Acta Mathematicae Applicatae Sinica, English Series
Vol. 38, No. 2 (2022) 235–253
https://doi.org/10.1007/s10255-022-1077-5
http://www.ApplMath.com.cn & www.SpringerLink.com

Acta Mathematicae Applicatae Sinica, English Series

Null-free False Discovery Rate Control Using Decoy Permutations

Kun HE[1,3], Meng-jie LI[2,3], Yan FU[2,3,†], Fu-zhou GONG[2,3], Xiao-ming SUN[1,3]

[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[2]CEMS, NCMIS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China ([†]E-mail: yfu@amss.ac.cn)
[3]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract The traditional approaches to false discovery rate (FDR) control in multiple hypothesis testing are usually based on the null distribution of a test statistic. However, all types of null distributions, including the theoretical, permutation-based and empirical ones, have some inherent drawbacks. For example, the theoretical null might fail because of improper assumptions on the sample distribution. Here, we propose a null distribution-free approach to FDR control for multiple hypothesis testing in the case-control study. This approach, named *target-decoy procedure*, simply builds on the ordering of tests by some statistic or score, the null distribution of which is not required to be known. Competitive decoy tests are constructed from permutations of original samples and are used to estimate the false target discoveries. We prove that this approach controls the FDR when the score function is symmetric and the scores are independent between different tests. Simulation demonstrates that it is more stable and powerful than two popular traditional approaches, even in the existence of dependency. Evaluation is also made on two real datasets, including an arabidopsis genomics dataset and a COVID-19 proteomics dataset.

Keywords multiple testing; false discovery rate; null distribution-free; p-value-free; decoy permutations; knockoff filter
2000 MR Subject Classification 62G10; 62H15

1 Introduction

1.1 Traditional Approaches to FDR Control

Multiple testing has become increasingly popular in the present big-data era. For example, a typical scenario of applying multiple testing in biomedical studies is to look for differentially expressed genes/proteins, from thousands of candidates, between two groups (i.e., cases and controls) of samples[13, 16]. Currently, controlling the false discovery rate (FDR), which is defined as the expected proportion of incorrect rejections among all rejections[6], is the predominant way to do multiple testing. FDR control procedures aim at selecting a subset of rejected hypotheses such that the FDR is no more than a given level.

Because a p-value is typically computed from the null distribution of a test statistic in each single test, the canonical approaches to FDR control for multiple testing at present are based on the p-values of all tests or at least the null distribution of the test statistic. Since Benjamini

Manuscript received September 29, 2020. Accepted on January 27, 2022.
This paper is supported by the National Key R&D Program of China (No. 2018YFB0704304), the National Natural Science Foundation of China (Nos. 32070668, 62002231, 61832003, 61433014) and the K.C. Wong Education Foundation.
[†]Corresponding author.

and Hochberg[6] proposed the first p-value based sequential procedure to control the FDR (BH procedure), many FDR control approaches have been developed, e.g., [5, 7, 8, 41, 44, 45].

A key problem faced by these approaches is how to obtain the proper null distribution. Popular null distributions, including the theoretical null, permutation null and empirical null, often suffer one way or another[16, 17]. The theoretical null, though widely used, might fail in practice for many reasons, such as improper mathematical assumptions or unobserved covariates[15, 16]. For example, for the Student's t-test, if the sample distribution is not normal, the t-value will not follow a t-distribution and the p-values calculated will not be uniform $(0, 1)$ distributed for true null hypotheses. The permutation null is also widely used. There are mainly two different permutation methods, i.e., the permutation tests and the pooled permutation[31]. The permutation tests are a class of widely used non-parametric tests to calculate p-values, and are most useful when the information about the data distribution is insufficient. However, the statistical power of permutation tests is limited by the sample size of a test[50]. Instead of estimating a null distribution for each test individually, the pooled permutation in multiple testing estimates an overall null distribution for all tests[19]. However, it has been found that pooling permutation null distributions across hypotheses can produce invalid p-values, since even true null hypotheses can have different permutation distributions[31].

To overcome the shortcomings of the theoretical and permutation null distributions, new methods were proposed to estimate an empirical null distribution from a large number of tests[16, 18, 19, 42]. For example, the empirical Bayes method estimates the empirical null distribution by decomposing the mixture of null and alternative distributions[16]. However, decomposing the mixture distribution is intrinsically a difficult problem. For example, if the empirical distribution has a strong peak, the decomposing may fail[48].

Moreover, the proportion of true null hypotheses has to be estimated either explicitly or implicitly to apply these FDR control methods. If this null proportion is ignored (e.g., assumed to be one as in the original BH procedure), the power of testing would be reduced. Since Storey[44] proposed the first approach, estimation of the null proportion has become a key component of current FDR methods to enhance the power, such as the Bayes and the empirical Bayes methods[16, 45, 46]. More accurate estimation of the null ratio has been of great interest in the field[32, 39, 53].

1.2 Our Approach to FDR Control

Here, we propose a new approach to FDR control in the case-control study, named *target-decoy procedure*, which is free of the null distribution and the null proportion. The procedure (simplified version) can be briefly described with the following steps. First, a target score and a number of decoy scores are calculated for each hypothesis test. These scores are used to measure the (dis)similarities of two groups of samples, and can be common statistics (e.g., t-value) or other scoring functions. The target score is calculated with regard to the original samples, while the decoy scores are calculated with regard to random permutations of the original samples. Then, based on the target score and decoy scores, a label and a final score are calculated for each test in a competitive manner. If the target score is more significant than half of the decoy scores, the test is labelled as target and the final score is set as the target score. Otherwise, the test is labelled as decoy and the final score is set as the decoy score with a specific rank that is mapped symmetrically from the rank of the target score. Next, the tests are sorted by their final scores in descending order (assuming larger scores are more significant), and for each test, a ratio of $(N_d + 1)/N_t$ is calculated, where N_d and N_t represent the numbers of decoy and target tests ranked above this test (included), respectively. At last, the lowest-ranked test that has a $(N_d + 1)/N_t$ ratio below the given FDR control level is localized, and all the hypotheses of target tests ranking no lower than this test are rejected.

The addition of one (+1 correction) to the number of decoy tests is essential to our approach.

We prove that the target-decoy procedure with such correction can rigorously control the FDR when the score function is symmetric and the scores are independent between different tests.

Our approach is exclusively based on the scores and labels of tests. The scoring function used is not limited to traditional p-value or test statistics which have clear null distributions, but can be in any free forms with some symmetry property. Therefore, our approach provides great flexibility and can be potentially more powerful than traditional approaches, the performance of which largely relies on the precision of p-values or the sample size of each test. Monte-Carlo simulations demonstrate that our approach effectively controls the FDR and is more powerful than two popular methods, i.e., the Bayes method[44–46] and the empirical Bayes method[16, 18, 19]. The performances of the three methods were also compared on two real biological datasets, including an arabidopsis genomics dataset and a COVID-19 proteomics dataset. Because our procedure is more straightforward and can be used with arbitrary score functions, we believe that it will have many practical applications.

Our approach was inspired by the widely used target-decoy database search approach to estimating the FDR of peptide identifications in tandem mass spectrometry-based proteomics[20]. In this approach, tandem mass spectra of peptides are searched against a database consisting of equal size of target and decoy protein sequences. The peptide-spectrum matches (PSMs) are scored and filtered by some score threshold. The FDR of selected PSMs is estimated by the ratio of the number of decoy matches to the number of target matches. Usually, the lowest score threshold is taken such that the estimated FDR is below a given level. Although this empirical target-decoy approach to FDR has been very effective in practice, its theoretical foundation was not established until we proved that a +1 correction to the number of decoy matches leads to rigorous FDR control under the assumption of independence between PSMs[27]. Our work in the context of mass spectrometry was initially submitted to journals in 2013 (unpublished) and was made public in 2015[28]. The extension to general multiple testing as presented here was first described in an earlier manuscript[29].

1.3 Related Works

The most related work to ours is the knockoff filter method proposed by Barber and Cands[2], which aims to control the FDR of variables selected via Lasso regression for a Gaussian linear model. In this method, knockoff variables, which are not associated with the response (conditioning on the original variables), are constructed and subjected to competition with the original variables (covariates). The basic rationale of knockoff filter in FDR control is identical to the target-decoy approach. First, knockoff is essentially synonymous with decoy in their roles. Second, the method used by knockoff filter to derive the rejection region, i.e., the FDR estimation formula with +1 correction and the procedure of selecting the score threshold, is exactly the same as the target-decoy approach. Third, after the proof of equal probabilities of a null variable obtaining a positive score (target label) or a negative score (decoy label), the proof of FDR control is the same mathematical problem addressed by the knockoff filter and the target-decoy approach, although their proving techniques are different. The main contribution of the knockoff filter is its sophisticated knockoff construction method that makes possible the proof of the aforementioned 'equal probabilities' for dependent variables. Knockoff filter allows the variables to be correlated with each other, but assumes the Gaussian noise in the linear model. In comparison, our approach (this paper) achieves FDR control for independent variables only, but puts no assumptions on the distribution of the variables. In addition, the original knockoff filter method required that the sample size (n) is no less than the number of variables (p) for FDR control.

Cands et al.[9] later re-framed the knockoff procedure and proposed the so-called model-X knockoffs method. Unlike the original linear model in which $X_{i,j}$ was treated as fixed (randomness was from the Gaussian noise), the model-X knockoffs method treats $X_{i,j}$ as random. It

assumes knowledge of the joint distribution of the covariates, and constructs knockoffs probabilistically instead of geometrically. This removes the restriction on sample size ($n \geq p$) and makes the method applicable to both linear and non-linear models. Although the construction of model-X knockoffs does not rely on the specific distribution forms of the original variables in principle, Gaussian distribution is the only one that can be implemented at present. Another limitation of the knockoff method is its high computational cost on knockoff construction, which involves complex matrix computation, such as eigenvalue computation and semidefinite programming.

In the current knockoff methods, only one knockoff copy is constructed for each original variable, and the probability of a null variable or its knockoff copy being selected is equal (0.5). In our target-decoy procedure, multiple decoy permutations are constructed for each original variable, which offers us the flexibility of setting different probabilities of producing target or decoy tests for true null hypotheses. This kind of multiple competition can enhance the power as we experimentally illustrated. Recently, Emery et al.[22] investigated the multiple competition problem in more depth. They presented two methods, namely max method and mirror method, for competition with the multiple decoys/knockoffs. The max method is most intuitive. It selects the variable (original or knockoff) with the highest score. Gimenez and Zou[26] also used the max method for multiple knockoffs. The mirror method is like what we do in our standard target-decoy procedure but is more flexible. It uses two adjustable rank cutoffs for target/decoy labelling, while we only use one adjustable cutoff for target labelling. Emery and Keich[23] also proposed methods to construct multiple knockoffs that offer both FDR control and enhanced power.

In recent years, the approach of FDR control using competitive decoys/knockoffs has attracted much attention from the field of statistics[3, 4, 24, 25, 34, 36, 37, 40]. No doubt, this success was owed to the publication of the knockoff method by Candes et al. However, it should be noticed that we first proposed the FDR estimation formula with the +1 correction, which is the key to FDR control, and gave the first proof of FDR control (in the context of mass spectrometry and under the independence assumption)[27, 28]. We also first introduced the multiple competition strategy [29]. These have been recognized by the community, e.g., [11, 12, 21, 22, 30, 35]. Thus, despite the similarity of our approach to the well-known knockoff method which has been published earlier[2], we still would like to introduce the target-decoy procedure to the community with our original notations and proofs[27–29]. As far as we know, the target-decoy approach to FDR control was first used and named in mass spectrometry-based proteomics as early as in 2007[20]. Therefore, we use the terminology *decoy* instead of *knockoff*. Moreover, compared to the knockoff method, our approach has different technical arguments, different motivations, and is verified with different simulation experiments and real data here.

Other related works include that Levitsky et al.[35] proposed an interpretation to the +1 correction based on the negative binomial distribution. However, this interpretation assumes that the number of null targets can be infinite and has uniform prior probability, and therefore, is not a rigorous interpretation. Storey et al.[46] also had a +1 correction in their pFDR estimation to achieve FDR control. However, this correction was made to the number of p-values greater than a fixed threshold λ, which amounts to the total number of decoys in our case. This is very different from the target-decoy/knockoff approach in which the +1 correction is made to the number of decoys/knockoffs in the rejection region.

Organization. The rest of the paper is organized as follows. Section 2 describes our target-decoy approach for FDR control. Section 2.1 discusses a general scenario of case-control study. The simplified and standard target-decoy procedures are presented in Sections 2.2 and 2.3, respectively. Section 2.4 provides an adaptive version of the target-decoy procedure. Section 2.5 establishes the theoretical foundation of our approach (Proofs are given in Supplementary

Material). Numerical results on independent and dependent variables are given in Section 3. Applications to real data are shown in Section 4. Section 5 concludes the paper and points out some directions worthy of further study.

2 The Target-decoy Approach

2.1 Problem Formulation

Consider a two-groups (case and control) study involving m random variables, $X_1, X_2, \cdots, X_m$. For each random variable X_j where $1 \leq j \leq m$, there are n random samples $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$, in which $X_{j,1}, X_{j,2}, \cdots, X_{j,n_1}$ are from the n_1 cases and $X_{j,n_1+1}, \cdots, X_{j,n}$ are from the $n_0 = n - n_1$ controls. For simplicity, the numbers of random samples (and similarly, cases and controls) are assumed to be the same for all random variables here, although our method does not rely on this assumption.

The goal is to search for random variables differently distributed between cases and controls. The null hypothesis for random variable X_j used here is the exchangeable hypothesis H_{j0}: the joint distribution of $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ is symmetric. In other words, the joint probability density function of $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ (or the joint probability mass function if $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ are discrete) satisfies $f_{X_{j,1},\cdots,X_{j,n}}(x_{j,1}, \cdots, x_{j,n}) = f_{X_{j,1},\cdots,X_{j,n}}(\pi_n(x_{j,1}, \cdots, x_{j,n}))$ for any possible $x_{j,1}, \cdots, x_{j,n}$ and any permutation π_n of $x_{j,1}, \cdots, x_{j,n}$. If $X_{j,1}, \cdots, X_{j,n}$ are independent, this hypothesis is equivalent to that $X_{j,1}, \cdots, X_{j,n}$ are identically distributed. Here we use the exchangeable hypothesis to deal with the case where $X_{j,1}, \cdots, X_{j,n}$ are correlated but still an exchangeable sequence of random variables[10].

Let $S(x_1, x_2, \cdots, x_n)$ be some scoring function satisfying

$$S(x_1, \cdots, x_n) = S(\pi_{n_1}(x_1, \cdots, x_{n_1}), \pi_{n_0}(x_{n_1+1}, \cdots, x_n))$$

for any possible $x_1, \cdots, x_n$, any permutation of n_1 elements $\pi_{n_1}(\cdot)$ and that of n_0 elements $\pi_{n_0}(\cdot)$. Note that most scoring functions evaluating the difference between $x_1, x_2, \cdots, x_{n_1}$ and $x_{n_1+1}, x_{n_1+2}, \cdots, x_n$ have the above symmetry property, including commonly used test statistics, e.g., the t-value as we used in this paper. Without loss of generality, we assume that larger scores are more significant. Note that neither the null distributions of scores nor the distributions of random variables are required to be known.

2.2 The Simplified Target-decoy Procedure

We first introduce the simplified version of our target-decoy procedure for FDR control. The intuition of the procedure is to let each random variable X_j be labelled as target or decoy with the same chance if the null hypothesis for X_j is true. At the same time, the chance of X_j being labelled as decoy is expected to be negligible if its null hypothesis is false (this assumption is not needed for FDR control). Thus, the number of target tests of the true null hypotheses beyond a threshold can be approximated by the number of decoy ones.

Algorithm 2.1. The simplified target-decoy procedure.

1. For each $1 \leq j \leq m$, calculate t scores including a target score and $t-1$ decoy scores. The target score is $S_j^T = S(X_{j,1}, X_{j,2}, \cdots, X_{j,n})$. Each decoy score is obtained by first sampling a permutation π_n of $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ randomly and then calculating the score as $S(\pi_n(X_{j,1}, X_{j,2}, \cdots, X_{j,n}))$. Sort these t scores in descending order. For equal scores, sort them randomly with equal probability.

2. For each test j, calculate a final score S_j and assign it a label $L_j \in \{T, D\}$, where T and D stand for target and decoy, respectively. Assume that the rank of S_j^T is i. If $i < (t+1)/2$,

let L_j be T and set S_j as S_j^T. If $i > (t+1)/2$, let L_j be D and set S_j as the score ranking $i - \lceil t/2 \rceil$. Otherwise, $i = (t+1)/2$, let L_j be T or D randomly and set S_j as S_j^T.

3. Sort the m tests in descending order of the final scores. Let $i_1, \cdots, i_m$ be such that $S_{i_1} \geq \cdots \geq S_{i_m}$ (with tied values randomly broken). Let $L_{(1)}, \cdots, L_{(m)}$ be the the corresponding labels $L_{i_1}, \cdots, L_{i_m}$, respectively.

4. If the specified FDR control level is α, let

$$K = \max\left\{k \Big| \frac{\#\{L_{(j)} = D,\ j \leq k\} + 1}{\#\{L_{(j)} = T,\ j \leq k\} \vee 1} \leq \alpha\right\} \tag{2.1}$$

and reject the hypothesis with rank j if $L_{(j)} = T$ and $j \leq K$.

Note that there is a +1 correction to the number of decoy tests in the numerator of equation (2.1), which is key to FDR control as shown in Section 2.5. This correction was first proposed in the context of proteomics for target-decoy based FDR control of peptide identifications[27, 28].

$X_{j,1}, \cdots, X_{j,6}$

4.75	1.36	5.24	1.06	−0.56	0.41
−0.23	−0.64	0.65	1.16	0.56	−0.95
−1.15	0.32	−0.43	0.05	−0.56	0.32
8.05	4.28	6.10	−1.29	−0.90	0.08
−2.36	−0.71	0.66	−0.37	−0.41	1.32
−0.51	0.78	2.51	−0.76	−0.16	−0.21

(a)

$\pi_6(X_{j,1}, \cdots, X_{j,6})$

1.36	5.24	−0.56	4.75	1.06	0.41
−0.23	−0.95	0.65	1.16	−0.64	0.56
−0.56	0.32	−1.15	0.32	−0.43	0.05
4.28	−0.90	0.08	−1.29	8.05	6.10
−0.41	0.66	−0.71	1.32	−2.36	−0.37
−0.21	2.51	−0.16	0.78	−0.76	−0.51

(b)

S_j^T	S_j^p	S_j	L_j
10.44	0.18	10.44	T
0.99	1.61	1.61	D
1.07	1.33	1.33	D
20.54	9.40	20.54	T
2.95	0.95	2.95	D
3.91	2.63	3.91	T

(c)

i_j	S_{i_j}	$L_{i_j}(L_{(j)})$	$\frac{\#\{L_{(j)}=D, j\leq k\}+1}{\#\{L_{(j)}=T, j\leq k\}\vee 1}$	
4	20.54	T	1.00	
1	10.44	T	0.50	
6	3.91	T	0.33	Rejected with α=0.25
5	2.95	T	0.25	
2	1.61	D	0.50	
3	1.33	D	0.75	

(d)

i_j	p-value	$\widehat{FDR}$	
4	0.01	0.60	None rejected with α=0.25
1	0.01	0.30	
6	0.40	0.80	
5	0.40	0.60	
2	0.80	0.96	
3	1.00	1.00	

(e)

Figure 2.1. An example of the simplified target-decoy procedure

An example of the simplified target-decoy procedure is shown in Figure 2.1. In it, $m = n = 6$, and $t = 2$. The first three columns of the data are from cases and the other three columns are from controls. The scoring function is $S(x_1, \cdots, x_6) = |x_1 + x_2 + x_3 - x_4 - x_5 - x_6|$. Obviously, the function satisfies the symmetry property defined in Section 2.1 but the null distribution is unknown. For each row, a target score S_j^T is first calculated for the original samples. Then, the procedure performs one permutation π_6 and calculates one decoy score $S(\pi_6(X_{j,1}, \cdots, X_{j,6}))$, since $t - 1 = 1$. If $S_j^T > S(\pi_6(X_{j,1}, \cdots, X_{j,6}))$, the final score S_j is set as S_j^T and L_j is set as T. Otherwise, if $S_j^T < S(\pi_6(X_{j,1}, \cdots, X_{j,6}))$, S_j is set as $S(\pi_6(X_{j,1}, \cdots, X_{j,6}))$ and L_j is set as D (Figure 2.1 (c)). The 6 tests are sorted in descending order of S_j to derive i_j, S_{i_j} and L_{i_j} (i.e., $L_{(j)}$). For example, i_1 is 4 because S_4 is maximal in all the final scores. Then, with $L_{(1)}, \cdots, L_{(6)}$, we can calculate $\frac{\#\{L_{(j)}=D, j\leq k\}+1}{\#\{L_{(j)}=T, j\leq k\}\vee 1}$ for each row k. If α is set as 0.25, we reject the first four hypotheses since $\frac{\#\{L_{(j)}=D, j\leq 4\}+1}{\#\{L_{(j)}=T, j\leq 4\}\vee 1} = 0.25$ and the formula is larger than 0.25 for any

$k > 4$ (Figure 2.1 (d)). For comparison, the Bayes method[46] is also applied to this example. In our setting, λ is set as 1/2. Meanwhile, the p-values are calculated with Wilcoxon rank sum test since the null-distributions of the random variables are unknown. If α is set as 0.25, no hypothesis is rejected with the Bayes method (Figure 2.1 (e)). Note that this toy example is for illustration purposes only. In most real-world applications, the Bayes method works well and the choice of $\alpha = 0.25$ is too loose.

The random permutation used in the procedure can be generated by simple random sampling either with or without replacement, just as in the permutation tests. Similarly, with larger sampling number $t-1$, the power of our approach will become slightly stronger as shown in Section 3. We can set t as $\min\{\binom{n}{n_0}, \tau\}$, where τ is the maximum number of permutations we would perform.

Unlike other FDR control methods, our approach does not depend on the null distribution. The number of permutations, $t-1$ can be much smaller than that used in permutation tests. In our simulations, $t-1$ was set as 49 or 1, while in the real data experiments, it was set as 19. Simulations demonstrate that the target-decoy approach can still control the FDR even if $t-1$ was set as 1, in which case little information was revealed about the null distribution.

2.3 The Standard Target-decoy Procedure

The +1 in the numerator of equation (2.1) is essential to accomplish FDR control. However, it has a side effect of reducing the power. This effect can be amplified under some conditions, e.g., when the number of false null hypotheses or the total number of hypotheses is small. To enhance the power, we introduce a parameter $r(>1)$ into the procedure. The intuition is to let each random variable X_j be labelled as target with probability $\frac{1}{2r}$ and as decoy with probability $\frac{1}{2}$ if the null hypothesis for X_j is true. Thus, the number of decoy tests beyond a threshold is about r times the number of target ones of the true null hypotheses, and then $\frac{1}{r}$ times the ratio of the number (added by one) of decoy tests to the number of target ones beyond a threshold can be used for FDR control.

For any fixed $1 \le r \le \binom{n}{n_0}$, the standard target-decoy procedure (we will omit the word *standard* below for simplicity) is as follows.

Algorithm 2.2. The target-decoy procedure (Steps 1,3 are identical to Algorithm 2.1 and are omitted here).

2. For each $1 \le j \le m$, let $\Lambda_j = i - P_j$ where i is the rank of S_j^T in the t scores, and P_j is a random draw from uniform$[0,1)$ distribution. Calculate a final score S_j and assign a label $L_j \in \{T, D, U\}$, where T, D and U stand for target, decoy and unused, respectively. If $\Lambda_j \le \frac{t}{2r}$, let $L_j = T$ and $S_j = S_j^T$. If $\frac{t}{2} < \Lambda_j \le t$, let $\Lambda_j^{'}$ be a random draw from uniform$\big(0, \frac{t}{2r}\big]$ distribution, L_j be D and S_j be the score ranking $\lceil \Lambda_j^{'} \rceil$-th. Otherwise, let L_j be U and S_j be $-\infty$.

4. If the specified FDR control level is α, let

$$K = \max\left\{k \Big| \frac{1}{r} \times \frac{\#\{L_{(j)} = D,\ j \le k\} + 1}{\#\{L_{(j)} = T,\ j \le k\} \vee 1} \le \alpha\right\} \tag{2.2}$$

and reject the hypothesis with rank j if $L_{(j)} = T$ and $j \le K$.

In Step 2, we introduce P_j to make that $\Pr(L_j = T) = \frac{1}{2r}$ and $\Pr(L_j = D) = \frac{1}{2}$ if the null hypothesis for random variable X_j is true. If $\Lambda_j \le \frac{t}{2r}$, X_j is labelled as target and S_j is set as S_j^T. If $\frac{t}{2} < \Lambda_j \le t$, X_j is labelled as decoy, and S_j is a random score sampled from the largest $\lceil \frac{t}{2r} \rceil$ scores. Otherwise, we have $\frac{t}{2r} < \Lambda_j \le \frac{t}{2}$, X_j is labelled as unused, and S_j is set as $-\infty$ such that it is at the end of the queue after Step 3.

Section 2.5 will show that the above target-decoy procedure controls the FDR for any fixed r. In practice, one can set the value of r empirically or simply set $r = 1$, which reduces the target-decoy procedure into its simplified version described in Section 2.2. Alternatively, an algorithm can be used to choose r adaptively for a given dataset as discussed in Section 2.4.

2.4 The Adaptive Target-decoy Procedure

The parameter r is for adjusting the probability that a true null hypothesis is labelled as T. On the one hand, equation (2.2) can be too conservative for a small r, e.g., 1 as in the simplified target-decoy procedure, because of the addition of 1 in the numerator if there are only a few false null hypotheses. For example, assume that the total number of tests is 80 and the FDR control level is 0.01. If r is set as 1, no hypothesis will be rejected, because the numerator of equation (2.2) is always no less than 1 and the fraction is greater than $1/80 > 0.01$. On the other hand, if r is too large, many false null hypotheses will be labelled as U or D, potentially decreasing the power of testing. Thus, r should be set appropriately in practice to enhance the power. Below, we provide an adaptive procedure to choose a suitable r for the given dataset and the FDR control level. The intuition of the adaptive procedure is to split the data into two parts, with one part used to choose r and the other for inference.

Algorithm 2.3. The adaptive target-decoy procedure.

1. Divide the samples of each random variable into two parts as follows. Choose a suitable n_2 which is smaller than n_0 and n_1 from some range, say $5 \le n_2 \le \min\{\lfloor n_0/2 \rfloor, \lfloor n_1/2 \rfloor\}$. For each random variable X_j where $1 \le j \le m$, randomly choose n_2 samples from $X_{j,1}, X_{j,2}, \cdots, X_{j,n_1}$ and $X_{j,n_1+1}, \cdots, X_{j,n}$, respectively. Let $X^1_{j,1}, X^1_{j,2}, \cdots, X^1_{j,2n_2}$ be these samples. The rest has $n_1 - n_2$ samples from the cases and $n_0 - n_2$ samples from the controls. Let $X^2_{j,1}, X^2_{j,2}, \cdots, X^2_{j,n-2n_2}$ be the rest samples.

2. Set t as $\binom{2n_2}{n_2}$ and perform the target-decoy procedure on $X^1_{j,1}, X^1_{j,2}, \cdots, X^1_{j,2n_2}$ where $1 \le j \le m$ for some range of r, say $R = \{1, 2, 5, 10, 15, 20, 25\}$. Let r_{max} be the one such that the most hypotheses are rejected by the target-decoy procedure.

3. Perform the target-decoy procedure on $X^2_{j,1}, X^2_{j,2}, \cdots, X^2_{j,n-2n_2}$ where $1 \le j \le m$ with $r = r_{max}$ and reject corresponding hypotheses.

2.5 Control Theorem

In this section, we will show that the target-decoy procedure controls the FDR. Let $H_j = 0$ and $H_j = 1$ denote that the null hypothesis for test j is true and false, respectively. Note that $H_1, H_2, \cdots, H_m$ are constants in the setting of hypothesis testing. Define Z_j for $1 \le j \le m$ as follows.

	$L_j = T$	$L_j = D$
$H_j = 0$	$Z_j = 1$	$Z_j = -1$
$H_j = 1$	$Z_j = 0$	$Z_j = -2$

Let $S_{(1)}, S_{(2)}, \cdots, S_{(m)}$ denote the sorted scores and $Z_{(1)}, Z_{(2)}, \cdots, Z_{(m)}$ denote the sorted sequence of $Z_1, Z_2, \cdots, Z_m$. Let $\vec{S}$ and $\overrightarrow{S_{\neq j}}$ denote $S_1, \cdots, S_m$ and $S_1, \cdots, S_{j-1}, S_{j+1}, \cdots, S_m$, respectively. Let $\overrightarrow{S_{(\cdot)}}$ and $\overrightarrow{S_{(\neq j)}}$ denote $S_{(1)}, \cdots, S_{(m)}$ and $S_{(1)}, \cdots, S_{(j-1)}, S_{(j+1)}, \cdots, S_{(m)}$, respectively. We define $\vec{s}$, $\overrightarrow{s_{\neq j}}$, $\overrightarrow{s_{(\cdot)}}$ and $\overrightarrow{s_{(\neq j)}}$ similarly. For example, we will use $\overrightarrow{s_{(\cdot)}}$ to denote a sequence of m constants, $s_{(1)}, \cdots, s_{(m)}$, which is one of the observed values of $S_{(\cdot)}$. We also define $\vec{L}, \vec{Z}, \vec{H}, \overrightarrow{L_{(\neq j)}}$, etc. Then we have the following three theorems.

Theorem 2.1. *In the simplified target-decoy procedure, if the m random variables are independent, then for any fixed $1 \le j \le m$ and any possible $\overrightarrow{s_{(\cdot)}}$ and $\overrightarrow{z_{(\neq j)}}$ we have*

$$\Pr\big(Z_{(j)} = -1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big) = \Pr\big(Z_{(j)} = 1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big).$$

Theorem 2.2. *In the target-decoy procedure, if the m random variables are independent, then for any fixed $1 \le j \le m$ and any possible $\overrightarrow{s_{(\cdot)}}$ and $\overrightarrow{z_{(\neq j)}}$ we have*

$$\Pr\big(Z_{(j)} = -1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big) = r \Pr\big(Z_{(j)} = 1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big).$$

Theorem 2.3. *Suppose that $S_{(1)}, S_{(2)}, \cdots, S_{(m)}, Z_{(1)}, Z_{(2)}, \cdots, Z_{(m)}$ are random variables satisfying $S_{(1)} \ge S_{(2)} \ge \cdots \ge S_{(m)}$ and $Z_{(1)}, Z_{(2)}, \cdots, Z_{(m)} \in \{-2, -1, 0, 1\}$, and r is a positive constant. For any $\alpha \in (0, 1]$, define*

$$K = \max\Big\{k \Big| \frac{1}{r} \times \frac{\#\{Z_{(j)} < 0,\ j \le k\} + 1}{\#\{Z_{(j)} \ge 0,\ j \le k\} \vee 1} \le \alpha\Big\}.$$

If there is no such k, let $K = 0$. If for any fixed j and any possible $\overrightarrow{s_{(\cdot)}}$ and $\overrightarrow{z_{(\neq j)}}$,

$$\Pr\big(Z_{(j)} = -1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big) = r \Pr\big(Z_{(j)} = 1 \big| \overrightarrow{S_{(\cdot)}} = \overrightarrow{s_{(\cdot)}}, \overrightarrow{Z_{(\neq j)}} = \overrightarrow{z_{(\neq j)}}\big),$$

then we have

$$\mathbb{E}\Big(\frac{\#\{Z_{(j)} = 1, j \le K\}}{\#\{Z_{(j)} \ge 0, j \le K\} \vee 1}\Big) < \alpha.$$

The proofs of these theorems are given in the Supplementary Materials. Theorem 2.3 indicates that the target-decoy procedure controls the FDR if the m random variables are independent.

Specially, all of the above theorems hold for the adaptive target-decoy procedure. Recall that the null hypothesis for random variable X_j used here is the exchangeable hypothesis H_{j0}: the joint probability density function of $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ satisfies $f_{X_{j,1},\cdots,X_{j,n}}(x_{j,1}, \cdots, x_{j,n}) = f_{X_{j,1},\cdots,X_{j,n}}(\pi_n(x_{j,1}, \cdots, x_{j,n}))$ for any possible $x_{j,1}, \cdots, x_{j,n}$ and any permutation π_n of $x_{j,1}$, $\cdots, x_{j,n}$. If H_{j0} is true, it is easy to see that $X_{j,1}^2, X_{j,2}^2, \cdots, X_{j,n-2n_2}^2$ are also exchangeable.

3 Simulation Studies

We used Monte-Carlo simulations to study the performance of our approach. The target-decoy procedure were compared with two popular traditional multiple testing methods, including the Bayes method[44–46] and the empirical Bayes method[16, 18, 19]. Simulations were conducted for both independent and dependent random variables. We mainly evaluated the performance of the simplified target-decoy procedure. To show the effectiveness of adjusting r, we also did a simulation on a small dataset and compared the adaptive target-decoy procedure with the simplified target-decoy procedure.

3.1 Simulation Setup

In the simulation, we considered the case-control studies in which the random variables follow the normal distribution or the gamma distribution. In addition to the normal distribution, we did simulation experiments for the gamma distribution because many random variables in real world are gamma-distributed. Recall that the case-control study consists of m random variables. For each random variable, there are n random samples, n_1 of which are from the

cases and the other $n_0 = n - n_1$ are from the controls. Let $X_{j,1}, X_{j,2}, \cdots, X_{j,n}$ be the n random samples for random variable X_j.

The observation values from the normal distribution were generated in a way similar to [7]. First, let $\zeta_0, \zeta_{11}, \cdots, \zeta_{1n}, \cdots, \zeta_{m1}, \cdots, \zeta_{mn}$ be independent and identically distributed random variables following the $N(0,1)$ distribution. Next, let $X_{j,i} = \sqrt{\rho}\zeta_0 + \sqrt{1-\rho}\zeta_{ji} + \mu_{ji}$ for $j = 1, \cdots, m$ and $i = 1, \cdots, n$. We used $\rho = 0, 0.4$ and 0.8, with $\rho = 0$ corresponding to independence and $\rho = 0.4$ and 0.8 corresponding to typical moderate and high correlation values estimated from real microarray data, respectively[1]. The values of μ_{ji} are zero for $i = n_1 + 1, n_1 + 2, \cdots, n$, the n_0 controls. For the n_1 cases where $i = 1, 2, \cdots, n_1$, the values of μ_{ji} are also zero for $j = 1, 2, \cdots, m_0$, the m_0 hypotheses that are true null. The values of μ_{ji} for $i = 1, 2, \cdots, n_1$ and $j = m_0 + 1, \cdots, m$ are set as follows. We let $\mu_{ji} = 1, 2, 3$ and 4 for $j = m_0 + 1, m_0 + 2, m_0 + 3, m_0 + 4$, respectively. Similarly, we let $\mu_{ji} = 1, 2, 3$ and 4 for $j = m_0 + 5, m_0 + 6, m_0 + 7, m_0 + 8$, respectively. This cycle was repeated to produce $\mu_{(m_0+1)1}, \cdots, \mu_{(m_0+1)n_1}, \cdots, \mu_{m1}, \cdots, \mu_{mn_1}$ for the false null hypotheses.

The observation values from the gamma distribution, which is characterized using shape and scale, were generated in the following way. First, let $\Gamma_0, \Gamma_{11}, \cdots, \Gamma_{1n}, \cdots, \Gamma_{m1}, \cdots, \Gamma_{mn}$ be independent random variables where Γ_0 follows the $\Gamma(k_0, 1)$ distribution and Γ_{ji} follows the $\Gamma(k_{ji}, 1)$ distribution for any $j = 1, \cdots, m$ and $i = 1, \cdots, n$. Next, let $X_{j,i} = \Gamma_{ji}$ for $j = 1, \cdots, m$ and $i = 1, \cdots, n$ in the simulation study for independent random variables and let $X_{j,i} = \Gamma_0 + \Gamma_{ji}$ for dependent random variables. To obtain reasonable correlation values, k_0 was set as 4 and k_{ji} was set as 1 for $i = n_1 + 1, n_1 + 2, \cdots, n$, the n_0 controls. For the n_1 cases where $i = 1, 2, \cdots, n_1$, k_{ji} was set as 1 for $j = 1, \cdots, m_0$, the m_0 hypotheses that are true null. The values of k_{ji} for $i = 1, 2, \cdots, n_1$ and $j = m_0 + 1, \cdots, m$ are set as follows. We let $k_{ji} = 2, 3, 4$ and 5 for $j = m_0 + 1, m_0 + 2, m_0 + 3, m_0 + 4$, respectively. Similarly, we let $k_{ji} = 2, 3, 4$ and 5 for $j = m_0 + 5, m_0 + 6, m_0 + 7, m_0 + 8$, respectively. This cycle was repeated to produce $k_{(m_0+1)1}, \cdots, k_{(m_0+1)n_1}, \cdots, k_{m1}, \cdots, k_{mn_1}$ for the false null hypotheses.

The specified FDR control level α was set as 5% or 10%. The total number of tests, m, was set as 10000. The proportion of false null hypotheses was 1% or 10%. The total sample size, n, was set as 20, consisting of the same numbers of cases and controls.

Three different approaches to FDRs were compared, including the Bayes method[44–46], the empirical Bayes method[16, 18, 19] and our target-decoy approach. The Bayes method and the empirical Bayes method are among the most remarkable multiple testing methods. To compare the power of these methods, we rejected the hypotheses against the specified FDR control level α. The rejection threshold, s, for the Bayes method was set as the largest p-value such that q-value(s) is no more than α[44, 45]. The rejection threshold, s, for the empirical Bayes method was set as the minimum z-value such that Efdr(s) is no more than α, where Efdr(s) is the expected fdr (local false discovery rate) of hypotheses with z-values no smaller than s[14, 15]. Specifically, the R packages "locfdr" version 1.1-8[14], and "qvalue" version 2.4.2[47] were used. Each simulation experiment was repeated for 1000 times. We calculated the mean number of rejected hypotheses to evaluate the power of each method. The realized FDR of rejected hypotheses was calculated as the mean of observed false discovery proportions (FDPs) in all repetitions. Note that the variance of the mean of FDPs of 1000 repetitions is one thousandth of the variance of FDPs. We also estimated the standard deviation of the mean of FDPs from the sample standard deviation of FDPs.

The p-values of the Bayes method and the z-values of the empirical Bayes method were calculated with the Student's t-test, Wilcoxon rank sum test or the Student's t-test with permutation. For the Student's t-test, we used the Welch's t-test, a two-sample unequal variances t-test, which is defined as follows,

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1^2/N_1 + s_2^2/N_2}}.$$

Here, $\overline{X}_i$, s_i, and N_i are the i-th sample mean, sample standard deviation and sample size, respectively. For the Student's t-test with permutation, we sampled the cases and the controls for each test, calculated the z-values for sampled data by t-test, and calculated the p-values with the null distribution of pooled z-values[38, 52]. The sampling number of permutations was set as 10[17].

For our target-decoy approach, the cases and the controls of each test were permuted for 49 times or only once, and the t-values and the test statistics of the Wilcoxon rank sum test were used. We did the one-permutation experiments where little information about the null distributions was revealed to demonstrate that our approach does not rely on the null distribution. Because the permutation is performed inherently in our target-decoy approach, the extra permutation is unnecessary.

We will use abbreviations to represent the experiments. For example, Bayes,permutation, Normal, 10%, $\rho = 0.8$ represents the simulation experiment where the Bayes method combined with the pooled permutation is used, the random variables follow the normal distribution, the proportion of false null hypotheses is 10% and the correlation values are 0.8. For our target-decoy approach, t-value, 49, Gamma, 1% represents the simulation experiment where the t-value is used as the score, 49 permutations are performed for each test, the random variables follow the gamma distribution and the proportion of false null hypotheses is as low as 1%.

3.2 Results on Independent Random Variables

Table 3.1. Realized FDRs with independent random variables. The realized FDRs were calculated as the means of FDPs and the standard deviations of the means are less than 0.0020. All the cases where realized FDRs exceed the control level α are labelled with $*$.

	Normal,1%		Normal,10%		Gamma,1%		Gamma,10%	
α	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
Bayes								
t-test	0.050	0.100	0.050	0.100	0.023	0.048	0.031	0.072
permutation	0.048	0.099	0.048	0.098	0.027	0.068	0.047	0.103*
rank-sum	0.039	0.088	0.039	0.087	0.045	0.087	0.042	0.083
Empirical Bayes								
t-test	0.044	0.092	0.040	0.084	0.006	0.013	0.008	0.023
permutation	0.039	0.078	0.039	0.086	0.048	0.124*	0.055*	0.119*
rank-sum	0.046	0.092	0.037	0.078	0.046	0.091	0.037	0.077
Target-decoy								
t-value,49	0.041	0.094	0.049	0.099	0.043	0.094	0.050	0.100
t-value,1	0.044	0.093	0.048	0.097	0.042	0.092	0.047	0.096
rank-sum,49	0.042	0.096	0.049	0.099	0.042	0.096	0.050	0.100
rank-sum,1	0.042	0.093	0.048	0.097	0.042	0.096	0.048	0.097

Figure 3.1 shows the realized FDRs of different methods with independent random variables while the specified FDR control level α was no more than 10%. Table 3.1 gives the realized FDRs while the specified FDR control level α was 5% or 10%. In all cases, the target-decoy approach controlled the FDR, and the realized FDRs were favourably close to α. The empirical Bayes and Bayes methods performed well when the random variables followed the normal distribution. However, they considerably overestimated the FDRs with t-test for the gamma distribution. With the pooled permutation, some of the realized FDRs exceeded α for the gamma distribution as marked by asterisks in the table. Of course, some small exceedances are not necessarily the

evidence of a fail of FDR control but may be due to Monte Carlo error. At last, the Wilcoxon rank-sum test coupled with Bayes or empirical Bayes occasionally overestimated the FDRs.

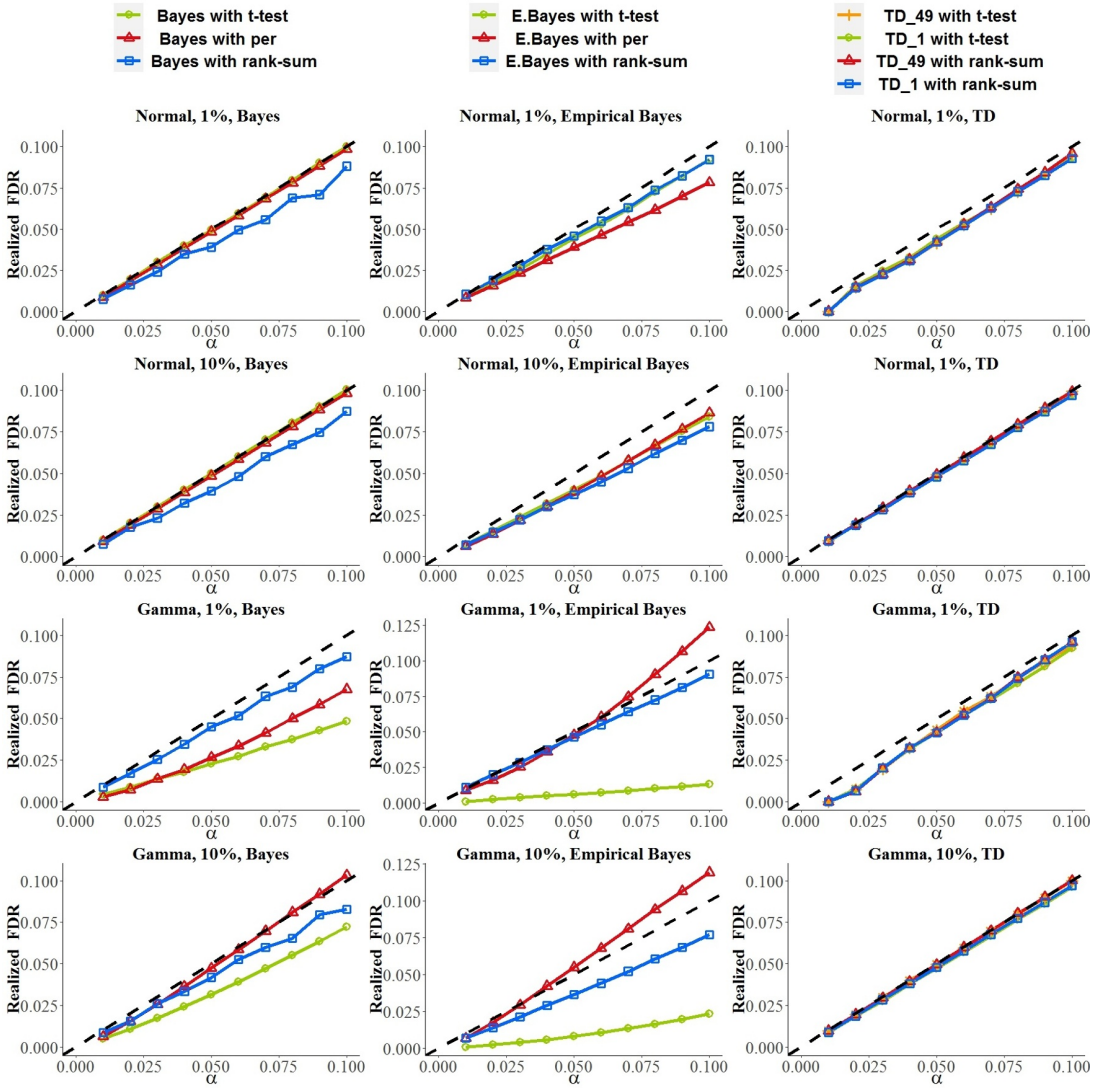

Figure 3.1. Realized FDRs with independent random variables. The realized FDRs were calculated as the means of FDPs

Table 3.2 shows the statistical powers of different methods with independent random variables. When the random variables followed the normal distribution, the powers of the three methods were overall comparable with each other. In the case of gamma distribution, the target-decoy approach was much more powerful than Bayes and empirical Bayes when t-test was used. Permutation based Bayes and empirical Bayes had higher power but at the cost of uncontrolled FDR. When the Wilcoxon rank-sum test was used, our approach was more powerful than the other two methods except the only case of Gamma, 1% and α=0.05.

In all the above experiments, the target-decoy approach successfully controlled the FDR and meanwhile it was remarkably powerful. Notably, the results obtained with 49 permutations or 1 permutation in the target-decoy approach were quite similar, indicating that the proposed

approach is not sensitive to the number of permutations.

Table 3.2. Power with independent random variables. All the cases where realized FDRs exceed α as shown in Table 3.1 are labelled with $*$.

	Normal,1%		Normal,10%		Gamma,1%		Gamma,10%	
α	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
Bayes								
t-test	71	80	845	937	40	50	687	798
permutation	71	80	842	933	41	55	737	861*
rank-sum	67	76	813	906	48	59	734	836
Empirical Bayes								
t-test	70	78	823	909	23	32	534	650
permutation	69	76	821	913	49	66*	755*	891*
rank-sum	69	77	806	889	47	59	715	823
Target-decoy								
t-value,49	69	79	843	935	45	60	743	853
t-value,1	69	79	841	931	45	60	736	845
rank-sum,49	67	77	834	926	42	60	755	872
rank-sum,1	66	77	831	922	42	60	751	865

3.3 Results on Dependent Random Variables

Table 3.3. Realized FDRs with dependent random variables. The realized FDRs were calculated as the means of FDPs and the standard deviations of the means of FDPs are less than 0.0021. All the cases where realized FDRs exceed α are labelled with $*$.

	Normal, $\rho = 0.4$				Normal,$\rho = 0.8$				Gamma			
	1%		10%		1%		10%		1%		10%	
α	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
Bayes												
t-test	0.052*	0.102*	0.050	0.100	0.050	0.101*	0.050	0.100	0.023	0.048	0.031	0.072
permutation	0.050	0.100	0.048	0.098	0.049	0.099	0.047	0.098	0.026	0.067	0.047	0.103*
rank-sum	0.046	0.088	0.044	0.085	0.038	0.092	0.039	0.083	0.043	0.085	0.042	0.082
Empirical Bayes												
t-test	0.047	0.097	0.044	0.090	0.048	0.100	0.047	0.097	0.006	0.013	0.008	0.023
permutation	0.042	0.083	0.043	0.093	0.041	0.084	0.045	0.099	0.048	0.123*	0.055*	0.121*
rank-sum	0.049	0.095	0.042	0.086	0.048	0.094	0.046	0.095	0.045	0.090	0.037	0.077
Target-decoy												
t-value,49	0.047	0.097	0.050	0.100	0.047	0.095	0.049	0.100	0.043	0.094	0.048	0.099
t-value,1	0.046	0.096	0.048	0.098	0.045	0.096	0.049	0.100	0.042	0.092	0.047	0.096
rank-sum,49	0.049	0.099	0.049	0.099	0.045	0.096	0.050	0.100	0.042	0.090	0.050	0.100
rank-sum,1	0.048	0.100	0.049	0.099	0.048	0.097	0.050	0.100	0.040	0.089	0.047	0.096

In this part, we present the simulation results for the simplified target-decoy procedure on dependent random variables. Table 3.3 shows the realized FDRs of different methods with dependent random variables while the specified FDR control level α was 5% or 10%. The results show that the t-test with empirical Bayes overestimated the FDRs for the gamma distribution. The realized FDRs of pooled permutation significantly exceeded α when the random variables

followed the gamma distribution. The Wilcoxon rank-sum test with Bayes or empirical Bayes overestimated the FDRs. The target-decoy approach controlled the FDR in all cases.

Table 3.4. Power with dependent random variables. The sample size is 20. All the cases where realized FDRs exceed α as shown in Table 3.3 are labelled with $*$.

	Normal,$\rho = 0.4$				Normal,$\rho = 0.8$				Gamma			
	1%		10%		1%		10%		1%		10%	
α	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
Bayes												
t-test	82*	90*	927	1016	101	108*	1047	1109	40	50	687	797
permutation	82	90	922	1012	101	108	1043	1106	42	55	737	861*
rank-sum	80	87	907	983	98	106	1031	1086	47	59	735	836
Empirical Bayes												
t-test	81	90	914	999	100	108	1043	1105	23	32	536	652
permutation	81	87	912	1003	99	106	1041	1108	49	66*	757*	893*
rank-sum	80	88	900	984	99	107	1040	1102	47	59	716	823
Target-decoy												
t-value,49	81	90	926	1015	100	108	1046	1109	44	60	741	852
t-value,1	81	89	923	1013	100	108	1046	1109	45	60	735	845
rank-sum,49	80	89	917	1007	99	107	1045	1108	42	59	756	870
rank-sum,1	80	89	916	1005	99	107	1044	1108	41	59	749	863

Table 3.4 shows the statistical power of different methods with dependent random variables. When the random variables followed the normal distribution, the Bayes method was less powerful than the target-decoy approach while the Wilcoxon rank-sum test was used. Though the Bayes method seems to be a little more powerful than the target-decoy approach while the t-test was used, the realized FDR of this method exceeded the specified FDR control level. The empirical Bayes method was less powerful than the Bayes method and our target-decoy approach in the Normal,10%, $\rho = 0.4$ experiments.

When the random variables followed the gamma distribution, the target-decoy approach was much more powerful than the Bayes and empirical Bayes methods, even if only one permutation was performed. Though the pooled permutation seems to be powerful, the FDRs were not controlled.

Similar to the results for independent random variables, the target-decoy approach performed significantly better than other methods for dependent random variables. It controlled the FDR in all cases without loss of statistical power.

3.4 Simulation for the Adaptive Procedure

To show the effectiveness of the adaptive target-decoy procedure for small datasets, a case-control study involving 200 random variables was simulated. The null hypotheses of 20 random variables were true and the others were false. For each random variable, there were 20 random samples, 10 of which were from the cases and the other 10 were from the controls. The observation values from the cases where the null hypotheses were false followed the $N(4, 1)$ distribution, and all the other observation values followed the $N(0, 1)$ distribution. All the observation values were independent. In the simulation, the cases and the controls of each test were permuted for 49 times and the t-values were used.

As shown in Table 3.5, the adaptive procedure controlled the FDR for all values of α, and

Table 3.5. Realized FDRs and power of the adaptive target-decoy procedure. The realized FDRs were calculated as the means of FDPs.

α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Simplified target-decoy procedure										
FDR	0	0	0	0.006	0.044	0.044	0.044	0.055	0.070	0.087
Power	0	0	0	1	21	21	21	21	22	22
Adaptive target-decoy procedure										
FDR	0.007	0.018	0.026	0.032	0.044	0.049	0.058	0.069	0.079	0.093
Power	13	18	18	19	18	20	21	21	21	22

its power was much larger than the simplified target-decoy procedure for small α.

4 Applications to Real Data

Table 4.1. Power of different methods for *Arabidopsis* microarray data

α	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
Bayes										
t-test	0	5	5	171	322	712	1108	1469	1875	2208
permutation	0	0	0	0	251	1266	2035	2816	3499	4150
rank-sum test	0	0	0	0	0	0	0	0	0	0
Empirical Bayes										
t-test	0	0	0	0	0	0	0	0	0	0
permutation	0	0	0	0	0	0	0	0	0	0
rank-sum test	*	*	*	*	*	*	*	*	*	*
Target-decoy										
t-value	0	0	0	1026	1481	1824	2204	2951	3506	3820
rank-sum test	0	0	0	0	0	0	0	0	0	0

*The R package, 'locfdr', crashed while the Wilcoxon rank-sum test is used.

We applied the target-decoy approach to two real biological datasets, including an arabidopsis genomics dataset and a COVID-19 proteomics dataset. Similar to the simulation experiments, the Bayes method, the empirical Bayes method and our target-decoy approach (the simplified procedure) are compared here. The p-values in the Bayes method and the z-values in the empirical Bayes method were calculated with the Student's t-test, Wilcoxon rank sum test, and the Student's t-test with permutation, respectively. For the Bayes method, two-tailed tests were used. For the empirical Bayes method, we first transformed the FDR control level to the threshold of local fdr and then identified differentially expressed genes/proteins according to the threshold. For the target-decoy approach, the absolute t-values and the test statistics of the Wilcoxon rank sum test were used.

4.1 An Application to Arabidopsis

To determine whether *Arabidopsis* genes respond to oncogenes encoded by the transfer-DNA (T-DNA) or to bacterial effector proteins codelivered by *Agrobacteria* into the plant cells, Lee et al.[33] conducted microarray experiments at 3 h and 6 d after inoculating wounded young *Arabidopsis* plants with two different *Agrobacterium* strains, C58 and GV3101. Strain GV3101 is a cognate of strain C58, which only lacks T-DNA, but possesses proteinaceous virulence (Vir) factors such as VirD2, VirE2, VirE3 and VirF[51]. Wounded, but uninfected, stalks were served as control. Here we just use the 6-d postinoculation data as an example (downloaded from http://www.ncbi.nlm.nih.gov/geo/, GEO accession: GSE14106). The data consisting of 22810

genes were obtained from the C58 infected and control stalks. Both infected and control stalks were with three replicates.

Because it is unknown which genes were really differentially expressed, the realized FDRs cannot be computed here. The power of these methods are compared. In fairness, the sampling numbers were set as $19 = \binom{6}{3} - 1$ in all the experiments, including the pooled permutation and the target-decoy approach. That is, all possible permutations were generated for each gene.

As shown in Table 4.1, no differentially expressed genes were found by the empirical Bayes method or the Wilcoxon rank-sum test. For the Bayes method, the t-test was more powerful than the pooled permutation for small α (≤ 0.05) while the pooled permutation was more powerful for large α (≥ 0.06). The target-decoy approach with t-test was most powerful for $0.04 \leq \alpha \leq 0.09$. The additional genes identified by the target-decoy approach are reliable, because similar numbers of genes, i.e., 785 genes for FDR 0.034, 1427 genes for FDR 0.050 and 2071 genes for FDR 0.065, were reported by a more specific analysis[49].

4.2 An Application to COVID-19

In a study to discover differentially expressed proteins that correlate with the COVID-19 disease, the serums of 118 subjects were sampled, including 28 severe COVID-19 patients, 37 nonsevere COVID-19 patients, 25 non-COVID-19 patients and 28 healthy subjects[43]. The proteins from these serum samples were analyzed with tandem mass spectrometry. From the resulting mass spectra, 791 proteins were successfully identified and quantified, and were subjected to subsequent statistical analysis.

To find differentially expressed proteins specific to the COVID-19 patients, the healthy subjects were first served as the control group and were compared with the other three groups, respectively. The FDR control level of 0.05 was used. The numbers of differentially expressed proteins found by the three FDR methods (Bayes, empirical Bayes and target-decoy) are listed in columns 2 to 4 of Table 4.2. Then, those proteins found in the severe or nonsevere COVID-19 patients but not in the non-COVID-19 patients were regarded as the final set of differentially expressed proteins related to the COVID-19 disease. The numbers of them are listed in column 5 of Table 4.2. As shown, the target-decoy method using t-test had reported 132 proteins, more than those reported by other methods.

Table 4.2. Power of different methods for COVID-19 data

	Severe	Nonsevere	Non-COVID-19	Final	Consistent
Bayes					
t-test	136	49	29	118	104
permutation	48	14	0	50	33
rank-sum test	154	65	66	115	84
Empirical Bayes					
t-test	129	70	36	121	91
permutation	101	67	69	121	27
rank-sum test	6	3	0	9	2
Target-decoy					
t-test	142	62	24	132	104
rank-sum test	0	0	0	0	0

In the original study by [43], 105 COVID-19 related proteins were reported with FDR controlled at 0.05 using the BH method[6]. Here, we compared the proteins found by the three methods with the 105 proteins. The numbers of consistent proteins were listed in column

6 of Table 4.2. Higher numbers probably indicate higher sensitivity and precision. It can be seen that both the target-decoy method and the Bayes method found 104 out of the 105 originally reported proteins when t-test was used. Moreover, the target-decoy method reported 28 additional proteins, which could also be COVID-19 related ones.

5 Conclusion

In this paper, we presented the target-decoy approach to FDR control for multiple hypothesis testing. This approach is free of estimating the null distribution or the null proportion, and can rigorously control the FDR for independent variables. Simulation studies demonstrated its ability in FDR control and higher power than two representative traditional methods. The higher power of the approach was also illustrated by two applications to real biomedical data.

In the target-decoy approach, the scores are only used to determine the labels and ranks of tests, and the statistical meaning of the scores is not required. Therefore, any test statistic can be used, regardless of whether or not its null distribution is known. This flexibility brings the potential to increase the power of multiple testing. In this paper, we only used the t-value and the test statistic of the Wilcoxon rank sum test for a fair comparison with the traditional FDR control methods. In the simulation study, the t-value is more powerful than the statistic of the Wilcoxon rank sum test for the normal distribution and is less powerful for the Gamma distribution. In the applications to real data, no differentially expressed genes or proteins were found by the Wilcoxon rank-sum test, but the t-value performed pretty well. Overall, the t-value is a good choice for the target-decoy procedure. Trying other statistics or engineering specific scoring functions for different types of data is a topic worthy of future research. For example, machine learning-derived feature importance scores can in principle be directly used in our approach.

The adaptive target-decoy procedure chooses an r by data splitting and thus reduces the size of the data used for inference. As shown in Section 3.4, the impact of size reduction is insignificant if the sample size of each variable is moderate. In this case, the adaptive target-decoy procedure can be much more powerful than the simplified procedure. However, if the sample size is very small, the size reduction may diminish the power greatly. How to choose r properly in that case deserves further study.

In this paper, FDR control was proved for independent variables, and only simulation evaluation was performed for dependent variables. The theoretic analysis under dependency will be our future work. Especially, whether permutation-based decoys can lead to FDR control under some kind of dependency is an interesting problem that needs to be addressed.

Moreover, our control theorem is based on the exchangeable hypothesis. This null hypothesis is stronger than the more popular hypothesis that the two groups have the same means. The performance of our approach for the 'equality of means' hypothesis needs further studies.

Finally, our approach can be extended to the pair-matched case-control study by adjusting Step 1 of the target-decoy procedure, i.e., randomly exchange the paired observed values just as the permutation tests for pair-matched study instead of permuting them. The other steps and analyses are the same.

Supplementary Materials. The supplementary material provides the proofs of theorems in the main text.

Software package. The R package for the target-decoy procedure can be downloaded from http://fugroup.amss.ac.cn/software/TDFDR/TDFDR.html.

Acknowledgments. The authors would like to thank Xiaoya Sun for her help in data analysis.

References

[1] Almudevar, A., Klebanov, L.B., Qiu, X., Salzman, P., Yakovlev, A.Y. Utility of correlation measures in analysis of gene expression. *NeuroRx*, 3: 384–395 (2006)

[2] Barber, R.F., Candès, E. J. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43: 2055–2085 (2015)

[3] Barber, R.F., Candès, E.J. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47: 2504–2537 (2019)

[4] Barber, R.F., Cands, E.J., Samworth, R.J. Robust inference with knockoffs. *The Annals of Statistics*, 48: 1409–1431 (2020)

[5] Basu, P., Cai, T.T., Das, K., Sun, W. Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113: 1172–1183 (2018)

[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)*, 57: 289–300 (1995)

[7] Benjamini, Y., Krieger, A.M., Yekutieli, D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika,* 93: 491–507 (2006)

[8] Benjamini, Y., Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics,* 29: 1165–1188 (2001)

[9] Candès, E., Fan, Y., Janson, L., Lv, J. Panning for gold: model-x knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80: 551–577 (2018)

[10] Chow, Y.S., Teicher, H. Probability theory: independence, interchangeability, martingales. Springer Science & Business Media, 2012

[11] Couté, Y., Bruley, C., Burger, T. Beyond target-decoy competition: Stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics. *Analytical Chemistry*, 92: 14898–14906 (2020)

[12] Danilova, Y., Voronkova, A., Sulimov, P., Kertsz-Farkas, A. Bias in false discovery rate estimation in mass-spectrometry-based peptide identification. *Journal of Proteome Research*, 18: 2354–2358 (2019)

[13] Diz, A.P., Carvajal-Rodríguez, A., Skibinski, D.O. Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular & Cellular Proteomics*, 10: M110–004374 (2011)

[14] Efron, B. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99: 96–104 (2004)

[15] Efron, B. Size, power and false discovery rates. *Annals of Statistics*, 35: 1351–1377 (2007)

[16] Efron, B. Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23: 1–22 (2008)

[17] Efron, B. Large-scale inference: empirical Bayes methods for estimation, testing, and prediction. Cambridge University Press, 2012

[18] Efron, B., Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23: 70–86 (2002)

[19] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96: 1151–1160 (2001)

[20] Elias, J.E., Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4: 207–214 (2007)

[21] Emery, K. Controlling the FDR through multiple competition. Ph. D. thesis, The University of Sydney, 2020

[22] Emery, K., Hasam, S., Noble, W.S., Keich, U. Multiple competition-based fdr control and its application to peptide detection. *International Conference on Research in Computational Molecular Biology*, 54–71 (2020)

[23] Emery, K., Keich, U. Controlling the fdr in variable selection via multiple knockoffs. *arXiv:1911.09442* (2019)

[24] Fan, Y., Demirkaya, E., Li, G., Lv, J. Rank: Large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*, 115: 362–379 (2020)

[25] Fan, Y., Lv, J., Sharifvaghefi, M., Uematsu, Y. Ipad: Stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, 115: 1822–1834 (2020)

[26] Gimenez, J.R., Zou, J. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. *Proceedings of Machine Learning Research*, 89: 2184–2192 (2019)

[27] He, K. Multiple hypothesis testing methods for large-scale peptide identification in computational proteomics. Master's thesis, University of Chinese Academy of Sciences, 2013

[28] He, K., Fu, Y., Zeng, W., Luo, L., Chi, H., Liu, C., Qing, L., Sun, R., He, S. A theoretical foundation of the target-decoy search strategy for false discovery rate control in proteomics. *arXiv:1501.00537* (2015)

[29] He, K., Li, M., Fu, Y., Gong, F., Sun, X. A direct approach to false discovery rates by decoy permutations. *arXiv:1804.08222* (2018)

[30] Keich, U., Tamura, K., Noble, W.S. Averaging strategy to reduce variability in target-decoy estimates of false discovery rate. *Journal of proteome research*, 18: 585–593 (2019)

[31] Kerr, K.F. Comments on the analysis of unbalanced microarray data. *Bioinformatics*, 25: 2035–2041 (2009)
[32] Langaas, M., Lindqvist, B.H., Ferkingstad, E. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67: 555–572 (2005)
[33] Lee, C.-W., Efetova, M., Engelmann, J.C., Kramell, R., Wasternack, C., Ludwig-Müller, J., Hedrich, R., Deeken, R. Agrobacterium tumefaciens promotes tumor induction by modulating pathogen defense in arabidopsis thaliana. *The Plant Cell*, 21: 2948–2962 (2009)
[34] Lei, L., Fithian, W. Power of ordered hypothesis testing. *International conference on machine learning*, 48: 2924–2932 (2016)
[35] Levitsky, L.I., Ivanov, M.V., Lobas, A.A., Gorshkov, M.V. Unbiased false discovery rate estimation for shotgun proteomics based on the target-decoy approach. *Journal of proteome research*, 16: 393–397 (2017)
[36] Li, J., Maathuis, M.H. Ggm knockoff filter: False discovery rate control for gaussian graphical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83: 534–558 (2021)
[37] Liu, W., Ke, Y., Liu, J., Li, R. Model-free feature screening and fdr control with knockoff features. *Journal of the American Statistical Association*, to appear (2020)
[38] Liu, W., Shao, Q. Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics*, 42: 2003–2025 (2014)
[39] Meinshausen, N., Rice, J. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34: 373–393 (2006)
[40] Romano, Y., Sesia, M., Cands, E. Deep knockoffs. *Journal of the American Statistical Association*, 115: 1861-1872 (2020)
[41] Sarkar, S.K. Some results on false discovery rate in stepwise multiple testing procedures. *Annals of statistics*, 30: 239–257 (2002)
[42] Scott, J.G., Berger, J.O. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38: 2587–2619 (2010)
[43] Shen, B., Yi, X., Sun, Y., Bi, X., Guo, T. Proteomic and metabolomic characterization of covid-19 patient sera. *Cell*, 182: 59–72 (2020)
[44] Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64: 479–498 (2002)
[45] Storey, J.D. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31: 2013–2035 (2003)
[46] Storey, J.D., Taylor, J.E., Siegmund, D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66: 187–205 (2004)
[47] Storey, J.D., Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100: 9440–9445 (2003)
[48] Strimmer, K. A unified approach to false discovery rate estimation. *BMC bioinformatics*, 9: 1–14 (2008)
[49] Tan, Y.-D., Xu, H. A general method for accurate estimation of false discovery rates in identification of differentially expressed genes. *Bioinformatics*, 30: 2018–2025 (2014)
[50] Tusher, V.G., Tibshirani, R., Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98: 5116–5121 (2001)
[51] Vergunst, A.C., van Lier, M.C., den Dulk-Ras, A., Hooykaas, P.J. Recognition of the agrobacterium tumefaciens vire2 translocation signal by the virb/d4 transport system does not require vire1. *Plant physiology*, 133: 978–988 (2003)
[52] Xie, Y., Pan, W., Khodursky, A.B. A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data. *Bioinformatics*, 21: 4280–4288 (2005)
[53] Yu, C., Zelterman, D. A parametric model to estimate the proportion from true null using a distribution for p-values. *Computational statistics & data analysis*, 114: 105–118 (2017)