# A practical and powerful method to control the generalized family wise error rate in multiple testing

Jeffrey Miecznikowski [1,2,3,*], David Gold [1,2,3], Lori A. Shepherd [1,2,3], Song Liu [1,2,3]

[1] *Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000* [2] *New York State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY*
[3] *Department of Biostatistics, Roswell Park Cancer Institute, Buffalo, NY*

## 1. SUMMARY

In a multiple testing setting, the investigator is faced with choosing what error to control and what method to use in controlling that error. Considerations include the assumptions that are made for each method and if the assumptions are acceptable in that setting. Failure to acknowledge these considerations can lead to drastically misleading results. Our recent study showed that, in applications where reduced multiplicity is encountered, the widely used Benjamini and Hochberg's (BH) false discovery rate (FDR) analysis is less robust than approaches controlling the number of false positives. In this manuscript we assess the current methods to control the probability of committing a fixed number of false positives and provide a new method. We provide theoretical proof that our proposed approach, KBIN, is more powerful than alternative approaches. We also conduct simulations and real data studies to evaluate the proposed finding. We expect that the KBIN method has promising applications in biomarker settings where the goal is to choose a set of significant biomarkers from among a panel of potential putative biomarkers.

KEY WORDS: multiple testing, false discovery rate, false positives, hypothesis testing

*to whom correspondence should be addressed

*August 12, 2009*

## 2. INTRODUCTION

The early adjustments for multiple testing are attributed to Boole's inequality which provides the proof for the "Bonferroni" method to control error when testing $N$ dependent or independent hypothesis on a given set of data [1]. Incremental improvements to the Bonferroni method are provided in [2, 3]. Until the early 1990s, many of the multiple testing procedures described in [4, 5] were adequate given the nature of their application. However, with the advent of modern computing and high dimension datasets, researchers needed more powerful methods to determine statistical significance in large scale experiments. Thus, statistical multiple testing procedures (MTPs) began to be reexamined to handle applications in high dimensional datasets such as those obtained through high-throughput genomic technologies, e.g. microarray analysis for differentially expressed genes and mass spectrometry to discover significant peptides between two conditions. These new methods controlled more liberal error quantities than the traditional family wise error rate (FWER).

In general, statistical methodologies in MTPs provide trade offs such as increasing the power to detect subtle changes at the expense of increasing the number of false positives. Common methods of controlling errors in an MTP setting is to either control the expected proportion of falsely rejected hypothesis (FDR) or control against the probability of committing a fixed number of false positives ($k$-FWER). An important aspect to consider when choosing an MTP procedure and controlling method is the number of tests under consideration. In general, with thousands of tests, it is reasonable to consider controlling the false discovery rate (FDR) using methods described in [6, 7, 8]. However, our recent study found that in applications with less than 1000 tests, FDR is less reliable than alternative approaches controlling $k$-FWER [9].

By controlling $k$-FWER, the researcher has a reliable and easily interpretable method of reporting results in a multiple testing situation. Methods to control $k$-FWER are commonly done by making a slight adjustment on the Bonferroni method which was designed to control the standard family wise error rate ($k = 1$ in $k$-FWER). However, the available methods to control this quantity are often too conservative and lack the statistical power to detect real significance. A $k$-FWER control method achieving improved power is highly desired. In this manuscript, we propose a new method to control $k$-FWER, called KBIN, which shows improved reliability over FDR control methods. The theoretic framework of KBIN and its comparison with other $k$-FWER control methods are presented in this manuscript. We demonstrate that KBIN is more powerful than other $k$-FWER methods including the adjustment to the Bonferroni method and the Holm method to control $k$-FWER.

The remainder of this paper is organized as follows. In Section 3, we briefly review the notation of $k$-FWER control and described the commonly used approaches for $k$-FWER control. In Section 4 we describe a recently proposed approach, KBIN, to re-address the $k$-FWER control and prove that it is more powerful than alternative approaches. We demonstrate KBIN's performance in Sections 5 and 6 with extensive simulations and real data studies. We conclude the article in Section 7 with some discussion.

## 3. *k*-FWER CONTROL METHODS

The $k$-FWER error control is a generalized version of the family wise error rate (FWER). Control of FWER, is to control the probability that there are 1 or more false discoveries. Notationally, (according to [10]) $\alpha$ control of FWER can be expressed as,

$$\mathbb{P}(V \geq 1) \leq \alpha$$

or equivalently,

$$\mathbb{P}(V = 0) \geq 1 - \alpha$$

where $V$ is the number of false discoveries and $\alpha$ is usually small, e.g. 0.05. More simply this can be expressed as FWER $\leq \alpha$. In $k$-FWER the equation becomes,

$$\mathbb{P}(V \geq k) \leq \alpha$$

where $k$ and $\alpha$ are usually determined prior to the analysis. Similar to FWER, control of $k$-FWER can be expressed as $k$-FWER $\leq \alpha$. Note, this formulation for $k$-FWER can be slightly impractical since researchers often phrase their response in terms of "committing no more than X" errors which means $k$ is set to "$X + 1$". Thus, occasionally, as in [11] $k$-FWER may be expressed as $\mathbb{P}(V > k) \leq \alpha$.

The following subsections detail several commonly used methods designed to control the $k$-FWER.

### *3.1. APPROACH BASED ON ADJUSTED BONFERRONI METHOD*

The adjusted Bonferroni method to control $k$-FWER is a generalized version of the Bonferroni correction which is designed to control FWER [10]. The Bonferroni correction is designed to control the FWER at level $\alpha$ by doing each individual test at level $\alpha/N$ where $N$ is the number of tests. The adjustment given in [10] to control $k$-FWER at $\alpha$ is to perform each test at level $k\frac{\alpha}{N}$. By doing each test at level $k\frac{\alpha}{N}$, this controls the probability against $k$ or more false positives to be no larger than $\alpha$, that is,

$$\mathbb{P}(V \geq k) \leq \alpha$$

The proof is supplied in [10] and is a generalization of the proof for the original Bonferroni method designed to control FWER.

### *3.2. APPROACH BASED ON HOLM METHOD*

A further method to control $k$-FWER using the Holm procedure is given in [10]. This method is based on the Holm method designed to control the FWER [2]. The Holm method is considered a "step-down" procedure [12] which, essentially, means the cutoff is based on considering the ranked vector of $p$-values. The following procedure describes the Holm method to control FWER. Let

$$\alpha_1 \leq \alpha_2 \leq \ldots, \leq \alpha_N \tag{1}$$

be constants defined by $\alpha_i = \alpha/(N - i + 1)$ and let the ordered $p$-values be denoted by $p_{(1)} \leq \cdots \leq p_{(N)}$ corresponding to hypotheses, $H_{(1)}, \cdots, H_{(N)}$. If $p_{(1)} > \alpha_1$,then reject no null hypothesis. Otherwise, if

$$p_{(1)} \leq \alpha_1, \ldots, p_{(r)} \leq \alpha_r, \tag{2}$$

then reject hypothesis $H_{(1)}, \ldots, H_{(r)}$ where the largest $r$ satisfying (2) is used. With this framework for FWER, the Holm method to control $k$-FWER is done by redefining $\alpha_i$ as

$$\alpha_i = \begin{cases} \frac{k\alpha}{N}, & i \leq k, \\ \frac{k\alpha}{N+k-i}, & i > k. \end{cases}$$

## 4. THE PROPOSED KBIN METHOD

We propose a novel approach, KBIN, to control $k$-FWER at level $\alpha$. With $N$ hypotheses under consideration and $k$ being the number of acceptable false positives, the KBIN procedure works by rejecting all hypothesis less than $p_{cut}$ where $p_{cut}$ is the supremum over the set of $p$'s that solve the equation below

$$F(k-1|N,p) \geq 1-\alpha \tag{3}$$

where $F$ is the cumulative density function for a Binomial distribution with $N$ trials and probability of success $p$. Note, $\alpha$ is usually chosen to be small, e.g. 0.05. The proof for KBIN is based on an assumption for the distribution of $p$-values in a multiple testing setting. Specifically, in a multiple testing procedure setting, it is common to assume that the distribution of the $p$ values follows a mixture distribution specified by $f(x) = 1-\pi+\pi f_1$, for $i = 1, ..., N$, where $0 \leq \pi \leq min(f(x)) \leq 1$ and $f_1$ is a well defined probability density function (PDF). In this situation the second component $f_1$ is assumed to be concave [13, 14, 15], and the mixture weight $\pi$ controls the percentage of hypotheses that follow the alternative. This is the standard beta-uniform (BUM) mixture distribution when performing $N$ tests [14]. Note, in a microarray setting $N$ could be the number of genes to be tested or the number of gene pathways tested. In a gene pathway setting $N$ could be several hundred while in a gene setting, $N$ could be several thousand.

Specifically we have the following theorem for KBIN,

**Theorem 4.1.** *For testing $N$ hypotheses, assume the $p$-values are independent and identically distributed following the PDF*

$$f(x|a,\lambda) = 1-\pi+\pi f_1(x)$$

*where $0 \leq \pi \leq min(f(x)) \leq 1$ and $f_1(x)$ is a well defined PDF. Consider a procedure that rejects any hypothesis, $H_i$, for which $p_i \leq p_{cut}$ where $p_{cut}$ is a solution to*

$$sup_p F(k-1|N,p) \geq 1-\alpha$$

*where sup denotes the supremum over the interval $[0,1]$ and $F$ is the cumulative density function for a binomial random variable with size $N$ and probability of success $p$. Then this procedure controls $k$-FWER such that $\mathbb{P}(V \geq k) \leq \alpha$ where $V$ is the number of false positives.*

**Proof** Assume $f(x|\pi) = 1-\pi+\pi f_1(x)$ represents a mixture distribution of two components, one arising from the null hypothesis and one arising from the alternative component ($f_1$). Thus $V$, the number of false positives, follows a binomial distribution with size $N$ and probability of success $(1-\pi)p_{cut}$, notationally, $V \sim Bin\,(N, (1-\pi)p_{cut})$. Consider the random variable $W$ where $W \sim Bin(N, p_{cut})$. By design, $W$ is stochastically greater than $V$, so we have,

$$\begin{aligned}
\mathbb{P}(V \geq k) &\leq \mathbb{P}(W \geq k) \\
&= 1 - \mathbb{P}(W \leq k-1) \\
&= 1 - F(k-1|N, P_{cut}) \\
&\leq 1 - (1-\alpha) \\
&= \alpha
\end{aligned}$$

The use of the binomial distribution to control $k$-FWER is the reason for calling this the "KBIN" method.

Figure 1 shows the KBIN $p$-value cut points as a function of $k$ and $\alpha$ to control $k$-FWER. For example, when $N = 250$ and the researcher wants to protect against 5 or more false discoveries, the KBIN $p$-value cut point is $p_{cut} = .007914$ since

$$\sum_{x=0}^{4} P(X = x) = .95$$

where $X \sim Bin(250, .007914)$. Table 1 is a table of the $p$-value cut points for a variety of $k$, $\alpha$ and $N$ values. As expected, for a fixed $N$, the $p$-value cut point increases as a function of either $k$ or $\alpha$, that is, as the researcher is willing to accept more false positives or a higher probability of false positives, the $p$-value cut point will be more liberal (larger). Figure 2 details some of the theoretical aspects of the KBIN method as a function of $N$ and $V$. Interestingly, when controlling $k$-FWER at small values ($< .1$), the number of false positives, rarely greatly exceeds $k$. For example, if $k$-FWER was controlled by KBIN with $k = 5$ and $\alpha = 0.20$, then $V$ is rarely larger than 6 and never larger than 8 (see Figure 2 (b,d)). This is one of the nice features of controlling $k$-FWER; when the number of false positives exceeds $k$, it is rarely much larger than $k$.

Figure 3 shows the difference in $p$-value cut points between the KBIN method and the adjusted Bonferroni method for $k$-FWER $= 0.05$. As $N$ increases, the difference in cut points between the two methods converge. Roughly speaking, for more than 1000 tests, the difference between using KBIN and the adjusted Bonferroni method is negligible. Coincidentally, in situations with more than 1000 tests, it may be more reasonable for the researcher to consider controlling the false discovery rate (FDR) rather than $k$-FWER.

Theorem 4.1 establishes that KBIN controls $k$-FWER at $\alpha$. We can further show that the KBIN method is more powerful than the adjusted Bonferroni method when performing 2 or more tests. The theoretical proof that KBIN is more powerful than the adjusted Bonferroni method to control $k$-FWER is accomplished by establishing several necessary lemmas.

**Lemma 4.2.** *The cumulative density function (CDF), $F(k-1|N, ak/N)$, of a binomial random count with $N$ trials and success probability $ak/N$, is an increasing function of $N$, for $k = 1$, and $0 < a < 1$ with $F(0|2, \frac{a}{2}) > 1 - a$.*

**Proof** The CDF evaluated at 0 is

$$\binom{N}{0}\left(\frac{a}{N}\right)^0\left(1-\frac{a}{N}\right)^N = \left(1-\frac{a}{N}\right)^N.$$

an increasing function of $N$. Further note that

$$\begin{aligned} F\left(0|2, \frac{a}{2}\right) &= \binom{2}{0}\left(\frac{a}{2}\right)^0\left(1-\frac{a}{2}\right)^2 \\ &= \left(1-\frac{a}{2}\right)^2 \\ &= 1 - a + \frac{a^2}{4} \\ &> 1 - a \end{aligned}$$

**Lemma 4.3.** *The CDF $F(k-1|N, ak/N)$ is an increasing function in $k$, given $N \geq max\{2, k\}$.*

**Proof** Intuitively, for any $N \geq k$, as $k$ increases, the probability of a success increases, as well as the probability that a random count takes probability less than $k$. To see this we re-express the CDF, letting $k = x$,

$$\begin{aligned} &\sum_{j=0}^{x-1}\binom{N}{j}\left(\frac{ax}{N}\right)^j\left(1-\frac{ax}{N}\right)^{N-j} \\ = \; &\left(1-\frac{ax}{N}\right)^N\left(1+\frac{ax}{1-\frac{ax}{N}}+\sum_{i=2}^{x-1}\binom{N}{i}\left(\frac{ax}{N}\right)^i\left(1-\frac{ax}{N}\right)^{-i}\right) \end{aligned}$$

Letting $k = x + 1$, we have,

$$\left(1 - \frac{a(x+1)}{N}\right)^N \left(1 + \frac{a(x+1)}{1 - \frac{a(x+1)}{N}} + \sum_{i=2}^{(x+1)-1} \binom{N}{i} \left(\frac{a(x+1)}{N}\right)^i \left(1 - \frac{a(x+1)}{N}\right)^{-i}\right).$$

Comparing the $k = x$ and $k = x + 1$ expressions, term by term, we see that each term is increasing, hence, the CDF $F(k - 1|N, ak/N)$ is an increasing function in $k$.

**Lemma 4.4.** *The CDF $F(k-1|N, ak/N)$ is a decreasing function of $N$, for $k = 2$, and $0 < a < 1$.*

**Proof** Consider the CDF, re-expressed as

$$\begin{aligned}
&\sum_{j=0}^{1} \binom{N}{j} \left(\frac{2a}{N}\right)^j \left(1 - \frac{2a}{N}\right)^{N-j} \\
= &\left(1 - \frac{2a}{N}\right)^N + 2a \left(1 - \frac{2a}{N}\right)^{N-1} \\
= &\left(1 - \frac{2a}{N}\right)^{N-1} \left(1 + 2a\left(1 - \frac{1}{N}\right)\right).
\end{aligned}$$

The first multiplicative term above is decreasing, while the second growing in $N$. The change in $2a(1 - 1/N)$ for an increment in $N$ is $a2/(N(N + 1))$. Note that

$$\left(1 - \frac{2a}{N+1}\right)^N - \left(1 - \frac{2a}{N}\right)^{N-1} > \left(1 - \frac{2a}{N}\right)^N - \left(1 - \frac{2a}{N}\right)^{N-1}.$$

The change in the expression on the left hand side of the inequality, with an increment in $N$, is greater than

$$-\frac{2a}{N}\left(1 - \frac{2a}{N}\right)^{N-1}.$$

Monotonic decay follows since,

$$\frac{1}{N+1} < \left(1 - \frac{2a}{N}\right)^{N-1}.$$

**Theorem 4.5.** *The k-BIN procedure is more powerful than the adjusted Bonferroni method when $N \geq 2$.*

**Proof** We will proceed by showing that

$$lim_{N\to\infty} F(k-1|N, ak/N) > 1-a. \tag{4}$$

for all $(N, k) : N \geq \max\{2, k\}$.

For fixed $\alpha, k, N$, let the KBIN based $p$-cutoff be denoted as $p_{KBIN}$. Then Equation 4 implies that $ak/N < p_{KBIN}$, since by definition,

$$F(k-1|N, p_{KBIN}) = 1-a.$$

From Lemmas 4.2, 4.3 and 4.4, it is sufficient to show that for $k = 2$, the above result holds. We can re-express the limit in Equation 4 for $k = 2$ as

$$\begin{aligned}
lim_{N\to\infty} F(k-1|N, ak/N) &= lim_{N\to\infty} \sum_{j=0}^{1} \binom{N}{j} \left(\frac{2a}{N}\right)^j \left(1-\frac{2a}{N}\right)^{N-j} \\
&= lim_{N\to\infty} \left(1-\frac{2a}{N}\right)^N + 2a\left(1-\frac{2a}{N}\right)^{N-1} \\
&= (1+2a)e^{-2a} \\
&> 1-a.
\end{aligned}$$

The last inequality is easily verified, checking that the minimum slope for $(1+2a)e^{-2a}$ is -0.74, while both expressions, $(1+2a)e^{-2a}$ and $1-a$, equal 1 when evaluated at $a = 0$, and 0 and $3e^{-2}$, respectively, when evaluated at $a = 1$.

## 5. SIMULATION STUDY

The BUM model is used in the simulations comparing KBIN, the adjusted Bonferroni method, and the Holm method. Two different simulation settings are provided using the standard BUM model. In Simulation 1, we let $N = 250$ and $\pi = 0.20$ and the distribution of p-values under the alternative hypothesis follows a beta distribution with parameters 1/2 and 2. For example, in a gene pathway test setting, this would correspond to the case where many pathways are enriched, although, the differential effects in gene expression underlying enrichment are moderate. In Simulation 2, we let $\pi = 0.05$ and the distribution of p-values for the alternative hypotheses follows a beta distribution with parameters .1 and 10. For the gene pathway example, this could correspond to the situation of a relatively few number of enriched gene pathways, but with large effect sizes in differential gene expression underlying the enrichment.

The results from Simulation 1 and Simulation 2 are shown in Figures 4 and 5, respectively. For both simulations, we see that the KBIN method has a larger true positive rate (TPR) compared to either the adjusted Bonferroni method or the Holm method (upper panels in Figures 4 - 5). Also, for both simulations, KBIN also has a larger false positive rate (FPR) than either the adjusted Bonferroni method or the Holm method (see, e.g. the lower panels in Figures 4 and 5). Table 2 shows the mean number of false discoveries for both simulations under different settings for $k$ and $\alpha$. In all situations, the KBIN method has the largest number of true discoveries and false discoveries, however, the mean number of false positives is always less than $k$. Thus, when controlling $k$-FWER at stringent levels (e.g. $\alpha = 0.05$), the researcher can expect to have less than $k$ false positives. Figures 6 - 7 display the estimated value of $k$-FWER for each method in each of the simulations. In both simulations, we see that the KBIN method is less conservative than either the adjusted Bonferroni method or the Holm method. In other words, the KBIN method is closer to exact control of $k$-FWER at $\alpha$ than either of the other methods. Figures 6 - 7 illustrate the key point; both the adjusted Bonferroni method and the Holm method are very conservative in controlling $k$-FWER, while the KBIN method is more liberal when controlling $k$-FWER. Therefore, KBIN is more powerful than competing methods, while still maintaining control of $k$-FWER.

## 6. REAL DATA STUDY

From Theorem 4.5, the KBIN method is more powerful than the adjusted Bonferroni method. Since Figure 3 shows the difference in power between KBIN and the adjusted Bonferroni method to be negligible for $N > 1000$, we choose examples with less than 1000 tests. We divide our examples based on the number of hypotheses (N) tested. In Section 6.1 we consider examples with less than 20 tests, Section 6.2 consists of an example containing 119 tests, while Section 6.3 consists of example with over 200 tests. In all situations, the KBIN method yields more discoveries than the adjusted Bonferroni method or the Holm method.

*6.1. Small N*

In the first example we consider the dataset used in [6]. This data examines the ability of thrombolysis with recombinant tissue type plasminogen activator (rt-PA) and anisoylated pasminogen streptokinase activator (APSAC) to reduce mortality in patients with myocardial infarction. As given in [6], the $p$-values from 15 comparisons are listed and the results when controlling the false discovery rate (FDR) are reported. Using the reported $p$-values, we can examine the results when controlling $k$-FWER rather than the FDR. Setting $k = 2$ and $\alpha = 0.01$, we have 4 discoveries with the KBIN method, and 2 discoveries each with the adjusted Bonferroni method and the Holm method. With the additional discoveries in the KBIN method, we can further validate more hypothesis than specified by either the adjusted Bonferroni or Holm method.

In the second example, we consider a hospital severity dataset provided in [16]. This dataset provides a meta-analysis of the incidence of adverse drug reactions (ADRs) in hospitalized patients from 39 perspective studies of US hospitals. From this analysis, the overall incidence of serious ADRs was given as 6.7 percent. Table 1 from [16] provides the summary data from 18 of the 39 perspective studies. Using a binomial test of proportions we can examine if each individual dataset provides significant evidence of ADRs being different than 6.7 percent. When controlling $k$-FWER at 0.05 with $k = 2$, we find 15 reported ADRs significantly different from 6.7 percent according to the KBIN algorithm, 12 datasets from the adjusted Bonferroni method and 15 datasets from the Holm method. The large number of datasets showing different rates of ADRs demonstrates the relatively large amount of variability in this meta-analysis.

*6.2. Moderate N*

In genetics, microRNAs (miRNAs) are a single strand of RNA molecules of approximately 20 nucleotides in length. They can regulate gene expression and, as such, have been examined as potential biomarkers for disease. A recently published dataset in [17] explores the ability of a flow cytometry bead based miRNA platform to classify basal versus luminal tumor subtypes in an breast cancer dataset. Their dataset contains approximately 120 miRNAs and was recently deposited in the Gene Expression Omnibus (GEO) online microarray repository [18]. From this dataset we can examine the number of miRNA discoveries obtained by controlling $k$-FWER under various settings for $k$ and $\alpha$. With 500 bootstraps, Figure 9 shows the mean number of significant miRNAs that differentiate cancer tissue subtypes when controlling $k$-FWER. In general, the KBIN method provides a larger mean number of target miRNA's that can differentiate between breast cancer subtypes.

*6.3. Large N*

In this example, we examine the number of significant genetic pathways associated with smokers versus non smokers using gene expression microarrays. This dataset is provided in [19]. For background on the science and nature of gene pathway analysis see [20, 21]. For this example, the pathways were obtained through the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG is one of the most complete and publicly available pathway databases, whose latest release contains approximately 210 curated non redundant pathways in humans [21]. The Gene Set Analysis (GSA) software was employed as a measure of significance for each pathway where the $p$-value for each pathway is determined through a permutation based approach.

Using 500 bootstrap version of the data in [22], we examined the (random) number of pathways discovered as enriched when controlling $k$-FWER using KBIN, the Adjusted Bonferroni and the Holm method. Figure 8 shows the mean number of pathways discovered under typical choices for $\alpha$ and $k$. From Figure 8, we conclude that the KBIN method tended to yield a larger mean number of discovered pathways than either the adjusted Bonferroni method or the adjusted Holm method. While all methods theoretically control $k$-FWER, KBIN had a larger mean number of discoveries versus either the adjusted Bonferroni method or Holm method.

## 7. DISCUSSION

In multiple testing situations, with greater than say, 1000 tests, it is reasonable to control the FDR and the asymptotic arguments supporting the methods designed to control FDR are reasonable. However, in situations with less than 1000 tests, it may be reasonable to consider controlling $k$-FWER and, in these situations, the available methods to control $k$-FWER are too conservative. The current methods to control $k$-FWER include the adjusted Bonferroni method, the Holm method, and also the relatively modern minP and maxT methods. The adjusted Bonferroni method is considered a fixed algorithm where the $p$-value cut point is not data dependent. The Holm method to determine the $p$-value cut point is data driven in the sense that the $N$ length vector of $p$-values is required in order to determine the significant of an individual hypothesis. Recently two data driven methods, minP and maxT, have been proposed to control $k$-FWER [23, 24, 25]. The methods require a bootstrap step employed to estimate the null distribution [22]. The methods are heavily data dependent because of the required bootstrap step. This is in contrast to the KBIN method, the Adjusted Bonferroni method, and the Holm method which merely requires the $N$-length vector of $p$-values. For this reason, the minP and maxT methods are not considered in the data applications or simulations in this manuscript.

A large body of research in the FDR literature involves estimating the proportion of true null hypothesis among the $N$ tests [7, 13, 8]. In general, these methods "borrow" strength across the dataset or use data splitting procedures (see, for example, [26]) to estimate the proportion of true null hypothesis. These techniques can be used to improve the KBIN method. Note, in the proof of Theorem 4.1, we employ a random variable that is stochastically greater than $V$ where $V \sim Bin(N, 1-\pi)p_{cut}$. Using the techniques to estimate the proportion of null hypothesis, we can estimate $\pi$ and thus develop more powerful $p$-value cut offs that will maintain control of $k$-FWER.

This manuscript proposes using KBIN, a new more powerful method to control $k$-FWER when the $p$-value distribution follows the BUM model. Of course, regardless of the situation, rejecting any $k-1$ will always control $k$-FWER, but, at the very least, will be too optimistic and naive since we assume that there are more real discoveries [10]. With the BUM model as the distribution for $p$-values, KBIN is more powerful than either the adjusted Bonferroni method or the Holm method for $k$-FWER. Advantages to using the KBIN method to control $k$-FWER include simplicity and ease of interpretation. Further, when control of $k$-FWER is violated, it is usually not an extreme violation. From the simulation study, if the number of false positives is controlled to be no more than 5, it never exceeds 10 false positives, and typically 6 to 7 false positives are realized with an $\alpha = 0.20$. This is assuring to the researcher since, as shown in [9], in gene pathway testing with FDR set to 0.10, the actual FDR may be as large as 0.40. Figure 3 hints at some of the asymptotic comparisons between KBIN and the adjusted Bonferroni method. As the number of tests increases to 1000 or more, the difference in $p$-value cut points between KBIN and the adjusted Bonferroni method become negligible. However, in situations where there are 1000 tests or less, the KBIN method gives a larger $p$-value cut point and thus is more powerful than the adjusted Bonferroni method. Thus,we advocate using KBIN in multiple testing settings where the researcher is interested in controlling $k$-FWER. This situation has many applications in gene pathway analysis and other potential biomarker correlation studies such as mice studies as well as potential financial and public health situations.

## References


[1] Abdi H. The bonferroni and sidak corrections for multiple comparisons. *The encyclopedia of measurement and statistics* 2007; .

[2] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; :65–70.

[3] Sidak Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 1967; :626–633.

[4] Miller R. *Simultaneous statistical inference*. Springer New York, 1981.

[5] Hochberg Y, Tamhane A. *Multiple comparison procedures*. Wiley New York, 1987.

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

[7] Efron B, Tibshirani R, Storey J, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**(456):1151–1160.

[8] Storey J, Tibshirani R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 2003; **100**(16):9440–9445.

[9] Gold D, Miecznikowski J, Liu S. Error control variability in pathway-based microarray analysis. *Bioinformatics* 2009; *In Press*.

[10] Lehmann E, Romano J. Generalizations of the familywise error rate. *Annals of Statistics* 2005; :1138–1154.

[11] Gentleman R, Carey V, Huber W, Dudoit S, Irizarry R. *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer Verlag, 2005.

[12] Lehmann E. *Testing statistical hypotheses*. Springer Verlag, 1997.

[13] Storey J. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 2002; :479–498.

[14] Pounds S, Morris S. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 2003; **19**(10):1236–1242.

[15] Hunt D, Cheng C, Pounds S. The beta-binomial distribution for estimating the number of false rejections in microarray gene expression studies. *Computational Statistics and Data Analysis* 2009; **53**(5):1688–1700.

[16] Lazarou J, Pomeranz B, Corey P. Incidence of adverse drug reactions in hospitalized patients a meta-analysis of prospective studies. *JAMA* 1998; **279**(15):1200–1205.

[17] Blenkiron C, Goldstein L, Thorne N, Spiteri I, Chin S, Dunning M, Barbosa-Morais N, Teschendorff A, Green A, Ellis I, *et al.*. MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology* 2007; **8**(10):R214.

[18] Edgar R, Domrachev M, Lash A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 2002; **30**(1):207.

[19] Spira A, Beane J, Shah V, Liu G, Schembri F, Yang X, Palma J, Brody J. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proceedings of the National Academy of Sciences* 2004; **101**(27):10 143–10 148.

[20] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, *et al.*. Gene Ontology: tool for the unification of biology. *Nature genetics* 2000; **25**(1):25–29.

[21] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic acids research* 2004; **32**(Database Issue):D277.

[22] Efron B, Tibshirani R. *An introduction to the bootstrap*. Chapman & Hall, 1997.

[23] Westfall P, Young S. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley-Interscience, 1993.

[24] Dudoit S, van der Laan M, Pollard K. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1):1040.

[25] van der Laan M, Dudoit S, Pollard K. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Statistical Applications in Genetics and Molecular Biology* 2004; **3**(1):1041.

[26] Rubin D, Dudoit S, van der Laan M. A method to increase the power of multiple testing procedures through sample splitting. *Statistical Applications in Genetics and Molecular Biology* 2006; **5**(1):1148.

| | **K = 5** | | | | | **K = 10** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N=25 | N=50 | N=100 | N=500 | N=1000 | N=25 | N=50 | N=100 | N=500 | N=10 |
| $\alpha = 0.01$ | 0.054 | 0.026 | 0.013 | 0.003 | 0.001 | 0.185 | 0.087 | 0.042 | 0.008 | 0.004 |
| $\alpha = 0.05$ | 0.082 | 0.04 | 0.02 | 0.004 | 0.002 | 0.236 | 0.113 | 0.055 | 0.011 | 0.005 |
| $\alpha = 0.10$ | 0.101 | 0.049 | 0.025 | 0.005 | 0.002 | 0.265 | 0.128 | 0.063 | 0.012 | 0.006 |
| $\alpha = 0.20$ | 0.126 | 0.062 | 0.031 | 0.006 | 0.003 | 0.303 | 0.149 | 0.074 | 0.015 | 0.007 |

Table 1: **KBIN Results:** Table of $p$-value cut points for KBIN under a variety of different values for $N$, $\alpha$, $K$.

| | $\alpha$ | **0.05** | | | | | **0.01** | **0.05** | **0.10** | **0.20** | **0.50** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **K** | **1** | **5** | **10** | **20** | **50** | **5** | | | | |
| Sim 1 | KBIN | 0.04 | 1.59 | 4.41 | 10.77 | 31.94 | 1.04 | 1.59 | 1.96 | 2.48 | 3.76 |
| | Adj. Bonf | 0.04 | 0.20 | 0.41 | 0.80 | 2.01 | 0.04 | 0.20 | 0.41 | 0.80 | 2.01 |
| Sim 2 | KBIN | 0.04 | 1.89 | 5.23 | 12.82 | 37.99 | 1.24 | 1.89 | 2.34 | 2.97 | 4.45 |
| | Adj. Bonf | 0.05 | 0.23 | 0.46 | 0.95 | 2.39 | 0.05 | 0.23 | 0.46 | 0.95 | 2.39 |

Table 2: **Simulation Results:** The mean number of false positives when $k$-FWER is controlled at different values of K and $\alpha$ using KBIN and the Adjusted Bonferroni method. Simulation 1 (Sim 1) has BUM parameters, $\pi = .2, N = 250$ with $f_1 = beta(1/2, 2)$ while Simulation 2 (Sim 2) has BUM parameters $\pi = 0.05, N = 250$ with $f_1 = beta(.1, 10)$. Roughly speaking, Sim 1 corresponds to a large number of discoveries but with small effect sizes, while Sim 2 corresponds to a small number of discoveries but with large effect sizes. In either case Adjusted Bonferroni is more conservative than KBIN and thus is less powerful than KBIN when controlling $k$-FWER.

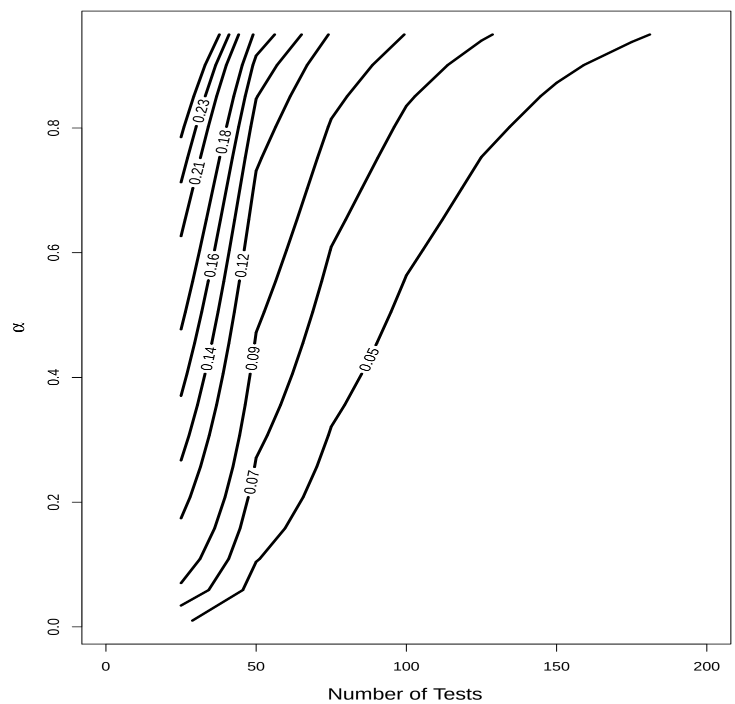



Figure 1: **KBIN Values**: Contour plot showing $\alpha$ on the y-axis, the number of tests on the x-axis and the contour lines show the p-value cutoffs for a KBIN procedure to control $k$-FWER at level $\alpha$.

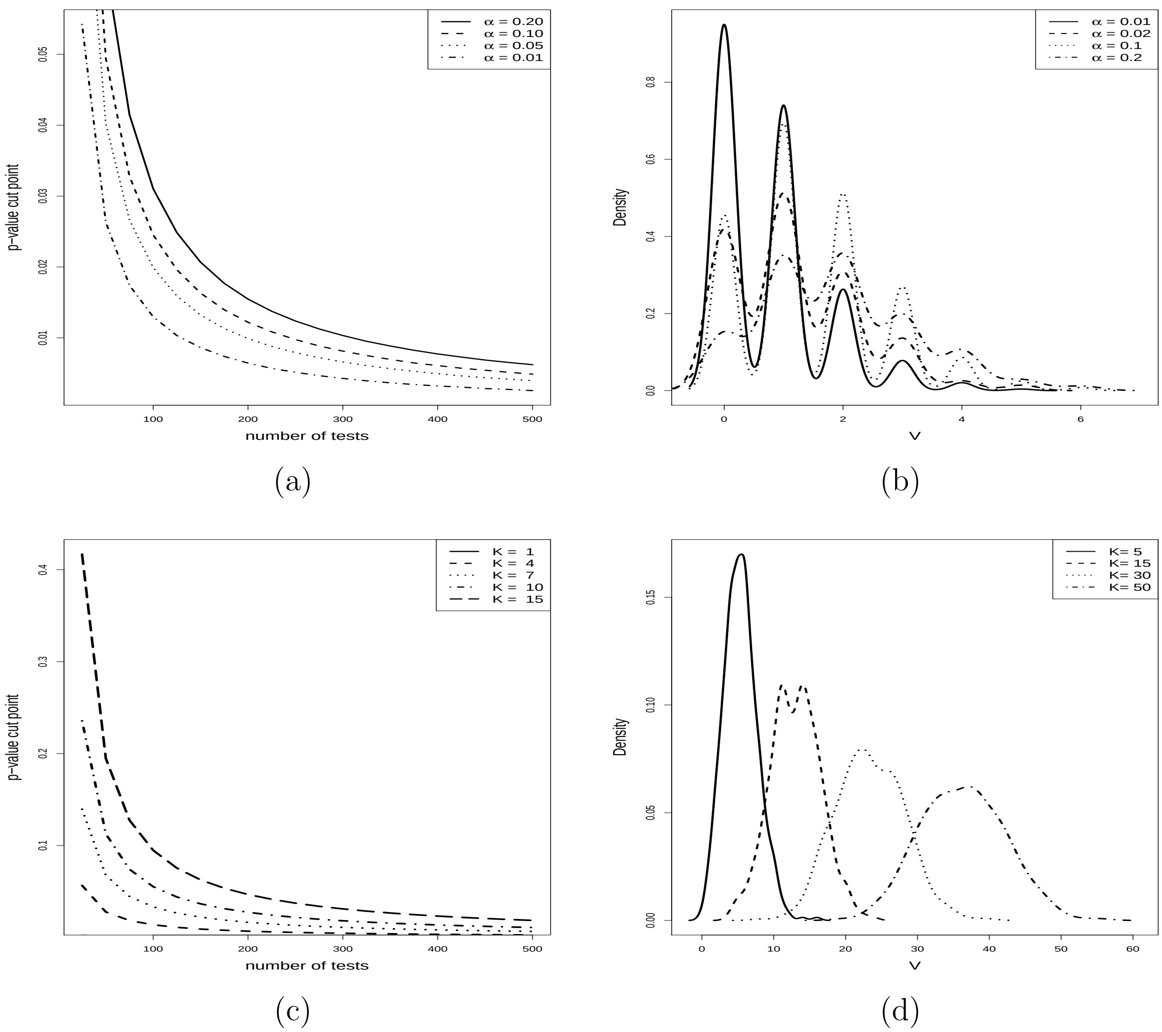



Figure 2: **KBIN Simulations**: (a) The KBIN $p$-value cut point values as a function of the number of tests. As expected, as the number of tests increases the cut point for the $p$-value decreases. As $\alpha$ increases, the $p$-value cut point also increases. (b) The distribution for $V$ when $k$ is fixed for a variety of values for $\alpha$. These curves show that when the number of false positives exceeds $k$, it is rarely much larger. (c) Curves showing the KBIN $p$-value cut point as a function of $k$, keeping $\alpha$ fixed at 0.05. (d) The distribution curves for the number of false positives when $\alpha$ is fixed at 0.05. Each curve corresponds to a different value of $k$.

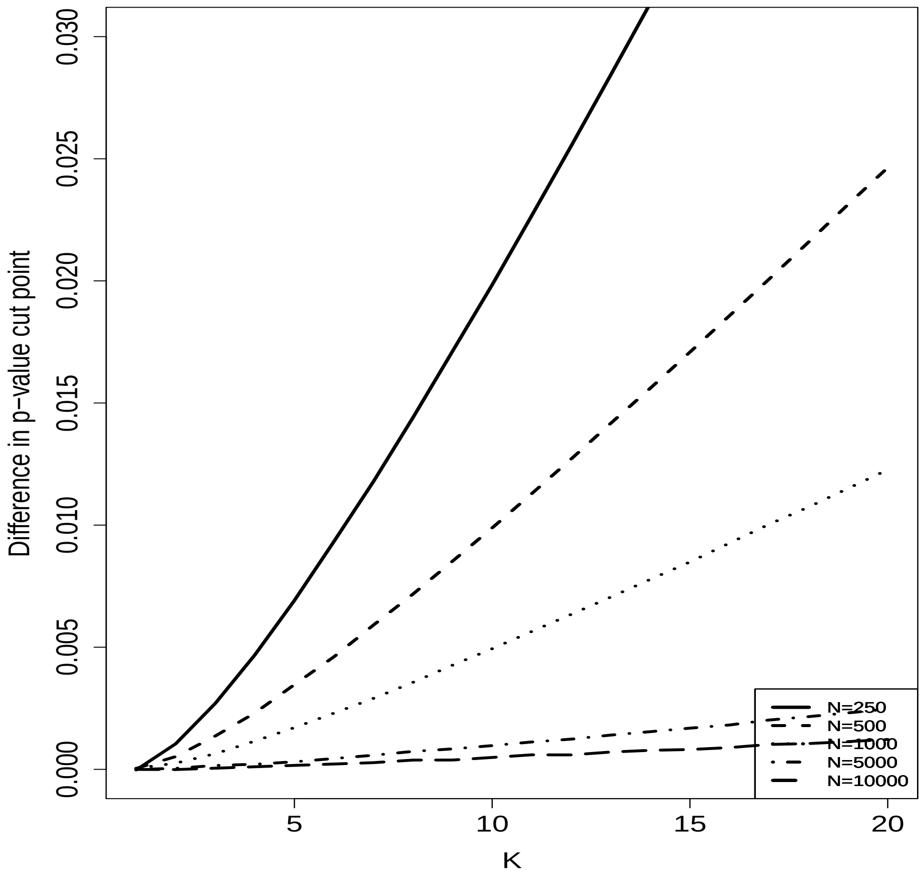



Figure 3: **P Cut point Difference**: The difference between the KBIN $p$-value cut point and adjusted Bonferroni $p$-value cut point as a function of $k$ when $\alpha$ is fixed at 0.05. For the number of tests ($N$) greater than 5000, the difference between the KBIN $p$-value cut point and the adjusted Bonferroni $p$-value cut point is negligible.

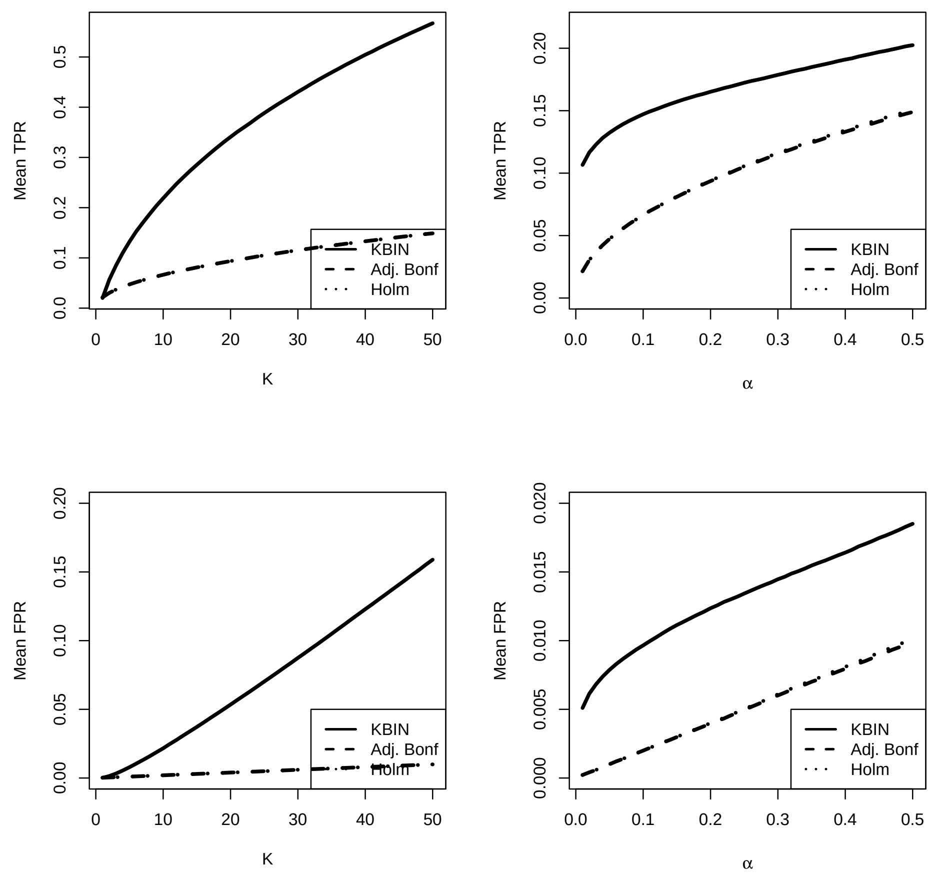



Figure 4: $k$**-FWER Simulation 1**: The summary statistics for Simulation 1 consisting of a large number of discoveries but with small effect sizes. The KBIN method (KBIN) is compared against the adjusted Bonferroni method (Adj. Bonf) and the Holm method (Holm). The mean true positive rate (TPR) is shown as a function of $k$ in the upper left panel, and as a function of $\alpha$ in the upper right panel. The mean false positive rate (FPR) is shown as a function of $k$ and $\alpha$ in the lower left and lower right panels, respectively. As expected, the KBIN method has a higher TPR at the expense of an increased FPR.

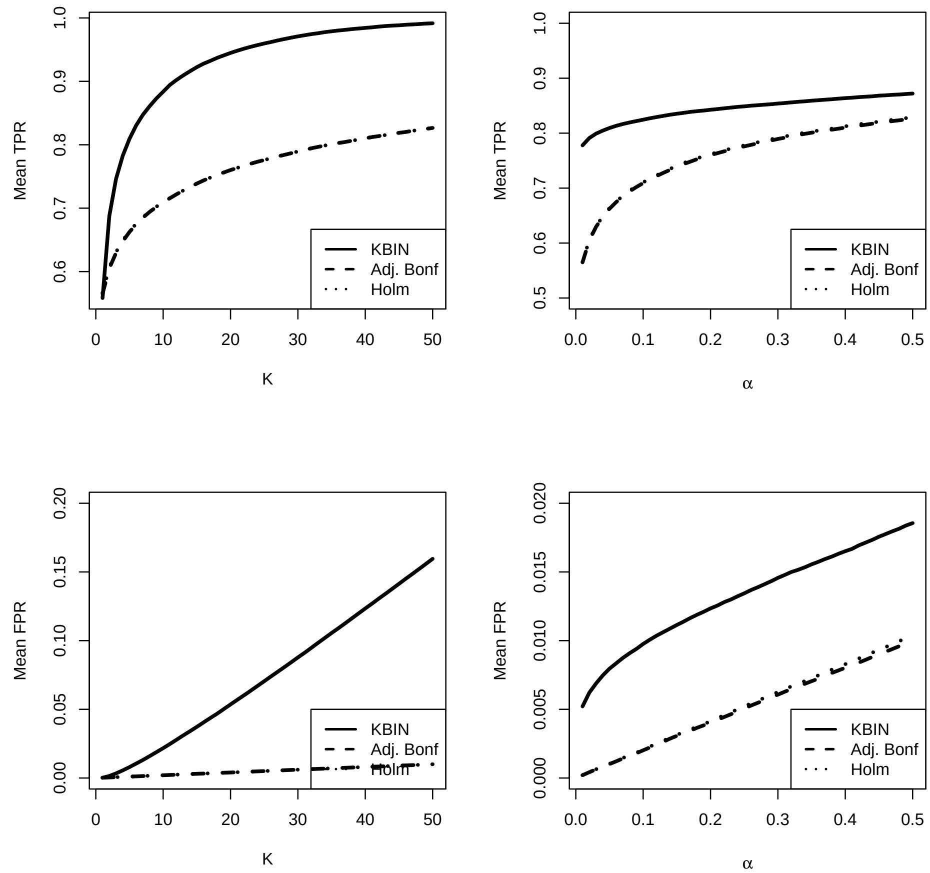
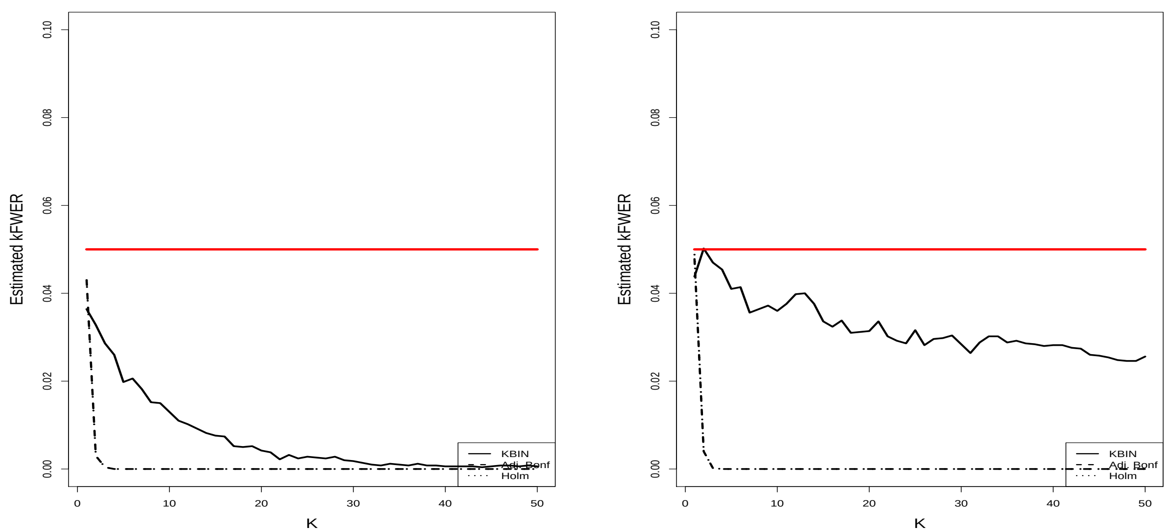
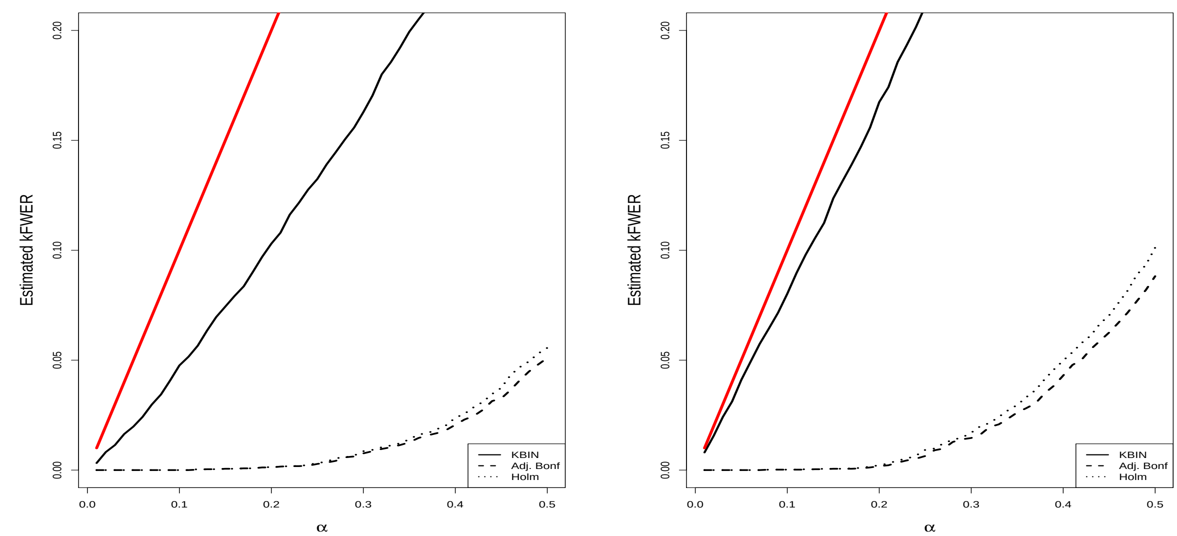



Figure 5: $k$**-FWER Simulation 2**: The summary statistics for Simulation 2 consisting of a small number of discoveries but with large effect sizes. The KBIN method (KBIN) is compared against the adjusted Bonferroni method (Adj. Bonf) and the Holm method (Holm). The mean true positive rate (TPR) is shown as a function of $k$ in the upper left panel, and as a function of $\alpha$ in the upper right panel. The mean false positive rate (FPR) is shown as a function of $k$ and $\alpha$ in the lower left and lower right panels, respectively. As expected, the KBIN method has a higher TPR at the expense of an increased FPR.




(a) (b)

Figure 6: **Estimated control of $k$-FWER** : (a) For Simulation 1, with $\alpha$ fixed, the estimated $k$-FWER shown as a function of $k$. (b) For Simulation 2, with $\alpha$ fixed, the estimated $k$-FWER shown as a function of $k$. Note, in these simulations, $k$-FWER was theoretically controlled at 0.05 (red line). The adjusted Bonferroni method (Adj. Bonf) and Holm method (Holm) are more conservative than KBIN causing a decreased power when compared to KBIN.




(a) (b)

Figure 7: **Estimated control of $k$-FWER**: (a) For Simulation 1, with $k$ fixed, the estimated $k$-FWER shown as a function of $\alpha$. (b) For Simulation 2, with $k$ fixed, the estimated $k$-FWER shown as a function of $\alpha$. Exact control of $k$-FWER would follow the red line. The KBIN method is more powerful than either the adjusted Bonferroni method (Adj. Bonf) or the Holm method (Holm).

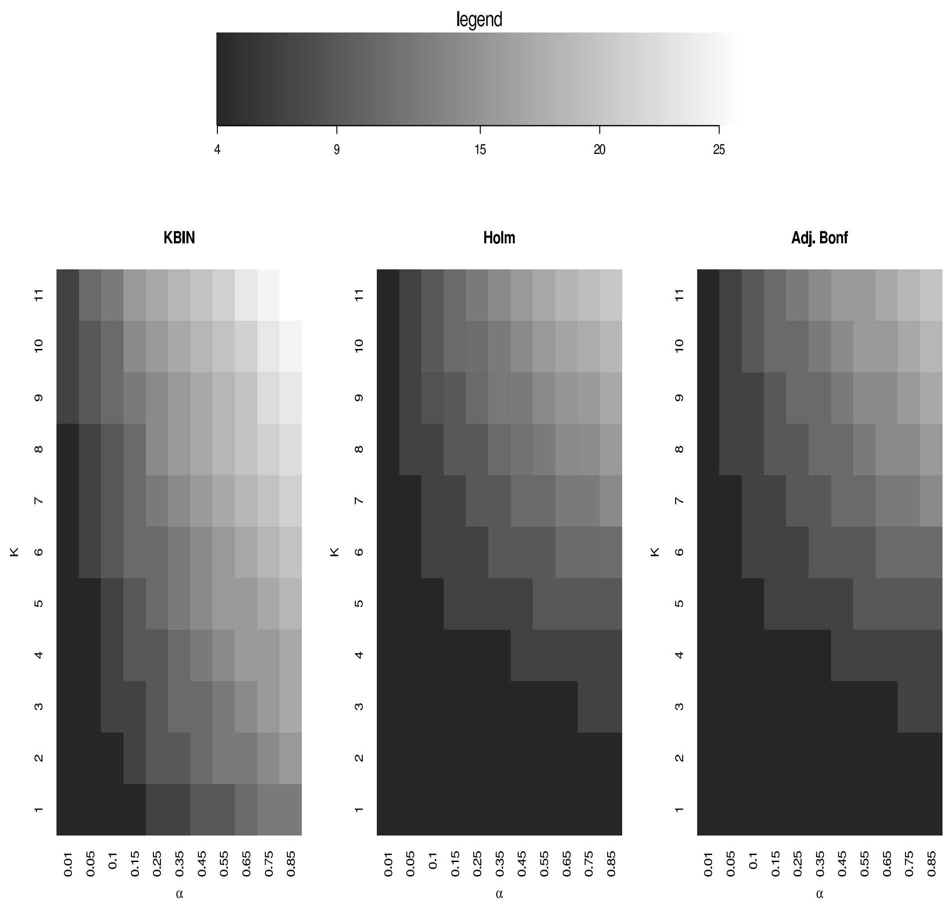



Figure 8: **Gene Pathway Analysis**: Gene pathway analysis showing the mean number of pathways discovered as determined by a bootstrap analysis for the dataset in [19]. The KBIN method (KBIN) provides a larger mean number of discoveries than either the Holm method (Holm) or the adjusted Bonferroni method (Adj. Bonf).

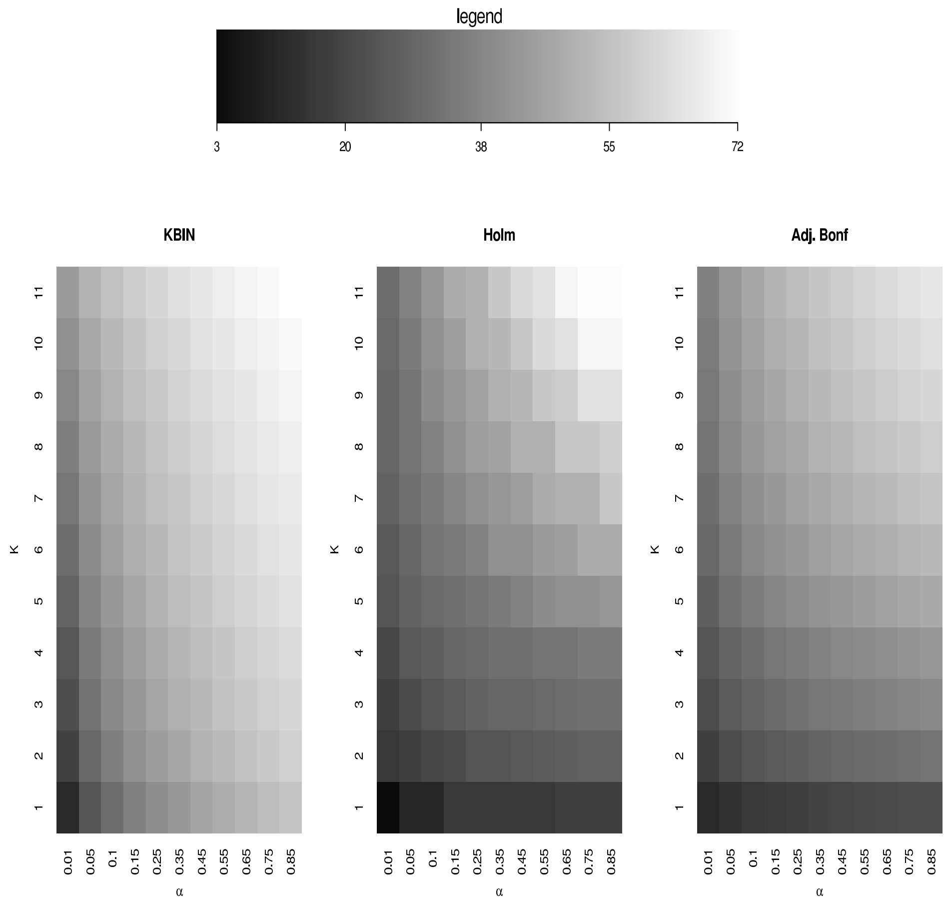



Figure 9: **Micro RNA Analysis**: The mean number of micro RNAs discovered as determined by a bootstrap analysis for the [17] dataset. The KBIN method (KBIN) provides a larger mean number of discoveries than either the Holm method (Holm) or the adjusted Bonferroni method (Adj. Bonf).