3 compares the FDRs based on the original permutation p -values and the FDRs based on the 99% conservatively adjusted permutation p -values. The FDRs before and after the conservative adjustment are clearly different. For example, there are 72 genes with FDRs < 0.05 before the adjustment but > 0.05 after the adjustment.

For the second data set, 462 permutations were performed for each of 22283 genes to generate 462×22283 permuted test values. Since there were only 13 genes with FDRs (based on the original permutation p -values) less than 0.05, we chose a less stringent 90% conservative adjustment for permutation p -values. Figure 4 compares the FDRs based on the original permutation p -values and the FDRs based on the 90% conservatively adjusted permutation p -values. There are still some differences between the FDRs before and after the conservative adjustment. For example, there are 2 genes with FDRs < 0.01 before the adjustment but > 0.01 after the adjustment and one gene with FDR < 0.05 before the adjustment but > 0.05 after the adjustment.

4 Discussion

In this study, we first discussed the relationship between the theory of permutation p -value and the theory of order statistics. Then, we studied the conservative property of permutation p -values. To reduce the likelihood of under-evaluation of p -values, we proposed to conservatively adjust permutation p -values. The adjustment requires no parametric assumption on the distribution of test statistic. The solution can be expressed by a normalized incomplete beta function. The related normal distribution approximation was also discussed. The method was illustrated through simulations and then applied to two microarray gene expression data sets.

Figures 1, 3 and 4 show that the adjustments of these “non-tiny” p -values are negligible. This is actually the case when the permutation procedure is used for traditional statistical analyses, in which data sets usually contain a small number of variables but a large number of samples. For these data, p -values are not required to be evaluated at a “tiny” level for significance. However, for genomics and proteomics data, p -values are required to be evaluated at a “tiny” level so that the issue of multiple hypothesis testing can be addressed. When sample sizes are relatively small, only a limited number of permutations can be obtained and “tiny” p -values may not be reliably evaluated. To reduce the likelihood of false positives, we proposed the conservative adjustment for permutation p -values.

We also analyzed the famous microarray data set for breast cancer study (Hedenfalk et al., 2001). This data set contains 3170 (after gene filtering) gene expression profiles for 7 BRCA1 and 8 BRCA2 samples. There are 6435 possible permutations for each gene. Because of this relatively large number of permutations, the effect of conservative adjustment is negligible. However, this complete permutation procedure requires more than 1GB memory when R is used for computations. Notice that the number of genes on a current microarray chip can be much higher (about 10k to 40k). For these microarray data, it will be difficult for most computers to perform all possible permutations if the sample size is greater than 5 in each group. In practice, we may have to limit the number of permutations for each gene (e.g. < 500) because of limited computer resources. Then, the conservative adjustment will not be negligible.

Figure 2 shows that the adjustments of adjacent permutation quantiles are dependent. It is necessary to pursue further studies with the consideration of dependence among order statistics so that more efficient adjustments for permutation p -values can be

achieved. Furthermore, p -values should also be adjusted in the situation of multiple hypothesis testing. Is there a more efficient way to incorporate both p -value adjustments?

Acknowledgement

This work was partially supported by a start-up fund from the George Washington University and a NIH grant DK-75004. The R codes are available at <http://home.gwu.edu/~ylai/research/permutation>.

References

- Adam, B.L., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., Cazares, L.H., Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z. and Wright, Jr. G.L. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, **62**, 3609-3614.
- Balakrishnan, N. and Cohen, A.C. (1991) Order statistics and inference. Academic Press, INC., San Diego, CA.
- 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*, **57**, 289-300.
- Berger, V.W. (2000) Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, **19**, 1319-1328.
- Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71-103.
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A. and Trent, J. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **344**, 539-548.
- Klebanov, L., Gordon, A., Xiao, Y., Land, H. and Yakovlev, A. (2006) A permutation test motivated by microarray data analysis. *Computational Statistics & Data Analysis*, **50**, 3619-3628.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences USA*, **100**, 9440-9445.