A Bayesian Informative Shrinkage Approach for Large-scale Multiple Hypothesis Testing (BISHOT): with Applications in Differential Analysis of Omics Data

Ya Su
Department of Statistical Sciences and Operational Research, Virginia Commonwealth University,
Richmond, VA 23284-3083, U.S.A., suyaf@vcu.edu
Mary Eunice Joy Z. Clark
Department of Statistical Sciences and Operational Research, Virginia Commonwealth University,
Richmond, VA 23284-3083, U.S.A., clarkmz@vcu.edu Chi Wang
Division of Cancer Biostatistics, Department of Internal Medicine and Markey Cancer Center,
University of Kentucky, Lexington, 40536, U.S.A., chi.wang@uky.edu

Abstract

A major goal of many omics studies is to identify differential features, e.g. differentially expressed genes, between experimental groups. When performing differential analysis for a given dataset, relevant information from another platform or species is often available. Incorporating such prior information can help identify features that show consistent differential patterns across platforms or species, which are more likely to reflect shared biological processes, and thereby enhance the robustness and generalizability of the findings. However, existing differential analysis methods typically analyze only the data from the current study and do not leverage prior knowledge about the magnitude or direction of changes from other platforms or species. We address this challenge, and the associated multiple testing problem, using a Bayesian framework that enables the incorporation of prior knowledge obtained from different platforms or species. We propose a new test statistic, Bayesian Credible Ratio (BCR), based on a heteroscedastic global local shrinkage prior, and a new multiple testing criterion, sign-adjusted FDR (SFDR), that emphasize information regarding the direction of the differentially features. We prove that BCR achieves the largest count of sign-based true positives among all legitimate SFDR-controlling methods. Simulation results offer numerical evidence of its advantage compared to an empirical Bayesian method. The approach is demonstrated through the analysis of RNAseq and single-cell RNAseq datasets.

Some Key Words: Bayesian large scale hypothesis tests; non-central global local shrinkage; differential gene expression; sign-adjusted classification risk

1 Introduction

The statistical differential analysis plays an important role in many omics studies that involve two or more experimental groups. Various statistical methods have been developed to identify differentially expressed genes from RNAseq or single cell RNAseq (scRNAseq) (Finak et al., 2015; Love et al., 2014; Robinson et al., 2010; Smyth, 2005; Wu et al., 2013), differentially bound sites from ChIPseq (Ross-Innes et al., 2012), differentially accessible peaks from ATACseq (Gontarz et al., 2020), or differentially abundant proteins/metabolites from mass spectrometry (Huang et al., 2020; Li et al., 2019) between experimental groups.

When conducting the differential analysis for a given omics dataset, there is often relevant information available from another platform or species. Such prior information enables investigators to distinguish features that exhibit consistent differential patterns across platforms or species from those specific to a single platform or species. Because features that are consistent across platforms or species likely reflect shared underlying biological processes (Vitorino, 2024), prioritizing them can enhance the robustness and generalizability of the findings (Altenbuchinger et al., 2017; Wang et al., 2022). A few such examples are as follows:

- Example 1: Hohmann et al. (2016) performed an RNAseq experiment on a human acute myeloid leukemia cell line, MV4, to investigate changes in the gene expression profile with the treatment of BI-7273, a BRD9 inhibitor. In addition to this experiment, the investigators also performed an RNAseq experiment on a mice cell line, RN2, using the same treatment. Hohmann et al. (2016) found that *MYC* expression was downregulated in BRD4 inhibitor–treated cells in both human (MV4) and mouse (RN2) cell lines. This consistent suppression of *MYC* across species provides strong evidence that BRD4 plays a critical role in sustaining *MYC* transcription.

- Example 2: Angelidis et al. (2019) conducted a scRNAseq experiment that profiled the cell types and gene expressions in old and young mice to study lung aging. An analysis of interest is to identify differentially expressed genes in type-2 pneumocytes, one of the most prevalent cell types in the samples, between old and young mice. In parallel, the authors also conducted a bulk RNAseq experiment on type-2 pneumocytes,

selected by flow cytometry sorting, from old and young mice. Angelidis et al. (2019) found significant agreement between the scRNAseq results and bulk RNAseq results, supporting the robustness of the age-related genes identified by scRNAseq.

- Example 3: Li et al. (2022) performed an ATACseq experiment to identify chromatin regions related to resistance to second-generation androgen receptor inhibitors, e.g. enzalutamide, by comparing enzalutamide-sensitive and resistant cell lines. Prior to this experiment, the investigators had already performed an RNAseq experiment on the same cell lines (Li et al., 2020). Li et al. (2022) focused on consensus genes that were both differentially expressed in RNAseq data and had differentially accessible promoter peaks in ATACseq data, thereby narrowing down the list of candidate genes and ultimately identifying a potential therapeutic target, GSTM2, for overcoming resistance to enzalutamide.

A common approach for leveraging prior information is to consider a two-step procedure, where the first step is to perform differential analysis on the current dataset alone to identify significant features, and the second step is to only keep the subset of significant features that are consistent with prior information from other platform/species (Chen et al., 2024; Li et al., 2022). However, since all features are analyzed together and treated equally in the first step, the significance levels of true consistently differential features can be reduced due to the presence of platform/species-specific differential features, e.g. a relatively small (but significant) effect claimed by both prior and current data, therefore consistent, can be declared as insignificant due to low signal noise ratio in the current data, e.g., in the analysis of MV4 cell line data in Section 5, the current standard approach fails to identify *MYC* which is claimed to be significant by BISHOT. In addition, downstream analyses, such as gene set enrichment analysis, may become complicated since ranking genes based on significance levels is difficult with the filtering based on prior information.

Despite the demand for a rigorous procedure to incorporate these prior information, most existing differential analysis methods perform the analysis based on the given dataset alone without allowing historical prior information to be supplied (Robinson et al., 2010; Smyth, 2005; Sun and McLain, 2012). Although leveraging historical data has been explored (Li

et al., 2017, 2016), the focus has primarily been on stabilizing the estimation of mean and variance parameters, either by treating historical data as informative priors (Li et al., 2016) or by grouping genes with similar variations from historical data (Li et al., 2017). To our knowledge, there is a lack of methods that explicitly incorporate prior knowledge about the magnitude or direction of changes, e.g. fold changes in gene expression, between experimental groups.

Another potential area of research is regarding multiple testing error control, given the number of genes in expression data grows substantially. The property of false discovery rate (FDR) has been extensively studied in the literature (Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2004; Storey, 2003, among others). Various methods consider alternatives to the traditional p-value based FDR control approach (Benjamini and Hochberg, 1995), by using some 'local' measure for evidence towards H_{ag} (Efron, 2008; Stephens, 2017; Sun and McLain, 2012), which quantifies individualized 'significance' under data heteroscedasticity and arguably enhances the power in large scale testing problems. More evidence regarding benefits of empirical Bayes based procedures in handling multiple comparisons can be found in Efron (2012), Kendziorski et al. (2003), Muralidharan (2010) and references therein. As a comparison method, we briefly describe Sun and McLain (2012) shortly.

For simplicity purpose, we will derive our method based on a linear regression model for RNAseq data. An extension to a more complicated model for scRNAseq data is provided in Section 6. For gene g ($g = 1, \ldots, G$) in subject i ($i = 1, \ldots, n$), let Y_{ig} be the normalized expression level in the form of $\log_2(\text{CPM} + 1)$, where CPM denotes counts per million. We consider a classical regression model that allows for heteroscedasticity for Y_{ig}:

$$Y_{ig} \sim \mathrm{N}(c_g + X_i^\top \boldsymbol{\beta}_g, \sigma_g^2). \tag{1}$$

Generally speaking, $X_i \in \mathbf{R}^p$ is a set of covariates for subject i, c_g is the intercept, σ_g^2 is the variance and $\boldsymbol{\beta}_g$ is a vector of coefficients for gene g. The methodology is illustrated for the application of comparing RNAseq gene expressions between two groups, e.g. control and treatment, corresponding to model (1) with $p = 1$ where $X_i \in \{0, 1\}$ represents a treatment group indicator, equivalently, β_g is the $\log_2$ fold change (LFC) quantifying the treatment effect for gene g.

The problem we are interested in is whether β_g is different from 0 for $g = 1, \ldots, G$. A natural statistical formulation is via hypothesis test. Specifically we consider the null hypotheses to be composite, that is

$$H_{0g} : \beta_g \in A \quad \text{vs} \quad H_{ag} : \beta_g \notin A. \tag{2}$$

In this context we consider $A = (-\epsilon, \epsilon)$, a small neighborhood containing 0. The composite null hypothesis ensures that identified differences are sufficiently large to be biologically meaningful, with the choice of ϵ informed by domain experts (McCarthy and Smyth, 2009).

A popular multiple hypothesis testing procedure is built upon some summary statistic which is normally distributed and centered on β_g, that is, $\widehat{\beta}_g \sim \mathrm{N}(\beta_g, \widetilde{\sigma}_g^2)$. A common shrinkage prior with a zero component is typically utilized for β_g, $g = 1, \ldots, G$. That is, $\beta_g \sim f(\cdot)$, with $f(x) = \pi_0 \delta_0(x) + (1 - \pi_0) g(x)$, where $\delta_0(x)$ is a point mass at 0, $g(x)$ is the distribution for non-zero signals and the probability of observing 0 is given by π_0. Sun and McLain (2012) is specifically designed for composite null and the proposed 'local' estimate is proven to be optimal within a family of estimators called MLRC in large scale hypothesis testing.

There are several potential limitations in the above and similar procedures. First, they do not apply when there is prior knowledge about β_g as those procedures make heavy use of a common prior distribution for β_g. Second, they rely on the existence of the summary statistics $\widehat{\beta}_g$ and its marginal distribution over β_g, which are crucial in estimating π_0 and $g(\cdot)$. Besides, estimations of π_0 and $g(\cdot)$ are shown to be unstable from our simulation studies especially for smaller sample sizes.

We introduce a new **B**ayesian **i**nformative **s**hrinkage multiple **h**yp**o**thesis **t**esting (BISHOT) procedure that addresses the question of elucidating consistent features between two data sources with proven error control. Specifically, we propose a new shrinkage prior distribution where the local parameter plays a key role in governing the prior influence and is numerically shown to be adaptive to the feature-level coherence from prior to present. In addition, a new test statistic, Bayesian Credible Ratio (BCR), based on the posterior distribution is proposed which accounts for the heteroscedasticity of genes in favor of one side of tail over the other. While conducting multiple tests, we also take the tail-favoring perspective and

propose a new criterion called the sign-adjusted FDR (SFDR). The calculation of BCRs for all genes combined naturally promotes parallelization across all genes. The decision process according to BCR is proven to attain the maximum number of true positives among all valid sign-adjusted false positive controlled procedures.

To the best of our knowledge, it is a difficult task for the available approaches to work with a more complicated model or when a vector $\boldsymbol{\beta}_g$ is present due to the lack of explicit distribution of $\widehat{\boldsymbol{\beta}}_g$ as well as multivariate deconvolution methods in the presence of unknown point mass on zero. The fully Bayesian framework we propose lays out a potential path to this setting without the need of establishing $\widehat{\boldsymbol{\beta}}_g$ and carefully modeling $f(\cdot)$ and thus it is free from estimating non-null probability and non-null distribution. We present our solution for a two component model for scRNAseq data in Section 6.

Outline of this paper is as follows. Section 2 describes the construction of the prior distribution and the test statistic named the Bayesian credible ratio, its connection with the decision rule. Section 3 introduces a classification risk-based framework for large-scale differential analysis, detailing threshold selection and establishing the procedure's optimality. Section 2 and Section 3 combined lead to the proposed BISHOT. Section 4 includes simulation results and comparisons with the above popular method. Applications to RNAseq and scRNAseq gene expression data are illustrated in Section 5 and 6. We conclude with several discussions in Section 7.

2 Bayesian credible ratio

We illustrate the proposed method using model (1), we later dive into a more complicated model applicable to scRNAseq data in Section 6. The likelihood function corresponding to (1) for a given gene feature g along with observed expressions $i = 1, \ldots, n$ is:

$$\prod_{i=1}^{n} \phi_{\sigma_g}(Y_{ig} - c_g - X_i^\top \beta_g),$$

by convention, ϕ_σ denotes the normal density with mean zero and standard deviation σ.

We propose a heteroscedastic global local shrinkage (HGLS) prior for β_g based on our

knowledge about β_g,

$$\Pi(\beta_g|\sigma_g^2, \lambda_g, \tau) \sim \mathrm{N}(h_g, \sigma_g^2\lambda_g^2\tau^2), \tag{3}$$

the input of non-central parameter h_g is reflective of the information given about β_g, in this work, h_g is a fixed value and set as the prior estimate for the treatment effect associated with feature g based on findings from earlier studies. The prior variance consists of the local parameter λ_g and the global parameter τ, in addition to the standard deviation σ_g from error distribution in (1).

HGLS for β_g builds upon the classical global and local prior by explicitly modeling the substantial heteroscedasticity present in gene expression data. Contrary to zero, it has a shrinkage effect towards h_g, the existing knowledge about β_g. The shrinkage is also relative to the variability of that specific gene feature captured by σ_g. The local shrinkage λ_g governs the prior influence and is learned in a data-driven manner based on the gene-specific coherence between historical and current data. As illustrated in Figure 5, λ_g increases with the discrepancy between historical and current treatment effects, thereby reducing the impact of historical values when they are not aligned with current findings.

In this work we adopt standard prior distributions for shrinkage parameters λ_g and τ from horseshoe. That is, $\lambda_g, \tau \sim$ Half-Cauchy$(0, 1)$. The heteroscedastic error variance σ_g^2 takes a conjugate prior, $\Pi(\sigma_g^2) \sim \mathrm{IG}(\xi_1, \xi_2)$ with ξ_1 and ξ_2 set according to the moments of empirical estimates of σ_g^2 given the observed gene expressions. The overall expression c_g is assigned a prior distribution $\Pi(c_g|\sigma_g^2) \sim \mathrm{N}(0, 10^2\sigma_g^2)$ in the general case.

Under the proposed prior (3) and model (1) for the expression data, we can derive the (conditional) posterior distribution of β_g, λ_g, σ_g^2 and τ given data $D_g = \{Y_{ig}, i = 1, \ldots, n\}$ and h_g, where λ_g and τ are sampled using slice sampling algorithm, similar to Bhattacharya et al. (2016). For better mixing of the chain, we sample β_g given λ_g and τ while integrating out σ_g. The conditional posterior distributions for all parameters are explicitly written in the Appendix A.

The proposed test procedure is based on the *tail evidence* from the posterior distribution of the targeting parameter. For the purpose of illustration, in the remainder of the section we simplify the notation from β_g and D_g to β and D correspondingly, e.g., posterior distribution

of β is denoted as $\Pr(\beta|D)$.

- We define the Bayesian credible ratio (BCR) for the hypothesis test $H_0 : \beta \in (-\epsilon, \epsilon)$ vs $H_a : \beta \notin (-\epsilon, \epsilon)$ by $C_\beta = C_{0,\beta}(1 - C_{a,\beta})^{-1}$, where $\mathrm{C}_{0,\beta} = \Pr(\beta \in (-\epsilon, \epsilon)|D)$ and $\mathrm{C}_{a,\beta} = \min\{\Pr(\beta \leq -\epsilon|D), \Pr(\beta \geq \epsilon|D)\}$ are the posterior probability for β over $(-\epsilon, \epsilon)$ and its smaller tail over $(-\infty, -\epsilon)$ or (ϵ, ∞).

- The null hypothesis is rejected if C_β is smaller than a threshold $0 \leq \lambda \leq 1$.

Given the model and the prior structures, the hypothesis test in (2) is analyzed by BCR based on the posterior samples of β_g. BCR compares evidence towards the null/alternative hypotheses by $C_{0,\beta}$ and $1 - C_{a,\beta}$. The notion of BCR is model and prior free and might be of interest for its own purpose. When hypotheses regarding all feature effects are considered, we suggest the threshold λ be selected to control the sign-adjusted FDR (SFDR) in the multiple testing regime, see details in Section 3.

2.1 BCR as a Bayesian decision rule

Generally speaking, a hypotheses testing problem $H_0 : \beta \in \Theta_0$ vs $H_1 : \beta \in \Theta_1$(or Θ_0^c) can be formulated as a Bayesian decision process (Berger, 2013). Let a_0/a_1 be the action of accepting of H_0/H_1, the loss function for a_i is $L(a_i, \beta)$, $i = 0, 1$. Consider a simple '$0 - K_i$' loss function, $L(a_i, \beta) = 0$ if $\beta \in \Theta_i$ and $L(a_i, \beta) = K_i$ if $\beta \notin \Theta_i$. The expected posterior loss for a_0 and a_1 is $K_0 P(\Theta_1|D)$ and $K_1 P(\Theta_0|D)$ correspondingly. The Bayes rule opts for the action which minimizes the expected posterior loss, that is, accepting H_0 when $P(\Theta_0|D) > K_0/(K_0 + K_1)$ and accepting H_1 otherwise.

When $\Theta_0 = (-\epsilon, \epsilon)$, we revise the above decision process as follows. First since the alternative set $\Theta_0^c = (-\infty, -\epsilon) \cup (\epsilon, \infty)$ is composed of two non-overlapping sets, we consider performing two hypotheses with alternatives being $H_1 : \beta \in \Theta_1 = (-\infty, -\epsilon)$ and $H_2 : \beta \in \Theta_2 = (\epsilon, \infty)$. In this circumstance, we define three actions a_0, a_1 or a_2 corresponding to accepting of H_0, H_1 or H_2. Suppose the '$0 - K_i$' loss functions are chosen for $a_i, i = 0, 1, 2$ under a constraint $K_1 = K_2$ (no distinction of loss under a_1 or a_2), the Bayes rule regarding H_0 and H_1 is accepting H_0 when $P(\Theta_0|D)K_1 > K_0 P(\Theta_1|D)$ and accepting H_1 otherwise;

the Bayes rule regarding H_0 and H_2 is accepting H_0 when $P(\Theta_0|D)K_1 > K_0P(\Theta_2|D)$ and accepting H_2 otherwise.

We propose the 'universal' Bayes rule regarding $H_0 : \beta \in \Theta_0$ vs $H_a : \beta \in \Theta_1 \cup \Theta_2$ as accepting H_0 when both decisions made above agree with action a_0 and accepting H_a when either of the above decisions says otherwise. Equivalently, the universal decision process is accepting H_0 when $P(\Theta_0|D) > (K_0/K_1)\max\{P(\Theta_1|D), P(\Theta_2|D)\}$ and accepting H_1 otherwise. Define $\lambda = K_0/(K_0 + K_1)$, $0 < \lambda < 1$, our decision can be further formulated as accepting H_0 if $C_\beta > \lambda$ or accepting H_1 if $C_\beta < \lambda$, where C_β is the BCR defined previously.

3 Optimal decision rule for differential analysis via sign-adjusted classification risk

In the presence of a large number of features, the issues of multiple testing emerge when hypotheses tests (2) are carried out for each feature individually. To design an optimal decision rule while controlling the marginal false discovery rate (mFDR), Sun and Cai (2007) shows the equivalence to study the weighted classification risk corresponding to $\beta \in (-\epsilon, \epsilon)$ or $\beta \notin (-\epsilon, \epsilon)$. The optimal decision rule is shown to be the minimizer to the classification risk, while the associated statistic is optimal in that it minimizes the false nondiscovery rate, subject to a constraint on the false discovery rate.

As detailed in Section 2.1, the Bayes decision rule governed by BCR, is based on two separate decisions regarding $\beta_g \leq -\epsilon$ or $\beta_g \geq \epsilon$. We will show that this Bayes rule is an estimator of the optimal decision/minimizer to a weighted classification risk with sign preferences for all features. Furthermore, it is proven to capture the maximum number of sign-adjusted expected true positives while controlling for the marginal SFDR (mSFDR). The proofs to all Theorems in this Section can be found in the Appendix A.

3.1 BCR as minimizer of classification risk utilizing sign preferences

Define

$$\omega_g = \begin{cases} 0, & \text{if } \beta_g \in (-\epsilon, \epsilon), \\ 1, & \text{if } \beta_g \leq -\epsilon, \\ 2, & \text{if } \beta_g \geq \epsilon. \end{cases}$$

We purposely treat $\beta_g \leq -\epsilon$ and $\beta_g \geq \epsilon$ separately since the combination of the data and the informative shrinkage prior should provide evidence (if any) favoring one of them rather than both.

Define $\delta_g = 1$ if the null hypothesis for feature g is rejected and 0 otherwise, $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ represent the concatenated vector of δ_g and ω_g for $g = 1, \ldots, G$. Let $\lambda \in [0, 1]$ represent the weight governing the false positive/negative losses. Define $\widetilde{\omega}_{0g} = \Pr(\omega_g = 0|\boldsymbol{y}_g) = \Pr(\beta_g \in (-\epsilon, \epsilon)|\boldsymbol{y}_g)$, $\widetilde{\omega}_{1g} = \Pr(\omega_g = 1|\boldsymbol{y}_g) = \Pr(\beta_g \leq -\epsilon|\boldsymbol{y}_g)$, $\widetilde{\omega}_{2g} = \Pr(\omega_g = 2|\boldsymbol{y}_g) = \Pr(\beta_g \geq \epsilon|\boldsymbol{y}_g)$, where $\boldsymbol{y}_g = \{Y_{1g}, \ldots, Y_{ng}\}$ is the vector of observations for feature g.

Consider first two loss functions to losses of detecting negative effects ($\omega_g = 1$) or positive effects ($\omega_g = 2$) correspondingly:

$$L_{1,\lambda}(\boldsymbol{\delta}, \boldsymbol{\omega}) = \sum_{g=1}^{G} [(1-\lambda) I\{\omega_g = 0\}\delta_g + \lambda I\{\omega_g = 1\}(1-\delta_g)],$$

$$L_{2,\lambda}(\boldsymbol{\delta}, \boldsymbol{\omega}) = \sum_{g=1}^{G} [(1-\lambda) I\{\omega_g = 0\}\delta_g + \lambda I\{\omega_g = 2\}(1-\delta_g)].$$

<u>Theorem</u> 1. For $k = 1, 2$, the minimizer of $EL_{k,\lambda}$ (the expected value of $L_{k,\lambda}$) is denoted by $\boldsymbol{\delta}_k$, which is the vector form of $\{\delta_{kg}, g = 1, \ldots, G\}$ where $\delta_{kg} = I\{(1-\lambda)\widetilde{\omega}_{0g} < \lambda\widetilde{\omega}_{kg}\}$. $EL_{k,\lambda}$ is defined over the probability law of $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$.

Next, we define a partition of all features with regards to the preference of positive or negative signs when rejecting. This partition exists under Assumption 1 below. Incorporating information about the sign of the parameter to construct decision, or to choose between the terms in $L_{1,\lambda}$ or $L_{2,\lambda}$, leads to a more conservative decision than ignoring the sign. It

is preferred in our setting when features are provided with additional confidence per prior knowledge (3).

Assumptions 1. For a fixed g, $g = 1, \ldots, G$, $\varphi_{a,b}(\lambda) = \int_a^b \Pi(\beta_g) P(\delta_g = 1|\beta_g) d\beta_g$, either (i) or (ii) holds: (i) $\varphi_{\epsilon,\infty}(\lambda) \leq \varphi_{-\infty,-\epsilon}(\lambda)$ for all $\lambda \in (0,1)$ (ii) $\varphi_{\epsilon,\infty}(\lambda) \geq \varphi_{-\infty,-\epsilon}(\lambda)$ for all $\lambda \in (0,1)$.

Assumption 1 requires the sign preference is invariant to λ, that is, case (i) and case (ii) holds universally for all any choice of λ. $\varphi_{-\infty,-\epsilon}(\lambda)$ has an equivalent form $E(\widetilde{\omega}_{1g}|\delta_g = 1) \cdot P(\delta_g = 1)$ and similarly $\varphi_{\epsilon,\infty}(\lambda) = E(\widetilde{\omega}_{2g}|\delta_g = 1) \cdot P(\delta_g = 1)$. Therefore, under case (i) or (ii) $\widetilde{\omega}_{1g}$ is expected to be larger or smaller when the decision is rejection.

Let $\mathcal{S}$ or $\mathcal{S}^c$ denote the set of features which belong to case (i) or (ii) correspondingly. Based on Assumption 1, loss in $L_{1,\lambda}$ tends to concern features in $\mathcal{S}$ than those in $\mathcal{S}^c$, this motivates us to prioritize $\omega_g = 1$ or $\omega_g = 2$ in each scenario and thus combine $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ into one. The combined decision we propose is $\delta_g^{OPT} = I\{(1-\lambda)\widetilde{\omega}_{0g} < \lambda\widetilde{\omega}_{1g}\}$, $g \in \mathcal{S}$ and $\delta_g^{OPT} = I\{(1-\lambda)\widetilde{\omega}_{0g} < \lambda\widetilde{\omega}_{2g}\}$, $g \in \mathcal{S}^c$. Denote $\boldsymbol{\delta}^{OPT}$ as the concatenated vector.

Theorem 2. $\boldsymbol{\delta}^{OPT}$ minimizes $EL_{3,\lambda}$ where

$$L_{3,\lambda}(\boldsymbol{\delta}, \boldsymbol{\omega}) = \sum_g (1-\lambda) I\{\omega_g = 0\}\delta_g + \sum_{g \in \mathcal{S}} \lambda I\{\omega_g = 1\}(1-\delta_g) + \sum_{g \in \mathcal{S}^c} \lambda I\{\omega_g = 2\}(1-\delta_g)$$

The optimal decision $\boldsymbol{\delta}^{OPT}$ has one missing piece: the set $\mathcal{S}$ is unknown and needs to be estimated. When $\delta_g = 1$, it is expected that $\widetilde{\omega}_{1g} \geq \widetilde{\omega}_{2g}$ if $g \in \mathcal{S}$ and $\widetilde{\omega}_{2g} \geq \widetilde{\omega}_{1g}$ if $g \in \mathcal{S}^c$. Therefore the optimal decision can be estimated empirically as $\delta_g = I\{(1-\lambda)\widetilde{\omega}_{0g} < \lambda \max(\widetilde{\omega}_{1g}, \widetilde{\omega}_{2g})\}$. According to the definitions of $\widetilde{\omega}_{1g}$ and $\widetilde{\omega}_{2g}$, one comes to an equivalent expression $\delta_g = I\{C_{0,\beta_g}(1 - C_{a,\beta_g})^{-1} < \lambda\}$ where $C_{0,\beta_g} = \widetilde{\omega}_{0g}$ and $C_{a,\beta_g} = \min\{\widetilde{\omega}_{1g}, \widetilde{\omega}_{2g}\}$. This establishes the relationship that BCR is an estimator of δ_g^{OPT}.

3.2 Selection of λ in the multiple testing regime

The choice of λ can be selected by controlling the SFDR at the desired level α. In practice,

$$\text{mSFDR}(\boldsymbol{\delta}) = \frac{E(\sum_{g=1}^G I\{\omega_g = 0\}\delta_g)}{\sum_{g \in \mathcal{S}} E(\delta_g I\{\omega_g \neq 2\}) + \sum_{g \in \mathcal{S}^c} E(\delta_g I\{\omega_g \neq 1\})} \tag{4}$$

which denotes the marginal SFDR, and it can be shown that mSFDR $= SFDR + O(G^{-1/2})$ using the same technique as Genovese and Wasserman (2002) for the asymptotic equivalence for mFDR and FDR. mSFDR is the mFDR based on sign of the parameters and has a natural link to the classification risk $L_{3,\lambda}$.

Although the explicit form of mSFDR($\boldsymbol{\delta}$) is generally unknown, we can estimate it using the following procedure (a slight modification of Section 3.2 in Sun and McLain (2012)):

$$\begin{aligned} E(I\{\omega_g = 0\}\delta_g) &= P(\beta_g \in A, C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})) \\ &= \int_A f(\beta_g) \int I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\} f(\boldsymbol{y}_g|\beta_g) d\boldsymbol{y}_g d\beta_g \\ &= \int I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\} f_g(\boldsymbol{y}_g) \int_A f(\beta_g|\boldsymbol{y}_g) d\beta_g d\boldsymbol{y}_g \\ &= \int I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\} C_{0,\beta_g} f_g(\boldsymbol{y}_g) d\boldsymbol{y}_g \end{aligned}$$

The numerator of mSFDR($\boldsymbol{\delta}$) can be unbiasedly estimated by $\sum_{g=1}^{G} I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\} C_{0,\beta_g}$. The same arguments hold true for the denominator of mSFDR($\boldsymbol{\delta}$), which can be estimated by $\sum_{g\in\mathcal{S}} I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\}\widetilde{\omega}_{1g} + \sum_{g\in\mathcal{S}^c} I\{C_{0,\beta_g} < \lambda(1 - C_{a,\beta_g})\}\widetilde{\omega}_{2g}$. Together with these estimators, we select $\lambda \in (0, 1)$ to be the largest such that $\widehat{\text{mSFDR}}(\boldsymbol{\delta}) < \alpha$.

3.3 Optimal property of $\boldsymbol{\delta}^{OPT}$

The discovery of true positives is one criterion to compare procedures given the same FDR α. Under additional knowledge about the sign of the signal (given rejection), as the proposed mSFDR, the true positives count is modified accordingly, that is, true positives count are refined with information about sign appropriately.

Definition 1. The sign-adjusted expected number of true positives is defined as follows:

$$\text{SETP}(\boldsymbol{\delta}) = \sum_{g\in\mathcal{S}} E(I\{\omega_g = 1\}\delta_g) + \sum_{g\in\mathcal{S}^c} E(I\{\omega_g = 2\}\delta_g).$$

Clearly, compared to the traditional ETP, SETP only counts when the parameter β_g is in the correct sign given rejection, which is divided up by either $\mathcal{S}$ or $\mathcal{S}^c$ (Remark 3.1). The following Theorem states that the proposed decision is optimal in achieving the greatest SETP under the same mSFDR threshold.

Theorem 3. Assume that λ is chosen such that mSFDR $= \alpha$ for a fixed α. The procedure $\boldsymbol{\delta}^{OPT}$ has the largest SETP among all valid mSFDR procedures.

4 Simulations

4.1 Simulation setup

Our simulation studies is based on model (1) which can be useful for comparing gene expressions in two groups, e.g., control and treatment. This is achieved by setting X_i to be an indicator (0 or 1) variable if ith expression in the treatment group, β_g thus becomes the treatment effect.

We select a simulation scheme where $100(1-p)\%$ of the genes share the same expression mechanisms between two groups. That is, β_g is set to zero except for $100p\%$ of the genes that are randomly selected. The proportion p is normally referred to as the proportion of differentially expressed (DE) genes. The following choices are used $p = 0.1$ or $p = 0.5$ corresponding to low or moderate presence of DE genes. Denote the set of DE genes as Ω, hence $\beta_g = 0$ when $g \notin \Omega$. For $g \in \Omega$, $\beta_g \sim \mathrm{N}(\log_2(2), 0.1^2)$ (up to a random sign placement). In the Appendix A, a separate simulation with expressions of DE genes generated from a unimodal distribution is conducted.

In model (3) the following values are used: $c_g = 4$, $\sigma_g = 1$ for a total $G = 5000$ genes. The sample size for treatment/control group is $n = 10$ or $n = 50$ with the total sample size for the two groups doubled. The rest of the hyperparameters required are set as follows. The hyperparameters ξ_1 and ξ_2 are determined such that the first and second moments for all (pooled) sample variances of the expressions (combining treatment and control group) match with those under the prior. We consider two options for ϵ in the hypothesis tests (2), $\log_2(1.5)$ or $\log_2(1.2)$ targeting at moderate or small treatment effects in DE analysis. The prior distributions for β_g is centered on h_g if $g \in \Omega$. For h_g, we purposely perturb the true value of β_g with $h_g \sim \mathrm{N}(\beta_g, 0.5^2)$, that is, the information regarding β_g is only within a vicinity of its true value. We generate $N = 50$ replicated data sets under each scenario and results are summarized in Section 4.2.

We provide a concise description of the method that is compared to BISHOT, denoted as SunSpike0. As mentioned in the introduction, the algorithm in Sun and McLain (2012) is theoretically-justified and suitable for composite null hypothesis. Since it relies on estimating the distribution of β_g with a zero component and a distribution for DE genes given $\widehat{\beta}_g \sim \mathrm{N}(\beta_g, \widetilde{\sigma}_g^2)$, it is not directly applicable to any data-generating model (when the distribution of $\widehat{\beta}_g$ does not have an explicit form as such) but can be adapted to model (1) if $\widehat{\beta}_g$ is selected to be the MLE and $\widetilde{\sigma}_g$ as the standard error. To the best of our knowledge the code underlying Sun and McLain (2012) is not publicly available. Alternatively, the `R` package `ashr` (Stephens, 2017) can extract the distribution of β_g assuming the distribution for DE genes is unimodal and symmetric with which we estimate the test statistic in Sun and McLain (2012). The resulting estimate then used in controlling mSFDR is compared with BISHOT in the simulation above. Similar findings are present in simulation II where the unimodal and symmetric distribution is used as the true distribution for β_g (the Appendix A).

4.2 Simulation results

Under a collection of SFDR nominal values ranging over $[0, 0.25]$, we validate the actual SFDR achieved using (4) with δ_g and the true ω_g. Figure 1 and 2 display the actual SFDR values averaged over 50 simulations against the corresponding nominal SFDR thresholds for the four scenarios given by n and p when $\epsilon = \log_2(1.2)$ (Figure 1) and $\epsilon = \log_2(1.5)$ (Figure 2). It can be seen that almost always BISHOT produces SFDR much closer to its nominal value while SunSpike0 tends to yield much higher FDR and not surprisingly the advantage becomes more prominent when the sample size is smaller and/or there is less DE genes (small n and/or small p). It is noteworthy that SunSpike0 behaves quite differently when the proportion of DE genes p and sample size n changes, conversely, BISHOT is dramatically more stable/robust among all scenarios given, thanks to the informative prior. Generally speaking, ϵ has an impact on the conservativeness of both methods, genes with small or zero β_g become more distinct from the genes with large β_g relative to a larger threshold hence the actual false discovery proportion tends to decrease when ϵ increases. Hence the desired conservativeness of the method can provide some guidance of choosing ϵ.

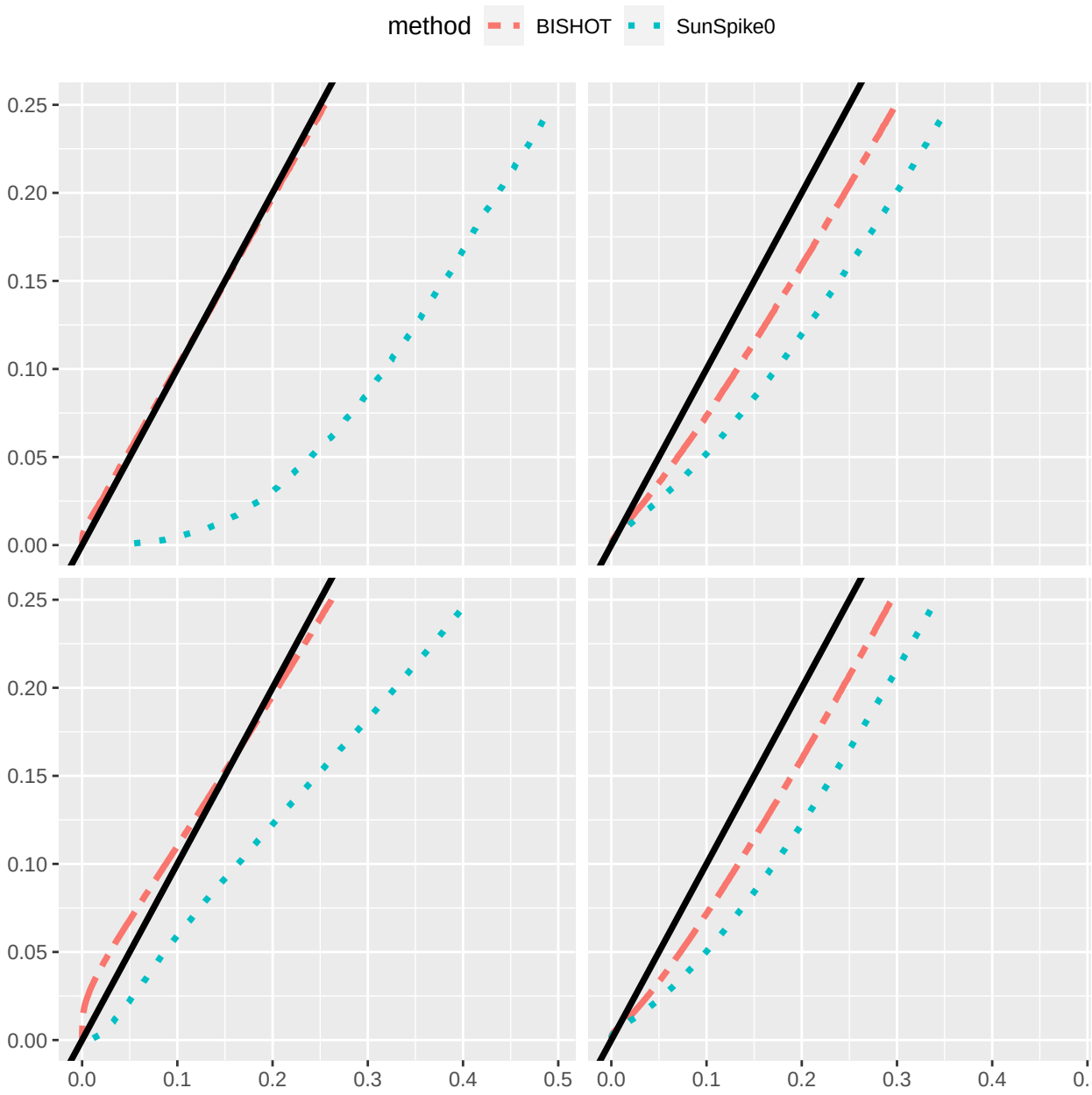

Figure 1: The actual signed false discovery rate (SFDR) when $\epsilon = \log_2(1.2)$ averaged over 50 simulated data using mSFDR threshold 0.001 to 0.25 corresponding to 10 (left panel) or 50 (right panel) samples in the treatment and control group. The top and bottom panel displays the case when $p = 0.1$ and $p = 0.5$. The solid black line corresponds to the $y = x$ line.

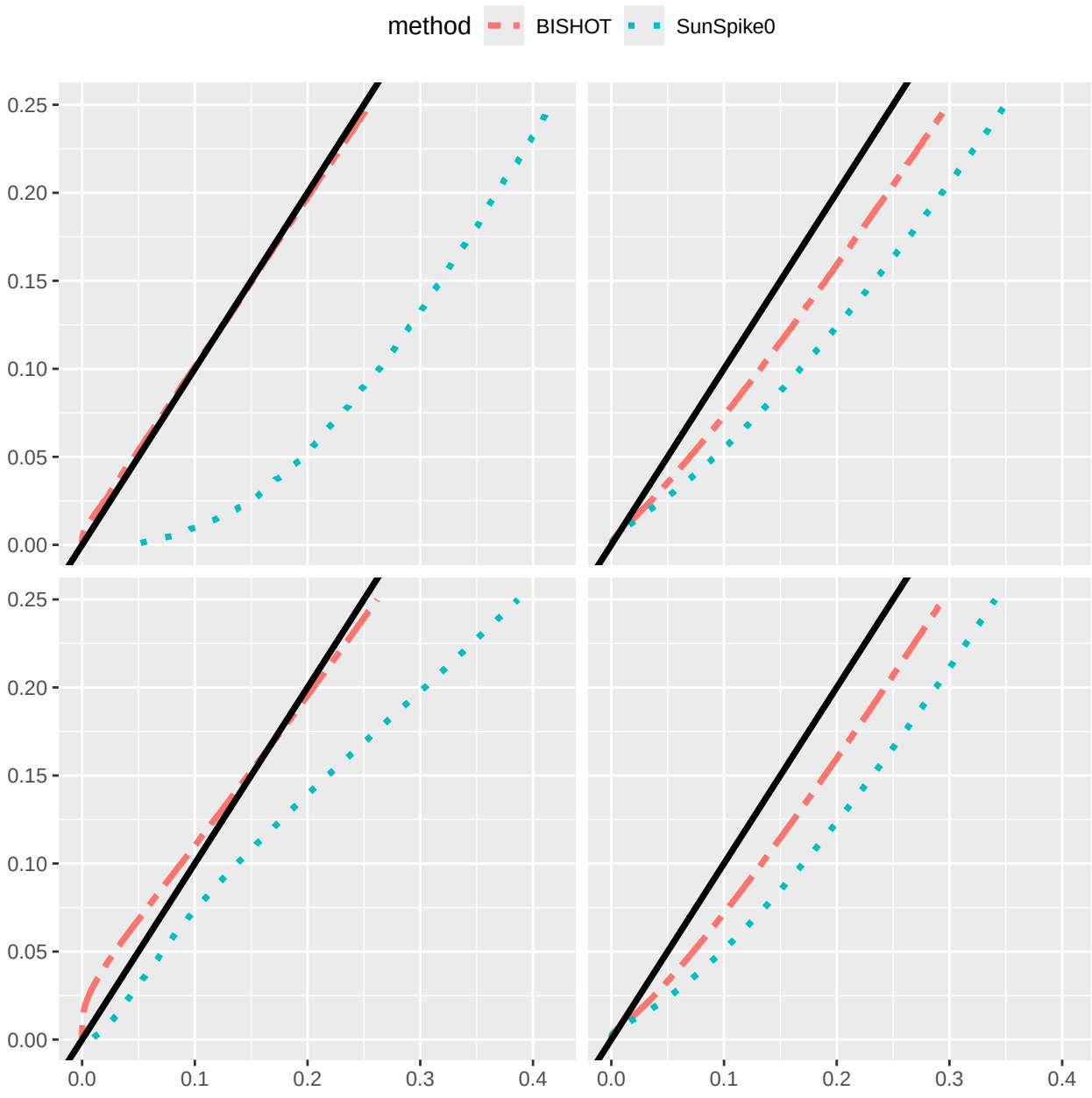

Figure 2: The actual signed false discovery rate (SFDR) when $\epsilon = \log_2(1.5)$ averaged over 50 simulated data using mSFDR threshold 0.001 to 0.25 corresponding to 10 (left panel) or 50 (right panel) samples in the treatment and control group. The top and bottom panel displays the case when $p = 0.1$ and $p = 0.5$. The dashed black line corresponds to the $y = x$ line.

We next compare how the top-ranked genes match the true DE genes. For this purpose, we select the first R genes, $\widehat{\Omega}_R$, ranked by their SFDR based q values (the minimum SFDR you can achieve by calling a given test significant) and report the proportion of those fall in Ω whose absolute values of β_g are greater than ϵ as well, that is, $p_R = \frac{\#\{g \in \widehat{\Omega}_R \cap \Omega : |\beta_g| < \epsilon\}}{R}$. The value R is selected from 1 to the total number of DE genes plus 100, that is, $Gp + 100$. The averaged proportion p_R for all simulations against R is shown in Figure 3 and 4 corresponding to $\epsilon = \log_2(1.2)$ and $\epsilon = \log_2(1.5)$. The R highest ranked genes by BISHOT are never falsely detected for even relatively large R (relative to the size of Ω), while SunSpike0 does not perform as good especially when expression sample size is small regardless of p and ϵ. Hence BISHOT is more robust and their top ranked genes are more accurate.

5 Transcriptional effects of BRD9 inhibition in acute myeloid leukemia

Acute myeloid leukemia (AML) is a type of blood cancer with a five-year survival rate of 31.9% (Surveillance Research Program, National Cancer Institute, 2025). Hohmann et al. (2016) investigated the potential of BRD9, a subunit of the SWI/SNF chromatin remodeling complex that promotes cell proliferation, to serve as a therapeutic target for AML. Hohmann et al. (2016) conducted two RNAseq experiments, one using murine RN2 cell line and the other using human MV4 cell line. In each experiment, cells were treated with BI-7273, a BRD9 inhibitor, or control with two replicates per group. RNAseq data were generated to elucidate the transcriptional effects of BRD9 inhibition.

We applied BISHOT to the Hohmann RNAseq data, which were transformed into $\log_2(\text{CPM} + 1)$ values. We focused on identifying DE genes from human MV4 cell line while using the RNAseq data from murine RN2 cell line as prior knowledge. Specifically, the parameter h_g was specified based on the LFC for gene g in the data from the murine RN2 cell line. Two options for ϵ are considered, $\log_2(1.2)$ or $\log_2(1.5)$, same as what are used in Section 4.2.

In Figure 5 we illustrate the effect of shrinkage by examining the relationship of the local

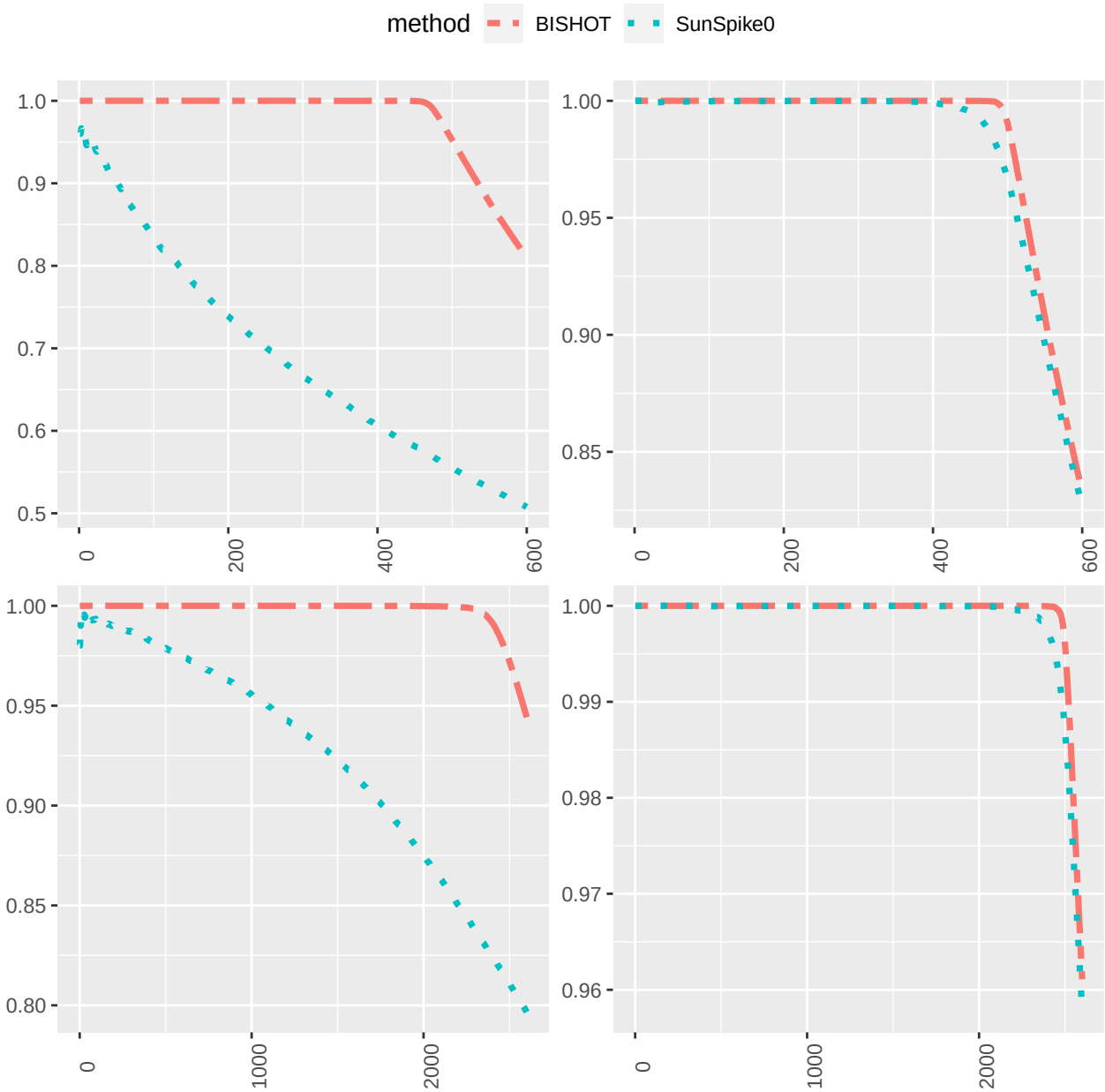

Figure 3: The proportion of true DE genes for the first R genes with smallest q values when $\epsilon = \log_2(1.2)$ averaged over 50 simulated data for $R = 1, \ldots, 5000p + 100$ corresponding to 10 (left panel) or 50 (right panel) samples in the treatment and control group. The top and bottom panel displays the case when $p = 0.1$ and $p = 0.5$.

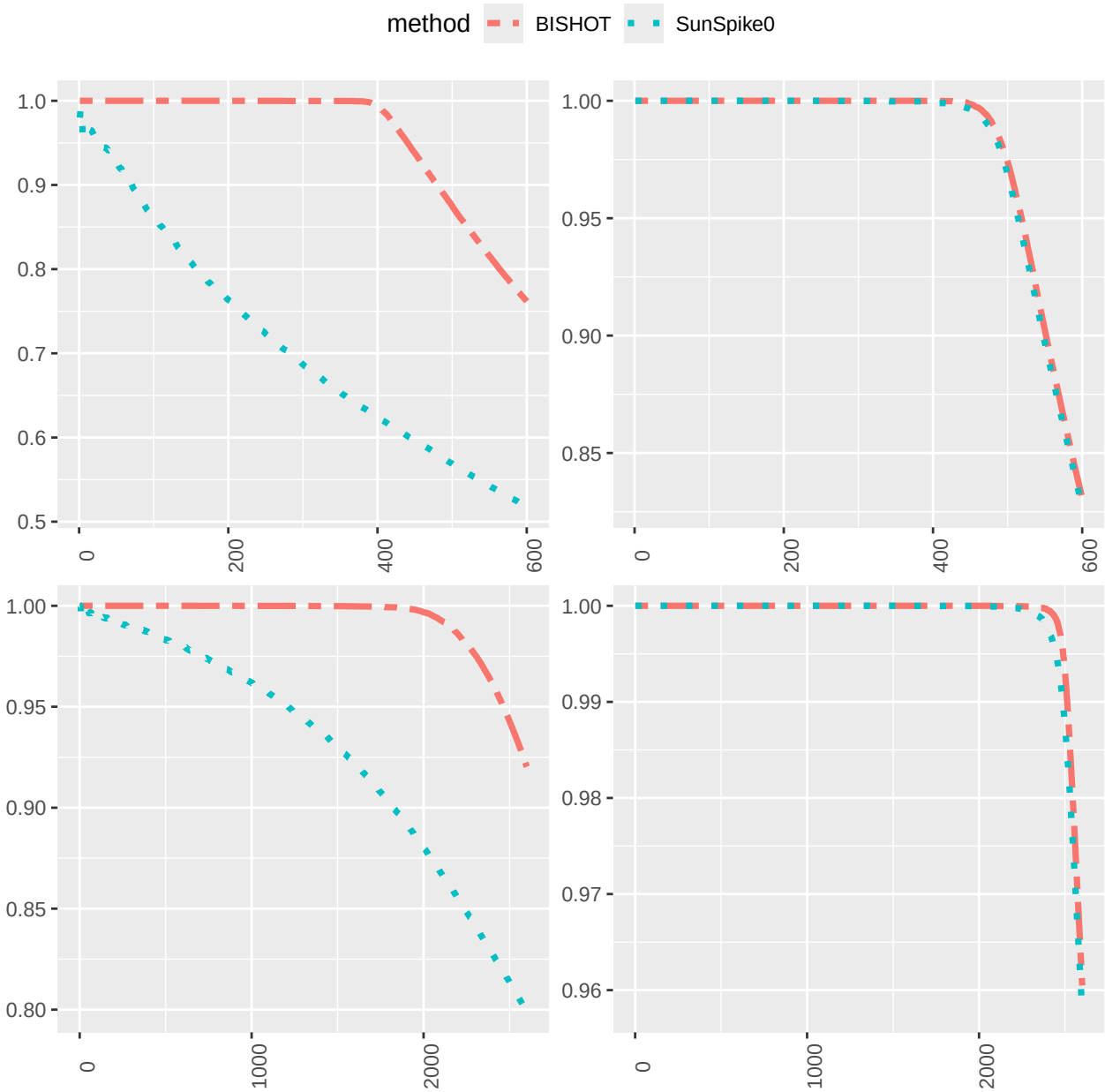

Figure 4: The proportion of true DE genes for the first R genes with smallest q values when $\epsilon = \log_2(1.5)$ averaged over 50 simulated data for $R = 1, \ldots, 5000p + 100$ corresponding to 10 (left panel) or 50 (right panel) samples in the treatment and control group. The top and bottom panel displays the case when $p = 0.1$ and $p = 0.5$.

shrinkage λ_g with the consistency between prior knowledge based on RN2 and what the MV4 data conveys about β_g, measured by the distance between h_g and the maximum likelihood estimator $\widehat{\beta}_g$. The shrinkage is in accordance to this consistency, to be more specific, it is prominent when $|\widehat{\beta}_g - h_g|$ is small and less so when $|\widehat{\beta}_g - h_g|$ is large.

Figure 6 presents the identified DE genes at SFDR < 0.05 based on BISHOT relative to LFCs present in the RN2 and MV4 data. Almost all the genes with consistently large absolute LFCs ($> \epsilon$) in both RN2 and MV4 (top-right and bottom-left regions) were identified as DE, except for a few genes with high variations. Conversely, no genes with consistently small LFCs ($< \epsilon$) in both RN2 and MV4 (middle region) were identified as significant. Hence, by sharing information, the prior data from RN2 reinforced the analysis of MV4 data in determining DE genes, thanks to the reduction of posterior uncertainty when information is consistent.

For genes with large LFCs in MV4 but relatively small LFCs in RN2 (vertical band outside the middle square), many were also identified as DE, demonstrating that our method prioritizes genes strongly supported by observed data, even when prior information is limited. This is important, as differences between prior data (e.g., mice) and the data under analysis (e.g., human) are expected, and the method should be able to detect features specific to the current dataset. Interestingly, many genes with large LFCs in RN2 but relatively small LFCs in MV4 (horizontal band outside the middle square) were not identified as differentially expressed, likely due to inconsistent LFCs or high variability in MV4. This result highlights that prior information alone is insufficient. Genes must be supported by the data under analysis to be considered significant.

Figure 6 also highlights the difference in DE results between BISHOT and SunSpike0. Genes that were called as DE by SunSpike0 only (green dots) tend to be the ones with absolute LFCs $> \epsilon$ in the MV4 but not in the RN2 data. In contrast, genes that were called as DE by BISHOT only (red dots) tend to be the ones with LFCs consistently greater than ϵ in both MV4 and RN2 data. In other words, BISHOT preferentially identifies signals that exhibit consistency between the two datasets. A Venn diagram comparing the number of DE genes between BISHOT and SunSpike0 is provided in the Appendix B.

We also compared the results from BISHOT with other popular methods available in `R`

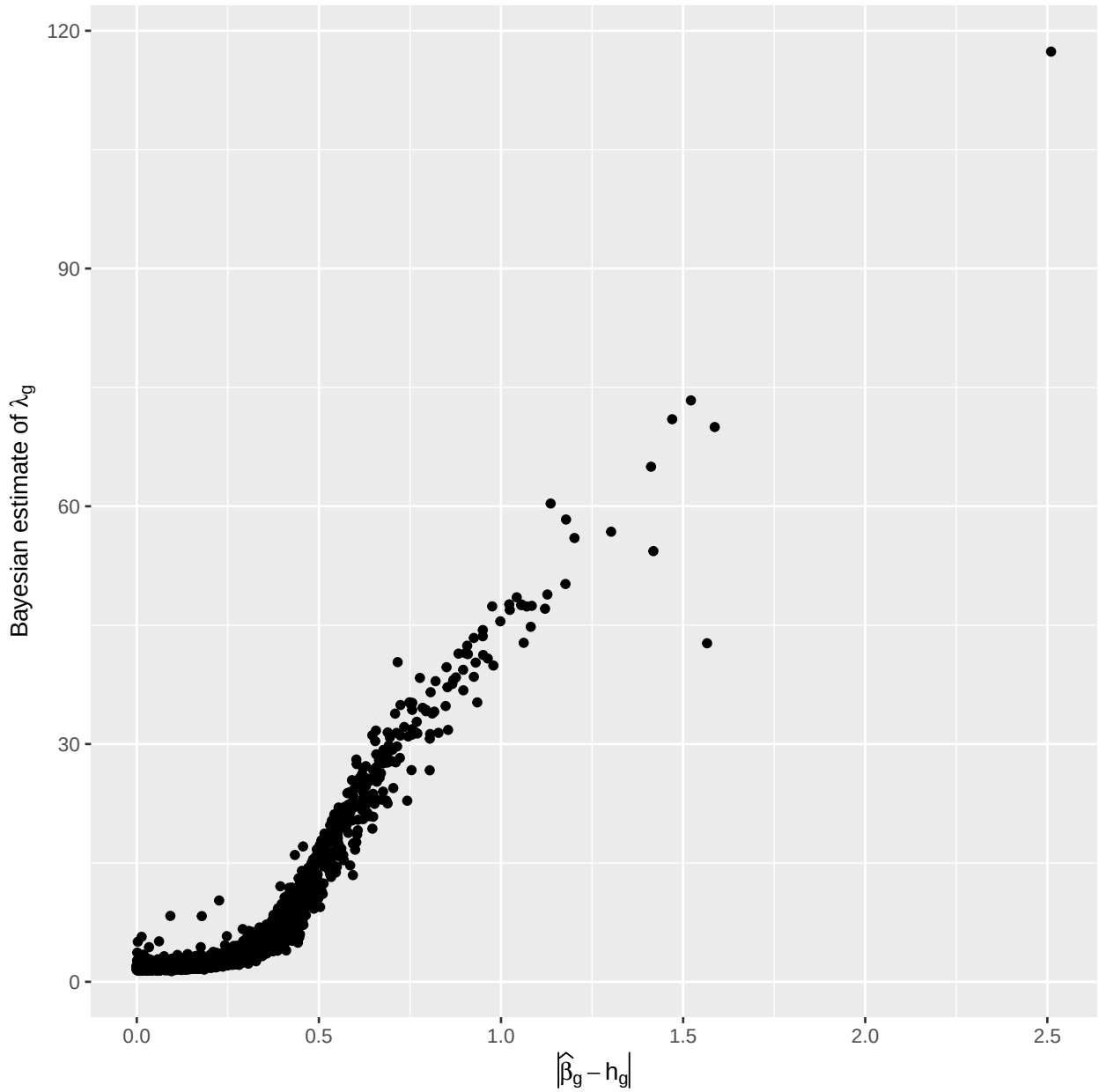

Figure 5: The relationship between the gene-specific shrinkage parameter and the distance between the two sources of information. The former is estimated by the posterior mean and the latter is by the absolute value of the difference between the Maximum likelihood estimator and the center of the prior distribution.

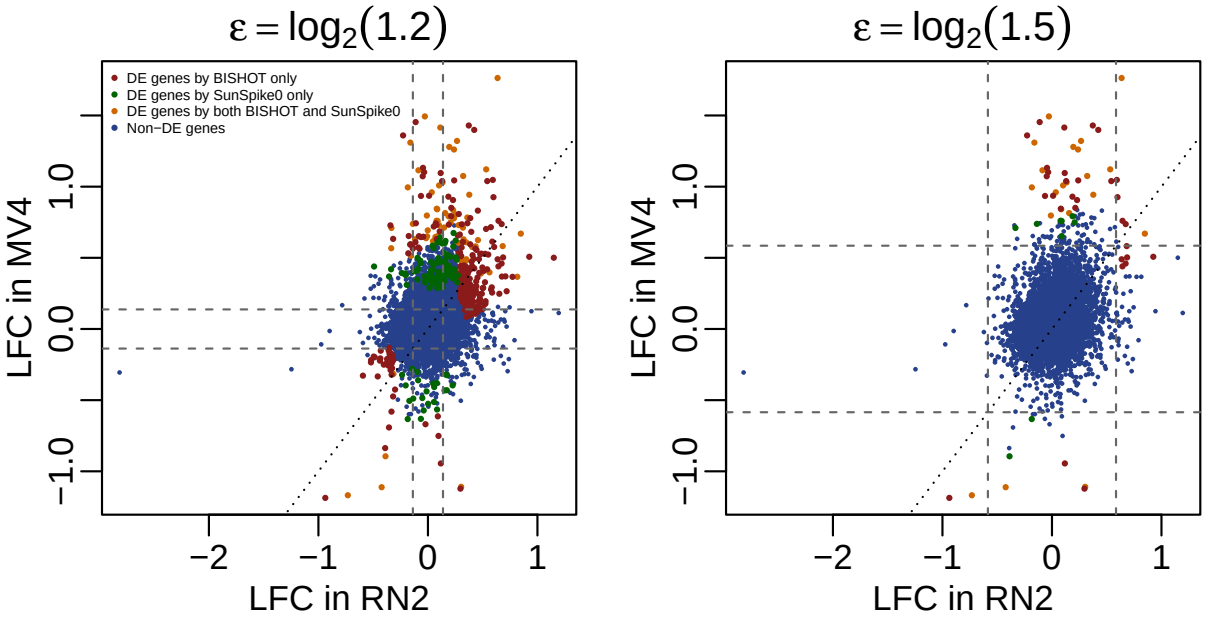

Figure 6: Gene expression LFCs between drug-treated and control groups based on RN2 (x-axis) and MV4 (y-axis) cell lines. Each dot represents a gene, colored according to DE results from the MV4 cell line. Left panel: $\epsilon = \log_2(1.2)$; right panel: $\epsilon = \log_2(1.5)$.

including limma (Smyth, 2005), voom (Law et al., 2014), and edgeR (Robinson et al., 2010). The limma (using the `lmFit` function in the `limma` package with $\log_2(\text{CPM} + 1)$ values as the input), voom (using the `voomLmFit` function in the `limma` package with raw counts as the input), and edgeR (using the `edgeR` package with raw counts as the input) methods did not identify any DE genes at FDR < 0.05.

A key gene of interest in Hohmann et al. (2016) is *MYC*, where the authors showed that BRD4 plays a critical role in sustaining *MYC* transcription. We compared the significant level of *MYC* gene across different differential analysis methods. BISHOT with an $\epsilon = \log_2(1.2)$ yielded a q-value of 0.02, whereas all other methods produced non-significant q-values (Figure 7). Because *MYC* had a moderate LFC of -0.33 in the MV4 data, it was missed by differential analysis methods relying solely on that dataset. However, because *MYC* also showed a LFC of -0.36 in the RN2 data, which was highly consistent with that in the MV4 data, BISHOT with an $\epsilon = \log_2(1.2)$ was able to leverage such prior information to identify the consistent down-regulation of *MYC*. In addition, we noted that BISHOT with an $\epsilon = \log_2(1.5)$ produced a non-significant q-value for *MYC* because the method emphasizes

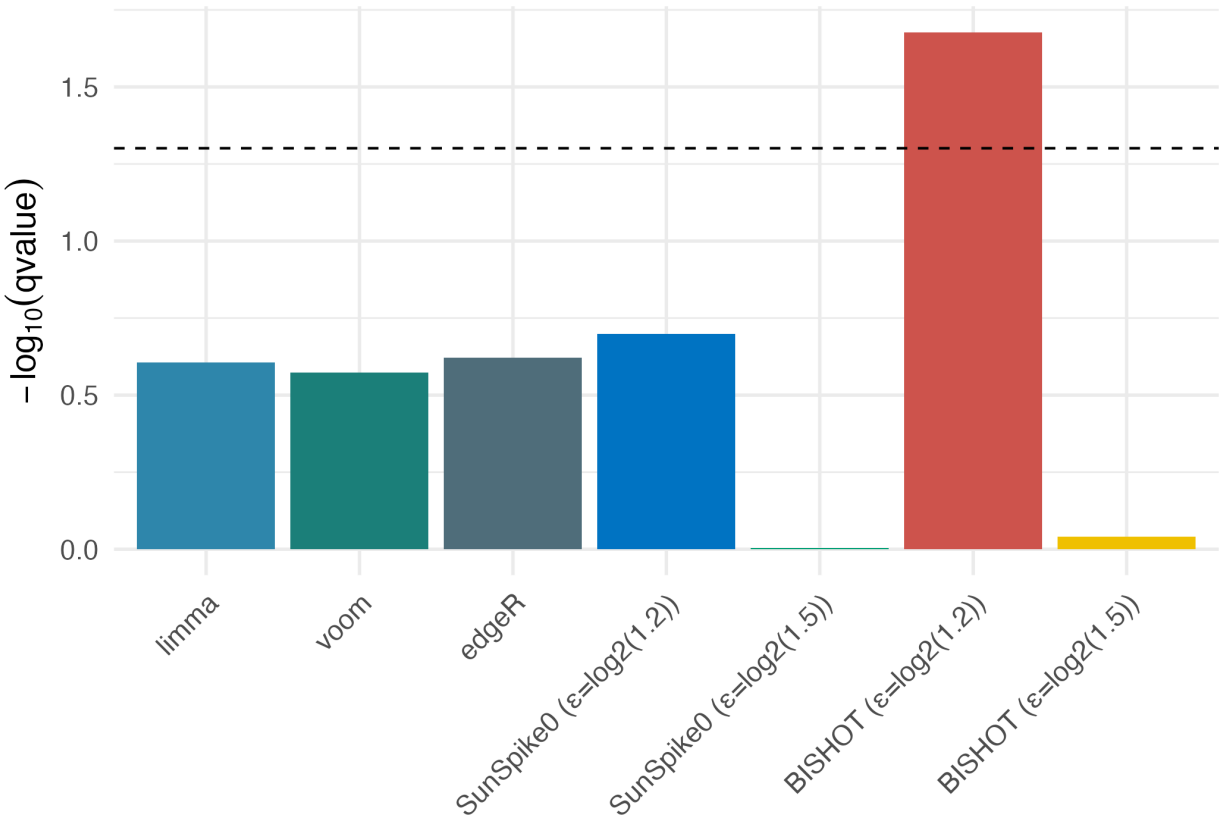

Figure 7: Comparison of the q-value for MYC across different methods. The dashed horizontal line indicates q-value=0.05.

genes with large fold changes and therefore did not capture MYC given its moderate fold change.

6 The scRNAseq analysis of the aging lung

Angelidis et al. (2019) conducted a scRNAseq study in old (24 months) and young (3 months) mice to investigate cell type and gene expression alterations in lung aging. One of the most prevalent cell types identified in the study was type-2 pneumocytes. In parallel, the authors also conducted a bulk RNAseq experiment on type-2 pneumocytes, selected by flow cytometry sorting, from old and young mice. We applied BISHOT to scRNAseq data for type-2 pneumocytes to identify differentially expressed genes between old and young mice. The flow-sorted bulk RNAseq data for type-2 pneumocytes were used as our prior knowledge.

We first extend our model to handle scRNAseq data, which contains dropouts. Let Y_{ig} be the $\log_2(\text{CPM}+1)$ expression value of gene g in cell i. Suppose Z_{ig} is a dropout indicator, where $Z_{ig} = 1$ if $Y_{ig} \neq 0$ and $Z_{ig} = 0$ if $Y_{ig} = 0$. We adopt the two-component model developed in Finak et al. (2015): a logistic regression for Z_{ig} and a classic regression model for Y_{ig} conditioning on $Z_{ig} = 1$,

$$\begin{aligned} \Pr(Z_{ig} = 1) &= \frac{1}{1 + \exp(-c_g^D - X_i^\top \beta_g^D)}, \\ Y_{ig}|Z_{ig} = 1 &\sim \mathrm{N}(c_g^C + X_i^\top \beta_g^C, \sigma_g^2). \end{aligned}$$

where corresponding to the probability component and the nonzero component, the parameters are superscripted by D or C. One would identify two treatment effects β_g^D and β_g^C representing differentiations in terms of log odds for the probability of nonzero and the expected nonzero expression.

The question becomes which genes should be claimed significant based on either β^D or β^C. Equivalently, we aim at a bivariate hypothesis for each gene, $H_{0g} : \boldsymbol{\beta}_g \in \boldsymbol{A}$ vs $H_{ag} : \boldsymbol{\beta}_g \notin \boldsymbol{A}$ with $\boldsymbol{\beta}_g = (\beta_g^D, \beta_g^C)^\top$, $\boldsymbol{A} = (-\epsilon_1, \epsilon_1) \times (-\epsilon_2, \epsilon_2)$. The choices of $\epsilon_1 \in \{\log(1.2), \log(1.5)\}$ and $\epsilon_2 \in \{\log_2(1.2), \log_2(1.5)\}$ correspond to a LFC or log odds ratio of 1.2 or 1.5 converted to the scale of Z or Y. To this end, we need to adapt the BCR and SFDR to the bivariate case. BCR is motivated to incorporate the information about sign of the parameter, in this

case, four scenarios regarding the signs for β_g^C and β_g^D are possible,

$$\omega_g = \begin{cases} 1, & \text{if } \beta_g^D \in (-\infty, -\epsilon_1) \text{ and } \beta_g^C \in (-\infty, -\epsilon_2), \\ 2, & \text{if } \beta_g^D \in (\epsilon_1, \infty) \text{ and } \beta_g^C \in (\epsilon_2, \infty), \\ 3, & \text{if } \beta_g^D \in (-\infty, -\epsilon_1) \text{ and } \beta_g^C \in (\epsilon_2, \infty), \\ 4, & \text{if } \beta_g^D \in (\epsilon_1, \infty) \text{ and } \beta_g^C \in (-\infty, -\epsilon_2). \end{cases}$$

so the bivariate BCR is defined as $\mathrm{C}_{0,\boldsymbol{\beta}_g}/(1 - \mathrm{C}_{a,\boldsymbol{\beta}_g})$ where $\mathrm{C}_{0,\boldsymbol{\beta}_g} = \Pr(\boldsymbol{\beta}_g \in \boldsymbol{A}|D)$ and $\mathrm{C}_{a,\boldsymbol{\beta}_g} = \min\{\Pr(\omega_g = 1|D), \Pr(\omega_g = 2|D), \Pr(\omega_g = 3|D), \Pr(\omega_g = 4|D)\}$. Similarly, mSFDR can be adjusted based on these four cases (expressions are omitted here for conciseness).

We applied BISHOT to the scRNAseq data while incorporating LFCs from bulk RNAseq (denoted as μ_g) in the prior, specifically, $\boldsymbol{h}_g = (\kappa\mu_g, \mu_g)^\top$ where κ is a parameter potentially used to specify the sign and size of β_g^D through μ_g. With $\epsilon_1 = \epsilon_2 = \log_2(1.5)$, we identified 436 DE genes at $SFDR = 0.05$. When $\epsilon_1 = \epsilon_2 = \log_2(1.2)$, this number increased to 2642. For comparison, we applied MAST (Finak et al., 2015) to the data using the `FindMarkers` function in the `R Seurat` package (Hao et al., 2024), which yielded 108 DE genes. Figure 8 left panel presents the overlaps in DE genes across these methods: BISHOT with larger $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)^\top$, identified 65 out of the 108 DE genes from MAST. For lower $\boldsymbol{\epsilon}$, BISHOT was able to identify almost all the DE genes from MAST. Notably, BISHOT identified 371 additional DE genes, within the intersection of genes identified using both thresholds, that were not identified by MAST.

To further investigate the biological interpretations of the DE genes, we performed Gene Ontology (GO) enrichment analysis based on the `GSEA` software (Subramanian et al., 2005). As shown in Figure 8 right panel, there were 60 significant (FDR<0.05) GO terms that were identified by BISHOT with both thresholds, but not by MAST. The Appendix B provides a list of these GO terms, many of which are related to lung aging. The endothelial barrier (GO:0061028) is crucial for lung function, regulating substance exchange between blood and tissue. Senescence of endothelial cells during aging compromises this barrier, increasing susceptibility to acute lung injury and acute respiratory distress syndrome (Barabutis et al., 2016; Najari Beidokhti et al., 2025). Mitochondria, essential for energy, senescence, apoptosis, and regeneration of type-2 pneumocytes, show impaired function with age due to al-

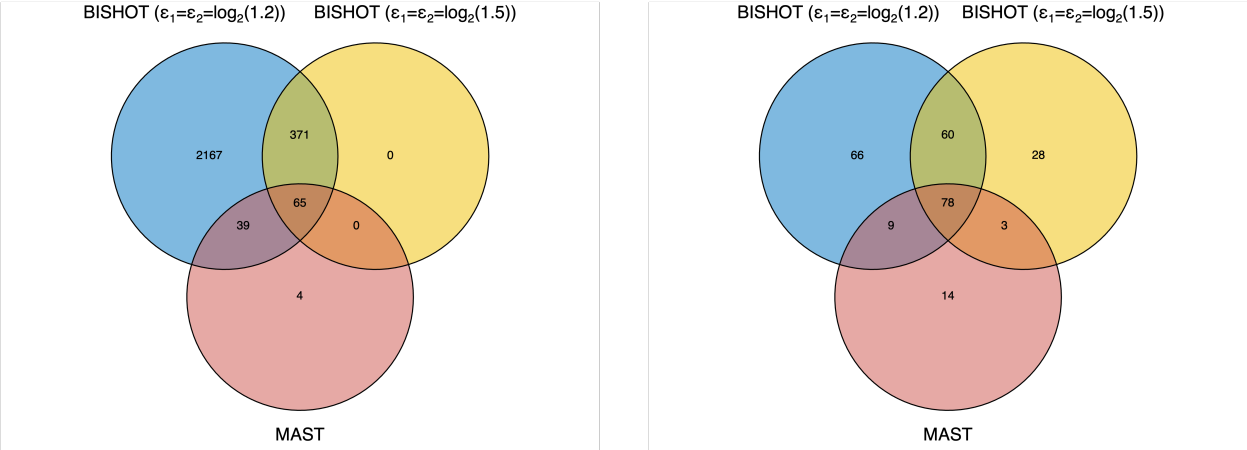

Figure 8: Venn diagrams comparing different methods in terms of identified DE genes (left panel) and enriched GO terms (right panel).

tered expression of mitochondrial genes and ribosomal subunits (GO:0140053, GO:0005763, GO:0005762) (Cloonan et al., 2020). Cell morphogenesis (GO:0000902) disruptions contribute to lung aging, marked by alveolar enlargement and structural changes known as the "senile lung" (Wang et al., 2024). Protein serine/threonine kinases (GO:0004674) influence age-related lung disease. For example, PINK1 regulates mitophagy, with its dysregulation linked to fibrosis and COPD (Mizumura et al., 2014). Overexpression of MAP kinase-interacting serine/threonine kinase 2 promotes lung cancer progression (Guo et al., 2017), while protein kinase B (Akt) signaling protects myofibroblast from apoptosis, making it a therapeutic target for lung fibrosis (Wang et al., 2022).

7 Discussions

As far as we know there is an absence in the literature where a multiple testing methodology was applied to $\boldsymbol{\beta}_g \in \mathbf{R}^p$ when $p > 1$. We have seen a success for $p = 2$ in the single cell data application in Section 6 on non-normally distributed data. This conveys a positive signal of the flexibility of BISHOT regarding model and parameter choices and therefore is of interest on a separate note to the multiple testing regime.

We consider a composite null hypothesis in this paper as desired by the practical problems we described. It is observed that the role of ϵ is connected with conservativeness in terms of

the actual SFDR. More thorough investigation regarding the combined effect of ϵ, the true distribution of β_g and the accuracy of h_g (these factors are what we believe contribute to the degree of conservativeness of our method) is beyond the scope of this paper and is left as potentially a future work.

BISHOT offers a general framework for Bayesian differential analysis that incorporates prior knowledge with relatively low computational cost and remarkable accuracy. In this paper, we illustrate the framework using normal linear and two-component models, though it is broadly applicable to other models. Extending BISHOT to other frequently used models, such as the negative binomial model (Robinson et al., 2010; Wu et al., 2013), is a direction for future research.

In this work, we limit ourselves to the specifications of the shrinkage parameters λ_g, τ based on the original horseshoe. As we explained in Section 5 there exists a close proximity between the notion of sparsity (in nonzero parameters) and the consistency between h_g and $\widehat{\beta}_g$. Given the well-known drawbacks for the horseshoe prior distributions, we can extend the proposed framework to be combined with more recent developments regarding the horseshoe prior (Bhadra et al., 2017), in order to tackle the various scenarios regarding the overall consistency between the two studies.

Acknowledgements

High Performance Computing resources provided by the High Performance Research Computing (HPRC) core facility at Virginia Commonwealth University (https://hprc.vcu.edu) were used for conducting the research reported in this work. This research was supported by the Biostatistics and Bioinformatics Shared Resource of the University of Kentucky Markey Cancer Center (P30CA177558).

Data availability

The AML data (Hohmann et al., 2016) used in Section 5 is available from GEO under accession number GSE79284. The lung aging data (Angelidis et al., 2019) used in Section 6

is available from Zenodo at https://doi.org/10.5281/zenodo.5048449, which is provided by Squair et al. (2021).

Supporting Information

Appendix A referenced in Sections 2 and 4 and Appendix B, referenced in Sections 5 and 6, is available with this paper at the journal website.

References

Altenbuchinger, M., Schwarzfischer, P., Rehberg, T., Reinders, J., Kohler, C. W., Gronwald, W., Richter, J., Szczepanowski, M., Masqué-Soler, N., Klapper, W., et al. (2017). Molecular signatures that can be transferred across different omics platforms. *Bioinformatics*, 33, i333–i340.

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T.-M., et al. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10, 963.

Barabutis, N., Verin, A., and Catravas, J. D. (2016). Regulation of pulmonary endothelial barrier function by kinases. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 311, 832–845.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12, 1105–1131.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103, 985–991.

Chen, J., Xu, J., Gou, L., Zhu, Y., Zhong, W., Guo, H., and Du, Y. (2024). Integrating transcriptomic and proteomic data for a comprehensive molecular perspective on the association between sarcopenia and osteoporosis. *Archives of Gerontology and Geriatrics*, 125, 105486.

Cloonan, S. M., Kim, K., Esteves, P., Trian, T., and Barnes, P. J. (2020). Mitochondrial dysfunction in lung ageing and disease. *European Respiratory Review*, 29, 200165.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23, 1–22.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16, 278.

Fu, L., Gang, B., James, G. M., and Sun, W. (2022). Heteroscedasticity-adjusted ranking and thresholding for large-scale multiple testing. *Journal of the American Statistical Association*, 117, 1028–1040.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 499–517.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061.

Gontarz, P., Fu, S., Xing, X., Liu, S., Miao, B., Bazylianska, V., Sharma, A., Madden, P., Cates, K., Yoo, A., et al. (2020). Comparison of differential accessibility analysis strategies for atac-seq data. *Scientific reports*, 10, 10150.

Guo, Z., Peng, G., Li, E., Xi, S., Zhang, Y., Li, Y., Lin, X., Li, G., Wu, Q., and He, J. (2017). Map kinase-interacting serine/threonine kinase 2 promotes proliferation, metastasis, and predicts poor prognosis in non-small cell lung cancer. *Scientific Reports*, 7, 10612.

Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42, 293–304.

Hohmann, A. F., Martin, L. J., Minder, J. L., Roe, J.-S., Shi, J., Steurer, S., Bader, G., McConnell, D., Pearson, M., Gerstberger, T., et al. (2016). Sensitivity and engineered resistance of myeloid leukemia cells to brd9 inhibition. *Nature Chemical Biology*, 12, 672–679.

Huang, Z., Lane, A. N., Fan, T. W., Higashi, R. M., Weiss, H. L., Yin, X., and Wang, C. (2020). Differential abundance analysis with Bayes shrinkage estimation of variance (dasev) for zero-inflated proteomic and metabolomic data. *Scientific Reports*, 10, 876.

Kendziorski, C., Newton, M., Lan, H., and Gould, M. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22, 3899–3914.

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15, 1–17.

Li, B., Li, Y., and Qin, Z. S. (2017). Improving hierarchical models using historical data with applications in high-throughput genomics data analysis. *Statistics in biosciences*, 9, 73–90.

Li, B., Sun, Z., He, Q., Zhu, Y., and Qin, Z. S. (2016). Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics*, 32, 682–689.

Li, C., Lanman, N. A., Kong, Y., He, D., Mao, F., Farah, E., Zhang, Y., Liu, J., Wang, C., Wei, Q., et al. (2020). Inhibition of the erythropoietin-producing receptor ephb4 antagonizes androgen receptor overexpression and reduces enzalutamide resistance. *Journal of Biological Chemistry*, 295, 5470–5483.

Li, C., Liu, J., He, D., Mao, F., Rao, X., Zhao, Y., Lanman, N. A., Kazemian, M., Farah, E., Liu, J., et al. (2022). Gstm2 is a key molecular determinant of resistance to sg-aris. *Oncogene*, 41, 4498–4511.

Li, Y., Fan, T. W., Lane, A. N., Kang, W.-Y., Arnold, S. M., Stromberg, A. J., Wang, C., and Chen, L. (2019). Sda: A semi-parametric differential abundance analysis method for metabolomics and proteomics data. *BMC bioinformatics*, 20, 1–10.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15, 1–21.

McCarthy, D. J. and Smyth, G. K. (2009). Testing significance relative to a fold-change threshold is a treat. *Bioinformatics*, 25, 765–771.

Mizumura, K., Cloonan, S. M., Nakahira, K., Bhashyam, A. R., Cervo, M., Kitada, T., Glass, K., Owen, C. A., Mahmood, A., Washko, G. R., et al. (2014). Mitophagy-dependent necroptosis contributes to the pathogenesis of copd. *The Journal of Clinical Investigation*, 124, 3987–4003.

Muralidharan, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics*, 4, 422–438.

Najari Beidokhti, M., Villalba, N., Ma, Y., Reynolds, A., Villamil, J. H., and Yuan, S. Y. (2025). Lung endothelial cell senescence impairs barrier function and promotes neutrophil adhesion and migration. *GeroScience*, pages 1–17.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26, 139–140.

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481, 389–393.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12, 5692.

Stephens, M. (2017). False discovery rates: a new deal. *Biostatistics*, 18, 275–294.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics*, 31, 2013–2035.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102, 15545–15550.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102, 901–912.

Sun, W. and McLain, A. C. (2012). Multiple testing of composite null hypotheses in heteroscedastic models. *Journal of the American Statistical Association*, 107, 673–687.

Surveillance Research Program, National Cancer Institute (2025). Seer*explorer: An interactive website for SEER cancer statistics.

Vitorino, R. (2024). Transforming clinical research: the power of high-throughput omics integration. *Proteomes*, 12, 25.

Wang, J., Hu, K., Cai, X., Yang, B., He, Q., Wang, J., and Weng, Q. (2022). Targeting pi3k/akt signaling for treatment of idiopathic pulmonary fibrosis. *Acta Pharmaceutica Sinica B*, 12, 18–32.

Wang, K. Y., Pupo, G. M., Tembe, V., Patrick, E., Strbenac, D., Schramm, S.-J., Thompson, J. F., Scolyer, R. A., Muller, S., Tarr, G., et al. (2022). Cross-platform omics prediction procedure: a statistical machine learning framework for wider implementation of precision medicine. *npj Digital Medicine*, 5, 85.

Wang, Y., Huang, X., Luo, G., Xu, Y., Deng, X., Lin, Y., Wang, Z., Zhou, S., Wang, S., Chen, H., et al. (2024). The aging lung: microenvironment, mechanisms, and diseases. *Frontiers in Immunology*, 15, 1383503.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14, 232–243.