

METHOD

Open Access



2dFDR: a new approach to confounder adjustment substantially increases detection power in omics association studies

Sangyoon Yi¹, Xianyang Zhang^{1*}, Lu Yang², Jinyan Huang³, Yuanhang Liu², Chen Wang², Daniel J. Schaid² and Jun Chen^{2*}

* Correspondence: zhangxiany@stat.tamu.edu; chenjun2@mayo.edu

¹Department of Statistics, Texas A&M University, College Station, TX 77843, USA

²Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA

Full list of author information is available at the end of the article

Abstract

One challenge facing omics association studies is the loss of statistical power when adjusting for confounders and multiple testing. The traditional statistical procedure involves fitting a confounder-adjusted regression model for each omics feature, followed by multiple testing correction. Here we show that the traditional procedure is not optimal and present a new approach, 2dFDR, a two-dimensional false discovery rate control procedure, for powerful confounder adjustment in multiple testing. Through extensive evaluation, we demonstrate that 2dFDR is more powerful than the traditional procedure, and in the presence of strong confounding and weak signals, the power improvement could be more than 100%.

Keywords: Confounding, False discovery rate, Association testing

Background

High-throughput genomic profiling technologies enable the interrogation of the biological system at different omics levels, generating enormous amounts of omics data [1]. One central task of statistical analysis of omics data is to test the association between omics features and a covariate of interest [2]. The associated omics features, once validated, can provide biological insights into health and disease, act as potential targets for intervention and serve as biomarkers for clinical applications [3, 4]. However, observational omics studies are subject to various types of confounding [5–7]. Confounding arises when the relationship between the primary variable and the omics feature is distorted by some other variable (confounder) due to its association with both. Demographic variables like age, gender, race, and obesity are frequent confounders in omics association studies. For example, in cancer studies, cancer patients are often older than benign controls [8]. In other diseases such as rheumatoid arthritis, the prevalence could differ by gender [9]. Since these demographic variables are known



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

to impact the omics profiles [7], their correlation with the variable of interest could confound the association analysis. Biological heterogeneity, such as the cell mixture in the tissue sample, is also a significant source of confounding. This is because the cell types have distinct omics profiles, and their composition could vary with the variable of interest. For example, the leukocyte composition in the peripheral blood often shifts in the disease condition, and therefore it could severely confound blood-based omics studies [5]. Finally, technical variation, or batch effect, which occurs when the biological samples are not processed or measured together, could strongly confound the associations of interest if the samples are not randomized into the batches [6]. Although the batch confounding can be avoided by careful study design, it is unfortunately prevalent in omics studies [6].

Confounding not only reduces the statistical power by introducing extra variability but also increases the chance of false findings if not properly accounted for. Standard statistical approaches to address confounding include stratification and regression, with the latter being most widely used due to its flexibility [10]. Although adjusting the confounders in the regression model controls false positives, it nevertheless reduces the statistical power. The need for multiple testing correction in omics association analysis further deteriorates statistical power [11]. In the presence of strong confounding, it is not unusual that no significant associations could be recovered after adjusting for confounders and multiple testing. Therefore, improving the power under confounding and multiple testing is a topic of critical importance and could potentially rescue an underpowered study.

Previous methods separate confounder adjustment from multiple testing. The standard approach, which fits a confounder-adjusted regression model for all omics features followed by multiple testing correction such as false discovery rate (FDR) control [12, 13], is used predominantly [14]. However, confounders may affect only a subset of omics features [15–18], and adjusting confounders for every omics feature will be an over-adjustment, leading to substantial power loss. To rescue the power, one naïve idea is to test the significance of the confounder first, and if it is not significant, we exclude the confounder in the regression model. Although this strategy substantially improves the power, controlling the type I error is difficult and heavily depends on the choice of the significance cutoff. We found that this strategy fails to control the type I error properly, even if we use a very lenient cutoff.

In this study, we take a different approach to this problem and integrate the confounder adjustment into multiple testing (FDR control) framework. The new approach uses the statistic from the unadjusted analysis to filter out omics features that are less likely to be associated with the covariate of interest or the confounder. FDR control is then performed based on the adjusted statistic on the remaining features. The challenge here is to account for the dependency between the unadjusted and adjusted statistic so that the FDR is controlled. We provide a robust and powerful procedure, two-dimensional false discovery rate control procedure (2dFDR), which is proved to offer asymptotic FDR control and dominate the power of the traditional procedure.

Results

Overview of the two-dimensional false discovery rate control procedure (2dFDR)

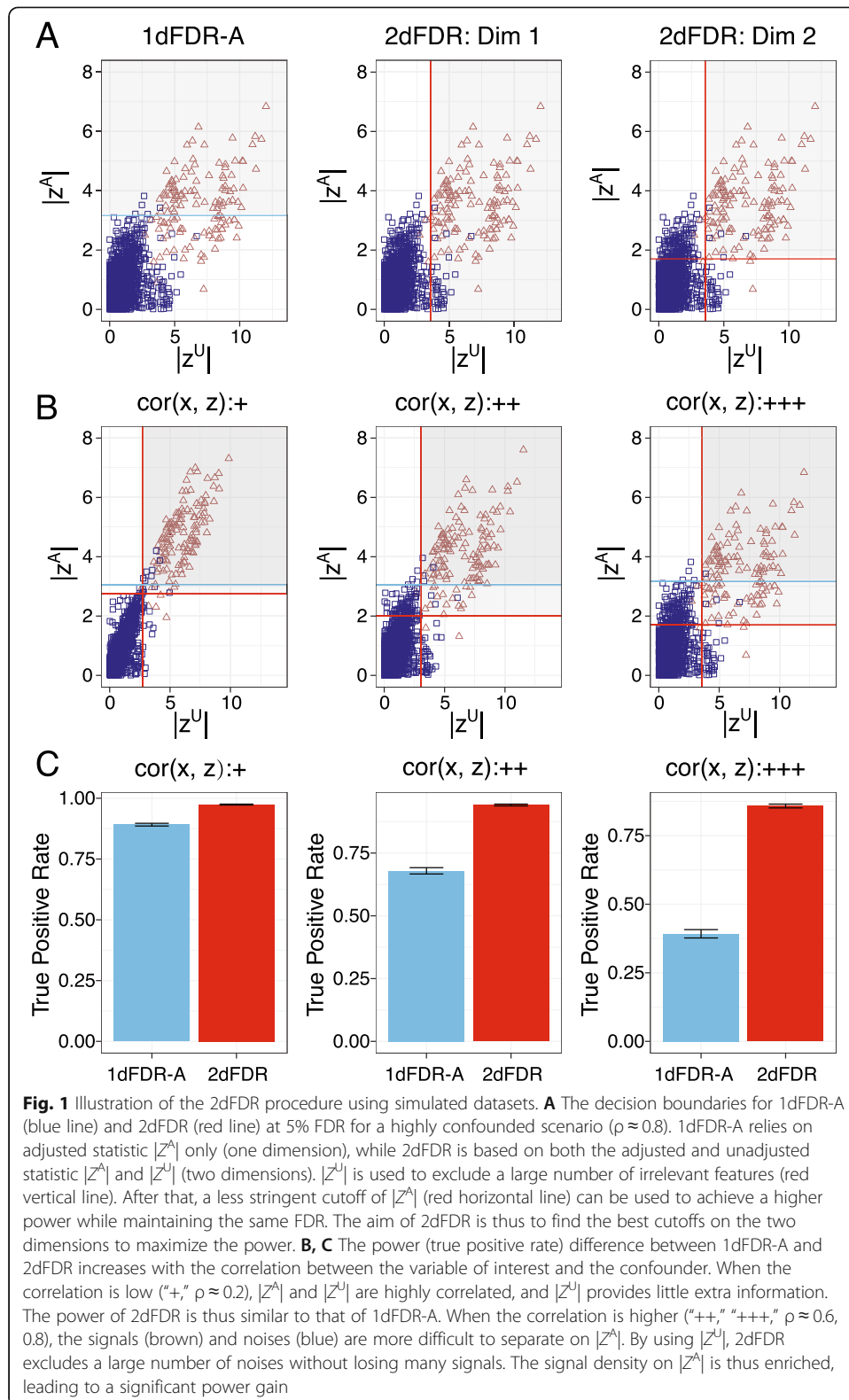
2dFDR is based on linear models with the measurement of the omics feature as the outcome, which is one popular modeling approach for functional omics data, and

assumes the confounders are known. It depends on the unadjusted and adjusted test statistics, denoted by Z_i^U and Z_i^A , respectively, from fitting both the unadjusted and the confounder-adjusted model to the i th omics feature. 2dFDR proceeds in two dimensions. In the first dimension, it uses the unadjusted statistic Z_i^U to screen out a large number of irrelevant features (noises) that are not associated with the covariate of interest or the confounder. In the second dimension, it uses the adjusted statistic Z_i^A to identify the true signals on the remaining features and control the FDR at the desired level. Although the unadjusted statistic is biased and captures the effects from both the covariate of interest and the confounder, it can be leveraged to increase the signal density and reduce multiple testing burden in the second dimension. Thus, 2dFDR boils down to selecting features with $|Z_i^U| \geq t_1$ (first dimension) and $|Z_i^A| \geq t_2$ (second dimension). The cutoffs t_1 and t_2 are chosen to achieve maximum power while controlling the FDR at the desired level. Figure 1A–C illustrate the idea using simulated data (Additional file 1: Note S1), where we plot Z_i^A against Z_i^U for confounded scenarios. The standard approach performs (one-dimensional) FDR control based on the adjusted statistic Z_i^A only (we refer it as 1dFDR-A). When there the correlation between the variable of interest and the confounder (denoted as “cor(x, z)”) is high, the signals (brown) and noises (blue) overlap much on Z_i^A due to loss of power with confounder adjustment (Fig. 1A). To achieve the desired FDR level, 1dFDR-A requires a high $|Z_i^A|$ cutoff (blue line). For 2dFDR, it first uses Z_i^U to exclude a large number of irrelevant features (vertical red line). Next, a much lower $|Z_i^A|$ cutoff (horizontal red 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 (Fig. 1B, C).

A particular challenge for this new approach is to address the dependency between the two dimensions. 2dFDR simultaneously selects t_1 and t_2 and considers the selection effect in the first dimension (“Methods” and Additional file 1: Note S2). 2dFDR can be viewed as a two-dimensional generalization of the classical Benjamini-Hochberg (BH) procedure [12], where we search for the cutoff values in a two-dimensional space. An intrinsic difficulty is to estimate the expected number of false rejections at a given t_1 and t_2 ; this is achieved by a non-parametric Empirical Bayes method [19] (Additional file 1: Note S2.3). We have conducted a thorough theoretical investigation of the proposed procedure and all the theoretical results are included in Additional file 1: Note S3 and S4. Under suitable assumptions, we show that 2dFDR provides asymptotic FDR control (Additional file 1: Note S3), and the power dominates the standard 1dFDR-A (Additional file 1: Note S4).

Simulation studies to evaluate FDR control and power

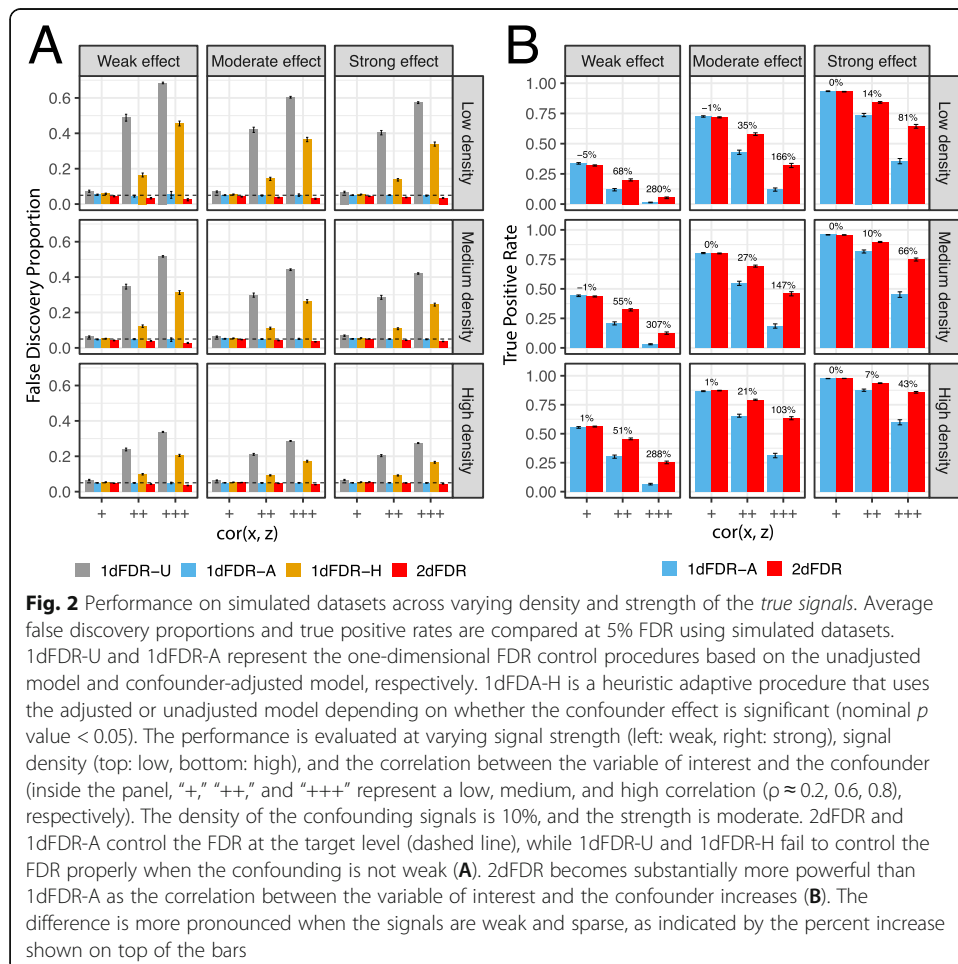
We demonstrate the power and robustness of 2dFDR using comprehensive simulations comparing to 1dFDR-U and 1dFDR-A, two one-dimensional FDR procedures based on the unadjusted and adjusted model, respectively. A heuristic strategy (1dFDR-H), which starts with the adjusted model and uses the unadjusted model if the effect of the confounder is not significant, was also compared. We refer the omics features affected by the variable of interest as “true signals” and the omics features affected by the



confounder as “confounding signals.” For both the true and confounding signals, we use “signal density” and “signal strength” to represent the percentage of features affected and their effect size, respectively.

We first study the performance of 2dFDR under varying signal density, signal strength, and $\text{cor}(x, z)$ (“Methods”). Both 1dFDR-A and 2dFDR controlled the FDR at the target level across settings, while 1dFDR-U and 1dFDR-H failed to control the FDR under a medium or high $\text{cor}(x, z)$ (Fig. 2A). 2dFDR was substantially more powerful than 1dFDR-A when $\text{cor}(x, z)$ was high and was comparable when $\text{cor}(x, z)$ was low (Fig. 2B). The power increase was more pronounced for weak and sparse signals, with a percent increase of more than 100% (Fig. 2B). This is particularly relevant for real applications, where weak signals and strong confounding are the most challenging situation that needs novel methodological developments.

We next study the effect of the confounding signals’ strength and density by varying their magnitudes (Fig. 3) while fixing the true signals’ strength and density. Similarly, 2dFDR maintained the FDR at the target level across settings (Fig. 3A) and was significantly more powerful when $\text{cor}(x, z)$ was medium or high (Fig. 3B). The power difference, however, decreased as the confounding signals became denser (top to bottom). When the confounder affected 50% of the features, 2dFDR could be less powerful than 1dFDR-A even when $\text{cor}(x, z)$ was high (Additional file 2: Figure S1). This is expected since if the confounder affects every omics feature, 1dFDR-A, which adjusts the confounder for every omics feature, is optimal. Higher strength of the confounding signals



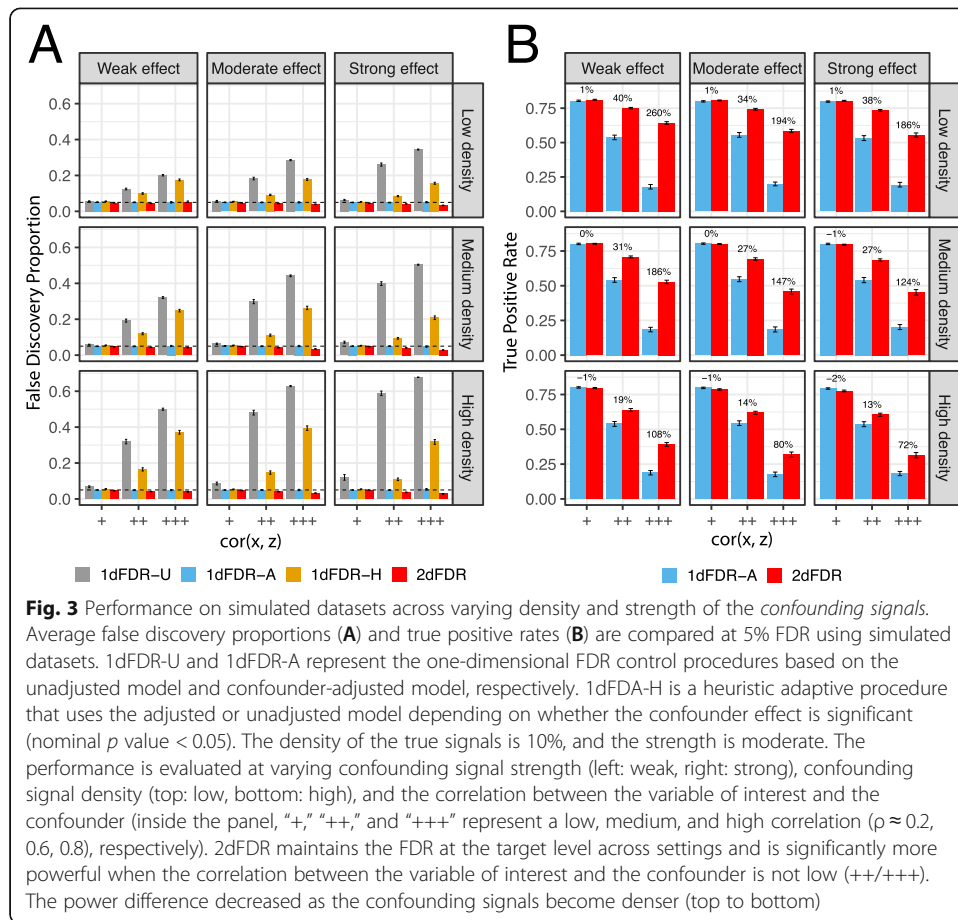


Fig. 3 Performance on simulated datasets across varying density and strength of the *confounding signals*. Average false discovery proportions (**A**) and true positive rates (**B**) are compared at 5% FDR using simulated datasets. 1dFDR-U and 1dFDR-A represent the one-dimensional FDR control procedures based on the unadjusted model and confounder-adjusted model, respectively. 1dFDA-H is a heuristic adaptive procedure that uses the adjusted or unadjusted model depending on whether the confounder effect is significant (nominal p value < 0.05). The density of the true signals is 10%, and the strength is moderate. The performance is evaluated at varying confounding signal strength (left: weak, right: strong), confounding signal density (top: low, bottom: high), and the correlation between the variable of interest and the confounder (inside the panel, “+,” “++,” and “+++” represent a low, medium, and high correlation ($\rho \approx 0.2, 0.6, 0.8$), respectively). 2dFDR maintains the FDR at the target level across settings and is significantly more powerful when the correlation between the variable of interest and the confounder is not low (++/+++). The power difference decreased as the confounding signals become denser (top to bottom)

(left to right) also reduced the power difference. The results remained the same if we simulated five confounders (Additional file 2: Figure S2).

We also studied the effect of colocation between the true and confounding signals, where the omics features were affected by both the variable of interest and the confounder (Additional file 2: Figure S3). We found that 2dFDR was more powerful than 1dFDR-A when the density of the confounding signals was low and $\text{cor}(x, z)$ was high. However, as the confounding signals became denser, 2dFDR could be less powerful than 1dFDR-A when $\text{cor}(x, z)$ was low.

Since 2dFDR is developed based on the assumption that the omics features are independent, it is important to study the robustness of 2dFDR to the correlations among omics features. We thus simulated block and autoregressive correlation structures (“Methods”), which were commonly observed for omics data. We found that 2dFDR was quite robust to these two correlation structures (Additional file 2: Figures S4 and S5), and the FDR was controlled near the target level. 2dFDR maintained the power in these scenarios.

2dFDR offers asymptotic FDR control, i.e., the FDR is proved to be controlled if the sample size and feature size are large. It is interesting to study the sample size and feature size where it breaks down. We thus simulated sample sizes of 50 and 25 (Additional file 2: Figure S6) and feature sizes of 500 and 100 (Additional file 2: Figure S7). We found that the performance 2dFDR remained robust and powerful at the sample

size of 50 and the feature size of 500. However, it became less powerful than the traditional procedure at the sample size of 25. The FDR also started to be inflated at the feature size of 100, especially when $\text{cor}(x, z)$ was high. We also found that increasing the sample size or feature size alone did not rescue the performance deterioration due to the other being small (Additional file 2: Figure S8). Overall, 2dFDR is robust up to a moderate sample size and feature size. For a very small sample or feature size, applying 2dFDR is not recommended.

2dFDR is computationally efficient and can scale up to a large sample and feature size. It can be run in parallel on each grid point to further increase its computational speed. Additional file 2: Figure S9 shows the computational time for running a simulation instance at different sample sizes and feature sizes. With $n = 800$, $m = 512k$ and one confounder, 2dFDR completes the analysis in 546 s using a search grid of 50×50 without parallelization on a MacBook Pro laptop. The memory requirement is the same as fitting regular linear regressions and requires only accommodating a matrix multiplication of $A_{p \times n} B_{n \times m}$, where p , n , and m are the number of covariates, sample size, and feature size, respectively.

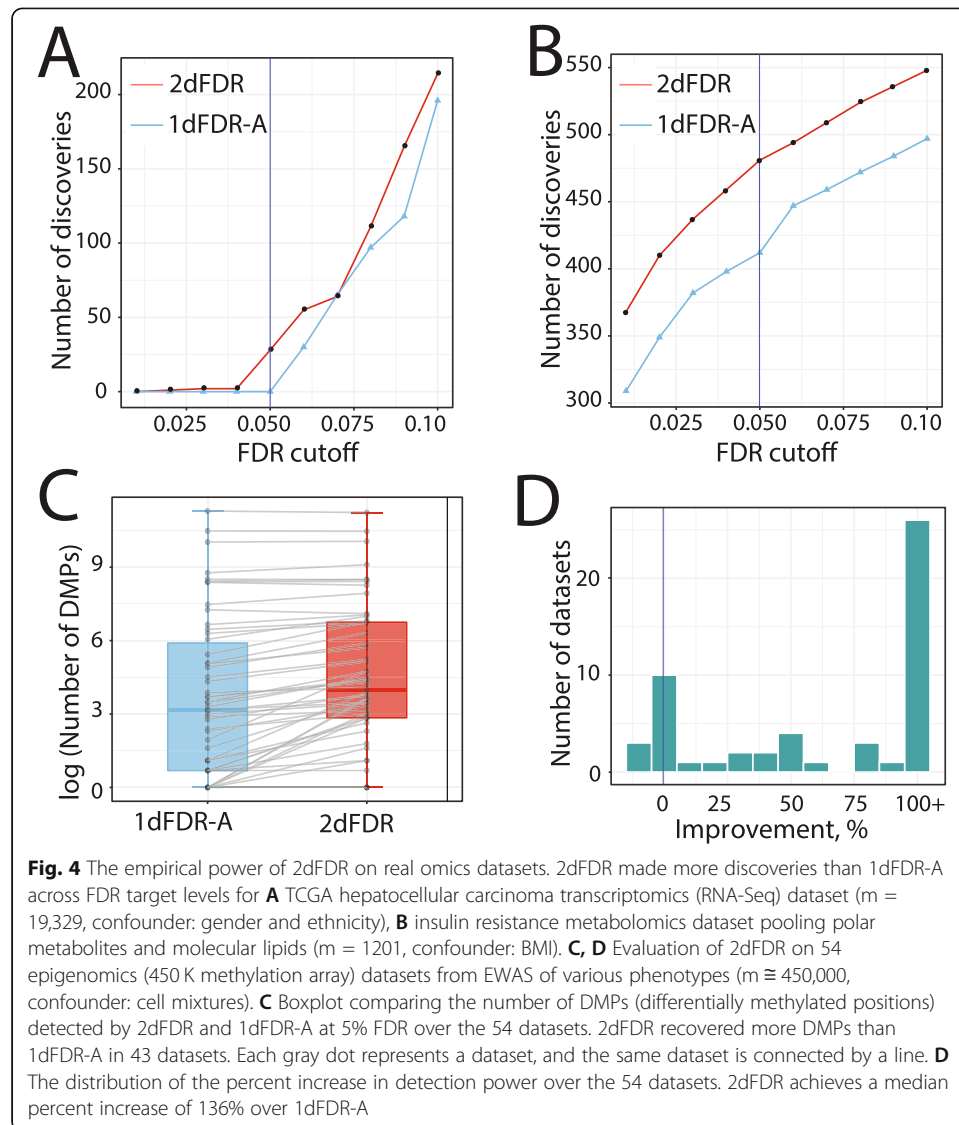
Evaluation of the detection power on real omics datasets

We apply 2dFDR to three different types of omics datasets to demonstrate its empirical power on real data. We compare to the traditional adjusted procedure 1dFDR-A based on the numbers of detected omics features at the same FDR level.

The first is a hepatocellular carcinoma transcriptomics dataset from TCGA [20] ($n = 342$, $m = 19,329$), which is used to detect gene expressions associated with human hepatitis B virus (HBV) infection [21]. Gender and ethnicity are confounders for this dataset and were adjusted in the model. 2dFDR detected more genes than 1dFDR-A across different FDR levels (Fig. 4A). At the standard 5% FDR level, 1dFDR-A failed to identify any HBV-associated genes, while 2dFDR successfully identified 27 genes.

The second is a metabolomics dataset [22, 23] ($n = 289$, $m = 1,201$), where the aim is to identify serum metabolites associated with insulin resistance (IR), accounting for the confounding effect of body mass index (BMI). Again, 2dFDR detected more IR-associated metabolites at different FDR levels (Fig. 4B). At 5% FDR, 2dFDR and 1dFDR-A recovered 481 and 412 metabolites, respectively. 2dFDR was able to identify the majority of the metabolites by 1dFDR-A (378 out of 412) and it also recovered 103 metabolites missed by 1dFDR-A.

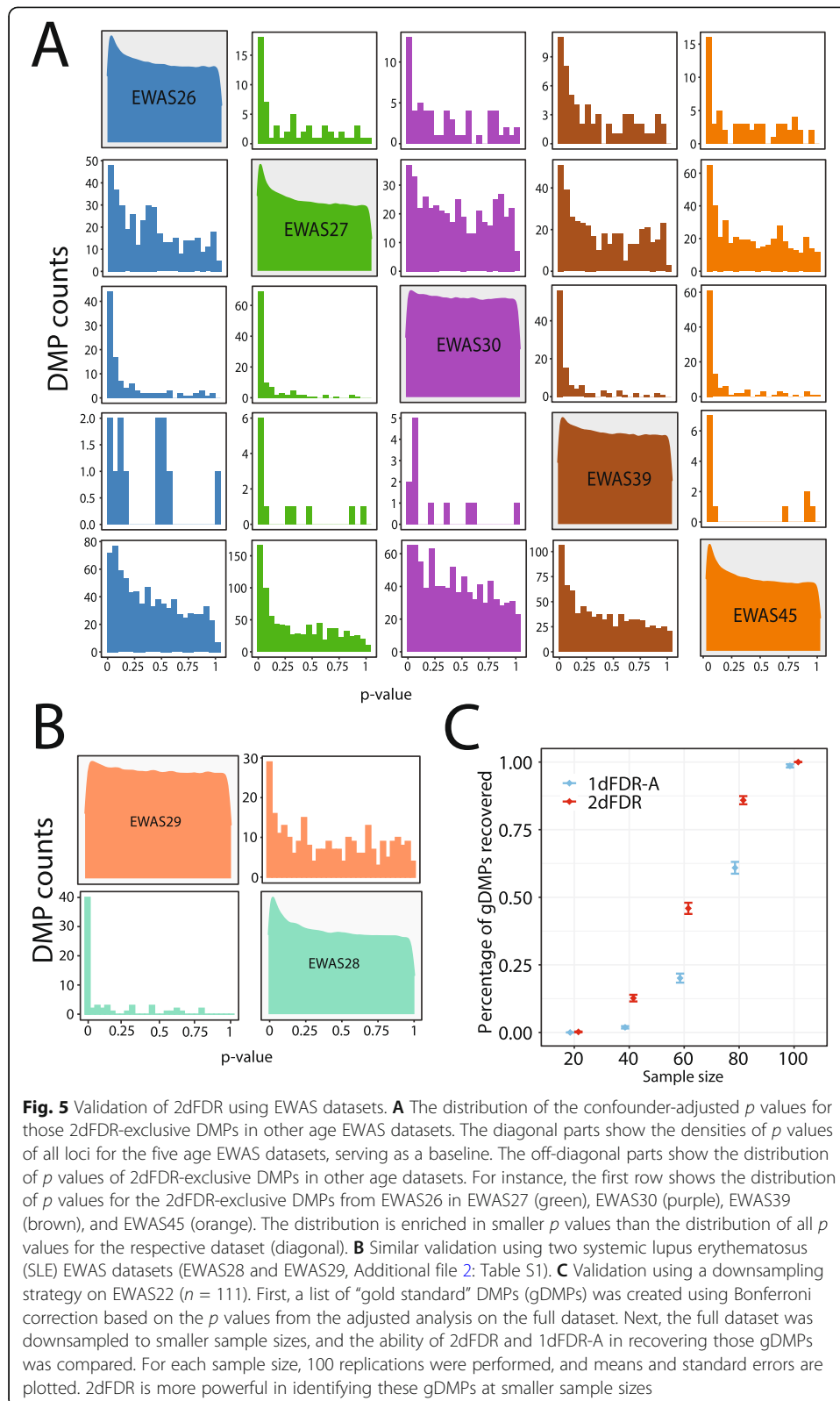
Finally, we benchmark 2dFDR using an extensive collection of epigenomics datasets from various epigenome-wide association studies (EWAS) using tissue samples [24] (Additional file 2: Table S1). The objective is to identify differentially methylated CpG positions (DMPs) associated with a condition of interest. Since a tissue sample contains a mixture of cell types, each with a distinct methylation profile, the covariation of their mixture proportion with the condition could strongly confound the associations of interest [5]. To capture the cell mixture, we used surrogate variable analysis (SVA), and the estimated surrogate variables were adjusted in the model [25]. For these EWAS datasets, 2dFDR detected significantly more DMPs than 1dFDR-A in most datasets with a median increase of 136% (Fig. 4C, D, Additional file 2: Table S1). Consistent with the simulations, the power improvement was more pronounced when the signals were



weak (lower part of the box plot in Fig. 4C). Moreover, 2dFDR was able to detect DMPs in six datasets, where 1dFDR-A failed to identify any.

Validation of the increased detection power on EWAS datasets

To validate the additional DMPs detected by 2dFDR, we resorted to the five age-related EWAS datasets (Additional file 2: Table S1) to see if the additional DMPs from one age dataset had evidence of support from the other four. This was achieved by examining the confounder-adjusted p value distribution of the DMPs detected by 2dFDR only (at 5% FDR) in the other four age datasets. If these DMPs from one age dataset were truly age-associated, we expect to see smaller p values for them in the other age datasets, compared to the p values of random CpG loci. Clearly, the distribution was enriched in small p values, indicating the plausibility of DMPs detected by 2dFDR (Fig. 5A). Validation based on the two SLE datasets reached a similar conclusion (Fig. 5B).



We further performed a downsampling analysis to validate the improved power of 2dFDR. We first curate a list of highly significant features using Bonferroni correction based on the p values from the adjusted analysis on the full dataset. Next, we down-sample the full dataset to smaller sizes and compare the ability of 2dFDR and 1dFDR-A in recovering these highly significant features as a way of power assessment. We illustrate this strategy using one EWAS dataset (EWAS22, $n = 111$), where the genome-wide methylation difference was compared between smokers and non-smokers in African American women using peripheral blood mononuclear cells [26]. With Bonferroni correction ($\alpha = 0.05$) to the p values based on the full dataset, we identified 10 differentially methylated CpG positions (DMPs) and these DMPs were treated as “gold standard” (we label them as “gDMPs”). We then subsampled the full dataset to sample sizes of 20, 40, 60, 80, and 100, and compared the power (recovery rate) of 2dFDR and 1dFDR-A in recovering these gDMPs. We observed that 2dFDR outperformed 1dFDR-A at nearly all sample sizes (Fig. 5C). The power improvement was more significant in the middle range (i.e., $n = 60$ and 80). At the sample size of 100, the recovery rates of both 2dFDR and 1dFDR-A both reached nearly 100%.

Discussion

Confounding and high-dimensionality are the two major statistical challenges in omics data analysis. Previous research separates these two problems, and methodological developments are focused on each of them. In this study, we integrate the confounder adjustment into multiple testing by performing two-dimensional false discovery rate control based on both the adjusted and unadjusted statistics. Although the unadjusted statistic is biased, it can be leveraged to enrich signals and reduce the multiple testing burden. The resulting procedure, 2dFDR, has proven to offer asymptotic FDR control and dominate the power of the traditional procedure based on the adjusted statistic only. Through simulations and real data applications, we demonstrate that 2dFDR is substantially more powerful than the traditional procedure. We also show that 2dFDR is robust to the typical correlation structures seen in omics data and performs well at moderate sample sizes and feature sizes.

The 2dFDR procedure is the most powerful when the correlation between the variable of interest and the confounder is high, and/or the signals are weak. This makes it a practically very useful approach since existing methods have limited power in these scenarios. 2dFDR also works best when the confounder only affects a subset of omics features. This is usually a reasonable assumption for many conditions, such as age and gender [15–18]. However, there could be situations where the assumption is violated. For example, strong batch effects could possibly affect a large number of omics features. In such a case, 2dFDR has a limited power advantage or could be less powerful than the traditional procedure (Additional file 2: Figure S1). One diagnostic approach is to calculate the percentage of the genomic variance explained (R^2) by the variable of interest and the confounder, respectively, using multivariate methods such as PERM ANOVA [27]. If the R^2 of the confounder is substantially larger than that of the variable of interest, it indicates that the confounder signals may be very dense. Another approach is to study the distribution of the p values of the confounder from the adjusted analysis. If we see a spike on the left end of the p value distribution, it also suggests dense confounder signals.

Since 2dFDR depends on the unadjusted statistic to filter features, when the confounder and variable of interest have opposite effects on the omics feature with similar magnitude, they will cancel out each other's effect, and the omics feature could be excluded erroneously in the first dimension. The optimal cutoff of $|Z_i^U|$ is thus determined based on the tradeoff between power reduction due to erroneously excluding these relevant features in the first dimension and power increase due to reducing the multiple testing burden and increasing the signal density in the second dimension. If the true signals can only be revealed after adjusting for the confounder, for example, when the true and confounding signals co-locate with opposite directions, unadjusted statistics will not be informative. In this case, the best cutoff on $|Z_i^U|$ should be 0 and 2dFDR is then reduced to the traditional 1dFDR-A. However, in finite samples, it may not always be possible to reduce 2dFDR to 1dFDR-A exactly and 2dFDR could be less powerful than 1dFDR-A in such situations (Additional file 2: Figure S10A, 10% true and confounding signals co-locate with opposite directions).

When the correlation between the variable of the interest and the potential confounder is small, adjusting the confounder will lead to power improvement if the confounder has large effects on the outcome and will not hurt the power much if the confounder has no effects on the outcome. Therefore, in such situation, performing adjusted analyses is a good choice. In fact, when $\text{cor}(x, z)$ is 0, the test statistic Z_i^U and Z_i^A are almost perfectly correlated (Additional file 2: Figure 10B), and performing two-dimensional FDR control based on Z_i^U and Z_i^A is almost equivalent to performing one-dimensional FDR control based on Z_i^A . Again, in finite samples, there could be some power loss for 2dFDR (Additional file 2: Figure 10B). Fortunately, $\text{cor}(x, z)$ can be known before the analysis. We thus do not recommend running 2dFDR when $\text{cor}(x, z)$ is small.

The proposed method belongs to the general topic of using auxiliary data to increase power in multiple testing, which has been an active research area recently [24, 28–32]. To be effective, the auxiliary data has to be informative of the probability of the null hypothesis or the statistical power. In our case, the unadjusted statistic can be considered as a particular type of auxiliary data, which informs the prior null probability (the smaller the statistic, the more likely the null hypothesis). However, as the auxiliary data (unadjusted statistic) and the primary data (adjusted statistic) are correlated even under the null hypothesis, existing structure-adaptive multiple testing methods are not directly applicable in our problem. 2dFDR explicitly addresses such correlation to enable asymptotic FDR control.

Although developed in the linear model setting, the idea of 2dFDR can be extended to the generalized linear model or generalized linear mixed effects model [33] using the Wald z-statistic. Another interesting extension is to adapt 2dFDR to the setup where the omics features are treated as covariates. For example, in genome-wide association studies (GWAS), the genetic variants are usually modeled as covariates. We expect that the same 2dFDR idea could be applied to GWAS to significantly increase its power, since population stratification, which occurs when samples come from genetically diverse underlying populations, is an important confounder for GWAS [34].

In summary, 2dFDR is a new approach to confounder adjustment under multiple testing. It is powerful, robust, and scalable. As a general methodology, we envision its broad applicability of 2dFDR in omics association studies.

Methods

The 2dFDR procedure

Here we give an overview of the method. Details could be found in Additional file 1: Note S2. Consider the following m linear models

$$E(Y_{ij}) = x_j \alpha_i + z_j^T \beta_i, 1 \leq j \leq n, 1 \leq i \leq m \tag{1}$$

where n and m are the sample size and feature size, respectively, $(Y_{i1}, \dots, Y_{in})^T \in \mathbb{R}^{n \times 1}$ is the measurement of the omics feature i , $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times 1}$ is the vector of the covariate of interest, $(z_1, \dots, z_n)^T \in \mathbb{R}^{n \times d}$ is the design matrix for the confounding factors, and $\alpha_i \in \mathbb{R}$, $\beta_i \in \mathbb{R}^{d \times 1}$ are the coefficients associated with the covariate and confounding factors respectively. Under the model, there are four different categories of features to consider: (A) Solely associated with the covariate of interest: $\alpha_i \neq 0, \beta_i = 0$; (B) Solely associated with the confounder: $\alpha_i = 0, \beta_i \neq 0$; (C) Associated with both the covariate of interest and the confounder: $\alpha_i \neq 0, \beta_i \neq 0$; (D) Not associated with either the covariate of interest or the confounder: $\alpha_i = 0, \beta_i = 0$. Our goal is to develop a multiple testing procedure for simultaneously testing m hypotheses $H_{0,i} : \alpha_i = 0$ vs $H_{1,i} : \alpha_i \neq 0$ ($i = 1, \dots, m$) in the presence of confounding effects. Let Z_i^A be the z-statistic for testing whether $\alpha_i = 0$ after adjusting for the confounding factors, and Z_i^U be the unadjusted version without adjusting for the confounding factors, i.e., setting $\beta_i = 0$ in the model. Given the thresholds $t_1, t_2 \geq 0$, the 2dFDR procedure can be described as follows:

Dimension 1: Signal enrichment. Use the unadjusted statistic Z_i^U to retain more promising features $S_1 = \{1 \leq i \leq m : |Z_i^U| \geq t_1\}$.

Dimension 2: Excluding false positives. For $i \in S_1$, reject $H_{0,i}$ with $Z_i^A \geq t_2$. As a result, the final set of discoveries is given by $S_2 = \{1 \leq i \leq m : |Z_i^U| \geq t_1, |Z_i^A| \geq t_2\}$.

Given $t_1, t_2 \geq 0$, the FDP is defined as

$$\text{FDP}(t_1, t_2) = \frac{\sum_{i=1}^m \mathbf{1}\{|Z_i^U| \geq t_1, |Z_i^A| \geq t_2, \alpha_i = 0\}}{1 \vee \sum_{i=1}^m \mathbf{1}\{|Z_i^U| \geq t_1, |Z_i^A| \geq t_2\}} \tag{2}$$

The key here is to select the thresholds t_1, t_2 to maximize the number of discoveries while controlling the FDR (expectation of FDP) at the desired level.

As the number of false rejections is unknown (the numerator of $\text{FDP}(t_1, t_2)$), we replace it by the expected number of false rejections. We can show that the expected number is given by $\sum_{i:\alpha_i=0} L(\mu_i, t_1, t_2) = P(|V_1 + \mu_i| \geq t_1, |V_2| \geq t_2)$ with μ_i being a nuisance parameter that depends on the confounding effect (the magnitude of β_i and the correlation between x and z) and (V_1, V_2) being bivariate mean-zero normal random variables, whose covariance can be calculated from the data (Additional file 1: Note S2.2). Note that the correlation between V_1 and V_2 is not equal to zero in general, which captures the dependence between the two dimensions.

An intrinsic difficulty here is that the expected number of false rejections depends on the effects of the confounding factors (i.e., μ_i) on each feature. As the number of features could be huge, it thus requires estimating a large number of nuisance parameters. To tackle this challenge, we adopt a Bayesian viewpoint by assuming that the nuisance parameters μ_i are generated from a common prior distribution G . The Bayesian viewpoint allows us to express the expected number of false rejections as a functional of the prior distribution. Therefore, we can translate the task into estimating the prior

distribution instead of the direct estimation of a large number of nuisance parameters. The prior distribution can be estimated via the general maximum likelihood empirical Bayes estimation [19, 35], which can be cast into a convex optimization problem (Additional file 1: Note S2.3). Denote \hat{G} the estimate of G . Then we can derive an approximate upper bound for $\text{FDP}(t_1, t_2)$, as follows:

$$\text{FDP}(t_1, t_2) \leq \frac{m^{-1} \sum_{i=1}^m L(\mu_i, t_1, t_2)}{m^{-1} \sum_{i=1}^m \mathbf{1}\{|Z_i^U| \geq t_1, |Z_i^A| \geq t_2\}} \approx \frac{\int L(\mu, t_1, t_2) d\hat{G}(\mu)}{m^{-1} \sum_{i=1}^m \mathbf{1}\{|Z_i^U| \geq t_1, |Z_i^A| \geq t_2\}} =: \widehat{\text{FDP}}(t_1, t_2). \quad (3)$$

For a desired FDR level $\alpha \in (0, 1)$, we choose the optimal threshold such that

$$(t_1^*, t_2^*) = \operatorname{argmax}_{(t_1, t_2) \in \mathcal{F}_\alpha} \sum_{i=1}^m \mathbf{1}\{|Z_i^U| \geq t_1, |Z_i^A| \geq t_2\}, \quad (4)$$

where $\mathcal{F}_\alpha = \{(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+ : \widehat{\text{FDP}}(t_1, t_2) \leq \alpha\}$. Finally, we select features with $|Z_i^U| \geq t_1^*, |Z_i^A| \geq t_2^*$. This procedure can be viewed as two-dimensional generalization of the Benjamini-Hochberg (BH) procedure [12]. It is well known that when the number of signals is a substantial proportion of the total number of hypotheses, the BH procedure will be overly conservative. To adapt to the signal density, we develop a modification of John Storey’s approach [13] in our setting (Additional file 1: Note S2.4).

Data simulation

We conduct comprehensive simulations to evaluate the finite-sample performance of the proposed method and compare it to competing methods. For genome-scale multiple testing, the number of hypotheses could range from thousands to millions. For demonstration purposes, we start with $m = 10,000$ features, $n = 100$ samples, one covariate of interest \mathbf{x} , and one confounder \mathbf{z} . To comprehensively evaluate the proposed method’s performance under different scenarios, we study the following important parameters:

- The correlation between the covariate of interest (\mathbf{x}) and the confounder (\mathbf{z}) (denoted as “ $\text{cor}(\mathbf{x}, \mathbf{z})$ ”);
- The density and strength of the true signals;
- The density and strength of the confounding signals;
- The degree of colocation of the true signals and confounding signals.

To induce the correlation between \mathbf{x} and \mathbf{z} , we let $\mathbf{x}_0 \sim N(0, 1)$, $\mathbf{x} = c\mathbf{x}_0 + N(0, 1)$, $\mathbf{z} = c\mathbf{x}_0 + N(0, 1)$. We set $c = 0.5, 1.25$ and 2 , which achieves a correlation (ρ) about $0.2, 0.6$, and 0.8 respectively, representing weak (“+”), moderate (“++”) and strong confounding (“+++”) level. For the multiple-confounder scenario, we simulate each \mathbf{z} in the same way.

Next, we generate

$$\mathbf{y}_i = \alpha_i \mathbf{x} + \beta_i \mathbf{z} + \boldsymbol{\epsilon}_i, \quad (i = 1, \dots, m) \quad (5)$$

where $\alpha_i, \beta_i \sim \frac{\pi}{2} \text{Unif}(-l - \delta, -l) + \frac{\pi}{2} \text{Unif}(l, l + \delta) + (1 - \pi) I_0$, I_0 is a mass at 0 and $\boldsymbol{\epsilon}_i \sim N(0, 1)$. For both the covariate of interest and the confounder, we simulate three levels of signal density ($\pi = 5\%, 10\%$, and 20% —representing low, medium, and high density) and three levels of signal strength ($\delta = 0.2$ and $l = 0.2, 0.3$ and 0.4 , representing weak, moderate, and strong effect). In the basic setup, α_i, β_i are independently drawn

from the distribution mentioned above. To study the effect of the collocation of the true and confounding signals (nonzero α_i and β_i), we further simulate two scenarios:

- Non-colocation: $\alpha_i \beta_i = 0$
- 50% colocation: $|S_{\alpha\beta}| = 0.5 \min\{|S_\alpha|, |S_\beta|\}$, where $S_{\alpha\beta} = \{i \mid \alpha_i \beta_i \neq 0\}$, $S_\alpha = \{i \mid \alpha_i \neq 0\}$, $S_\beta = \{i \mid \beta_i \neq 0\}$.

In addition, we investigate the robustness of the proposed method to correlated errors. Specifically, we study:

- First-order autoregressive (AR(1)) correlation structure, where $\rho(\epsilon_i, \epsilon_k) = 0.6^{|i-k|}$.
- Block correlation structure, where we simulate 100 blocks with within-block correlation 0.6.

We compare 2dFDR to the following methods:

- 1dFDR-U: linear regression with the covariate of interest without adjusting for the confounder.
- 1dFDR-A: linear regression with the covariate of interest adjusting for the confounder, which is the traditional procedure.
- 2dFDR-H: a heuristic hybrid procedure, which first runs “1dFDR-A,” and if the confounder is not significant (nominal $p < 0.05$), “1dFDR-U” is used.

All the methods use the q-value approach for FDR control [13], following the computation of feature-wise p values. We evaluate the performance based on FDR control (false discovery proportion) and power (true positive rate) at a target FDR level of 5%. Results are averaged over 100 simulation runs (for small feature sizes, 1000 simulation runs are performed to reduce variability). Both the means and 95% CIs are reported in the bar plots.

Real datasets

Transcriptomics dataset

The dataset consists of 342 RNA-Seq samples from patients with hepatocellular carcinoma from The Cancer Genome Atlas (TCGA) [20]. We use this dataset to identify gene expressions associated with chronic infection of the hepatitis B virus (HBV). HBV status was examined by counting the percentage of reads mapped to HBV genome [36], resulting in 103 and 239 HBV-positive and HBV-negative cases. Ethnicity and gender are confounders in this analysis since they are correlated with the HBV status (OR 0.051 and 2.67, respectively, $p < 0.001$) and are known to affect the gene expressions. The raw FASTQ files were processed through Mayo’s internal MAP-RSeq pipeline (Version 3.0) [37]. The gene counts were generated by FeatureCounts using the gene definitions files from Ensembl v78 [38]. Quality control was carried out using RSeqQC [39], and a total of 19,329 genes were included. Transcript per million (TPM) was calculated, quantile normalized, and log-transformed before analysis.

Metabolomics dataset

The dataset came from a study of the impact of the gut microbiome on host serum metabolome and insulin sensitivity in non-diabetic Danish adults [22]. It consists of measurements of 1201 metabolites (325 serum polar metabolites and 876 serum molecular lipids) on 289 serum samples using mass spectrometry. The cleaned dataset was downloaded from <https://bitbucket.org/hellekp/clinical-micro-meta-integration> [23]. We use this data set to identify insulin resistance (IR)-associated metabolites. IR was estimated by the homeostatic model assessment [22]. Body mass index (BMI) is a confounder for this dataset since it is highly correlated with IR (Spearman's $\rho = 0.67$) and is known to affect the serum metabolome. Two samples without IR measurement were excluded. For metabolites with zero measurements, zeros were replaced by half of the minimal nonzero value. Log transformation was performed to make the data more symmetrically distributed before analysis.

Epigenomics datasets

The datasets came from 51 epigenome-wide association studies (EWAS) of various phenotypes using Infinium Human Methylation 450 K BeadChip. They were collected and processed as previously described [24]. A total of 54 datasets with binary or continuous phenotypes and sample sizes larger than 100 were included in the evaluation (Additional file 2: Table S1). Since these EWAS studies all used tissue samples, which consist of a mixture of different cell types, each with a distinct methylation profile, the shift in the cell mixture proportions with regard to the phenotype of interest could strongly confound the association analysis [5]. For the peripheral blood sample, it consists of different leukocyte subtypes, whose composition usually changes with the onset of disease as a host immune defense mechanism. Since the cell mixture proportions were not directly available for these datasets, we used the surrogate variable analysis [25] to infer the latent factors (also called surrogate variables) that capture the cell mixtures. Specifically, the “isva” R package [40] was used to estimate the number of surrogate variables based on random matrix theory, and the “SmartSVA” R package [18] was used to compute the surrogate variables. The inferred surrogate variables were highly correlated with the phenotypes (a median R^2 0.49, Additional file 2: Table S1) and were adjusted in the regression analysis. All the analysis was performed on the methylation M-values [41], and 5% FDR was used to identify differentially methylated CpG positions (DMPs).

Abbreviations

EWAS: Epigenome-wide association studies; FDR: False discovery rate; BH: Benjamini-Hochberg step-up procedure; DMPs: Differentially methylated CpG positions; HBV: Hepatitis B virus; BMI: Body mass index; IR: Insulin resistance; SLE: Systemic lupus erythematosus; SVA: Surrogate variable analysis; TCGA: The Cancer Genome Atlas; 2dFDR: Two-dimensional false discovery rate control procedure

Supplementary information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02418-8>.

Additional file 1: Note S1. Simulation Setup for Fig. 1a-c. **Note S2.** Full Method Description. **Note S3.** Asymptotic FDR Control. **Note S4.** Power analysis.

Additional file 2: Figure S1. Performance comparison when 50% of the features are affected by the confounder. **Figure S2.** Performance on simulated datasets across varying density (top to bottom) and strength (left to right) of the confounding signals when there are five confounders. **Figure S3.** Performance across varying density (top to bottom) and strength (left to right) of the confounding signals when the confounding and true signals do not

overlap (“NoCoLoc”) and when the true and confounding signals have extensive overlap (“CoLoc”). **Figure S4.** Performance across varying density (top to bottom) and strength (left to right) of the confounding signals when the errors have a block correlation structure. **Figure S5.** Performance across varying density (top to bottom) and strength (left to right) of the confounding signals when the errors have the first-order auto-regressive (AR(1)) correlation structure. **Figure S6.** Performance comparison across varying density (top to bottom) and strength (left to right) of the confounding signals under smaller sample sizes. **Figure S7.** Performance comparison across varying density (top to bottom) and strength (left to right) of the confounding signals under smaller feature sizes. **Figure S8.** Performance comparison across varying sample size (top to bottom) and feature size (left to right). **Figure S9.** The computation time (in seconds) under different sample sizes and feature sizes based on simulated datasets with one confounder, medium density and strength of the true and confounding signals, and a medium confounding level. **Figure S10.** The decision boundaries of 2dFDR and 1dFDR-A under two unfavorable scenarios for 2dFDR. **Table S1.** EWAS datasets used in the evaluation of the empirical power of 2dFDR. **Additional file 3.** Review history.

Acknowledgements

Not applicable.

Review history

The review history is available as Additional file 3.

Peer review information

Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

J.C. and X.Z. conceived, designed, and implemented the method together. X.Z. and S.Y. performed the theoretical analysis. J.C., X.Z., and S.Y. developed the software and performed the simulations. J.C. and L.Y. performed the real data analysis with the help from C.W., Y.L., J.Y., and D.S. contributed expertise to improve the manuscript. J.C. and X.Z. led the writing of the manuscript with contributions from all other authors. All author(s) read and approved the final manuscript.

Funding

This work was supported by the Center for Individualized Medicine at Mayo Clinic (Chen), National Science Foundation DMS-1830392 and NSF DMS-1811747 (Zhang), and NIH 1 R21 HG011662 (Chen&Zhang).

Availability of data and materials

Codes and data to reproduce the presented results are available at the repository Zenodo (<https://doi.org/10.5281/zenodo.4977446>) [42] and at GitHub (<https://github.com/jchen1981/tdfdr>) [43] under GNU General Public License v2.0. Detailed documentation and tutorial are available on the GitHub page. Transcriptomics dataset is from [20] and metabolomics dataset from [22]. The epigenomics datasets are listed in Additional file 2: Table S1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing financial interests.

Author details

¹Department of Statistics, Texas A&M University, College Station, TX 77843, USA. ²Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN 55905, USA. ³State Key Laboratory of Medical Genomics, Shanghai Institute of Hematology, National Research Center for Translational Medicine, Rui-Jin Hospital, Shanghai Jiao Tong University, Shanghai 200025, China.

Received: 28 October 2020 Accepted: 24 June 2021

Published online: 13 July 2021

References

1. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol.* 2017;18(1):83. <https://doi.org/10.1186/s13059-017-1215-1>.
2. Balding D, Moltke I, Marioni J, editors. *Handbook of Statistical Genomics*. 4th ed: New Jersey: Wiley; 2019.
3. Majewski IJ, Bernards R. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nat Med.* 2011;17(3):304–12. <https://doi.org/10.1038/nm.2311>.
4. Ziegler A, Koch A, Krockenberger K, Großhennig A. Personalized medicine using DNA biomarkers: a review. *Hum Genet.* 2012;131(10):1627–38. <https://doi.org/10.1007/s00439-012-1188-9>.

5. Liang L, Cookson WOC. Grasping nettles: cellular heterogeneity and other confounders in epigenome-wide association studies. *Hum Mol Genet.* 2014;23(R1):R83–8. <https://doi.org/10.1093/hmg/ddu284>.
6. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet.* 2010;11(10):733–9. <https://doi.org/10.1038/nrg2825>.
7. Boheler KR, Volkova M, Morrell C, Garg R, Zhu Y, Margulies K, et al. Sex- and age-dependent human transcriptome variability: Implications for chronic heart failure. *Proc Natl Acad Sci U S A.* 2003;100(5):2754–9. <https://doi.org/10.1073/pnas.0436564100>.
8. Walther-Antônio MRS, Chen J, Multinu F, Hokenstad A, Distad TJ, Cheek EH, et al. Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Med.* 2016;8(1):122. <https://doi.org/10.1186/s13073-016-0368-y>.
9. Chen J, Wright K, Davis JM, Jeraldo P, Marietta EV, Murray J, et al. An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. *Genome Med.* 2016;8(1):43. <https://doi.org/10.1186/s13073-016-0299-7>.
10. McNamee R. Regression modelling and other methods to control confounding. *Occup Environ Med.* 2005;62(7):500–6. <https://doi.org/10.1136/oem.2002.001115>.
11. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med.* 2014;33(11):1946–78. <https://doi.org/10.1002/sim.6082>.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met.* 1995;57:289–300.
13. Storey JD. A direct approach to false discovery rates. *J Roy Stat Soc B Met.* 2002;64(3):479–98. <https://doi.org/10.1111/1467-9868.00346>.
14. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004;3:3–25.
15. Lu T, Pan Y, Kao S-Y, Li C, Kohane I, Chan J, et al. Gene regulation and DNA damage in the ageing human brain. *Nature.* 2004;429(6994):883–91. <https://doi.org/10.1038/nature02661>.
16. Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, et al. Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol.* 2013;14:1–12.
17. Gershoni M, Pietrokovski S. The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* 2017;15:1–15.
18. Chen J, Behnam E, Huang J, Moffatt MF, Schaid DJ, Liang L, et al. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics.* 2017;18(1):413. <https://doi.org/10.1186/s12864-017-3808-1>.
19. Jiang W, Zhang C-H. General maximum likelihood empirical Bayes estimation of normal means. *Ann Stat.* 2009;37:1647–84.
20. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013;45:1113–20.
21. Lamontagne J, Mell JC, Bouchard MJ. Transcriptome-wide analysis of hepatitis B virus-mediated changes to normal hepatocyte gene expression. *PLoS Pathog.* 2016;12(2):e1005438. <https://doi.org/10.1371/journal.ppat.1005438>.
22. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyötyläinen T, Nielsen T, Jensen BAH, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature.* 2016;535(7612):376–81. <https://doi.org/10.1038/nature18646>.
23. Pedersen HK, Forslund SK, Gudmundsdottir V, Petersen AØ, Hildebrand F, Hyötyläinen T, et al. A computational framework to integrate high-throughput “-omics” datasets for the identification of potential mechanistic links. *Nat Protoc.* 2018;13(12):2781–800. <https://doi.org/10.1038/s41596-018-0064-z>.
24. Huang J, Bai L, Cui B, Wu L, Wang L, An Z, et al. Leveraging biological and statistical covariates improves the detection power in epigenome-wide association testing. *Genome Biol.* 2020;21:1–19.
25. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 2007;3(9):1724–35. <https://doi.org/10.1371/journal.pgen.0030161>.
26. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. *BMC Genomics.* 2014;15:1–13.
27. Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc Natl Acad Sci U S A.* 2006;103(51):19430–5. <https://doi.org/10.1073/pnas.0609333103>.
28. Lei L, Fithian W. AdaPT: an interactive procedure for multiple testing with side information. *J Roy Stat Soc B Met.* 2018;80(4):649–79. <https://doi.org/10.1111/rssb.12274>.
29. Zhang X, Chen J. Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *J Am Stat Assoc.* 2020. <https://doi.org/10.1080/01621459.2020.1783273>.
30. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods.* 2016;13(7):577–80. <https://doi.org/10.1038/nmeth.3885>.
31. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* 2019;20(1):118. <https://doi.org/10.1186/s13059-019-1716-1>.
32. Zhou H, Zhang X, Chen J. Covariate Adaptive Family-wise Error rate control for genome-wide association studies. *Biometrika.* 2021:asaa098.
33. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc.* 1993;88:9–25.
34. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet.* 2006;2(12):e190. <https://doi.org/10.1371/journal.pgen.0020190>.
35. Koenker R, Mizera I. Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J Am Stat Assoc.* 2014;109(506):674–85. <https://doi.org/10.1080/01621459.2013.869224>.
36. Baheti S, Tang X, O'Brien DR, Chia N, Roberts LR, Nelson H, et al. HGT-ID: an efficient and sensitive workflow to detect human-viral insertion sites using next-generation sequencing data. *BMC Bioinformatics.* 2018;19:1–11.
37. Kalari KR, Nair AA, Bhavsar JD, O'Brien DR, Davila JI, Bockol MA, et al. MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics.* 2014;15:224.
38. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* 2014;30(7):923–30. <https://doi.org/10.1093/bioinformatics/btt656>.

39. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics*. 2012;28(16):2184–5. <https://doi.org/10.1093/bioinformatics/bts356>.
40. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*. 2011;27(11):1496–505. <https://doi.org/10.1093/bioinformatics/btr171>.
41. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. <https://doi.org/10.1186/1471-2105-11-587>.
42. Yi S, Zhang X, Yang L, Huang J, Liu Y, Wang C, Schaid DJ, Chen J. A new approach to confounder adjustment substantially increases detection power in omics association studies. Zenodo. <https://doi.org/10.5281/zenodo.4728278> (2021).
43. Yi S, Zhang X, Yang L, Huang J, Liu Y, Wang C, Schaid DJ, Chen J. A new approach to confounder adjustment substantially increases detection power in omics association studies. Github. <https://github.com/jchen1981/TDFDR> (2021).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

