Genetics and population analysis

A general framework for powerful confounder adjustment in omics association studies

Asmita Roy¹, Jun Chen ²*, Xiyang Zhang ¹*

¹department of Statistics, Texas A&M University, 155 Ireland Street, College Station, TX 77840, United States
²Division of Computational Biology, Mayo Clinic, 200 1st St. SW, Rochester, MN 55905, United States
*Corresponding authors. Department of Statistics, Texas A&M University, 155 Ireland St., TX 77840, United States. E-mail: zhangxiang@stat.tamu.edu (X.Z.); Division of Computational Biology, Mayo Clinic, 200 1st St. SW, MN 55905, United States. E-mail: chen.jun2@mayo.edu (J.C.)

Abstract

Motivation: Genomic data are subject to various sources of confounding, such as demographic variables, biological heterogeneity, and batch effects. To identify genomic features associated with a variable of interest in the presence of confounders, the traditional approach involves fitting a confounder-adjusted regression model to each genomic feature, followed by multiplicity correction.

Results: This study shows that the traditional approach is suboptimal and proposes a new two-dimensional false discovery rate control framework (2DFDR+) that provides significant power improvement over the conventional method and applies to a wide range of settings. 2DFDR+ uses marginal independence test statistics as auxiliary information to filter out less promising features, and FDR control is performed based on conditional independence test statistics in the remaining features. 2DFDR+ provides (asymptotically) valid inference from samples in settings where the conditional distribution of the genomic variables given the covariate of interest and the confounders is arbitrary and completely unknown. Promising finite sample performance is demonstrated via extensive simulations and real data applications.

Availability and implementation: R codes and vignettes are available at https://github.com/asmita112358/tdfdr.np.

1 Introduction

One central theme of genomic data analysis is identifying genomic features associated with a variable of interest, such as disease status. Due to the constraint of clinical sample collection, the variable of interest is often correlated with other variables, which may confound the associations of interest. For example, when identifying microbiome biomarkers for endometrial cancer, the age of patients acts as a confounder, as older patients tend to have more malignant tumors, and the female reproduction tract microbiome changes with age. Controlling for confounders is crucial for successful validation, cost reduction, and faster translation of discoveries to clinical tests. However, confounder adjustment in genomicscale association tests exacerbates the low statistical power and inflates the type I error rate.

The traditional way of confounder adjustment for high-dimensional association tests is to adjust for confounders for each genomic feature and correct the individual association P-values for multiple testing using FDR control (Benjamini and Hochberg 1995, Storey 2002). However, adjusting confounders for every feature can lead to power loss when confounders only affect a subset of features. To address this, a two-dimensional false discovery rate control procedure (2DFDR) based on linear models was proposed Yi et al. (2021). The 2DFDR procedure screens out irrelevant features using unadjusted statistics (from fitting the unadjusted linear models to each omics feature) and identifies true signals using adjusted (from fitting the confounder-adjusted linear models to each omics feature) statistics, controlling the FDR at the desired level.

Although 2DFDR is an improvement over previous methods, its implementation has limitations. It is only applicable to normally distributed outcomes, while omics studies generate different outcome types. Extending 2DFDR to handle binary or discrete outcomes is desired. Additionally, type I error inflation can occur in certain scenarios, particularly with high confounding effects or a small number of features. Improving the type I error control of 2DFDR in these cases would enhance its robustness and reliability.

To address these limitations, we propose a general framework called 2DFDR+ for integrating confounder adjustment into multiple testing. The 2DFDR+ framework extends the original 2DFDR in several aspects.

1) It relaxes the linear model assumption of Yi et al. (2021), accommodating various outcome types such as continuous, binary, and count outcomes.
2) The marginal test statistic acts as an auxiliary statistic in this method, which improves power by screening out noise.
3) It provides a unified approach to approximate the joint distribution of test statistics under the null hypothesis, eliminating the need for the case-by-case derivation of the joint (asymptotic) distribution of the conditional and marginal independence test statistics of 2DFDR.
4) It enables different methods for estimating the conditional distribution of the covariate given the confounders, including permutation/bootstrap, simulation, MCMC, and conditional GAN.
5) It improves FDR control by explicitly modeling the relationship between the variable of interest and confounders, especially in the presence of strong confounding effects.

Theoretical analysis demonstrates that 2DFDR+ provides asymptotic FDR control and retains the same number of rejections as the corresponding 1D procedure. A unique feature of 2DFDR+ is that it lets the data decide the usefulness/informativeness of the auxiliary statistic. If the auxiliary statistic provides helpful information, 2DFDR+ has significant power gain, otherwise, it reduces to the corresponding 1D procedure. When the FDR is controlled at level \( q \), we can show that in the worst-case scenario, the asymptotic power loss for 2DFDR+ compared to the 1D procedure is at most \( q \); see Supplementary Section S1.

Section 2 describes the problem setup and a two-dimensional (2d) rejection region based on a primary statistic for testing the conditional independence between the omics feature and the covariate of interest given the confounders and an auxiliary statistic for testing the marginal independence between the omics feature and the covariate. Section 3 introduces an oracle FDR-controlling procedure, where the conditional distribution of the covariate given the confounders is assumed to be known. Sections 5 and 6 are devoted to numerical studies and real data analyses, respectively.

2 Problem statement and 2d rejection region

We formulate the feature selection problem by allowing the omics variables to depend on the covariate of interest and confounders arbitrarily. To state the problem and the procedure carefully, suppose we have \( n \) i.i.d. samples \( \{(X_i, Y_i, Z_i)\}_{i=1}^n \) with \( Y_i = (Y_{i1}, \ldots, Y_{im})^\top \) from a population, each of the form \( (X, Y, Z) \), where \( X \in \mathbb{R}^d \), \( Y = (Y_1, \ldots, Y_m) \in \mathbb{R}^m \) and \( Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d \). Here \( Y \) represents a vector of omics features, \( X \) is the covariate of interest, and \( Z \) denotes the set of the confounders. We aim to discover as many as possible omics features \( Y_j \) that are dependent of \( X \) conditionally on the confounders \( Z \). We formulate this as the problem of testing

\[
H_{0j} : Y_j \perp \!\!\!\!\perp X | Z \quad \text{against} \quad H_{1j} : Y_j \not\perp \!\!\!\!\perp X | Z
\]

for \( 1 \leq j \leq m \). To tackle this problem, one must adjust for the confounders and the multiplicity in testing. The burden from both adjustments could lead to potential power loss, especially when the confounding effect is strong.

Our idea to resolve this issue is to use two statistics jointly, namely a primary statistic for testing the conditional independence specified in \( H_{0j} \) and an auxiliary statistic for testing the marginal independence \( Y_j \perp X \), for deciding whether or not to reject \( H_{0j} \). The purpose of using the auxiliary statistic is to enrich signals, reduce the multiple testing burden, and thus enhance the multiple testing power. As marginal dependence does not necessarily imply conditional dependence (e.g., \( Y_j \) and \( X \) are both functions of \( Z \)), the use of auxiliary statistics could lead to selection bias and requires proper adjustment in the selection of cut-off values. One of our goals is to carefully design a way to simultaneously select the cut-off values for the primary statistic and the auxiliary statistic to control the FDR at the desired level.

As a motivation, we consider \( m \) independent generalized linear models:

\[
f(Y_j | X, Z, \mathbf{z}_j, \beta_j, \phi_j) = \exp \left\{ \frac{\theta_j Y_j - b(\theta_j)}{\phi_j} + c(Y_j, \phi_j) \right\},
\]

\[
g(\mathbb{E}[Y_j]) = g(b(\theta_j)) = X^\top \mathbf{z} + Z^\top \beta_j,
\]

for \( 1 \leq j \leq m \), where \( g \) is a known link function, \( \theta_j \) is the canonical parameter, \( \phi_j \) is the dispersion parameter, \( c(Y_j, \phi_j) \) is some function of \( (Y_j, \phi_j) \) and \( \mathbf{z}_j \in \mathbb{R}^p, \beta_j \in \mathbb{R}^d \) are the coefficients associated with the covariate of interest and confounders, respectively. Under the above model, there are four different categories to consider as follows:

- A. Associated with both the covariate of interest and confounders: \( \mathbf{z}_j \neq 0, \beta_j \neq 0 \);
- B. Solely associated with the covariate of interest: \( \mathbf{z}_j \neq 0, \beta_j = 0 \);
- C. Solely associated with the confounders: \( \mathbf{z}_j = 0, \beta_j 
eq 0 \);
- D. Not associated with either the covariate of interest or confounders: \( \mathbf{z}_j = 0, \beta_j = 0 \).

We note that: (i) \( \mathbf{z}_j = 0 \) if and only if \( Y_j \perp X | Z \); (ii) when \( \beta_j = 0 \) (Categories B and D), testing the conditional independence boils down to testing the marginal independence \( Y_j \perp X \). In a general setting, these four categories can be described as: (A) \( Y_j \perp X | Z \); (B) \( Y_j \perp Z | X \); (C) \( Y_j \perp X | Z \); (D) \( Y_j \perp X \).

As a way to enrich signals, we use a marginal independence test to screen out the omics features in Category D and further use a conditional independence test to pick out the true signals from Categories A and B. More precisely, we let \( T^{CM} \) and \( T^{IM} \) be two test statistics computed based on the samples \( \{(X_i, Y_i, Z_i)\}_{i=1}^n \) for testing the conditional independence \( Y_j \perp X | Z \) and the marginal independence \( Y_j \perp X \), respectively. In the discussions below, we assume that a large positive value of \( T^{IM} \) provides evidence against marginal (conditional) independence. The readers are referred to Supplementary Section S4 for some examples of conditional and unconditional independence tests. Given the thresholds, \( t_1, t_2 \geq 0 \), the 2d procedure can be described as follows:

**Dimension 1.** Use the marginal independence test statistics to determine a preliminary set of features
\[
\mathcal{D}_1 = \{1 \leq j \leq m : T^{IM}_j \geq t_1\}.
\]
**Dimension 2.** Reject \( H_{0j} \) for \( T^{CM}_j \geq t_2 \) and \( j \in \mathcal{D}_1 \). As a result, the final set of discoveries is given by
\[
\mathcal{D}_2 = \{1 \leq j \leq m : T^{CM}_j \geq t_1, T^{IM}_j \geq t_2\}.
\]

Although marginal dependence does not imply conditional dependence, it can be leveraged to increase the signal density and reduce multiple testing burden in the second dimension. More precisely, the usefulness of the marginal dependence test is due to

1) the marginal dependence test statistics screen out a large number of noises in Category D and thus ease the multiple testing burden in the second dimension;
the marginal dependence test statistics are more effective in detecting signals from Category B as the conditional dependence test causes over-adjustment, reducing the signal strength.

We illustrate the rational behind 2DFDR through the following example. A detailed description of the method is provided in the next section.

**Example 1.** Consider the following data generating process:

\[ Y_j \sim \text{Bernoulli}(p_j), \quad \log \left( \frac{p_j}{1 - p_j} \right) = \alpha_j X + \beta_j Z, \]  

where \( X = (\rho Z + \epsilon) / \sqrt{\rho^2 + 1} \) with \( Z \) and \( \epsilon \sim N(0, 1) \), independently. \( \rho \in \{0.1, 0.5, 1\} \) represents weak (+), medium (++), and strong (+++) confounding levels. \( \alpha_j, \beta_j \sim \text{Unif}(-0.7, -0.5) + 0.15 \times \text{Unif}(0.5, 0.7) + 0.7 \times \delta_0 \), independently, where \( \delta_0 \) denotes a point mass at 0. \( T^C_0 \) is the t-statistic for testing \( H_{0,j} : \alpha_j = 0 \) under the logistic model (1), and \( T^M_0 \) is the t-statistic for testing \( H_{0,j} \) under the reduced model by forcing \( \beta_j = 0 \) in model (1). In Fig. 1, we plot the marginal \( (T^M_0) \) statistic against the conditional \( (T^C_0) \) statistic for various confounded scenarios. The standard approach performs (1D) FDR control based on the conditional statistic \( (T^C_0) \) only (we refer it as 1DFDR). When the correlation between the variable of interest and the confounder (denoted as \( \text{cor}(X, Z) = \rho / \sqrt{\rho^2 + 1} \in \{0.1, 0.45, 0.71\} \)) is high, the signals (green) and noises (red) overlap much on \( T^C_0 \). To achieve the desired FDR level, 1DFDR requires a high cutoff (black line). For 2DFDR+, it first uses \( T^M_0 \) to exclude a large number of irrelevant features (horizontal blue line). Next, a lower cutoff (vertical blue line) is used to achieve the same FDR level. As a result, it achieves significant power improvement, and the improvement increases with the correlation between the variable of interest and the confounder.

### 3 Oracle procedure

We introduce an oracle FDR-controlling procedure, where we assume that the conditional distribution of \( X \) given \( Z \), denoted by \( P_{X|Z} \) below, is known. Supplementary Section S3 introduces several ways of estimating this conditional distribution from the observations.

#### 3.1 Estimating the false discovery proportion

Our goal here is to develop a principled way of finding the cutoff values \((t_1, t_2)\) such that the FDR is controlled at a desired level while the number of rejections is as large as possible. Let \( M_0 = \{1 \leq j \leq m : H_{0,j} \text{ is true}\} \) and \( m_0 = |M_0| \) be the set and the number of true null hypotheses, respectively.
Write $\tilde{X} = (X_1, \ldots, X_n)^T$, $\tilde{Y}_j = (Y_{1j}, \ldots, Y_{nj})^T$ and $\tilde{Z} = (Z_1, \ldots, Z_n)^T$. Based on the 2d rejection region, the false discovery proportion (FDP) is given by

$$\text{FDP}(t_1, t_2) = \frac{\sum_{j=1}^m P_0(T_j^M \geq t_1, T_j^C \geq t_2)}{1 \vee \sum_{j=1}^m 1\{T_j^M \geq t_1, T_j^C \geq t_2\}},$$

(2)

where $a \vee b = \max\{a, b\}$ for $a, b \in \mathbb{R}$. Note that the FDP is zero when no rejection is made. We replace the numerator in the definition of FDP($t_1, t_2$) by its conditional expectation with respect to $X$ given $Y_j$ and $Z_j$, which leads to the following approximate upper bound on the FDP:

$$\text{FDP}(t_1, t_2) \approx \frac{\sum_{j=1}^m P_0(T_j^M \geq t_1, T_j^C \geq t_2|\tilde{Y}_j, \tilde{Z})}{1 \vee \sum_{j=1}^m 1\{T_j^M \geq t_1, T_j^C \geq t_2\}},$$

(3)

$$:= \text{FDP}_{\text{oracle}}(t_1, t_2),$$

(4)

where $P_0(\tilde{Y}_j, \tilde{Z})$ denotes the conditional probability under the null hypothesis $H_{0j}$. The upper bound FDP$_{\text{oracle}}(t_1, t_2)$ relies on the conditional distribution $P_{X|Z}$. To find a feasible conservative estimator of the FDP, it remains to estimate the conditional probabilities in the numerator of FDP$_{\text{oracle}}(t_1, t_2)$. To this end, we write $T_j^M = T_j^M(X, \tilde{Y}_j)$ and $T_j^C = T_j^C(X, \tilde{Y}_j, \tilde{Z})$ to emphasize their dependence on the samples. As $P(X|\tilde{Y}_j, \tilde{Z}) = P(X|Z)$ under $H_{0j}$, and $P(X|Z) = \prod_{i=1}^n P_{X|Z}(X|Z)$, we have under $H_{0j}$ that

$$P_0(T_j^M(\tilde{X}, \tilde{Y}_j)) \geq t_1, T_j^C(\tilde{X}, \tilde{Y}_j, \tilde{Z}) \geq t_2|\tilde{Y}_j, \tilde{Z}) = \mathbb{E}[1\{T_j^M(\tilde{X}, \tilde{Y}_j) \geq t_1, T_j^C(\tilde{X}, \tilde{Y}_j, \tilde{Z}) \geq t_2\}|\tilde{Y}_j, \tilde{Z}].$$

$$= \int 1\{T_j^M(\tilde{x}, \tilde{y}_j) \geq t_1, T_j^C(\tilde{x}, \tilde{y}_j, \tilde{z}) \geq t_2\}dP_{X|Z}(x|z),$$

where $\tilde{x} = (x_1, \ldots, x_n)^T$ with $x_i \in \mathbb{R}$, which can be calculated once we know the conditional distribution $P_{X|Z}$. One way to approximate $P_0(T_j^M \geq t_1, T_j^C \geq t_2|\tilde{Y}_j, \tilde{Z})$ is via Monte Carlo simulation. Specifically, we generate $X_{j, b} \overset{\text{iid}}{\sim} P_{X|Z}(\cdot|Z_i), i = 1, 2, \ldots, n, b = 1, 2, \ldots, B$. In practice, this distribution is unknown. Details on how to estimate this can be found in Supplementary Section S3. Denote by $T_{j, b}^M$ and $T_{j, b}^C$ the marginal and conditional independence test statistics computed based on $(X_{j, b}, Y_j, Z)$ with $X_{j, b} = (X_{1j, b}, \ldots, X_{nj, b})$, respectively. We propose to estimate $P_0(T_j^M \geq t_1, T_j^C \geq t_2|\tilde{Y}_j, \tilde{Z})$ by

$$\hat{F}_{h,b}(t_1, t_2) := \frac{1}{B} \sum_{b=0}^B 1\{T_{j, b}^M \geq t_1, T_{j, b}^C \geq t_2\}$$

with $(T_{j, b}^M, T_{j, b}^C) = (T_j^M, T_j^C)$. Hence a conservative estimate for the FDP is given by

$$\hat{\text{FDP}}(t_1, t_2) := \frac{1}{1 \vee \sum_{j=1}^m 1\{T_j^M \geq t_1, T_j^C \geq t_2\}}.$$

3.2 Finding the optimal cut-off

We now introduce a greedy approach to select the cut-offs. For a desired FDR level $q$, we first define

$$F_q = \{(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+ : \hat{\text{FDP}}(t_1, t_2) \leq q\}$$

as the feasible set that contains all the cut-off values controlling the FDP estimate at the level $q$. We then select the optimal cut-off as the one delivering the most number of rejections from the feasible set:

$$(t_1^*, t_2^*) = \arg\max_{(t_1, t_2) \in F_q} \sum_{j=1}^m 1\{T_j^M \geq t_1, T_j^C \geq t_2\}.$$ 

Finally, we reject all the hypotheses $H_{0j}$ such that

$$T_j^M \geq t_1^* \text{ and } T_j^C \geq t_2^*.$$ 

A summary of the method can be found in Algorithm 1. We remark that the parameter $\nu$ controls the searching accuracy of the procedure.

Remark 1. In the Supplementary Material, we describe a variant of the 2d procedure (2d-FWER+) to control the family-wise error rate (FWER). Simulation studies suggest that 2d-FWER+ provides reliable FWER control in finite sample.

3.3 Estimating the null proportion

Following the idea in Storey (2002), we can further improve the power of our method by estimating the proportion of null
A general framework

on the estimated conditional distribution, we have

\[
\sum_{i=1}^{m} \mathbb{P}(T_i^C \leq \lambda) = \sum_{i=1}^{m} \mathbb{P}(T_i^C \leq \lambda) \geq \pi_0 \mathbb{P}_0(T_i^C \leq \lambda),
\]

where the approximation is due to the law of large numbers. Therefore, we propose to estimate the null proportion \(\pi_0\) by

\[
\hat{\pi}_0(\lambda) = \frac{1}{B} \sum_{b=0}^{B-1} \left\{ T_{i,b} \leq \lambda \right\},
\]

\[
F_{i,B}(\lambda) := \frac{1}{B} \sum_{b=0}^{B-1} \left\{ T_{i,b} \leq \lambda \right\}.
\]

We can then implement the 2DFDR+ based on the following estimate of the FDP:

\[
\text{FDP}_+(t_1, t_2) = \frac{\hat{\pi}_0(\lambda) \sum_{i=1}^{m} T_{i,B}(t_1, t_2)}{\sum_{i=1}^{m} \left\{ T_i^M \geq t_1, T_i^C \geq t_2 \right\}}.
\]

We can then implement the 2DFDR+ based on the following estimate of the FDP:

\[
\text{FDP}_+(t_1, t_2) = \frac{\hat{\pi}_0(\lambda) \sum_{i=1}^{m} T_{i,B}(t_1, t_2)}{\sum_{i=1}^{m} \left\{ T_i^M \geq t_1, T_i^C \geq t_2 \right\}}.
\]

which can be regarded as John Storey’s version of the 2DFDR+ procedure.

4 FDR control

To state the main theorem, we use a list of assumptions defined and justified in the Supplementary Material. The theorem below establishes the asymptotic FDR control of the 2DFDR+ procedure.

**Theorem 1.** Under Assumptions 1–3 in the Supplementary Material and as \(B \to +\infty\),

\[
\limsup_{n,m,B \to +\infty} \mathbb{E} \left[ \text{FDP}(t_1, t_2) \right] \leq q,
\]

where \(\text{FDP}(t_1, t_2)\) is defined in Equation (2).

In practice, \(\mathbb{P}_{XZ}\) is often unknown and has to be estimated from the data; see Supplementary Section S3 for more details. The following result shows that under suitable assumptions on the estimated conditional distribution, \(\mathbb{P}_{XZ}\), the FDR can still be controlled at the target level.

**Corollary 1.** Let \(\tilde{F}_{i,B}(t_1, t_2)\) be the corresponding value of \(F_{i,B}(t_1, t_2)\) based on \(\{X_{i,b} : 1 \leq i \leq n, 1 \leq b \leq B\}\) sampled from \(\mathbb{P}_{XZ}\). Define

\[
\tilde{\text{FDP}}(t_1, t_2) = \frac{\sum_{i=1}^{m} \tilde{F}_{i,B}(t_1, t_2)}{\sum_{i=1}^{m} \tilde{F}_{i,B}(t_1, t_2)} \left\{ T_i^M \geq t_1, T_i^C \geq t_2 \right\},
\]

and the corresponding cutoffs as:

\[
(\hat{t}_1, \hat{t}_2) = \operatorname{argmax}_{(t_1, t_2) \in \mathbb{R}^2} \sum_{i=1}^{m} \mathbb{1}\{T_i^M \geq t_1, T_i^C \geq t_2\}.
\]

Then under Assumptions 1–5 in the Supplementary Material,

\[
\limsup_{n,m,B \to +\infty} \mathbb{E} \left[ \text{FDP}(\hat{t}_1, \hat{t}_2) \right] \leq q.
\]

Additional theoretical results on FWER control and power analysis for the algorithm have been relegated to the Supplementary Material.

5 Numerical studies

5.1 Simulation setting

We conduct comprehensive simulations to evaluate the performance of 2DFDR+ and compare it to competing methods.

We vary the following factors in the simulations:

1) Degree of confounding: Let \(\rho\) determine the strength of association between \(X\) and \(Z\), \(\rho \in \{0.1, 1.1, 1.5\}\) roughly corresponds to weak (+), medium (++), and strong (+++) confounding, respectively. See Supplementary Section S8 for the role of \(\rho\) in each simulated model.

2) Signal density: \(\pi \in \{5\%, 10\%, 20\%\}\) represents low, medium, and high signal density, respectively.

3) Signal effect: \(l \in \{0.2, 0.3, 0.4\}\) represents weak, moderate, and strong effect, respectively.

We report the empirical FDR and power averaged over 100 simulation runs for all possible combinations of the three factors.

5.2 Competing methods

We compare the finite sample performance of the following seven methods.

1) MS-1DFDR: The 1D procedure based on the \(t\)-statistics for testing \(z_i = 0\) under the full model (see the detailed descriptions of each data generating model in Supplementary Section S8). The 1D procedure is essentially the same as the 2DFDR+ procedure, except that instead of a 2d rejection region, we are searching for a cutoff along a single dimension, namely that of the conditional statistic. The statistics used in this 1D procedure is the model-based statistic, i.e. the \(z\)-statistic (or \(t\)-statistic, depending on the model) corresponding to the coefficient of \(X\) for a full model fit.

2) RV-1DFDR: The 1D procedure based on the conditional RV coefficient. To account for the potential nonlinearity in the underlying relationship between \(X\) and \(Z\) (and similarly, \(Y\) and \(Z\)), the residuals obtained from a cubic spline regression of \(X\) on \(Z\) (and similarly, \(Y\) on \(Z\)) have been used in the calculation of the conditional RV coefficient.
3) HSIC-1DFDR: The 1D procedure based on the cHSIC described in Supplementary Section S4.2.2.

4) 2DFDR: The 2DFDR procedure proposed in Yi et al. (2021), which is based on linear models with the measurement of the omics feature as the outcome and the covariate of interest and confounders as the predictors.

5) MS-2DFDR+: The proposed 2DFDR+ procedure with $T_C^j$ and $T_M^j$ being the t-statistics for testing $a_j = 0$ under the full model and reduced model as described in Supplementary Section S4.1.

6) RV-2DFDR+: The proposed 2DFDR+ procedure with $T_C^j = cRV(X, Y; Z)$ and $T_M^j = RV(X, Y)$ which denote the sample estimates of the conditional and the unconditional RV coefficients, respectively. As before, to account for the potential nonlinearity in the underlying relationship between X and Z (and similarly, Y and Z), the residuals obtained from a cubic spline regression of X on Z (and similarly, Y on Z) have been used in the calculation of the conditional RV coefficient.

7) HSIC-2DFDR+: The proposed 2DFDR+ procedure with $T_C^j = cHSIC(X, Y; Z)$ and $T_M^j = HSIC(X, Y)$, where we set $\epsilon = 0.001$ and used the Gaussian kernel with the bandwidth parameter chosen using the median heuristic (Garreau et al. 2017).

The 1D procedure can be viewed as a special case of the corresponding 2d procedure by forcing the cutoff of the auxiliary statistic to be zero. As the 2d procedure is searching over a larger rejection region (by allowing the cutoff of the auxiliary statistic to be greater than zero and meanwhile lowering the cutoff for the primary statistic), the proposed 2d procedure is guaranteed to make more rejections in finite sample.

5.3 Data generating processes

Throughout the simulations, we set the sample size $n$ to be 100, and the number of hypotheses $m$ (i.e. the number of features) to be 1000. In the main paper, we consider two cases, remaining cases and detailed models have been discussed in Supplementary Sections S8 and S9.

- **Linear/Nonlinear models with continuous X and Z.**

  $Y_j = x_j f(X) + \beta_j g(Z) + \epsilon_j$, $\epsilon_j \sim N(0, 1)$, $j = 1, \ldots, m$. \hspace{1cm} (6)

X and Z are associated with each other through the following model:

$X \sim N(\rho b(Z), 1)$, $Z \sim N(0, 1)$, \hspace{1cm} (7)

where $\rho$ controls the degree of confounding and $b : \mathbb{R} \rightarrow \mathbb{R}$ is a possibly nonlinear function. We rescale X to dissociate any possible entanglement between signal strength and the degree of confounding. This type of simulation setup has been used in Models 1–4 to explore the effect of the relations among X, Y, and Z on the FDR and power. The empirical FDR and power of RV-1DFDR, HSIC-1DFDR, 2DFDR, RV-2DFDR+, and HSIC-2DFDR+ are summarized in Figs 2 (nonlinear) and 4 (linear) and again in Supplementary Figs S2 and S3. MS-1DFDR and MS-2DFDR+ have not been included under this scenario because the statistics associated with these procedures are directly proportional to the statistics in RV-1DFDR and RV-2DFDR+, respectively.

- **Binary response.** The following logistic regression model has been considered:

  $Y_j \sim Bernoulli(p_j), \quad \log \left( \frac{p_j}{1 - p_j} \right) = x_j X + \beta_j Z$. \hspace{1cm} (8)

with $X \sim N(0, 1)$ and $Z \sim N(0, 1)$. We implement the MS-1DFDR, RV-1DFDR, MS-2DFDR+, and RV-2DFDR+, and report the results in Fig. 3. In MS-2DFDR+, $T_C^j$ is the statistic for testing $a_j = 0$ under the full model (8) and $T_M^j$ is the statistic for testing $a_j = 0$ by forcing $\beta_j = 0$ in Equation (8).

In Supplementary Section S9, we report some additional numerical results under the following scenarios: (1) Linear/ nonlinear models with discrete X and continuous Z; (2) Linear models with discrete X and Z; (3) Count response; (4) FWER control; (5) global null; (6) dependent errors, and (7) separating the effects of the densities of the signal of interest and the confounder signal.

5.4 Simulation results

We now discuss the major simulation findings under the scenario described in the previous section. Full simulation results are summarized in Figs 2–4 and Supplementary Figs S2–S16. For continuous X and Z, when the underlying models between Y and (X, Z), and X and Z are both linear (see Fig. 4), all the methods provide tight FDR control except for the RV-2DFDR which has slight FDR inflation in some instances when the confounding effect is strong. In contrast, the proposed RV-2DFDR+, which is equivalent to MS-2DFDR+, controls the FDR at the target level across all cases, indicating more robustness of the proposed method than the original 2DFDR. In terms of power, we observe that the power decreases as the confounding effect becomes stronger for all procedures. The 2d procedure is comparable to the 1D counterpart when the confounding effect is weak but is substantially more powerful when the confounding effect is strong. We also observe that RV-2DFDR+ is comparable to 2DFDR and is more powerful than HSIC-2DFDR+. When the underlying model is nonlinear, 2DFDR suffers from severe FDR inflation (Fig. 2a). In contrast, 2DFDR+ controls the FDR at the target level across different cases. Among the 2DFDR+ variants, RV-2DFDR+ delivers the highest power in most cases. Similar to the linear case, the power decreases (Fig. 2b) as the degree of confounding increases. The 2d procedure is again comparable to its 1D counterpart when the degree of confounding is weak, but the power improvement is more apparent as the degree of confounding increases, especially under low signal density and weak effect, which is often the case in real-world data. For binary outcome (Fig. 3), 2DFDR is not applicable. As before, RV-2DFDR+ and MS-2DFDR+ show substantial improvement in power over their 1D counterparts while keeping the FDR under control. The power decreases as the degree of confounding increases, but the increase in power is higher for lower signal density.
Figure 2. Empirical FDR (a) and power (b) for HSIC-1DFDR, RV-1DFDR, 2DFDR, HSIC-2DFDR+, RV-2DFDR+ for $Y_i = \gamma_i X^3 + \beta_i \epsilon^2 + \epsilon_i$ and $X \sim N(\mu Z^2, 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.
Figure 3. Empirical FDR (a) and power (b) for MS-1DFDR, RV-1DFDR, MS-2DFDR+, RV-2DFDR+ for $Y_i \sim \text{Bernoulli}(1 + \frac{e^{-f(X, Z)}}{C_0})$, where $f_j(X, Z) = \alpha_j X + \beta_j Z$, $X \sim N(\mu X, 1)$ and $Z \sim N(0, 1)$. Error bars represent the 95% CIs. The horizontal line in (a) indicates the target FDR level of 0.05.
Figure 4. Empirical FDR (a) and power (b) for HSIC-1DFDR, RV-1DFDR, 2DFDR, HSIC-2DFDR+, RV-2DFDR+ under the model $Y_j = \alpha_j X + \beta_j \epsilon + \epsilon_j$ and $X \sim N(\mu X, 1)$, where $Z \sim N(0, 1)$. Error bars represent the 95% CIs and the horizontal line in (a) indicates the target FDR level of 0.05.
6 Real data analysis

6.1 Microbiome data

In the first example, we analyze a microbiome dataset in the R package GUniFrac. Here, we use the data from the left oro-pharynx of 32 nonsmokers and 28 smokers \((n = 60)\). The microbiome composition was profiled using 16S rRNA gene-targeted sequencing and processed using the QIIME bioinformatics pipeline (D’Argenio et al. 2014), resulting in a count table recording the frequencies of 856 detected OTUs (operational taxonomic units). Sex is a confounding factor in this dataset, with more smokers in males (odds ratio equals 2.3). The aim here is to identify smoking-associated OTUs while adjusting sex.

For illustration purposes, the OTU abundances were treated as both continuous and binary outcomes. The results for the binary outcomes are given in the Supplementary Material. We first filtered out the OTUs occurring in less than 10% of the subjects, which resulted in a total of 174 OTUs. The OTU abundance data were then transformed using a center log-ratio transformation, adding a pseudocount of 0.5. The numbers of rejections for varying levels of FDR (ranging from 0 to 0.2) were calculated for the following methods: Benjamini–Hochberg (BH; Benjamini and Hochberg 1995) procedure, 2DFDR, MS-2DFDR+, RV-2DFDR+, MS-1DFDR, RV-1DFDR. The BH procedure was applied to the \(P\)-values corresponding to the tests of significance of the coefficients of insulin resistance (IR) in a linear regression model with the IR and body mass index (BMI) being the predictors. The numbers of rejections at different FDR levels are shown in Fig. 5, panel 1. The trend is consistent with the simulations, where we have observed that the 2DFDR procedure is more powerful than the corresponding 1DFDR procedure and RV-2DFDR+ makes the highest number of rejections. In addition, we produced a Venn diagram (Supplementary Fig. S17) of the rejected features for each method at the FDR level 0.10 to visualize the degree to which the rejected features in various methods overlap. We find that at the level 0.1, MS-2DFDR+ successfully identifies all the seven features identified by the 2DFDR procedure and five additional features.

6.2 Metabolomics data

Next, we consider an Insulin Resistance dataset (Pedersen et al. 2018) where the goal is to identify serum metabolites associated with IR while controlling the effect of the BMI of the individual. A group of 289 nondiabetic Danish adults was recruited for the study, where their IR was estimated by homeostatic model assessment (HOMA-IR) (Matthews et al. 1985). Untargeted metabolome profiles were generated on fasting serum samples, producing measurements on 325 polar metabolites and 876 molecular lipids (collectively called serum metabolites, \(m = 1201\)). The BMI of a subject is a confounding factor as the IR of a subject is largely influenced by the BMI (correlation coefficient \(= 0.57\)). In this example, 2DFDR discovers the largest number of metabolites (481 at 5% FDR), followed by RV-2DFDR+ (432 metabolites at 5% FDR). Both are a significant improvement over RV-1DFDR (333), HSIC-1DFDR (323), and the BH procedure (377). The comparison of the number of rejections versus FDR level for all methods is displayed in Fig. 5, panel 2.

Again, the result generally agrees with the findings from the simulation studies. While 2DFDR is the most powerful in this example, its inflated type I error rate observed in many nonlinear simulation setups raises some concern about the reliability of the rejections solely found by itself. Supplementary Figure S18 shows the Venn diagram of the serum metabolites detected by the different methods and their degree of overlap at FDR = 0.05 is provided. It is interesting to note that while RV-2DFDR+ and 2DFDR have detected 403 metabolites in common, the BH procedure has significantly fewer overlapping metabolites with either of these methods.

6.3 Gene expression data

Finally, we consider a Pouchitis dataset (Morgan et al. 2015), where the goal is to identify gene expressions associated with

![Figure 5](image-url)
patient outcomes in a cohort with ileal pouch-anal anastomosis (IPAA) surgery in the past one year, adjusting for potential confounders such as antibiotics use and sex. The expression levels of 19 908 genes measured in the J-pouch for $n = 74$ candidates were considered. The conditioning variables were sex, smoking status, and antibiotic use in the previous month. The variable of interest is the disease outcome, including FAP (Familial Adenomatous Polyposis), No Pouchitis, Acute Pouchitis, Chronic Pouchitis, and Crohn’s Disease like Inflammation. As the variable of interest is nominal, we did not use the RV coefficients in this case. Figure 5, panel 3 shows the number of genes identified as associated with the disease outcome conditioning on sex, smoking status, and antibiotic usage. At the FDR level of 0.05, the 2DFDR+ identifies the maximum number of genes (2345), followed by MS-1DFDR (1811) and BH procedure (1640), respectively.

### 7 Conclusion

We have proposed a general framework (2DFDR+) for performing multiple hypotheses testing while adjusting for confounding effects. Within this new framework, the conditional distribution of the omics features given the variable of interest and confounders can be arbitrary and completely unknown. The framework is flexible by allowing the joint use of any conditional and marginal independence tests, continuous/binary/count/multivariate responses, and various ways of modeling the conditional distribution of the variable of interest given the confounders. As a general methodology, 2DFDR+ can be applied to multiple types of omics data. In view of the numerical results, we recommend using RV-2DFDR+ (based on the spline-transformed variables) under most scenarios due to its robustness and efficiency. In cases where the RV-based statistics are not applicable, for instance, when either of X, Y or Z are categorical, or when Y is discrete (e.g. originating from a Poisson or Negative Binomial distribution), the model-based statistics are recommended. Table 1 summarizes the statistics that we recommend using under different scenarios.

### Supplementary data

Supplementary data are available at Bioinformatics online.

### Conflict of interest

None declared.

### Funding

This work was supported by the National Institute of Health [R21HG011662, R0GM144351 to J.C. and X.Z.; National Science Foundation [DMS21113359 to X.Z., DMS21113360 to J.C.]; and Mayo Clinic Center for Individualized Medicine to J.C.

### Data availability

The Smoking Microbiome Data is publicly available in the R package GUniFrac. The remaining datasets and the code is available at https://github.com/asmita112358/tdfdr.np.

### References

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol 1995;57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

D’Argenio V, Casaburi G, Precone V et al. Comparative metagenomic analysis of human gut microbiome composition using two different bioinformatic pipelines. Biomed Res Int 2014;2014:1–10. https://doi.org/10.1155/2014/325340

Garreau D, Jitkrittum W, Kanagawa M. Large sample analysis of the median heuristic. 2017;7.

Matthews DR, Hosker JP, Rudenski AS et al. Homeostasis model assessment: insulin resistance and beta-cell function from fasting plasma glucose and insulin concentrations in man. Diabetologia 1985;28:412–9. https://doi.org/10.1007/BF00280883

Morgan XC, Kabakchiev B, Waldron L et al. Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol 2015;16:67. https://doi.org/10.1186/s13059-015-0637-x

Pedersen HK, Forslund SK, Gudmundsdottir V et al. A computational framework to integrate high-throughput ‘omics’ datasets for the identification of potential mechanistic links. Nat Protoc 2018;13:2781–800. https://doi.org/10.1038/s41596-018-0066-z

Storey JD. A direct approach to false discovery rates. J R Stat Soc B Stat Methodol 2002;64:479–98.

Yi S, Zhang X, Yang L et al. 2dFDR: a new approach to confounder adjustment substantially increases detection power in omics association studies. Genome Biol 2021;22:208–18.

---

**Table 1. Recommended statistics under various scenarios.**

| Y   | X   | Z   | $T^B$ and $T^C$                      |
|-----|-----|-----|-------------------------------------|
| C   | C   | C   | RV and cRV                          |
| CtzD| C   | C   | Model-based statistics (GLM)        |
| CtzD| C   | C   | Model-based statistics (GLM)        |
| C   | Ct  | Ct  | Model-based statistics (ANOVA)      |
| C   | Ct  | C   | Model-based statistics (ANCOVA)     |
| Ct  | Ct  | C   | $z^2$ and CMH-statistics            |

* C, continuous variable; Ct, categorical variable; D, discrete variable. 