Title
Statistical criteria and procedures for controlling false positives with applications to biological and biomedical data analysis

Author
Chen, Yiling

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Statistical criteria and procedures for controlling false positives with applications to biological and biomedical data analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Yiling Chen

2021

ABSTRACT OF THE DISSERTATION

Statistical criteria and procedures for controlling false positives with applications to biological and biomedical data analysis

by

Yiling Chen

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2021

Professor Jingyi Jessica Li, Chair

The need to control rates of false positives is prevalent in biological and biomedical data analysis. Two statistical conceptualizations of rates of false positives—type I error and false discovery rate (FDR)— are widely used in these analyses. For example, in automated cancer detection from transcriptomics data, practitioners often need to control type I error—the conditional probability of making a false positive as healthy—because false negatives could lead to severe consequences such as delayed treatment or even life loss. In contrast, a false positive leads to less serious consequences. Another example is the widely-used FDR control in multiple-testing problems such as differential expression genes identification from RNA sequencing data. Because discoveries are often subject to laborious and expensive downstream validation, researchers want to control the FDR—the expected proportion of false discoveries among discoveries—to save validation costs; in comparison, missing true discoveries is often less concerning. Despite existing efforts, controlling rates of false positives remain challenging. This dissertation aims to address them in three projects.

My first project involves prioritizing type I error of feature selection for binary classification problems. Binary classification problems are prevalent in biomedical data analysis: for example, the aforementioned automated cancer detection where the response is binary: with or without cancer. In those cases, type I error control, i.e., false positive rate control, is critical so that the chance of missing cancer patients is under a reasonable level, a

consideration neglected by existing model selection methods. In Chapter 2, we develop a novel model selection criterion, Neyman-Pearson Criterion (NPC), that prioritizes the type I error in binary classification. The theoretical model selection property of NPC is studied for non-parametric plug-in methods. A real data study on breast cancer detection using DNA methylation data suggests that NPC is a practical criterion that can reveal novel clinical biomarkers for cancer diagnosis with both high sensitivity and specificity.

My second project focuses on FDR control in high-throughput data analysis from two conditions. High-throughput data analysis commonly involves the identification of "interesting" features (e.g., genes, genomic regions, and proteins), whose values differ between two conditions. To ensure the reliability of such analysis, existing bioinformatics tools primarily use the FDR as the criterion, the control of which typically requires p-values. However, obtaining valid p-values is often hard or even impossible because of limited sample sizes in high-throughput data. In Chapter 3, we propose Clipper, a p-value-free FDR control framework for high-throughput data with two conditions. Through comprehensive simulation and real-data benchmarking, Clipper outperforms existing generic FDR control methods and specific bioinformatics tools designed for various tasks, including differentially expressed gene identification from RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data.

My third project focuses on FDR control in aggregating peptides identified by multiple database search algorithms from mass spectrometry data. The state-of-the-art shotgun proteomics analysis relies on database search algorithms to identify peptides and proteins in biological samples. A key step in this process is peptide identification, which is done via matching mass spectra that code the sequence information of a peptide against protein databases that contain known protein sequences. Numerous database search algorithms have been developed over time, each with distinct advantages in peptide identification. To utilize this, in Chapter 4 we develop a statistical framework, Aggregation of Peptide Identification Results (APIR), for combining peptide matching results from multiple database search algorithms with FDR control. We demonstrate using benchmark data that APIR achieves higher detection sensitivity than individual search algorithms do while maintaining FDR control.

Extensive real data studies show that APIR can uncover additional biologically meaningful proteins and post-translational modifications that are otherwise undetected by individual search algorithms.

The dissertation of Yiling Chen is approved.

Mark Stephen Handcock

Chad Hazlett

Xin Tong

Ying Nian Wu

Jingyi Jessica Li, Committee Chair

University of California, Los Angeles

2021

To my cats Xiao Hui and Xiao Bai without whom this dissertation would have been completed one month earlier.

To my boyfriend Siyuan who upon reading the previous line requested to be mentioned in this line.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Throughout my doctoral studies, I have received a great amount of support from my mentors, my collaborators, and my friends.

I would like to express my deepest gratitude to my advisor, Dr. Jingyi Jessica Li, who has been extremely supportive of my research and professional development. Her pursuit of scientific rigor, her dedication to studying important questions, and her passion for the interdisciplinary field of statistics and biology have always inspired me. I am grateful for her mentoring throughout my doctoral studies at UCLA.

I would like to thank Dr. Chad Hazlett, Dr. Mark Stephen Handcock, Dr. Ying Nian Wu, and Dr. Xin Tong for serving on my doctoral committee and for providing valuable feedback on my research. I would like to extend my appreciation to my collaborators Dr. Leo Wang from City of Hope and his research assistants MeiLu McDermott and Kyla Wyoshner, who made major contributions to the last part of this dissertation. I am especially grateful to Dr. Wei Li from UC Irvine for his mentorship in the last year of my doctoral studies.

I thank all current members of Jessica's group for their helpful discussions and suggestions about my research. Among them, I want to especially thank Xinzhou Ge, who is not only a collaborator but also a dear friend.

My sincere thanks also go to my families and friends for their love and support. I thank my longtime friends, Michelle Meng and Yiting Huang, for always being there when I need them. I also thank my parents, especially my mother, for her unwavering belief in me and her unconditional support of my career choice. Most importantly, I want to thank my boyfriend, Siyuan Huang, for his encouragement and support during the toughest moments of this journey and our cats Xiao Hui and Xiao Bai for their love and company.

VITA

2012-2016	B.S. in Mathematics, UCLA
2016-2021	Graduate Research Assistant, Department of Statistics, UCLA
2018-2019	Biomedical Big Data Training Grant matching funds, UCLA
2019	The Most Promising Statistician Award, UCLA

PUBLICATIONS

(* indicates equal contribution.)

Li, W. V., **Chen, Y.**, & Li, J. J. (2017). TROM: A testing-based method for finding transcriptomic similarity of biological samples. Statistics in biosciences, 9(1), 105-136.

Chen, Y., Tong, X. & Li, J. J. (2018). Budget-Constrained Feature Selection for Binary Classification: a Neyman-Pearson Approach. International Chinese Statistical Association (ICSA) China Conference (2018).

Ge, X.*, **Chen, Y. E.***, Song, D., McDermott, M., Woyshner, K., Manousopoulou, A., Li, W., Wang, L. D., & Li, J. J. (2020). Clipper: p-value-free FDR control on high-throughput data from two conditions. (manuscript)

Lyu, J., Li, J. J., Su, J., Peng, F., **Chen, Y. E.**, Ge, X., & Li, W. (2020). DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. Science advances, 6(46), eaba6784.

Chen, Y. E., McDermott, M., Woyshner, K., Wang, L. D., & Li, J. J. (2021). APIR: a flexible and powerful FDR control framework for aggregating peptide identification results for proteomics data. (manuscript)

CHAPTER 1

Introduction

High-throughput technologies are widely used to provide numerical measurements of system-wide biological features. Common examples include RNA sequencing (RNA-seq), which allows for genome-wide profiling of transcriptome landscapes, Hi-C, which captures genome-wide chromatin interaction regions, and mass spectrometry (MS)-based shotgun proteomics, which globally profiles complex protein mixtures. These high-throughput technologies have accelerated the generation of massive data, which has led to important scientific discoveries over the years [1–3].

High-throughput datasets often contain biological features measured under more than one condition, for example, experimental versus control or across developmental stages. Typical analyses of such data include both prediction and statistical inference. In prediction analyses, researchers often want to predict condition labels from biological features. A typical example is automated disease diagnosis, which aims to train classifiers that use measurements of biological features (such as gene expression levels measured by RNA-seq) from patients to predict their health conditions (such as healthy patients versus diseased patients). We refer to these problems as classification problems. In inference analyses, the most common type aims to identify "interesting" biological features that exhibit an elevated or differential measurement across conditions. For example, differential expression analysis based on RNA-seq data aims to identify genes that show different expression levels between conditions. Because such problems are formulated to test the mean difference between conditions for each biological feature, we refer to them as multiple testing problems. My dissertation will focus on high-throughput biomedical data with two conditions. Accordingly, we will restrict ourselves to discussing binary classification problems and (multiple) testing problems across

two conditions.

Although rooted in two different cultures, binary classification and (multiple) hypothesis testing share a similar goal: to construct a binary decision rule from the data for uncovering binary truths. In binary classification, the binary decision rule is learned by applying classification algorithms on the training data with known condition labels, and the binary truth is whether a new data point comes from one condition or the other. In classical hypothesis testing, which involves a single pair of null and alternative hypotheses, the binary decision rule is constructed by comparing the observed test-statistic with its theoretical distribution under the null hypothesis, and the binary truth is whether the null hypothesis or the alternative hypothesis holds. As an extension of the classical hypothesis testing, a multiple testing problem involves multiple pairs of null and alternative hypotheses and constructs one rule to decide whether each null hypothesis holds.

Given their common binary nature, both binary classification and (multiple) hypothesis testing share two types of errors. By convention, we use "positive" to refer to class 1 in binary classification and the fact that the alternative hypothesis holds and "negative " to refer to class 0 in binary classification and the fact that the null hypothesis holds. Then the aforementioned two types of errors are false positives, which means misclassifying a class 0 data point in binary classification and false rejecting the null in hypothesis testing, and false negatives, which means misclassifying a class 1 data point in binary classification and mistakenly not rejecting the null in hypothesis testing.

My dissertation focuses on the control of false positives in high-throughput data analysis with two conditions. There are two conceptualizations of the rate of false positives: type I error and false discovery rate (FDR). Type I error is termed under the classical hypothesis testing framework involving a single pair of the null and the alternative hypotheses. It is defined as the conditional probability of rejecting the null hypothesis, conditioning on that the null hypothesis is true. In multiple testing, family-wise error rate (FWER) was proposed as an analogy of type I error and is defined as the conditional probability of making one or more false rejection given that all null hypotheses are true. However, researchers have found that in practice controlling FWER (under a small value such as 0.05) is too stringent

and often leads to zero power. Motivated by this, FDR was proposed by Benjamini and Hochberg [4] in 1995 and is nowadays the most popular statistical criterion in multiple testing problems. FDR is defined as the expected proportion of true null hypotheses among the "discoveries". Here "a discovery" refers to a rejected null hypothesis; because the null hypothesis typically represents a biologically uninteresting event (e.g., a gene has the same expression level between cancer and healthy patients) that researchers would like to reject, rejecting a null hypothesis is also called making a discovery, hence the name false discovery rate.

1.1 Type I error control in feature selection under the binary classification framework

The first part of my dissertation involves type I error control in feature selection under the binary classification framework. It is motivated by automated disease diagnosis using high-throughput data; our goal is to build a binary classifier that differentiates patients with disease from healthy controls.

To realize this goal, we need to address two key issues specific to our problem. The first one lies in the asymmetric cost between making a false positive and making a false negative. From the patient's point of view, diagnosing a healthy patient as diseased often entails nothing more serious than some further health examinations. In contrast, diagnosing a diseased patient as healthy could lead to delayed treatment or even life loss, especially in cases of cancer diagnosis. Hence, the trained classifier needs to prioritize the more severe error over the less severe one. The second issue is the high cost of collecting high-throughput data, making it financially infeasible for clinical diagnosis to be made based on high-throughput data. Fortunately, existing low-throughput technologies, which measure only a few biological features, are cheaper and more practical. Therefore, to leverage the financial advantage of low-throughput technologies, it is necessary to identify from high-throughput data a small number of features for building a practical classifier.

Although numerous feature selection methods have been proposed in the field of statistics

and machine learning, none addresses the asymmetry in our problem [5–12]. To fill this gap, in Chapter 2 we propose Neyman-Pearson Criterion (NPC), a model selection criterion that prioritizes type I error in binary classification. NPC exploits the aforementioned similarity between hypothesis testing and binary classification and translates type I error control, a concept typically used in hypothesis testing, to binary classification. Specifically, suppose we code the more severe condition as 0 and the less severe condition as 1. By analogy, we define type I error in binary classification as the conditional probability of misclassifying a datapoint as 1 given that it comes from condition 0. NPC is a model-free criterion that adapts to nearly all classification algorithms. Given a candidate feature set, a classification method, and a user-specified type I error threshold, we split the data into two parts, build a binary classifier with type I error control on one part and calculate an NPC value as the empirical type II error on the other part. Consequently, NPC allows users to select among multiple candidate feature sets the one with the smallest type II error while controlling the type I error under the same threshold. We studied the theoretical model selection property of NPC for non-parametric plug-in methods. A real data study on breast cancer detection using DNA methylation data suggests that NPC is a practical criterion that can reveal novel clinical biomarkers for cancer diagnosis with both high sensitivity and specificity.

1.2 P-value free FDR control in multiple testing problems

The second part of my dissertation focuses on FDR control in multiple testing problems based on high-throughput data with two conditions. Examples of such problems include differentially expressed gene identification from RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data. Numerous bioinformatics tools with FDR control have been developed to perform such analyses; most of them rely on valid high-resolution p-value calculations for FDR control. Specifically, p-values are first calculated, one per biological feature (e.g., a gene), and are thresholded using predominantly the Benjamini-Horchberg (BH) procedure [4], the Storey's q-values [13] or other FDR control methods [14–17]. However, the calculation of p-values

requires either distribution assumptions, which are often questionable, or a large number of replicates, which are often unachievable in biological data. Due to these limitations, bioinformatics tools often output ill-posed p-values, which consequently leads to unreliable FDR control. Therefore, p-value-free FDR control is desirable, as it would make high-throughput data analysis more transparent and thus improve the reproducibility of scientific research.

In Chapter 3, we propose Clipper, a model-free and p-value-free FDR control framework for analyzing high-throughput data with two conditions [18]. Clipper applies to both enriched and differential features from high-throughput biological data of diverse types. In comprehensive simulation and real-data benchmarking, Clipper outperforms existing generic FDR control methods and specific bioinformatics tools designed for various tasks, including differentially expressed gene identification from RNA-seq data, differentially interacting chromatin region identification from Hi-C data, and peptide identification from mass spectrometry data. Our results demonstrate Clipper's flexibility and reliability for FDR control and its broad applications in high-throughput data analysis.

1.3 FDR control in aggregating multiple sets of high-throughput discoveries in the context of shotgun proteomics data

The third part of my dissertation focuses on FDR control in aggregating multiple high-throughput discoveries generated from shotgun proteomics data. Shotgun proteomics refers to a proteomics technique that aims to identify proteins in complex mixtures using a combination of high-performance liquid chromatography and tandem mass spectrometry (MS). During the experimental flows, the protein mixtures are digested into peptides. The resulting peptide mixture is separated and measured by tandem MS as mass spectra. Each mass spectrum encodes the chemical composition of a peptide and can be used to identify the amino acid sequence, detect the post-translational modifications, and quantify the abundance of the given peptide.

Since the development of shotgun proteomics, numerous database search algorithms have been developed to automate mass spectrum interpretation. A database search algorithm

takes as input the mass spectra from MS analysis and a protein database; it identifies for each mass spectrum the "best" matching peptide sequence from the database, where the "best" is defined based on individual algorithm's internal matching score. We call the resulting match a "peptide-spectrum match" (PSM). False PSMs could occur when mass spectra are matched to wrong peptide sequences due to issues such as low-quality spectra, data-processing errors, and incomplete protein databases, causing problems in the downstream protein identification and quantification. Therefore, a common goal of database search algorithms is to simultaneously control the FDR and maximize the number of identified PSMs, so as to maximize the number of proteins identified in a proteomics study.

It has been observed that different search algorithms capture distinct PSMs, which motivates the development of aggregation methods that combines the peptide identification results from multiple search algorithms. However, existing aggregation methods suffer two major drawbacks: the limited compatibility with various database search algorithms and the lack of guarantee of power increase [19–23]. To fill this gap, in Chapter 4 we propose Aggregation of Peptide Identification Results (APIR), a flexible and powerful FDR-control framework for aggregating peptides identified by multiple database search algorithms from mass spectrometry data. APIR is based on a simple intuition: given multiple disjoint sets of discoveries, each with the empirical false discovery rate under a user-specified value q, their union also has the empirical false discovery rate under q. Evaluation of APIR on a complex protein standard shows that APIR achieves higher detection sensitivity than individual search algorithms while maintaining FDR control. Real data studies show that APIR uncovers disease-related proteins that are missed by individual search algorithms.

1.4 Summary

During my doctoral study, I have developed three statistical methods that control false positives as a leading author. The details of these projects will be described in Chapter 2–4 of this dissertation. In addition, I have other collaborative work covering a wide range of topics, including developing Transcriptome Overlap Measure for comparing transcriptomes within

or between different species [24], discovering oncogenes and tumor suppressor genes using genetic and epigenetic features [25], and building an atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability [26]. These collaborative projects are omitted from this dissertation; interested readers can refer to the original papers for details.

CHAPTER 2

Type I error control in feature selection under the binary classification framework

2.1 Introduction

With the advance of high-throughput sequencing technologies, numerous genomic datasets have been generated and made publicly available to enable automated disease diagnosis and help improve human understanding of disease mechanisms. Automated disease diagnosis based on genomic data is naturally a binary classification problem, where the binary response variable indicates a subject's disease status, and the predictors are genomic features. However, genomic features are high dimensional (often of the order $10^4 - 10^7$), posing tremendous challenges for biomedical researchers to understand what features are most important for disease diagnosis. Therefore, model selection criteria are in great needs to provide informative disease-predictive feature subsets for downstream experiments.

In binary classification (class 0 vs. class 1), three practical evaluation criteria for accuracy include: the overall classification error (i.e., the probability that an observation is misclassified), the type I error (i.e., the conditional probability of misclassifying a class 0 observation into the class 1, or 1−specificity), and the type II error (i.e., the conditional probability of misclassifying a class 1 observation into the class 0, or 1−sensitivity). The overall classification error is a weighted sum of the type I error and the type II error, where the weights are the marginal probabilities of the class 0 and the class 1, respectively. Most classification methods aim to minimize the overall classification error [27, 28]. However, under two common application scenarios, this objective is no longer desirable. The first scenario is the "asymmetric importance scenario," where the consequence of making one type

of error (e.g., the type I error) far exceeds that of making the other type of error (e.g., the type II error). For instance, in the automated diagnosis of a severe disease, if we code the diseased and healthy status as the classes 0 and 1 respectively, then the type I error is the conditional probability of misclassifying a diseased patient as healthy, and the type II error is the conditional probability of misclassifying a healthy patient as diseased. Having a large type I error will clearly lead to a much more severe consequence than having a large type II error, because misclassifying a diseased patient as healthy will result in delayed treatment or even life loss [29–32] .

The second scenario, the "imbalanced size scenario," is where the two classes have great disparity in their proportions [33]. For example, a rare disease occurs only in 0.1% of the human population [34]. If we code the rare diseased class as class 0, then classifiers trained to minimize the overall classification error will likely lead to an undesirably large type I error [35]. In these two scenarios, the overall classification error fails to serve the purpose both as an optimization criterion and as an evaluation metric.

In this paper, we refer to the objective of minimizing the overall classification error as the *classical classification paradigm*, and refer to the model selection criterion that compares the overall classification error on a hold-out set as the *classical criterion*. To clarify, the term "model" in "model selection" refers to a feature subset instead of a probabilistic model. An alternative paradigm, the *Neyman-Pearson (NP) classification paradigm*, has been developed in the literature [36–41] to address the two above-mentioned scenarios. The NP classification paradigm specifically targets a prioritized control on the type I error: the type I error is controlled with high probability under a user-specified level α, usually a small value (e.g., .05), and the type II error is minimized under this constraint. Motivated by the NP classification paradigm, we propose a model selection criterion, the *Neyman-Pearson Criterion (NPC)*, which evaluates feature subsets by implementing a prioritized control on the type I error. In the automated diagnosis of a severe disease, the NPC is advantageous over the classical criterion, because the latter might select genes that result in a low overall classification error but an undesirably large type I error.

Does the NPC select a model different from the one selected by the classical criterion?

Intuitively, the "best" q features under the NP paradigm are not necessarily the same as the "best" q features under the classical paradigm. Under the classical paradigm, the "best" clearly means to have the lowest overall classification error; while under the NP paradigm, the "best" means to have the lowest type II error (subject to one type I error upper bound). Motivated by the fact that the classical criterion ranks models (i.e., feature subsets) based on the overall classification error on hold-out data, we design the NPC to rank models based on the type II error on hold-out data.

Similar to the classical criterion, the NPC belongs to the *validation set approach* to model selection for binary predictive problems. The validation set approach refers to a group of the techniques that hold out a labeled dataset unused for training classifiers, and that evaluate the trained classifiers on this set based on a certain criterion. The default evaluation criterion for binary classification is the classical criterion. We propose to use the NPC as a substitute when the prediction errors have asymmetric importance. A related more data-efficient variant of the validation set approach is the *cross validation*, where we randomly splits data into k folds, and for each $k-1$ folds, we train a classifier and report the average performance of the k classifiers on the left-out fold. We recommend implementing NPC in a way similar in spirit to cross validation. Besides the validation set and cross validation approaches, many other model selection approaches exist in the literature. The approach at the other end of the spectrum is to modify some fit measure evaluated on *training data*, and they include AIC [5], BIC [6], Mallows's C_p [7], LASSO [8], SCAD [9], MCP [10], Elastic net [11], and Group LASSO [12], among others. Another class of common model selection approaches concern model space search strategies, including exhaustive search, forward stepwise selection, backward stepwise selection, marginal screening [42–49], and interactive screening [50–54].

Previous work on NP classification has laid a good algorithmic and theoretic foundation for our new model selection criterion. In particular, [41] developed an umbrella algorithm that adapts popular binary classification methods (e.g., logistic regression, support vector machine, and random forest) to the NP paradigm, enabling application of the NP paradigm in a wide spectrum of real-world scenarios, and providing the algorithmic support for the

NPC. On the theoretic side, [39] and [40] developed conditions for plug-in NP classifiers to satisfy *NP oracle inequalities*, which was proposed in [38] as the theoretical criterion to evaluate the performance of NP classifiers. The formulated conditions and intermediate results in these works will lend support to establishing the model selection property of the NPC.

The development of the NPC as a practical criterion addresses the great needs of identifying a small number of genetic features to predict cancer with a high sensitivity in automated diagnosis. For malignant cancers with low survival rates, the priority is to achieve a high sensitivity, or equivalently a low false negative rate. The NP paradigm is naturally aligned with this high sensitivity requirement, and the feature selection criterion should be based on the specificity, which is exactly the goal of the NPC.

This article is organized as follows. In Section 2.2, we review the NP classification paradigm and use Gaussian examples to analytically illustrate that the classical criterion and the NPC can select different models on the population level. In Section 2.3, we introduce the NPC based on a finite sample. In Section 2.4, we explore the model selection properties of the NPC for plug-in NP classifiers. Section 2.5 contains simulation studies to verify the numerical performance of the NPC. Section 2.6 provides an in-depth real data study to show that the NPC identifies gene markers, among the overall predictive ones, to achieve a high specificity in cancer diagnosis. We conclude with a discussion in Section 2.7. All the proofs of lemmas, propositions, and theorems are relegated to the Appendix.

2.2 Background and motivation

We first introduce some mathematical notations to facilitate our discussion. Let $(\boldsymbol{X}, Y)$ be a pair of random observations where $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ is a vector of features and $Y \in \{0, 1\}$ indicates the class label of $\boldsymbol{X}$. A *classifier* $\phi : \mathcal{X} \to \{0, 1\}$ maps from the feature space to the label space. A *loss function* assigns a cost to each misclassified instance $\phi(\boldsymbol{X}) \neq Y$, and the *risk* is defined as the expectation of this loss function with respect to the joint distribution of $(\boldsymbol{X}, Y)$. We adopt in this work a commonly used loss function, the 0-1 loss function:

$\mathbb{1}\left(\phi(\boldsymbol{X}) \neq Y\right)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Let $\mathbb{P}$ and $\mathbb{E}$ denote the generic probability distribution and expectation, whose meaning depends on specific contexts. Then the risk is the overall classification error $R(\phi) = \mathbb{E}\left[\mathbb{1}\left(\phi(\boldsymbol{X}) \neq Y\right)\right] = \mathbb{P}\left(\phi(\boldsymbol{X}) \neq Y\right)$, which can be decomposed as:

$$\begin{aligned} R(\phi) &= \mathbb{E}\left[\mathbb{1}\left(\phi(\boldsymbol{X}) \neq Y\right)\right] = \mathbb{P}\left(\phi(\boldsymbol{X}) \neq Y\right) \\ &= \mathbb{P}(Y=0)\mathbb{P}\left(\phi(\boldsymbol{X}) \neq Y \mid Y=0\right) + \mathbb{P}(Y=1)\mathbb{P}\left(\phi(\boldsymbol{X}) \neq Y \mid Y=1\right) \\ &= \mathbb{P}(Y=0)R_0\left(\phi\right) + \mathbb{P}(Y=1)R_1\left(\phi\right) , \end{aligned}$$

where $R_j\left(\phi\right) := \mathbb{P}\left(\phi(\boldsymbol{X}) \neq Y \mid Y=j\right)$, $j = 0$ and 1. The notations $R_0(\cdot)$ and $R_1(\cdot)$ denote the (population) type I and type II errors respectively. While the classical classification paradigm aims to mimic the *classical oracle classifier* φ^* that minimizes the overall classification error,

$$\varphi^* = \underset{\varphi:\mathbb{R}^d \to \{0,1\}}{\arg\min}\ R\left(\varphi\right) ,$$

the Neyman-Pearson (NP) classification paradigm aims to mimic the *α-level NP oracle classifier*

$$\varphi_\alpha^* = \underset{\varphi: R_0(\varphi) \leqslant \alpha}{\arg\min}\ R_1(\varphi) , \tag{2.1}$$

where α is a user-specified type I error upper bound. It is well known that $\varphi^*(\cdot) = \mathbb{1}(\eta(\cdot) > 1/2)$, where $\eta(x) = \mathbb{E}(Y|X = x)$ is the regression function [55]. On the other hand, the famous Neyman-Pearson Lemma (Lemma 1) and a correspondence between classification and statistical hypothesis testing show that φ_α^* in (2.1) can be constructed by thresholding $p_1(\cdot)/p_0(\cdot)$, where p_1 and p_0 denote the class conditional probability density functions of the features $\boldsymbol{X}$.

Lemma 1 (Neyman-Pearson Lemma [56])**.** *Let P_0 and P_1 be probability distributions possessing densities p_0 and p_1 respectively. Let P be the probability distribution of a random feature vector $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^d$. The null and alternative hypotheses are $H_0 : P = P_0$ and*

$H_1 : P = P_1$. *Let* $s^*(\cdot) = p_1(\cdot)/p_0(\cdot)$. *For a given level* $\alpha \in (0,1)$, *let* $C^*_\alpha \in \mathbb{R}$ *be such that*

$$P_0\left(s^*(\boldsymbol{X}) > C^*_\alpha\right) \leqslant \alpha \quad \textit{and} \quad P_0\left(s^*(\boldsymbol{X}) \geqslant C^*_\alpha\right) \geqslant \alpha\,.$$

When $P_0\left(s^*(\boldsymbol{X}) = C^*_\alpha\right) = 0$, *the most powerful test of level* α *is*

$$\varphi^*_\alpha(\boldsymbol{X}) := \mathbb{1}\left(s^*(\boldsymbol{X}) > C^*_\alpha\right)\,. \tag{2.2}$$

Hypothesis testing bears a strong similarity with binary classification if we consider P_0 and P_1 as the conditional feature distributions of the classes 0 and 1 respectively. Rejecting H_0 based on the observed $s^*(\boldsymbol{X})$ is equivalent to classifying $\boldsymbol{X}$ as the class 1, and not rejecting H_0 is equivalent to classifying $\boldsymbol{X}$ as the class 0. The Neyman-Pearson Lemma (Lemma 1) states that the test φ^*_α maximizes the power at a significance level α. When considered equivalently as a classifier, φ^*_α in (2.2) is also the α-level NP oracle classifier.

2.2.1 Neyman-Pearson criterion (NPC) on the population level

Before introducing the sample-based version of NPC in the next section, we first define the classical criterion and the NPC on the population level. We show that these two criteria lead to different choices of feature subsets (i.e., models) under certain scenarios, and that NPC may choose different feature subsets at different α values. Denote respectively by φ^*_A and $\varphi^*_{\alpha A}$ the classical oracle classifier and the α-level NP oracle classifier that only use features indexed by $A \subseteq \{1, \ldots, d\}$. In other words, φ^*_A achieves

$$R\left(\varphi^*_A\right) = \min_{\varphi_A} \mathbb{P}(\varphi_A(\boldsymbol{X}) \neq Y)\,,$$

in which $\varphi_A : \mathcal{X} \subseteq \mathbb{R}^d \to \{0,1\}$ is any map that first projects $\boldsymbol{X} \in \mathbb{R}^d$ to its $|A|$-dimensional sub-vector $\boldsymbol{X}_A$, comprising of the coordinates of $\boldsymbol{X}$ from the index set A, and then maps from $\boldsymbol{X}_A \in \mathbb{R}^{|A|}$ to $\{0,1\}$.

In contrast, $\varphi^*_{\alpha A}$ achieves

$$R_1\left(\varphi^*_{\alpha A}\right) = \min_{\substack{\varphi_A \\ \mathbb{P}(\varphi_A(\boldsymbol{X})\neq Y|Y=0)\leqslant\alpha}} \mathbb{P}(\varphi_A(\boldsymbol{X}) \neq Y|Y=1)\,. \tag{2.3}$$

By the Neyman-Pearson lemma, $\varphi^*_{\alpha A}(\boldsymbol{x}) = \mathbb{1}\left(p_{1A}(\boldsymbol{x}_A)/p_{0A}(\boldsymbol{x}_A) > C^*_{\alpha A}\right)$ for some $C^*_{\alpha A}$. Among candidate feature subsets indexed by $A_1, \cdots, A_K$, *the population-level classical criterion* selects an A_i that achieves the smallest among $\left\{R\left(\varphi^*_{A_1}\right), \ldots, R\left(\varphi^*_{A_K}\right)\right\}$. In contrast, for a given level α, *the population-level NPC* selects an A_i that achieves the smallest among $\left\{R_1\left(\varphi^*_{\alpha A_1}\right), \ldots, R_1\left(\varphi^*_{\alpha A_K}\right)\right\}$. As a concrete illustration, suppose that we want to compare two features $\boldsymbol{X}_{\{1\}}, \boldsymbol{X}_{\{2\}} \in \mathbb{R}$ (Usually, these features are denoted by X_1 and X_2, but we opt to use $\boldsymbol{X}_{\{1\}}$ and $\boldsymbol{X}_{\{2\}}$ to be consistent with the notation $\boldsymbol{X}_A$), whose class conditional distributions are Gaussian as follows:

$$\begin{aligned} \boldsymbol{X}_{\{1\}} \,|\, (Y=0) &\sim \mathcal{N}(-5, 2^2)\,, & \boldsymbol{X}_{\{1\}} \,|\, (Y=1) &\sim \mathcal{N}(0, 2^2)\,, \\ \boldsymbol{X}_{\{2\}} \,|\, (Y=0) &\sim \mathcal{N}(-5, 2^2)\,, & \boldsymbol{X}_{\{2\}} \,|\, (Y=1) &\sim \mathcal{N}(1.5, 3.5^2)\,, \end{aligned} \tag{2.4}$$

and the class priors are equal, i.e., $\mathbb{P}(Y=1) = .5$. It can be calculated that $R\left(\varphi^*_{\{1\}}\right) = .106$ and $R\left(\varphi^*_{\{2\}}\right) = .113$. Therefore, $R\left(\varphi^*_{\{1\}}\right) < R\left(\varphi^*_{\{2\}}\right)$ and feature 1 is better than feature 2 under the classical criterion. Under NPC, the comparison is more subtle. If we set $\alpha = .01$, $R_1\left(\varphi^*_{\alpha\{1\}}\right) = .431$ is *larger* than $R_1\left(\varphi^*_{\alpha\{2\}}\right) = .299$. However, if we set $\alpha = .20$, $R_1\left(\varphi^*_{\alpha\{1\}}\right) = .049$ is *smaller* than $R_1\left(\varphi^*_{\alpha\{2\}}\right) = .084$. Figure 2.1 illustrates the NP oracle classifiers in this toy example.

The example above gives clues to a general phenomenon that the ranking of feature subsets under NPC may differ for distinct α values. For some values (e.g., $\alpha = .20$ in the example), the classical criterion and NPC agree on the ranking, while for others (e.g., $\alpha = .01$ in the example), they disagree. Under special cases however, we can derive conditions under which NPC gives an α-invariant feature subset ranking. In the following, we derive such a condition for Gaussian distributions.

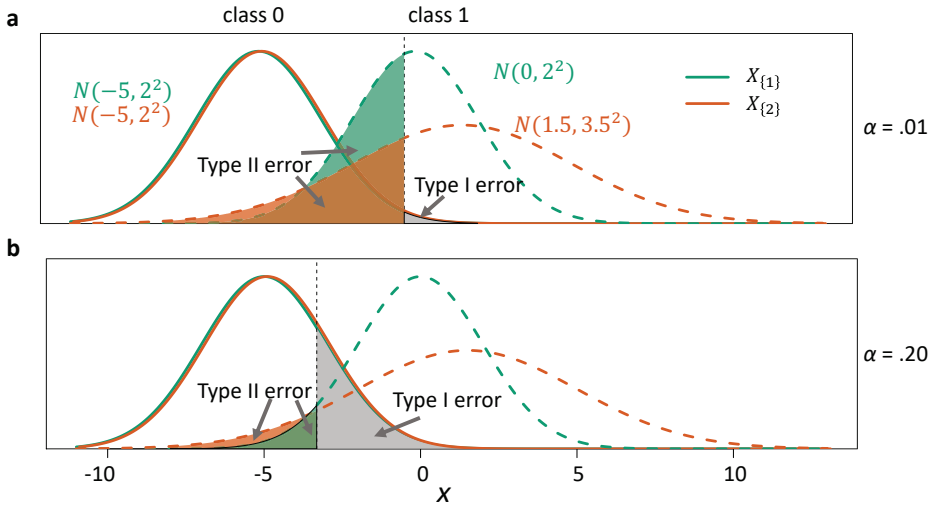

Figure 2.1: A toy example in which feature ranking under NPC changes as α varies. a: $\alpha = .01$. The NP oracle classifier based on feature 1 (or feature 2) has the type II error .431 (or .299). b : $\alpha = .20$. The NP oracle classifier based on feature 1 (or feature 2) has the type II error .049 (or .084).

Lemma 2. *Suppose that two features* $\boldsymbol{X}_{\{1\}}$ *and* $\boldsymbol{X}_{\{2\}}$ *have class-conditional densities*

$$\boldsymbol{X}_{\{1\}}|(Y=0) \sim \mathcal{N}\left(\mu_1^0, (\sigma_1)^2\right), \qquad \boldsymbol{X}_{\{1\}}|(Y=1) \sim \mathcal{N}\left(\mu_1^1, (\sigma_1)^2\right),$$
$$\boldsymbol{X}_{\{2\}}|(Y=0) \sim \mathcal{N}\left(\mu_2^0, (\sigma_2)^2\right), \qquad \boldsymbol{X}_{\{2\}}|(Y=1) \sim \mathcal{N}\left(\mu_2^1, (\sigma_2)^2\right).$$

That is, each feature has the same class-conditional variance under the two classes. For $\alpha \in (0,1)$, *let* $\varphi^*_{\alpha\{1\}}$ *or* $\varphi^*_{\alpha\{2\}}$ *be the level-*α *NP oracle classifier using only the feature* $\boldsymbol{X}_{\{1\}}$ *or* $\boldsymbol{X}_{\{2\}}$ *respectively, and let* $\varphi^*_{\{1\}}$ *or* $\varphi^*_{\{2\}}$ *be the corresponding classical oracle classifier. Then we have simultaneously for all* α,

$$sign\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right) - R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\} = sign\left\{R\left(\varphi^*_{\{2\}}\right) - R\left(\varphi^*_{\{1\}}\right)\right\} = sign\left\{\frac{|\mu_1^1 - \mu_1^0|}{\sigma_1} - \frac{|\mu_2^1 - \mu_2^0|}{\sigma_2}\right\},$$

where $sign(\cdot)$ *is the sign function.*

Lemma 3 shows a sufficient condition for a multi-dimensional Gaussian setting such that the ranking between two feature subsets is invariant to the level α under NPC and agrees with that under the classical criterion.

Lemma 3. *Let* A_1, $A_2 \subseteq \{1, \ldots, d\}$ *be two index sets. For a random vector* $\boldsymbol{X} \in \mathbb{R}^d$, *let* $\boldsymbol{X}_{A_1}$ *and* $\boldsymbol{X}_{A_2}$ *be sub-vectors of* $\boldsymbol{X}$ *comprising of coordinates with indexes in* A_1 *and* A_2

respectively, and assume they follow the class conditional distributions:

$$\boldsymbol{X}_{A_1} \,|\, (Y=0) \sim \mathcal{N}(\boldsymbol{\mu}_1^0, \boldsymbol{\Sigma}_1)\,, \qquad \boldsymbol{X}_{A_1} \,|\, (Y=1) \sim \mathcal{N}(\boldsymbol{\mu}_1^1, \boldsymbol{\Sigma}_1)\,,$$
$$\boldsymbol{X}_{A_2} \,|\, (Y=0) \sim \mathcal{N}(\boldsymbol{\mu}_2^0, \boldsymbol{\Sigma}_2)\,, \qquad \boldsymbol{X}_{A_2} \,|\, (Y=1) \sim \mathcal{N}(\boldsymbol{\mu}_2^1, \boldsymbol{\Sigma}_2)\,,$$

where $\boldsymbol{\mu}_j^i \in \mathbb{R}^{|A_j|}$, $i = 0, 1$, $j = 1, 2$ *denotes the mean vector and* $\boldsymbol{\Sigma}_j \in \mathbb{R}^{|A_j| \times |A_j|}$ *denotes the covariance matrix. For* $\alpha \in (0, 1)$, *let* $\varphi^*_{\alpha A_1}$ *and* $\varphi^*_{\alpha A_2}$ *be the* α*-level NP oracle classifiers using features indexed by* A_1 *and* A_2 *respectively, and let* $\varphi^*_{A_1}$ *and* $\varphi^*_{A_2}$ *be the corresponding classical oracle classifiers. Then we have for all* α,

$$sign\left(R_1\left(\varphi^*_{\alpha A_2}\right) - R_1\left(\varphi^*_{\alpha A_1}\right)\right) = sign\left(R\left(\varphi^*_{A_2}\right) - R\left(\varphi^*_{A_1}\right)\right)\,,$$

where $sign(\cdot)$ *is defined in Lemma 2.*

The conclusion in Lemma 3 is an exception rather than the rule. In general, the best feature subsets under the classical criterion and NPC do not necessarily agree on the population level. This suggests that the classical criterion on the sample level, i.e., the empirical risk on a hold-out set, is not suitable for model selection with asymmetric error control objectives. This issue motivates us to develop a new practical model selection criterion under the NP paradigm: NPC on the sample level.

2.3 Methodology

To enable the implementation of the model selection criterion NPC on the sample level, it is necessary to have flexible construction of NP classifiers.

2.3.1 Algorithmic foundation: construction of NP classifiers

Motivated by Lemma 1, [39] used a plug-in approach to construct NP classifiers, which satisfy the NP oracle inequalities [38] under low-dimensional settings (i.e., when d is small). Under the feature independence assumption, [40] extended the NP plug-in classifiers to

accommodate high-dimensional features. From a practical perspective, [41] developed an umbrella algorithm that adapts the *scoring-type classification methods* (e.g., logistic regression, support vector machine and random forest) to the NP paradigm so that they achieve a high probability control on the type I error under the pre-specified level α. A scoring-type classification method needs two components: a scoring function $s(\cdot)$ and a threshold C, to construct a classifier of the form $\phi_C(\cdot) = \mathbb{1}(s(\cdot) > C)$. A **good scoring-type classification method**, i.e., a method better than random guesses, should satisfy that

$$1 - \mathbb{P}(s(\boldsymbol{X}) \leqslant C | Y = 1) > \mathbb{P}(s(\boldsymbol{X}) > C | Y = 0), \forall C \in \mathbb{R}.$$

In other words, as C varies, the receiver operating characteristic (ROC) curve of this classification method is above the main diagonal line in the ROC space, which indicates random guesses. In other words, a good scoring-type classification method should satisfy

$$1 - R_1(\phi_C) > R_0(\phi_C), \forall C \in \mathbb{R}. \tag{2.5}$$

Most commonly used classification methods satisfy this property.

To construct an NP classifier using the NP umbrella algorithm [41], we first use a mixture of class 0 and class 1 observations to train a scoring function $\hat{s}$, and then set a threshold $\widehat{C} \in \mathbb{R}$ based on the left-out class 0 observations to obtain a classifier $\mathbb{1}\left(\hat{s}(\cdot) > \widehat{C}\right)$. Concretely, suppose we have a training dataset $\mathcal{S} = \mathcal{S}^0 \cup \mathcal{S}^1$, where $\mathcal{S}^0 = \left\{\boldsymbol{X}_1^0, \ldots, \boldsymbol{X}_m^0\right\}$ are i.i.d. class 0 observations, $\mathcal{S}^1 = \left\{\boldsymbol{X}_1^1, \ldots, \boldsymbol{X}_n^1\right\}$ are i.i.d. class 1 observations, and $\mathcal{S}^0$ is independent of $\mathcal{S}^1$. These sample sizes m and n are considered as fixed numbers in our methodology development. We randomly divide class 0 observations $\mathcal{S}^0$ for B times into two halves $\mathcal{S}_{\text{ts}}^{0(b)} = \left\{\boldsymbol{X}_1^{0(b)}, \ldots, \boldsymbol{X}_{m_1}^{0(b)}\right\}$ and $\mathcal{S}_{\text{lo}}^{0(b)} = \left\{\boldsymbol{X}_{m_1+1}^{0(b)}, \ldots, \boldsymbol{X}_{m_1+m_2}^{0(b)}\right\}$, where $m_1 + m_2 = m$, the subscripts "ts" and "lo" stand for *train-scoring* and *left-out* respectively, and the superscript $b \in \{1, \ldots, B\}$ indicates the b-th random split on class 0 observations. The default option in the NP umbrella algorithm takes an equal-sized split of the class 0 sample, that is, $m_1 = \lfloor m/2 \rfloor$. To do model selection under the NP paradigm, in addition to splitting the class 0

observations, we also randomly split class 1 observations $\mathcal{S}^1$ into $\mathcal{S}^{1(b)}_{\text{ts}} = \left\{\boldsymbol{X}^{1(b)}_1, \ldots, \boldsymbol{X}^{1(b)}_{n_1}\right\}$ and $\mathcal{S}^{1(b)}_{\text{lo}} = \left\{\boldsymbol{X}^{1(b)}_{n_1+1}, \ldots, \boldsymbol{X}^{1(b)}_{n_1+n_2}\right\}$, where $n_1 + n_2 = n$. Note that we do not need to split the class 1 observations in the NP umbrella classification algorithm. However, for model selection purposes, we must make the split to get a hold-out set to evaluate the type II error of the trained classifier. We will make a default option $n_1 = \lfloor n/2 \rfloor$. While $\mathcal{S}^{1(b)}_{\text{ts}}$ is used to train the scoring function, we leave out $\mathcal{S}^{1(b)}_{\text{lo}}$ to evaluate the type II error performance of the trained NP classifier, which will serve as the basis of our new model selection criterion NPC.

To construct an NP classifier given a scoring-type classification method, the NP umbrella algorithm first trains a scoring function $\hat{s}^{(b)}(\cdot)$ on $\mathcal{S}^{0(b)}_{\text{ts}} \cup \mathcal{S}^{1(b)}_{\text{ts}}$. Second, the algorithm applies $\hat{s}^{(b)}(\cdot)$ to $\mathcal{S}^{0(b)}_{\text{lo}}$ to obtain scores $\left\{T^{(b)}_i = \hat{s}^{(b)}\left(\boldsymbol{X}^{0(b)}_{m_1+i}\right), i = 1, \ldots, m_2\right\}$, which are sorted in an increasing order and denoted by $\left\{T^{(b)}_{(i)}, i = 1, \ldots, m_2\right\}$. Third, for a user-specified type I error upper bound $\alpha \in (0, 1)$ and a violation rate $\delta_1 \in (0, 1)$ which refers to the probability of the type I error of the trained classifier exceeding α, the algorithm chooses the order

$$k^* = \min_{k=1,\ldots,m_2} \left\{ k : \sum_{j=k}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j} \leqslant \delta_1 \right\} .$$

When $m_2 \geqslant \frac{\log \delta_1}{\log(1-\alpha)}$, k^* exists, and the umbrella algorithm chooses the threshold of the estimated scoring function as

$$\widehat{C}^{(b)}_\alpha = T^{(b)}_{(k^*)} ,$$

where a subscript "α" is added on $\widehat{C}^{(b)}$ to indicate the user-specified type I error upper bound. The resulting NP classifier is thus

$$\hat{\phi}^{(b)}_\alpha(\cdot) = \mathbb{1}\left(\hat{s}^{(b)}(\cdot) > \widehat{C}^{(b)}_\alpha\right) . \tag{2.6}$$

Proposition 1 in [41] proves that the probability that the type I error of the classifier

$\hat{\phi}_\alpha^{(b)}(\cdot)$ in (2.6) exceeds α is no more than δ_1:

$$\mathbb{P}\left(R_0(\hat{\phi}_\alpha^{(b)}) > \alpha\right) \leqslant \sum_{j=k^*}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j} \leqslant \delta_1\,, \tag{2.7}$$

for every $b = 1, \ldots, B$. When $T_i^{(b)}$ has a continuous distribution, the first inequality in (2.7) becomes an equality. Finally, the B NP classifiers, $\hat{\phi}_\alpha^{(1)}, \ldots, \hat{\phi}_\alpha^{(B)}$, will be combined into an ensemble classifier by majority voting. The number of splits, B, is often chosen to be greater than one to increase the stability and reduce the type II error. For details of the NP umbrella algorithm, we refer interested readers to [41].

2.3.2 NPC on the sample level

The construction of NP classifiers depends on users' choice of classification method, which could be the plug-in approach [39, 40, 57] or a more general scoring-type classification method adaptable to the NP umbrella algorithm [41]. In the following, we consider the problem of comparing models (i.e., feature subsets) for a given scoring-type classification method.

For a feature index set $A \subseteq \{1, \ldots, d\}$, we follow the NP umbrella algorithm described in section 2.3.1 and construct B NP classifiers, where the b-th NP classifier is based on training data ($\mathcal{S}_{\text{ts}}^{0(b)}, \mathcal{S}_{\text{lo}}^{0(b)}$ and $\mathcal{S}_{\text{ts}}^{1(b)}$) and a given classification method. We denote these B NP classifiers as $\hat{\phi}_{\alpha A}^{(1)}, \ldots, \hat{\phi}_{\alpha A}^{(B)}$ and evaluate their type II error performance on corresponding left-out class 1 sets $\mathcal{S}_{\text{lo}}^{1(1)}, \ldots, \mathcal{S}_{\text{lo}}^{1(B)}$ respectively. Our *sample-level NPC* for model A at level α, denoted by $\text{NPC}_{\alpha A}$, computes the average of these type II errors:

$$\begin{aligned} \text{NPC}_{\alpha A} &:= \frac{1}{B}\sum_{b=1}^{B} \text{NPC}_{\alpha A}^{(b)}\,, \\ \text{with } \text{NPC}_{\alpha A}^{(b)} &:= \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\left[1 - \hat{\phi}_{\alpha A}^{(b)}\left(\boldsymbol{X}_i^{1(b)}\right)\right] = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} \mathbb{1}\left(\hat{s}_A^{(b)}\left(\boldsymbol{X}_{iA}^{1(b)}\right) \leqslant \widehat{C}_{\alpha A}^{(b)}\right)\,, \end{aligned} \tag{2.8}$$

where $\hat{s}_A^{(b)}(\cdot)$ is the scoring function trained on $\mathcal{S}_{\text{ts}}^{0(b)} \cup \mathcal{S}_{\text{ts}}^{1(b)}$ using only the features indexed by A, and $\widehat{C}_{\alpha A}^{(b)}$ is the threshold estimated using the procedure described in section 2.3.1. The detailed implementation of NPC is described in Algorithm 1.

Correspondingly, we define the *sample-level classical criterion* for model A as

$$\mathrm{CC}_A := \frac{1}{B}\sum_{b=1}^{B}\mathrm{CC}_A^{(b)}, \tag{2.9}$$
$$\text{with } \mathrm{CC}_A^{(b)} := \frac{1}{m_2+n_2}\left\{\sum_{i=n_1+1}^{n_1+n_2}\left[1-\hat{\phi}_A^{(b)}\left(\boldsymbol{X}_i^{1(b)}\right)\right]+\sum_{i'=m_1+1}^{m_1+m_2}\hat{\phi}_A^{(b)}\left(\boldsymbol{X}_{i'}^{0(b)}\right)\right\},$$

where $\hat{\phi}_A^{(b)}(\cdot)$ is a classifier trained on $\mathcal{S}_{\mathrm{ts}}^{0(b)} \cup \mathcal{S}_{\mathrm{ts}}^{1(b)}$.

We also define the standard errors of $\mathrm{NPC}_{\alpha A}$ and CC_A as

$$\mathrm{se}(\mathrm{NPC}_{\alpha A}) := \sqrt{\frac{\sum_{b=1}^{B}\left(\mathrm{NPC}_{\alpha A}^{(b)}-\mathrm{NPC}_{\alpha A}\right)^2}{B(B-1)}}, \tag{2.10}$$

$$\mathrm{se}(\mathrm{CC}_{\alpha A}) := \sqrt{\frac{\sum_{b=1}^{B}\left(\mathrm{CC}_{\alpha A}^{(b)}-\mathrm{CC}_{\alpha A}\right)^2}{B(B-1)}}. \tag{2.11}$$

2.3.3 Method-specific NP oracle

Because $\mathrm{NPC}_{\alpha A}$ depends on the choice of classification methods, it does not necessarily converge to $R_1(\varphi_{\alpha A}^*)$ asymptotically, unless we use the plug-in approach and make certain assumptions on the class-conditional densities. Here we define the *method-specific NP oracle classifier* to address this concern.

Given a scoring-type classification method and a type I error upper bound $\alpha \in (0,1)$, let $\mathcal{M}$ denote the set of possible scoring functions for this method. We denote the "best" scoring function in $\mathcal{M}$ by $s : \mathcal{X} \to \mathbb{R}$, in the sense that $\mathbb{1}(s(\cdot) > 1/2)$ minimizes the overall (population) classification error among all $\mathbb{1}(h(\cdot) > 1/2)$ for all $h \in \mathcal{M}$. We refer to $s(\cdot)$ as the *method-specific optimal scoring function*. We define a *method-specific oracle classifier* with a threshold $C \in \mathbb{R}$ as

$$\phi_C(\boldsymbol{X}) := \mathbb{1}\left(s(\boldsymbol{X}) > C\right).$$

We denote by $\mathcal{C} := \{\phi_C : C \in \mathbb{R}\}$ the collection of method-specific oracle classifiers given

Algorithm 1 Implementation of the Neyman-Pearson Criterion (NPC)

input:

training set: $\mathcal{S} = \mathcal{S}^0 \cup \mathcal{S}^1$, where $\mathcal{S}^0 = \left\{\boldsymbol{X}_1^0, \ldots, \boldsymbol{X}_m^0\right\}$ are i.i.d. class 0 observations, and $\mathcal{S}^1 = \left\{\boldsymbol{X}_1^1, \ldots, \boldsymbol{X}_n^1\right\}$ are i.i.d. class 1 observations

feature index set $A \subseteq \{1, \ldots, d\}$

left-out class 0 sample size: m_2

left-out class 1 sample size: n_2

type I error upper bound $\alpha \in [0, 1]$

type I error violation rate $\delta_1 \in (0, 1)$

number of random splits $B \in \mathbb{N}$ on $\mathcal{S}^0$ and $\mathcal{S}^1$

function $\mathrm{NPC}(\mathcal{S}^0, \mathcal{S}^1, A, m_2, n_2, \alpha, \delta_1, B)$

  for k in $\{1, \ldots, m_2\}$ **do** ▷ *for each order* k

    $v(k) \leftarrow \sum_{j=k}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j}$ ▷ *calculate the violation rate*

    return $v(k)$

  $k^* \leftarrow \min\{k \in \{1, \ldots, m_2\} : v(k) \leqslant \delta_1\}$ ▷ *pick the order whose corresponding violation rate is under* δ_1

  for b in $1, \ldots, B$ **do**

    $\mathcal{S}_{\mathrm{lo}}^{0(b)} \leftarrow \texttt{subsample}(\mathcal{S}^0, m_2)$ ▷ $\mathcal{S}_{lo}^{0(b)} = \left\{\boldsymbol{X}_{m_1+1}^{0(b)}, \ldots, \boldsymbol{X}_{m_1+m_2}^{0(b)}\right\}$ *and* $m_1 = m - m_2$

    $\mathcal{S}_{\mathrm{lo}}^{1(b)} \leftarrow \texttt{subsample}(\mathcal{S}^1, n_2)$ ▷ $\mathcal{S}_{lo}^{1(b)} = \left\{\boldsymbol{X}_{n_1+1}^{1(b)}, \ldots, \boldsymbol{X}_{n_1+n_2}^{1(b)}\right\}$ *and* $n_1 = n - n_2$

    $\mathcal{S}_{\mathrm{ts}}^{0(b)} \leftarrow \mathcal{S}^0 \backslash \mathcal{S}_{\mathrm{lo}}^{0(b)}$ ▷ $\mathcal{S}_{ts}^{0(b)} = \left\{\boldsymbol{X}_1^{0(b)}, \ldots, \boldsymbol{X}_{m_1}^{0(b)}\right\}$

    $\mathcal{S}_{\mathrm{ts}}^{1(b)} \leftarrow \mathcal{S}^1 \backslash \mathcal{S}_{\mathrm{lo}}^{1(b)}$ ▷ $\mathcal{S}_{ts}^{1(b)} = \left\{\boldsymbol{X}_1^{1(b)}, \ldots, \boldsymbol{X}_{n_1}^{1(b)}\right\}$

    $\hat{s}_A^{(b)} \leftarrow \texttt{classification algorithm}(\mathcal{S}_{\mathrm{ts}}^{0(b)} \cup \mathcal{S}_{\mathrm{ts}}^{1(b)}, A)$ ▷ *train a scoring function* $\hat{s}_A^{(b)}$ *on* $\mathcal{S}_{ts}^{0(b)} \cup \mathcal{S}_{ts}^{1(b)}$ *using features with indexes in* A *only*

    $\mathcal{T}^{(b)} = \left\{t_1^{(b)}, \ldots, t_{m_2}^{(b)}\right\} \leftarrow \left\{\hat{s}_A^{(b)}\left(\boldsymbol{X}_{(m_1+1)A}^{0(b)}\right), \ldots, \hat{s}_A^{(b)}\left(\boldsymbol{X}_{(m_1+m_2)A}^{0(b)}\right)\right\}$ ▷ *apply* $\hat{s}_A^{(b)}$ *to* $\mathcal{S}_{lo}^{0(b)}$ *to obtain a set of threshold candidates*

    $\left\{t_{(1)}^{(b)}, \ldots, t_{(m_2)}^{(b)}\right\} \leftarrow \texttt{sort}(\mathcal{T}^{(b)})$ ▷ *sort elements in* $\mathcal{T}$ *in an increasing order*

    $\widehat{C}_{\alpha A}^{(b)} \leftarrow t_{(k^*)}^{(b)}$ ▷ *find the threshold corresponding to the chosen order* k^*

    $\hat{\phi}_{\alpha A}^{(b)}(\boldsymbol{X}) = \mathbb{1}\left(\hat{s}_A^{(b)}(\boldsymbol{X}_A) > \widehat{C}_{\alpha A}^{(b)}\right)$ ▷ *construct an NP classifier based on the scoring function* $\hat{s}_A^{(b)}$ *and the threshold* $\widehat{C}_{\alpha A}^{(b)}$

    $\left\{\hat{y}_1^{(b)}, \ldots, \hat{y}_{n_2}^{(b)}\right\} \leftarrow \left\{\hat{\phi}_{\alpha A}^{(b)}(\cdot), \cdot \in \mathcal{S}_{\mathrm{lo}}^{1(b)}\right\}$ ▷ *apply the trained classifier* $\hat{\phi}_{\alpha A}^{(b)}(\cdot)$ *to* $\mathcal{S}_{lo}^{1(b)}$

    $\mathrm{NPC}_{\alpha A}^{(b)} = \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbb{1}\left(\hat{y}_i^{(b)} \neq 1\right)$ ▷ *compute an empirical type II error by calculating the proportion of misclassified observations in* $\mathcal{S}_{lo}^{1(b)}$

    return $\mathrm{NPC}_{\alpha A}^{(b)}$

  return $\mathrm{NPC}_{\alpha A} = \frac{1}{B} \sum_{b=1}^{B} \mathrm{NPC}_{\alpha A}^{(b)}$

a classification method. Restricted to $\mathcal{C}$, we define the *method-specific α-level NP oracle classifier* as

$$\underset{\phi\in\mathcal{C}:R_0(\phi)\leqslant\alpha}{\arg\min}\ R_1(\phi)\,. \tag{2.12}$$

The notational dependency on the classification method is suppressed.

Given a scoring-type classification method, let $s(\cdot)$ be the method-specific optimal scoring function. We denote by F the cumulative distribution function of $s(\boldsymbol{X})|(Y=0)$. If we set the *method-specific NP threshold* $C_\alpha := F^{-1}(1-\alpha) = \inf\{x : F(x) \geqslant 1-\alpha\}$, then the classifier $\phi_{C_\alpha}(\cdot) = \mathbb{1}\left(s(\cdot) > C_\alpha\right)$ is the *method-specific α-level NP oracle classifier*, which was defined in (2.12).

Restricting to a feature subspace $A \subseteq \{1,\ldots,d\}$, the *method-specific population NPC for* A is $R_1\left(\phi_{\alpha A}\right)$, where $\phi_{\alpha A}(\boldsymbol{X}) := \phi_{C_{\alpha A}}(\boldsymbol{X}) := \mathbb{1}\left(s_A(\boldsymbol{X}_A) > C_{\alpha A}\right)$, in which $\boldsymbol{X}_A$ is the $|A|$-dimensional sub-vector of $\boldsymbol{X}$ comprising of coordinates index by A, s_A is the method-specific optimal scoring function for the feature subspace in $\mathbb{R}^{|A|}$, and $C_{\alpha A}$ is defined for the feature subspace in $\mathbb{R}^{|A|}$, similar to the C_α for the full feature space $\mathcal{X} \subseteq \mathbb{R}^d$.

2.4 Theoretical properties

This section investigates the model selection property for NPC. Concretely, we are interested in the answer to this question: among K candidate models $A_1,\ldots,A_K$, is it guaranteed with high probability that NPC selects the best model? We consider K as a fixed number in the following theory development. We also assume in this section that the number of random splits $B = 1$ in NPC, and for simplicity we suppress the super index (b) in all notations in this section and in the Appendix proofs. While NPC is adaptive to any scoring-type classification methods, we focus our theoretical investigation on the non-parametric plug-in approach. We discuss ideas regarding how to investigate other classification methods in the discussion section.

2.4.1 Definitions and key assumptions

We assume that the feature dimensionality d is fixed and moderate as in [39]. Following [58], for any multi-index $\boldsymbol{t} = (t_1, \cdots, t_d)^\mathsf{T} \in \mathbb{N}^d$ and $\boldsymbol{x} = (x_1, \cdots, x_d)^\mathsf{T} \in \mathbb{R}^d$, we define $|\boldsymbol{t}| = \sum_{i=1}^d t_i$, $\boldsymbol{t}! = t_1! \cdots t_d!$, $\boldsymbol{x}^{\boldsymbol{t}} = x_1^{t_1} \cdots x_d^{t_d}$, $\|\boldsymbol{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$, and the differential operator $D^{\boldsymbol{t}} = \frac{\partial^{t_1 + \cdots + t_d}}{\partial x_1^{t_1} \cdots \partial x_d^{t_d}}$. For all the theoretical discussions, we assume the domain of class conditional densities p_0 and p_1 is $[-1, 1]^d$. For $A \subseteq \{1, \ldots, d\}$, denote by P_{0A} and of P_{1A} the probability distributions of $\boldsymbol{X}_A|(Y = 0)$ and $\boldsymbol{X}_A|(Y = 1)$, with densities p_{0A} and of p_{1A} respectively. Throughout this paper, we only consider nonempty subset of $\{1, \ldots, d\}$.

Definition 1 (Hölder function class)**.** *Let $\beta > 0$. Denote by $\lfloor \beta \rfloor$ the largest integer strictly less than β. For a $\lfloor \beta \rfloor$-times continuously differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, we denote by $g_{\boldsymbol{x}}$ its Taylor polynomial of degree $\lfloor \beta \rfloor$ at a value $\boldsymbol{x} \in \mathbb{R}^d$:*

$$g_{\boldsymbol{x}}^{(\beta)}(\cdot) = \sum_{|\boldsymbol{t}| \leqslant \lfloor \beta \rfloor} \frac{(\cdot - \boldsymbol{x})^{\boldsymbol{t}}}{\boldsymbol{t}!} D^{\boldsymbol{t}} g(\boldsymbol{x}) .$$

For $L > 0$, the $\left(\beta, L, [-1, 1]^d\right)$-Hölder function class, denoted by $\Sigma\left(\beta, L, [-1, 1]^d\right)$, is the set of $\lfloor \beta \rfloor$-times continuously differentiable functions $g : \mathbb{R}^d \to \mathbb{R}$ that satisfy the following inequality:

$$\left| g(\boldsymbol{x}) - g_{\boldsymbol{x}}^{(\beta)}(\boldsymbol{x}') \right| \leqslant L \|\boldsymbol{x} - \boldsymbol{x}'\|^\beta , \quad \text{for all } \boldsymbol{x}, \boldsymbol{x}' \in [-1, 1]^d .$$

Definition 2 (Hölder density class)**.** *The $\left(\beta, L, [-1, 1]^d\right)$-Hölder density class is defined as*

$$\mathcal{P}_\Sigma\left(\beta, L, [-1, 1]^d\right) = \left\{ p : p \geqslant 0, \int p = 1, p \in \Sigma\left(\beta, L, [-1, 1]^d\right) \right\} .$$

The following β-valid kernels are multi-dimensional analog of univariate higher order kernels.

Definition 3 (β-valid kernel)**.** *Let $K(\cdot)$ be a real-valued kernel function on $\mathbb{R}^d$ with the support $[-1, 1]^d$. For a fixed $\beta > 0$, the function $K(\cdot)$ is a β-valid kernel if it satisfies (1) $\int |K|^l < \infty$ for any $l \geqslant 1$, (2) $\int \|\boldsymbol{u}\|^\beta |K(\boldsymbol{u})| d\boldsymbol{u} < \infty$, and (3) in the case $\lfloor \beta \rfloor \geqslant 1$, $\int \boldsymbol{u}^{\boldsymbol{t}} K(\boldsymbol{u}) d\boldsymbol{u} = 0$ for any $\boldsymbol{t} = (t_1, \ldots, t_d) \in \mathbb{N}^d$ such that $1 \leqslant |\boldsymbol{t}| \leqslant \lfloor \beta \rfloor$.*

One example of β-valid kernels is the product kernel whose ingredients are kernels of order β in 1 dimension:

$$\widetilde{K}(\boldsymbol{x}) = K(x_1)K(x_2)\cdots K(x_d)\mathbb{1}(\boldsymbol{x} \in [-1,1]^d)\,,$$

where K is a 1-dimensional β-valid kernel and is constructed based on Legendre polynomials. Such kernels have been considered in [59]. When a β-valid kernel K is constructed out of Legendre polynomials, it is also Lipschitz and bounded. For simplicity, we assume that all the β-valid kernels considered in the theory discussion are constructed from Legendre polynomials.

Definition 4 (Margin assumption)**.** *A function $f(\cdot)$ satisfies the margin assumption of the order $\bar{\gamma}$ at the level C, with respect to the probability distribution P of a random vector $\boldsymbol{X}$, if there exist positive constants $\bar{C}$ and $\bar{\gamma}$, such that for all $\delta \geqslant 0$,*

$$P\left(|f\left(\boldsymbol{X}\right) - C| \leqslant \delta\right) \leqslant \bar{C}\delta^{\bar{\gamma}}\,.$$

The above condition for densities was first introduced in Polonik [60], and its counterpart in the classical binary classification was called margin condition ([61]), which is a low noise condition. Recall that the set $\{\boldsymbol{x} : \eta(\boldsymbol{x}) = 1/2\}$ is the decision boundary of the classical oracle classifier, and the margin condition in the classical paradigm is a special case of Definition 4 by taking $f = \eta$ and $C = 1/2$. Unlike the classical paradigm where the optimal threshold $1/2$ on regression function β is known, the optimal threshold level in the NP paradigm is unknown and needs to be estimated, suggesting the necessity of having sufficient data around the decision boundary to detect it. This concern motivated [39] to formulate a detection condition that works as an opposite force to the margin assumption, and [40] improved upon it and proved its necessity in bounding the excess type II error of an NP classifier. To establish the model selection property of NPC, a bound on the excess type II error is an intermediate result, so we also need this assumption for our current work.

Definition 5 (Detection condition [40])**.** *A function $f(\cdot)$ satisfies the detection condition*

of the order $\underline{\gamma}$ *at the level* (C, δ^*) *with respect to the probability distribution* P *of a random vector* $\boldsymbol{X}$*, if there exists a positive constant* $\underline{C}$*, such that for all* $\delta \in (0, \delta^*)$*,*

$$P\left(C \leqslant f\left(\boldsymbol{X}\right) \leqslant C + \delta\right) \geqslant \underline{C}\delta^{\underline{\gamma}}\,.$$

2.4.2 A uniform deviation result of scoring functions in sub feature space

For $A \subseteq \{1, \ldots, d\}$ and $|A| = l$, estimate p_{0A} and p_{1A} respectively from $\mathcal{S}^0_{\text{ts}}$ and $\mathcal{S}^1_{\text{ts}}$ by kernel density estimators,

$$\hat{p}_{0A}(\boldsymbol{x}_A) = \frac{1}{m_1 h^l_{m_1}} \sum_{i=1}^{m_1} K_A\left(\frac{\boldsymbol{X}^0_{iA} - \boldsymbol{x}_A}{h_{m_1}}\right) \qquad \text{and} \qquad \hat{p}_{1A}(\boldsymbol{x}_A) = \frac{1}{n_1 h^l_{n_1}} \sum_{i=1}^{n_1} K_A\left(\frac{\boldsymbol{X}^1_{iA} - \boldsymbol{x}_A}{h_{n_1}}\right),$$

where h_{m_1} and h_{n_1} denote the bandwidths, and $K_A(\boldsymbol{u}_A) = \int K(\boldsymbol{u}_A, \boldsymbol{u}_{A^c}) d\boldsymbol{u}_{A^c}$. We are interested in deriving a high probability bound for $\|\hat{p}_{1A}(\boldsymbol{x}_A)/\hat{p}_{0A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)/p_{0A}(\boldsymbol{x}_A)\|_\infty$. Lemma 1 in [39] will be called upon to establish high probability bounds for $\|\hat{p}_{0A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)\|_\infty$ and $\|\hat{p}_{1A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)\|_\infty$. But to use that lemma, we need to translate the conditions on the full feature space to the subspaces.

Condition 1. *Suppose that the densities satisfy*

(i) There exists a positive constant $\mu_{\min}$ *such that* $p_{0A} \geqslant \mu_{\min}$ *for all* $A \subseteq \{1, \ldots, d\}$*.*

(ii) There is a positive constant L *such that* $p_{0A}, p_{1A} \in \mathcal{P}_\Sigma(\beta, L, [-1, 1]^{|A|})$ *for all* $A \subseteq \{1, \ldots, d\}$*.*

Lemma 4. *Let* $K(\cdot)$ *be a* β*-valid kernel on* $\mathbb{R}^d$ *with the support* $[-1, 1]^d$ *(Definition 3). Let* $\boldsymbol{u} = (v, \boldsymbol{w})$ *where* $v \in \mathbb{R}$ *and* $\boldsymbol{w} \in \mathbb{R}^{d-1}$*. Then* $K'(\boldsymbol{w}) := \int K(v, \boldsymbol{w}) dv$ *is a* β*-valid kernel function on* $\mathbb{R}^{d-1}$ *with the support* $[-1, 1]^{d-1}$*.*

Proposition 1. *Assume condition 1 and let the kernel* K *be* β*-valid and* L'*-Lipschitz. Let* $A \subseteq \{1, \ldots, d\}$*. Take the bandwidths* $h_{m_1} = \left(\frac{\log m_1}{m_1}\right)^{\frac{1}{2\beta+\ell}}$ *and* $h_{n_1} = \left(\frac{\log n_1}{n_1}\right)^{\frac{1}{2\beta+\ell}}$*, where*

$l = |A|$. *For any* $\delta_3 \in (0,1)$, *if sample size* $m_1 = |\mathcal{S}_{ts}^0|$ *and* $n_1 = |\mathcal{S}_{ts}^1|$ *satisfy*

$$\sqrt{\frac{\log(2m_1/\delta_3)}{m_1 h_{m_1}^\ell}} < 1 \wedge \frac{\mu_{\min}}{2C_0}, \sqrt{\frac{\log(2n_1/\delta_3)}{n_1 h_{n_1}^\ell}} < 1, \quad n_1 \wedge m_1 \geqslant 2/\delta_3,$$

where $C_0 = \max_{A\subseteq\{1,\ldots,d\}}\{\sqrt{48c_{1A}} + 32c_{2A} + 2Lc_{3A} + L'_A + L + \tilde{C}_A \sum_{1\leqslant|\boldsymbol{q}|\leqslant\lfloor\beta\rfloor} \frac{1}{\boldsymbol{q}!}\}$, *in which* $c_{1A} = \|p_{0A}\|_\infty \|K_A\|^2$, $c_{2A} = \|K_A\|_\infty + \|p_{0A}\|_\infty + \int |K_A||\boldsymbol{t}|^\beta d\boldsymbol{t}$, $c_{3A} = \int |K_A||\boldsymbol{t}|^\beta d\boldsymbol{t}$, $L'_A = 2^{d-l}L'$ *and* $\tilde{C}_A$ *is such that* $\tilde{C}_A \geqslant \sup_{1\leqslant|\boldsymbol{q}|\leqslant\lfloor\beta\rfloor} \sup_{\boldsymbol{x}_A\in[-1,1]^l} |p_{0A}^{(\boldsymbol{q})}(\boldsymbol{x}_A)|$. *Then there exists a positive constant* $\widetilde{C}$ *that does not depend on* A, *such that we have with probability at least* $1-\delta_3$,

$$\|\hat{p}_{1A}(\boldsymbol{x}_A)/\hat{p}_{0A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)/p_{0A}(\boldsymbol{x}_A)\|_\infty \leqslant \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\beta/(2\beta+\ell)} + \left(\frac{\log n_1}{n_1}\right)^{\beta/(2\beta+\ell)}\right].$$

2.4.3 Concentration of $\mathrm{NPC}_{\alpha A}$ around $R_1(\varphi_{\alpha A}^*)$

To establish the model selection property, an essential step is to develop a concentration result of $\mathrm{NPC}_{\alpha A}$ around $R_1(\varphi_{\alpha A}^*)$, where $\varphi_{\alpha A}^*$ was defined in (2.3). Since we have fixed the plug-in kernel density classifiers, $\hat{\phi}_{\alpha A}(\boldsymbol{x}) = \mathbb{1}(\hat{s}_A(\boldsymbol{x}_A) > \widehat{C}_{\alpha A}) = \mathbb{1}(\hat{p}_{0A}(\boldsymbol{x}_A)/\hat{p}_{1A}(\boldsymbol{x}_A) > \widehat{C}_{\alpha A})$ denotes the NP classifier, where $\widehat{C}_{\alpha A}$ is determined by the NP umbrella classification algorithm. We always assume that the cumulative distribution function of $\hat{s}_A(\boldsymbol{X}_A)$, where $\boldsymbol{X} \sim P_0$, is continuous.

Lemma 5. *Let* $\alpha, \delta_1, \delta_2 \in (0,1)$. *If* $m_2 = |\mathcal{S}_{lo}^0| \geqslant \frac{4}{\alpha\delta_1}$, *the classifier* $\hat{\phi}_{\alpha A}$ *satisfies with probability at least* $1-\delta_1-\delta_2$,

$$\left|R_0(\hat{\phi}_{\alpha A}) - R_0(\varphi_{\alpha A}^*)\right| \leqslant \xi, \tag{2.13}$$

where

$$\xi = \sqrt{\frac{\lceil d_{\alpha,\delta_1,m_2}(m_2+1)\rceil (m_2+1-\lceil d_{\alpha,\delta_1,m_2}(m_2+1)\rceil)}{(m_2+2)(m_2+1)^2\delta_2}} + d_{\alpha,\delta_1,m_2} + \frac{1}{m_2+1} - (1-\alpha),$$

$$d_{\alpha,\delta_1,m_2} = \frac{1 + 2\delta_1(m_2+2)(1-\alpha) + \sqrt{1+4\delta_1(m_2+2)(1-\alpha)\alpha}}{2\{\delta_1(m_2+2)+1\}},$$

and $\lceil z \rceil$ *denotes the smallest integer larger than or equal to* z*. Moreover, if* $m_2 \geqslant \max(\delta_1^{-2}, \delta_2^{-2})$*, we have* $\xi \leqslant (5/2)m_2^{-1/4}$.

Lemma 5 and a minor modification of proof for Proposition 2.4 in [40] (which provides an upper bound for the excess type II error) give rise to the following proposition. Essentially, the same upper bound works for both $\left|R_1(\hat{\phi}_{\alpha A}) - R_1(\varphi^*_{\alpha A})\right|$ and $R_1(\hat{\phi}_{\alpha A}) - R_1(\varphi^*_{\alpha A})$.

Proposition 2. *Let* $\alpha, \delta_1, \delta_2 \in (0,1)$*. Assume that the density ratio* $s_A(\cdot) = p_{1A}(\cdot)/p_{0A}(\cdot)$ *satisfies the margin assumption of order* $\bar{\gamma}$ *at level* $C^*_{\alpha A}$ *(with constant* $\bar{C}$*) and detection condition of order* $\underline{\gamma}$ *at level* $(C^*_{\alpha A}, \delta^*)$ *(with constant* $\underline{C}$*), both with respect to distribution* P_{0A}*. If* $m_2 \geqslant \max\{\frac{4}{\alpha\delta_1}, \delta_1^{-2}, \delta_2^{-2}, (\frac{2}{5}\underline{C}{\delta^*}^{\underline{\gamma}})^{-4}\}$*, the excess type II error of the classifier* $\hat{\phi}_{\alpha A}$ *satisfies with probability at least* $1 - \delta_1 - \delta_2$*,*

$$
\begin{aligned}
&\left|R_1(\hat{\phi}_{\alpha A}) - R_1(\varphi^*_{\alpha A})\right| \\
&\leqslant 2\bar{C}\left[\left\{\frac{|R_0(\hat{\phi}_{\alpha A}) - R_0(\varphi^*_{\alpha A})|}{\underline{C}}\right\}^{1/\underline{\gamma}} + 2\|\hat{s}_A - s_A\|_\infty\right]^{1+\bar{\gamma}} + C^*_{\alpha A}|R_0(\hat{\phi}_{\alpha A}) - R_0(\varphi^*_{\alpha A})| \\
&\leqslant 2\bar{C}\left[\left(\frac{2}{5}m_2^{1/4}\underline{C}\right)^{-1/\underline{\gamma}} + 2\|\hat{s}_A - s_A\|_\infty\right]^{1+\bar{\gamma}} + C^*_{\alpha A}\left(\frac{2}{5}m_2^{1/4}\right)^{-1}.
\end{aligned}
$$

Theorem 1. *Let* α*,* δ_1*,* δ_2*,* δ_3*,* $\delta_4 \in (0,1)$*, and* $l = |A|$*. In addition to the assumptions of Propositions 1 and 2, assume* $n_2 \geqslant \left(\log \frac{2}{\delta_4}\right)^2$*, then we have with probability at least* $1 - \delta_1 - \delta_2 - \delta_3 - \delta_4$*,*

$$
|NPC_{\alpha A} - R_1(\varphi^*_{\alpha A})| \leqslant \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + \left(\frac{\log n_1}{n_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + m_2^{-(\frac{1}{4}\wedge\frac{1+\bar{\gamma}}{\underline{\gamma}})} + n_2^{-\frac{1}{4}}\right],
$$

for some positive constant $\widetilde{C}$ *that does not depend on* A*.*

Under smoothness and regularity conditions and sample size requirements, Theorem 1 shows the concentration of $\mathrm{NPC}_{\alpha A}$ around $R_1(\varphi^*_{\alpha A})$ with probability at least $1 - \delta_1 - \delta_2 - \delta_3 - \delta_4$. The user-specified violation rate δ_1 represents the uncertainty that the type I error of an NP classifier $\hat{\phi}_{\alpha A}$ exceeds α, leading to the underestimation of $R_1(\varphi^*_{\alpha A})$; δ_2 accounts

for possibility of unnecessarily stringent control on the type I error, which results in the overestimation of $R_1(\varphi^*_{\alpha A})$; δ_3 accounts for the uncertainty in training scoring function $\hat{s}_A(\cdot)$ on a finite sample; and δ_4 represents the uncertainty of using leave-out class 1 observations $\mathcal{S}^1_{\text{lo}}$ to estimate $R_1(\hat{\phi}_{\alpha A})$. Note that while the δ_1 parameter serves both as the input of the NPC algorithm and as a restriction to the sample sizes, other parameters δ_2, δ_3 and δ_4 only have the latter role. Just like the constant C_0 in Proposition 1, the generic constant $\widetilde{C}$ in Theorem 1 can be provided explicitly, but it would be too cumbersome to do so.

2.4.4 NPC model selection property for nonparametric plug-in methods

Theorem 2. *Let α, δ_1, δ_2, δ_3, $\delta_4 \in (0,1)$, and $A_1, \ldots, A_K \subseteq \{1, \ldots, d\}$. Assume that A_1 is the best among $\{A_1, \ldots, A_K\}$ under the population-level NPC by some margin $g > 0$, that is,*

$$\min_{A \in \{A_2, \ldots, A_K\}} R_1\left(\varphi^*_{\alpha A}\right) - R_1\left(\varphi^*_{\alpha A_1}\right) > g\,.$$

In addition to the assumptions in Theorem 1, assume m_1, m_2, n_1, n_2 satisfy that

$$\widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}} + \left(\frac{\log n_1}{n_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+d}} + m_2^{-(\frac{1}{4} \wedge \frac{1+\bar{\gamma}}{\bar{\gamma}})} + n_2^{-\frac{1}{4}}\right] < \frac{g}{2}\,,$$

where $\widetilde{C}$ is the generic constant in Theorem 1. Then with probability at least $1 - K(\delta_1 + \delta_2 + \delta_3 + \delta_4)$, $NPC_{\alpha A_1} < \min_{j=2,\ldots,K} NPC_{\alpha A_j}$, that is, NPC selects the best model.

2.5 Simulation studies

We verify the practical performance of NPC on the sample level in two simulation studies. First, we demonstrate that NPC and the classical criterion select the best feature differently in the toy example in Figure 2.1. Second, we show that NPC selects the best feature subset (with a pre-specified size) that minimizes the population type II error with high probability, in an exhaustive best subset selection when the total number of features is small.

Table 2.1: The frequency that each of the two features is selected as the better feature by each criterion among 1000 samples in the toy example (Figure 2.1).

Criteria \ Features	Feature 1	Feature 2
NPC ($\alpha = .01$)	2.2%	97.8%
NPC ($\alpha = .20$)	98.7%	1.3%
Classical Criterion	74.9%	25.1%

2.5.1 The toy example on the sample level

To verify that NPC and the classical criterion select their corresponding best feature, found in the toy example (Figure 2.1) in Section 2.2.1, with high probability on the sample level, we design the following simulation study.

We simulate $1,000$ random samples of size $n = 2,000$ from the distribution defined in Equation (2.4), which contains two features. This sample size is chosen to guarantee the type I error control of NP classifiers at $\alpha = .01$. We apply the sample-level NPC (with $\delta = .05$) and classical criterion defined in Equations (2.8) and (2.9) to each sample to select the better feature. For each feature, we use the plug-in density ratio as the classification scoring function $\hat{s}(\cdot)$, where kernel density estimators based on a Gaussian kernel and bandwidths selected by the `R` function `bw.nrd0()` are used to plug in the class conditional densities.

The result summarized in Table 2.1 shows that NPC with $\alpha = .01$ selects feature 2 with high probability, while the classical criterion and NPC with $\alpha = .20$ select feature 1 with high probability. Recall our finding on the population level: feature 2 is the better feature when NPC at $\alpha = .01$ is used as the criterion, while feature 1 is the better feature based on the classical criterion and NPC at $\alpha = .20$ (Section 2.2.1). This result is a numerical support of Theorem 2.

2.5.2 Best subset selection on the sample level

We next demonstrate the performance of the sample-level NPC on selecting the best feature subset when d, the total number of features, is small. We design the following simulation

setting, where $d = 5$ and the subsets of interest is of size 2.

$$\boldsymbol{X} \mid (Y = 0) \sim \mathcal{N}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0)\,, \qquad \boldsymbol{X} \mid (Y = 1) \sim \mathcal{N}(\boldsymbol{\mu}^1, \boldsymbol{\Sigma}^1)\,, \tag{2.14}$$

where $\boldsymbol{\mu}^0 = (-5, -6, -5, -3, -3)^\mathsf{T}$, $\boldsymbol{\mu}^1 = (0, 1, 1.5, -2, -2)^\mathsf{T}$,

$$\boldsymbol{\Sigma}^0 = \begin{bmatrix} 4 & .8 & 0 & 0 & 0 \\ .8 & 4 & .8 & 0 & 0 \\ 0 & .8 & 4 & 0 & 0 \\ 0 & 0 & 0 & 7.355418 & 2.578002 \\ 0 & 0 & 0 & 2.578002 & 7.229438 \end{bmatrix}, \boldsymbol{\Sigma}^1 = \begin{bmatrix} 4 & .8 & 0 & 0 & 0 \\ .8 & 4 & .8 & 0 & 0 \\ 0 & 1.4 & 12.25 & 0 & 0 \\ 0 & 0 & 0 & 7.355418 & 2.578002 \\ 0 & 0 & 0 & 2.578002 & 7.229438 \end{bmatrix},$$

and the class prior is $\mathbb{P}(Y = 1) = .7$.

We simulate $1,000$ random samples of size $n = 2,500$ from this distribution. We apply the sample-level NPC (with $\delta = .05$) and classical criterion defined in Equations (2.8) and (2.9) to each sample to select the best feature subset with size 2. For each feature subset, we use the plug-in density ratio as the classification scoring function $\hat{s}(\cdot)$, where two-dimensional kernel density estimators constructed by the `R` function `kde()` in the `ks` package are used to plug in the class conditional densities.

The result summarized in Table 2.2 shows that NPC with $\alpha = .01$ selects $\{2, 3\}$ in 57.3% of the samples and $\{1, 2\}$ and $\{1, 3\}$ for 31.6% and 9.4% of the time, respectively. In contrast, the classical criterion selects $\{1, 2\}$ in two thirds of the samples and $\{1, 3\}$ in the other one third samples.

On the population level, we approximate the population-level NPC and classical criterion on each feature subset by applying their corresponding sample-level criteria to an independent large sample with size 2×10^6 from the same distribution in (2.14). The NPC is minimized at the feature subset $\{2, 3\}$ with a value .05176152, while the feature subsets $\{1, 2\}$ and $\{1, 3\}$ achieve the second and third smallest NPC values of .05378549 and .10764869, respectively. Given the small gap between the population-level NPC values of $\{2, 3\}$ and $\{1, 2\}$, it is reasonable that the sample-level NPC selects these two feature subsets with high

Table 2.2: The frequency that each two-feature subset is selected as the best feature subset by each criterion among 1000 samples.

Criteria \ Feature Subsets	$\{1,2\}$	$\{1,3\}$	$\{2,3\}$	$\{2,4\}$	$\{2,5\}$
NPC ($\alpha = .01$)	31.6%	9.4%	57.3%	0.7%	1.0%
Classical Criterion	67.0%	33.0%	0.0%	0.0%	0.0%

probability. On the other hand, the classical criterion is minimized at the feature subset $\{1,2\}$ with a value .02167862, while the feature subsets $\{2,3\}$ and $\{1,3\}$ achieve the second and third smallest classical criterion values of .02318542 and .04059764, respectively. This result confirms that when the population-level NPC and classical criterion prefer different best feature subsets, the sample-level NPC, given a reasonably large sample size, chooses the NPC-preferred feature subset with high probability.

2.6 Real data application: selection of DNA methylation features for breast cancer prediction

We use a real dataset containing genome-wide DNA methylation profiles of 285 breast tissues measured by the *Illumina HumanMethylation450* microarray technology. This dataset includes 46 normal tissues and 239 breast cancer tissues. Methylation levels are measured at $468,424$ CpG probes in every tissue [62]. We download the preprocessed and normalized dataset from the Gene Expression Omnibus (GEO) [63] with the accession number GSE60185. The preprocessing and normalization steps are described in detail in [62]. To facilitate the interpretation of our analysis results, we further process the data as follows. First, we discard a CpG probe if it is mapped to no gene or more than one genes. Second, if a gene contains multiple CpG probes, we calculate its methylation level as the average methylation level of these probes. This procedure leaves us with $19,363$ genes with distinct methylation levels in every tissue. We consider the tissues as data points and the genes as features, so we have a sample with the size $n = 285$ and the number of features $d = 19,363$.

If we would like to predict whether a patient has breast cancer based on the methylation levels of genes in her breast tissue, we face a binary classification problem under the

"asymmetric scenario," where misclassifying a patient with cancer to be healthy leads to more severe consequences than the other way around. Hence, we code the 239 breast cancer tissues as the class 0 and the 46 normal tissues as the class 1. Then in this cancer diagnosis problem, controlling the more severe false negative rate is equivalent to controlling the type I error under the NP paradigm.

Under a high-dimensional scenario where $d \gg n$, we apply the penalized logistic regression with an ℓ_1 penalty to generate candidate feature subsets along the solution path as the tuning parameter of the ℓ_1 penalty decreases. We obtain 41 candidate feature subsets. For each feature subset, we evaluate four criteria: Akaike information criterion (AIC), Bayesian information criterion (BIC), the sample-level classical criterion (Equation (2.9)), and the sample-level NPC with $\alpha = .05$ and $\delta = .05$ (Equation (2.8)). For the latter two criteria that require sample splitting, we randomly split the sample into equal-sized training and left-out data for $B = 100$ times.

Figure 2.2 displays the trends of the four criteria on the candidate feature subsets along the solution path. The minimum AIC is achieved at the 40^{th} feature subset containing 30 genes, while BIC suggests choosing the 27^{th} feature subset that contains 20 genes. The sample-level classical criterion has small values ($0-.03$) for all candidate feature subsets and thus does not lead to a clear choice of feature subset. The sample-level NPC exhibits the most interesting trend: it has small values at the 2^{nd}-4^{th} feature subsets but a sharp rise at the 5^{th} subset, suggesting that the difference between the 4^{th} and 5^{th} feature subsets greatly alters the type II errors of the corresponding NP classifiers. The difference is the addition of the gene *ZNF646* to the 5^{th} feature subset.

To investigate the effect of *ZNF646* on the sample-level NPC trend, we remove this gene from all subsequent feature subsets (if it is in those subsets) and re-evaluate the four criteria. The results are shown in Figure 2.3, where the trends of AIC, BIC and the sample-level classical criterion remain largely the same, while the rise of the sample-level NPC is delayed to the 9^{th} feature subset with *ZNF646* removed. Again, by inspecting the genes included in the 8^{th} and 9^{th} feature subsets, we find that their only difference is the addition of the gene *ERAP1*.

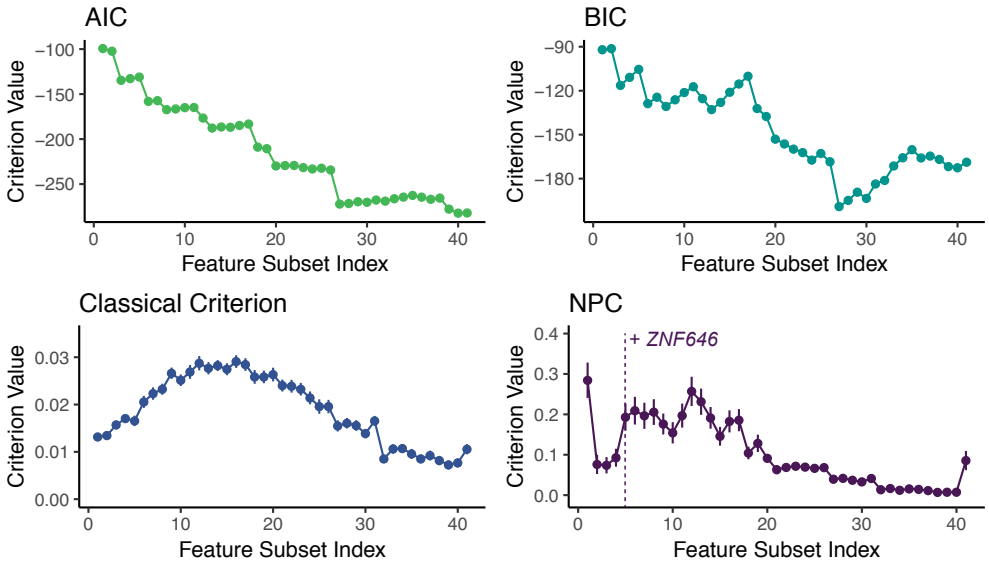

Figure 2.2: Four evaluation criteria on 41 candidate feature subsets identified by ℓ_1-penalized logistic regression from the breast cancer methylation data [62]. A larger feature subset index corresponds to a smaller value of the tuning parameter of the ℓ_1 penalty, which in most cases leads to a larger candidate feature subset. Compared with the 4^{th} subset, the 5^{th} subset contains an additional gene *ZNF646*. For the sample-level classical criterion and NPC (with $\alpha = .05$ and $\delta = .05$), each error bar shows the $\pm$ one standard error, defined in Equations (2.11) and (2.10), respectively.

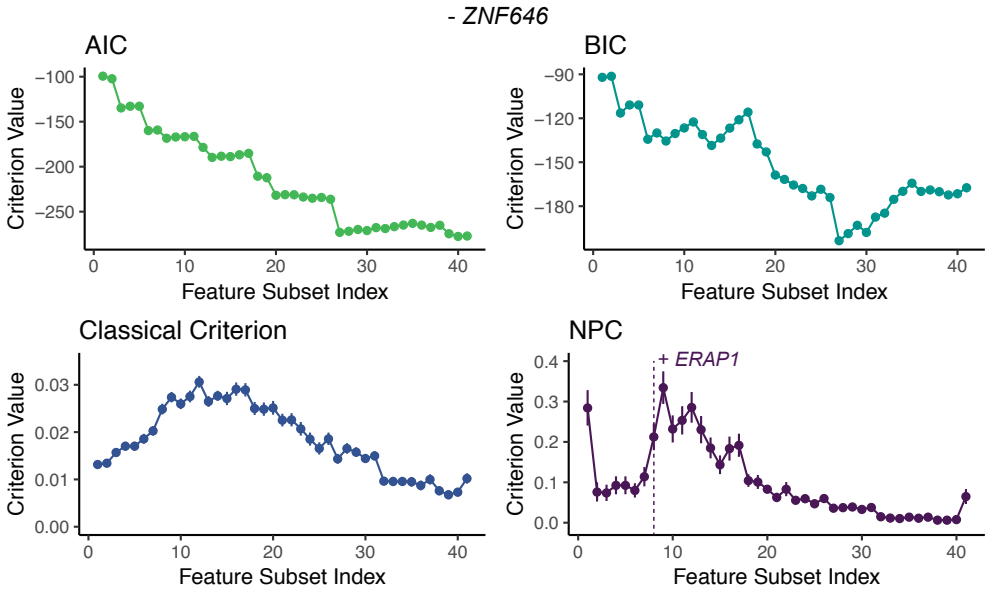

Figure 2.3: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the gene *ZNF646* removed. Other information is the same as in Figure 2.2.

Hence, we further remove *ERAP1* from all feature subsets that already exclude *ZNF646* and re-evaluate the four criteria. The results in Figure 2.4, show that the new rise in the sample-level NPC is due to the addition of the pseudogene *LOC121952* (also known as *METTL21EP*). By removing it and repeating our procedure, we find the genes *GEMIN4* and

BATF and the microRNA *MIR21* that subsequently inflate the sample-level NPC (Figures 2.5, 2.6 and 2.7). After removing all of these six genes (including pseudogenes and microRNAs), we observe (Figure 2.8) that the sample-level NPC no longer exhibits an obvious rise in its trend.

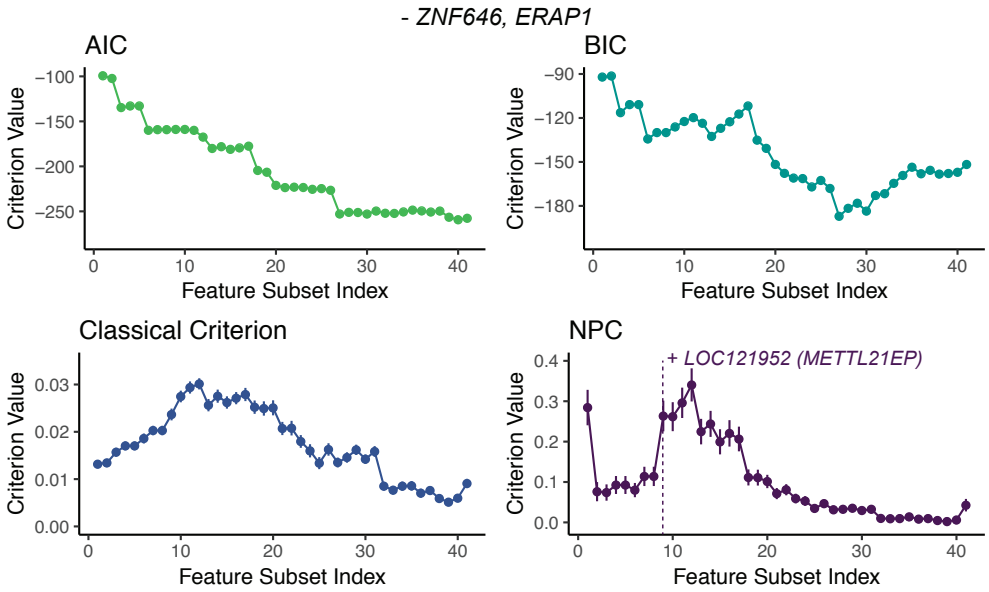

Figure 2.4: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the genes *ZNF646* and *ERAP1* removed. Other information is the same as in Figure 2.2.

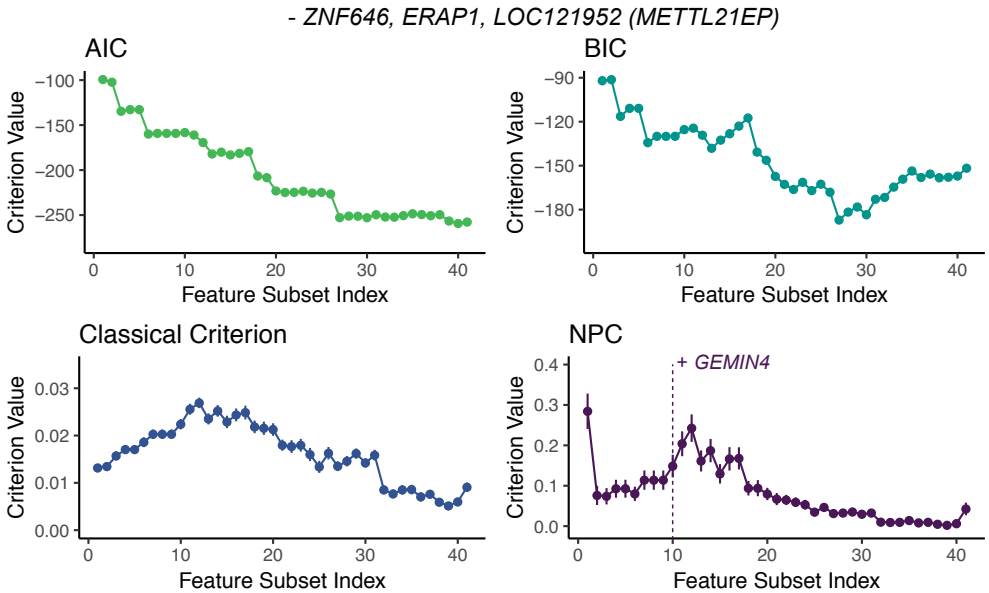

Figure 2.5: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the genes *ZNF646*, *ERAP1*, and *LOC121952 (METTL21EP)* removed. Other information is the same as in Figure 2.2.

Our results suggest that the inclusion of these six genes in feature subsets deteriorates the

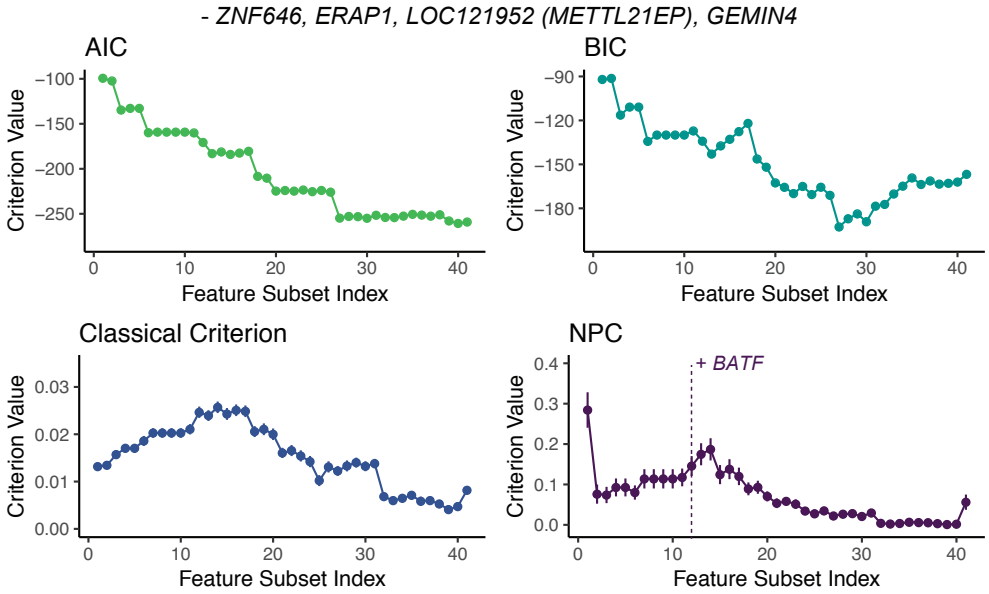

Figure 2.6: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the genes *ZNF646*, *ERAP1*, *LOC121952 (METTL21EP)*, and *GEMIN4* removed. Other information is the same as in Figure 2.2.

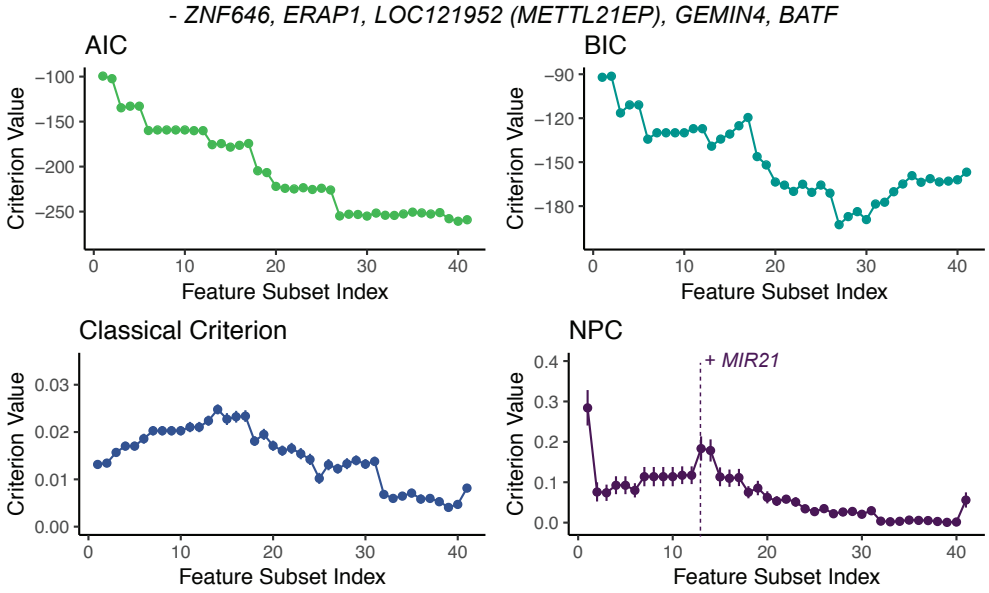

Figure 2.7: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the genes *ZNF646*, *ERAP1*, *LOC121952 (METTL21EP)*, *GEMIN4*, and *BATF* removed. Other information is the same as in Figure 2.2.

type II errors of NP classifiers with $\alpha = .05$ and $\delta = .05$. In other words, these NP classifiers enforce a 95% high sensitivity in detecting breast cancer (with at least 95% probability across samples), and a significant increase in the type II error suggests that the addition of these genes decreases the specificity of these NP classifiers.

To understand this finding, we investigate the functions of these six genes we remove, as

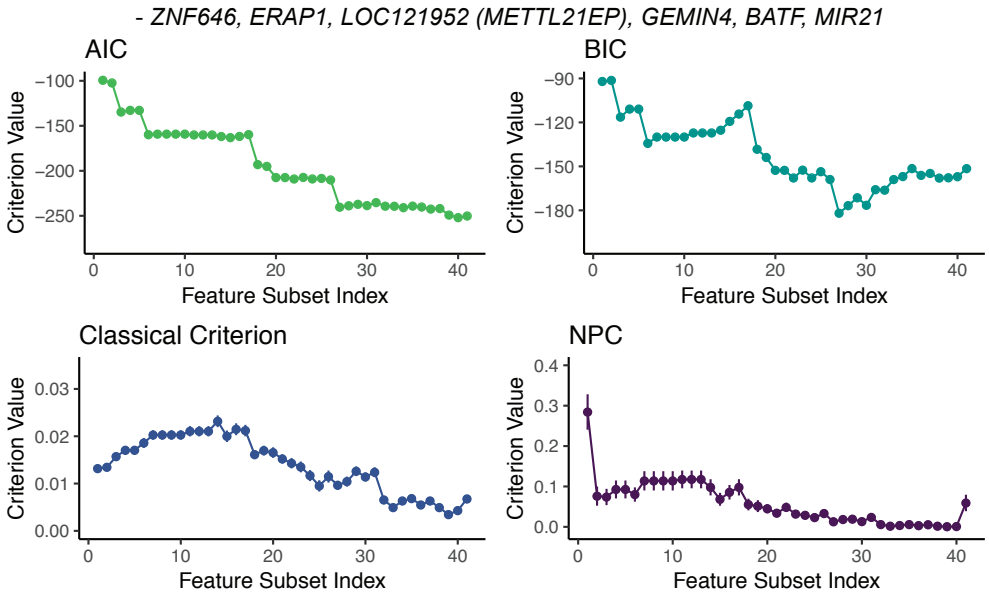

Figure 2.8: Four evaluation criteria on the 41 candidate feature subsets in Figure 2.2 with the genes *ZNF646*, *ERAP1*, *LOC121952 (METTL21EP)*, *GEMIN4*, *BATF*, and *MIR21* removed. Other information is the same as in Figure 2.2.

well as the other 35 genes in the candidate feature subsets. Among these total of 41 genes, there are 36 protein-coding genes (4 removed), 4 microRNAs (1 removed), and 1 pseudogene (1 removed).

We look up the functions of these 36 protein-coding genes in the Human Protein Atlas database (https://www.proteinatlas.org), which contains a Pathology Atlas where breast cancer relevance is specifically listed. Out of these 36 protein-coding genes, 9 genes do not yet have available protein expression data in breast cancer, so we only consider the remaining 27 genes, which include the 4 genes (*ZNF646*, *ERAP1*, *GEMIN4* and *BATF*) we remove from the candidate feature subsets. For these four removed genes, we find that only *ERAP1* and *GEMIN4* exhibit protein expression in breast cancer. *ZNF646* only has protein expression in testis cancer, and the *BATF* protein has not been detected in any cancers in this database, suggesting that excluding them from breast cancer diagnostic features is reasonable. For the other 23 genes in candidate feature subsets, we find that 20 of them are expressed in proteins in breast cancer, and the other three genes (*SPARCL1*, *GCNT4* and *CYP2S1*) have been reported with protein expression in ovarian cancer. Given that ovarian cancer and breast cancer are highly correlated in heredity [64], we hypothesize that they are also related to breast cancer diagnosis, and our hypothesis is supported by clinical research findings [65–69].

We also investigate the functions of the four microRNAs (*MIR195*, *MIR375*, *MIR21* and *MIR451*), all of which have been reported to be associated with the diagnosis, prognosis, and therapy of breast cancer [70–74]. and the pseudogene *LOC121952* (*METTL21EP*). However, we find that including *MIR21* as a predictive feature would decrease the specificity of breast cancer detection from 88.3% to 81.7% when the sensitivity is set to be high (Figure 2.7). Moreover, the pseudogene *LOC121952* (*METTL21EP*) we remove is related to DNA methylation and has not been reported to be associated with breast cancer.

Details of the above functional analysis results are summarized in the *Supplementary Excel File*. The sample-level NPC trend also shows that, in breast cancer detection with a 95% sensitivity on this dataset, a specificity higher than 90% is achievable with only three gene markers: *HMGB2*, *MIR195* and *SPARCL1*. Especially, the inclusion of *SPARCL1* significantly increases the specificity from around 70% to more than 90%. Therefore, *SPARCL1* is a potentially powerful marker for breast cancer detection when high sensitivity id desirable.

To summarize, our real data analysis shows that the sample-level NPC provides a useful and practical criterion for identifying genetic features, among the overall predictive ones, to achieve high specificity in highly-sensitive cancer diagnosis.

2.7 Discussion

In this work, we develop a new model selection criterion: Neyman-Pearson Criterion (NPC), which is tailored for asymmetric binary classification under the NP paradigm. NPC appeals to biomedical practitioners who are interested in identifying cancer drivers but are often constrained by experimental budgets for downstream validation. As experimental costs grow linearly with the number of candidate genomic features, an effective procedure that ranks models of (up to) certain sizes and accounts for the asymmetry in prediction errors, is clearly desirable for making scientific discoveries. In disease diagnosis where a high sensitivity (or a low false negative rate) is desirable, NPC serves as the first available criterion to select a model that achieve a best specificity among candidate models while maintaining a high sensitivity, a perspective different from existing ones. Apart from biomedical sciences, NPC is

also widely applicable to engineering applications such as network security control, financial applications such as loan screening, and social applications such as prediction of regional and international conflicts.

In the theoretical investigation of NPC, we focused on studying plug-in methods with nonparametric assumptions and bounded feature spaces. We leave the investigation of other scoring-type classification methods for future studies. The main idea is to replace the concentration result in Proposition 1 by a deviation result between the estimated scoring function and the method-specific optimal scoring function. Moreover, to accommodate unbounded feature spaces, we need to adopt the conditional versions of the margin assumption and the detection condition, similar to those in [57].

Same as other model selection criteria including AIC and BIC, NPC has to be combined with a proper model space search strategy when the candidate model space is very large. For example, when the number of features $d = 10$, there are $2^{10} = 1024$ feature subsets to search through. This is feasible for modern laptops, but when $d = 40$, an exhaustive search over all 2^{40} feature subsets is overwhelming, not mentioning that large-scale genomic datasets often have d in the order of 10^4. When d is large, forward stepwise selection, which incrementally adds one feature at a time, is often used in practice to reduce the number of candidate models to $d(d+1)/2$. When d far exceeds the sample size n (i.e., under the so-called ultra-high dimensional settings), screening techniques are often used. For example, marginal screening computes some relation between the response and each feature, one at a time, and keeps the most "informative" features.

The current implementation of NPC relies on the NP umbrella classification algorithm in [41], which was derived assuming independent observations. This is unwarranted in, for example, financial time series data. For future studies, it would be interesting to generalize the NP umbrella classification algorithm and NPC for dependent data.

2.8 Acknowledgments

This chapter is partially based on my joint work with Dr. Jessica Li and Dr. Xin Tong [75].

2.9 Supplementary materials: Proofs

Proof of Lemma 2

First we realize that the following three statements are equivalent:

(1) Feature importance ranking under the NP paradigm is invariant to α;

(2) Feature importance ranking under the classical paradigm is invariant to π_0;

(3) Feature importance ranking under the NP paradigm for $\forall \alpha \in (0,1)$ is the same as feature importance ranking under the classical paradigm $\forall \pi_0 \in (0,1)$.

We explore conditions for statement (1) to hold. We will divide our analysis into four scenarios (i)-(iv) regarding distribution means.

Scenario (i): suppose $\mu_1^0 \leqslant \mu_1^1$ and $\mu_2^0 \leqslant \mu_2^1$. Let $c_1, c_2 \in \mathbb{R}$ be such that

$$1-\alpha = \Phi\left(\frac{c_1-\mu_1^0}{\sigma_1}\right), \quad 1-\alpha = \Phi\left(\frac{c_2-\mu_2^0}{\sigma_2}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of $\mathcal{N}(0,1)$. Then the NP oracle classifier using the feature $\boldsymbol{X}_{\{1\}}$ or $\boldsymbol{X}_{\{2\}}$ can be written as

$$\varphi^*_{\alpha\{1\}}(\boldsymbol{X}) = \mathbb{1}\left(\boldsymbol{X}_{\{1\}} > c_1\right) \quad \text{or} \quad \varphi^*_{\alpha\{2\}}(\boldsymbol{X}) = \mathbb{1}\left(\boldsymbol{X}_{\{2\}} > c_2\right).$$

These oracle classifiers have type II errors

$$R_1\left(\varphi^*_{\alpha\{1\}}\right) = \Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right), \quad R_1\left(\varphi^*_{\alpha\{2\}}\right) = \Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right).$$

The chain of equivalence holds

$$
\begin{aligned}
& R_1\left(\varphi^*_{\alpha\{2\}}\right) \geqslant R_1\left(\varphi^*_{\alpha\{1\}}\right) \\
\Leftrightarrow \quad & \Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right) \geqslant \Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right) \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^1}{\sigma_2} \geqslant \frac{c_1-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^0}{\sigma_2}+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant \frac{c_1-\mu_1^0}{\sigma_1}+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \Phi^{-1}(1-\alpha)+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant \Phi^{-1}(1-\alpha)+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{\mu_1^1-\mu_1^0}{\sigma_1}-\frac{\mu_2^1-\mu_2^0}{\sigma_2} \geqslant 0\,,
\end{aligned}
$$

Therefore,

$$
\operatorname{sign}\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right)-R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\}=\operatorname{sign}\left\{\frac{\mu_1^1-\mu_1^0}{\sigma_1}-\frac{\mu_2^1-\mu_2^0}{\sigma_2}\right\}.
$$

Scenario (ii): suppose $\mu_1^0 > \mu_1^1$ and $\mu_2^0 > \mu_2^1$. Let c_1, $c_2 \in \mathbb{R}$ be such that

$$
\alpha=\Phi\left(\frac{c_1-\mu_1^0}{\sigma_1}\right), \quad \alpha=\Phi\left(\frac{c_2-\mu_2^0}{\sigma_2}\right),
$$

then the NP oracle classifier using the feature $\boldsymbol{X}_{\{1\}}$ or $\boldsymbol{X}_{\{2\}}$ can be written as

$$
\varphi^*_{\alpha\{1\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{1\}}<c_1\right) \quad \text{or} \quad \varphi^*_{\alpha\{2\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{2\}}<c_2\right).
$$

These oracle classifiers have type II errors

$$
R_1\left(\varphi^*_{\alpha\{1\}}\right)=1-\Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right), \quad R_1\left(\varphi^*_{\alpha\{2\}}\right)=1-\Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right).
$$

The chain of equivalence holds

$$
\begin{aligned}
& R_1\left(\varphi^*_{\alpha\{2\}}\right) \geqslant R_1\left(\varphi^*_{\alpha\{1\}}\right) \\
\Leftrightarrow \quad & \Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right) \leqslant \Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right) \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^1}{\sigma_2} \leqslant \frac{c_1-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^0}{\sigma_2}+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \leqslant \frac{c_1-\mu_1^0}{\sigma_1}+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \Phi^{-1}(\alpha)+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \leqslant \Phi^{-1}(\alpha)+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{\mu_1^1-\mu_1^0}{\sigma_1}-\frac{\mu_2^1-\mu_2^0}{\sigma_2} \leqslant 0\,.
\end{aligned}
$$

Hence,

$$
\operatorname{sign}\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right)-R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\}=\operatorname{sign}\left\{-\frac{\mu_1^1-\mu_1^0}{\sigma_1}+\frac{\mu_2^1-\mu_2^0}{\sigma_2}\right\}\,.
$$

Scenario (iii): suppose $\mu_1^0 \leqslant \mu_1^1$ and $\mu_2^0 > \mu_2^1$. Let c_1, $c_2 \in \mathbb{R}$ be such that

$$
1-\alpha=\Phi\left(\frac{c_1-\mu_1^0}{\sigma_1}\right)\,,\quad \alpha=\Phi\left(\frac{c_2-\mu_2^0}{\sigma_2}\right)\,,
$$

then the NP oracle classifier using the feature $\boldsymbol{X}_{\{1\}}$ or $\boldsymbol{X}_{\{2\}}$ can be written as

$$
\varphi^*_{\alpha\{1\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{1\}}>c_1\right)\quad \text{or}\quad \varphi^*_{\alpha\{2\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{2\}}<c_2\right)\,.
$$

These oracle classifiers have type II errors

$$
R_1\left(\varphi^*_{\alpha\{1\}}\right)=\Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right)\,,\quad R_1\left(\varphi^*_{\alpha\{2\}}\right)=1-\Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right)\,.
$$

Because $\Phi(a)+\Phi(-a)=1$ for any $a \in \mathbb{R}$ and $\Phi^{-1}(1-\alpha)=-\Phi^{-1}(\alpha)$ for all $\alpha \in (0,1)$, we

have the chain of equivalence

$$
\begin{aligned}
& R_1\left(\varphi^*_{\alpha\{2\}}\right) \geqslant R_1\left(\varphi^*_{\alpha\{1\}}\right) \\
\Leftrightarrow \quad & 1-\Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right) \geqslant \Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right) \\
\Leftrightarrow \quad & -\frac{c_2-\mu_2^1}{\sigma_2} \geqslant \frac{c_1-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & -\frac{c_2-\mu_2^0}{\sigma_2}-\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant \frac{c_1-\mu_1^0}{\sigma_1}+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & -\Phi^{-1}(\alpha)-\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant \Phi^{-1}(1-\alpha)+\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{\mu_1^1-\mu_1^0}{\sigma_1}+\frac{\mu_2^1-\mu_2^0}{\sigma_2} \geqslant 0\,.
\end{aligned}
$$

Hence,

$$
\operatorname{sign}\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right)-R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\}=\operatorname{sign}\left\{\frac{\mu_1^1-\mu_1^0}{\sigma_1}+\frac{\mu_2^1-\mu_2^0}{\sigma_2}\right\}\,.
$$

Scenario (iv): suppose $\mu_1^0 > \mu_1^1$ and $\mu_2^0 \leqslant \mu_2^1$. Let c_1, $c_2 \in \mathbb{R}$ be such that

$$
\alpha=\Phi\left(\frac{c_1-\mu_1^0}{\sigma_1}\right)\,,\quad 1-\alpha=\Phi\left(\frac{c_2-\mu_2^0}{\sigma_2}\right)\,,
$$

then the NP oracle classifier using the feature $\boldsymbol{X}_{\{1\}}$ or $\boldsymbol{X}_{\{2\}}$ can be written as

$$
\varphi^*_{\alpha\{1\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{1\}}<c_1\right)\quad \text{or}\quad \varphi^*_{\alpha\{2\}}(\boldsymbol{X})=\mathbb{1}\left(\boldsymbol{X}_{\{2\}}>c_2\right)\,.
$$

These oracle classifiers have type II errors

$$
R_1\left(\varphi^*_{\alpha\{1\}}\right)=1-\Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right)\,,\quad R_1\left(\varphi^*_{\alpha\{2\}}\right)=\Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right)\,.
$$

Because $\Phi(a)+\Phi(-a)=1$ for any $a \in \mathbb{R}$ and $\Phi^{-1}(\alpha)=-\Phi^{-1}(1-\alpha)$ for all $\alpha \in (0,1)$, we

have the chain of equivalence

$$
\begin{aligned}
& R_1\left(\varphi^*_{\alpha\{2\}}\right) \geqslant R_1\left(\varphi^*_{\alpha\{1\}}\right) \\
\Leftrightarrow \quad & \Phi\left(\frac{c_2-\mu_2^1}{\sigma_2}\right) \geqslant 1-\Phi\left(\frac{c_1-\mu_1^1}{\sigma_1}\right) \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^1}{\sigma_2} \geqslant -\frac{c_1-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \frac{c_2-\mu_2^0}{\sigma_2}+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant -\frac{c_1-\mu_1^0}{\sigma_1}-\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & \Phi^{-1}(1-\alpha)+\frac{\mu_2^0-\mu_2^1}{\sigma_2} \geqslant -\Phi^{-1}(\alpha)-\frac{\mu_1^0-\mu_1^1}{\sigma_1} \\
\Leftrightarrow \quad & -\frac{\mu_1^1-\mu_1^0}{\sigma_1}-\frac{\mu_2^1-\mu_2^0}{\sigma_2} \geqslant 0\,.
\end{aligned}
$$

Hence,

$$
\operatorname{sign}\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right)-R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\}=\operatorname{sign}\left\{-\frac{\mu_1^1-\mu_1^0}{\sigma_1}-\frac{\mu_2^1-\mu_2^0}{\sigma_2}\right\}\,.
$$

Finally to sum up scenarios (i)-(iv), we conclude that

$$
\operatorname{sign}\left\{R_1\left(\varphi^*_{\alpha\{2\}}\right)-R_1\left(\varphi^*_{\alpha\{1\}}\right)\right\}=\operatorname{sign}\left\{\frac{|\mu_1^1-\mu_1^0|}{\sigma_1}-\frac{|\mu_2^1-\mu_2^0|}{\sigma_2}\right\}\,.
$$

Proof of Lemma 3

By the Neyman-Pearson Lemma (Lemma 1), we can write out NP oracles $\varphi^*_{\alpha A_1}(\cdot)$ and $\varphi^*_{\alpha A_2}(\cdot)$ as follows:

$$
\varphi^*_{\alpha A_1}(\boldsymbol{X})=\mathbb{1}\left(\left(\boldsymbol{\mu}_1^1-\boldsymbol{\mu}_1^0\right)^{\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{X}_{A_1}>c_1\right)\,,\quad \varphi^*_{\alpha A_2}(\boldsymbol{X})=\mathbb{1}\left(\left(\boldsymbol{\mu}_2^1-\boldsymbol{\mu}_2^0\right)^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{X}_{A_2}>c_2\right)\,,
$$

where

$$
c_1=\frac{1}{2}\left(T_1-\boldsymbol{\mu}_1^{0\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^0+\boldsymbol{\mu}_1^{1\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^1\right)\,,\quad c_2=\frac{1}{2}\left(T_2-\boldsymbol{\mu}_2^{0\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^0+\boldsymbol{\mu}_2^{1\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^1\right)\,,
$$

in which T_i, $i = 1, 2$, is the threshold on log density ratio. The T_i's vary with α and are determined as in the Neyman-Pearson Lemma. Note that

$$\left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X}_{A_1} \,|\, (Y = 0) \in \mathbb{R}\,, \quad \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X}_{A_1} \,|\, (Y = 1) \in \mathbb{R}\,,$$

follow Gaussian distributions with the same variance $\left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)$. Similarly,

$$\left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{X}_{A_2} \,|\, (Y = 1) \in \mathbb{R}\,, \quad \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{X}_{A_2} \,|\, (Y = 0) \in \mathbb{R}\,,$$

follow Gaussian distributions with the same variance $\left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)$.

Let $\widetilde{\boldsymbol{X}}_{\{1\}} = \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{X}_{A_1}$ and $\widetilde{\boldsymbol{X}}_{\{2\}} = \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{X}_{A_2}$. Denote by

$$\begin{aligned}
\tilde{\mu}_1^1 &= \mathbb{E}\left(\widetilde{\boldsymbol{X}}_{\{1\}} | Y = 1\right) = \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1^1\,, \\
\tilde{\mu}_1^0 &= \mathbb{E}\left(\widetilde{\boldsymbol{X}}_{\{1\}} | Y = 0\right) = \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1^0\,, \\
\tilde{\mu}_2^1 &= \mathbb{E}\left(\widetilde{\boldsymbol{X}}_{\{2\}} | Y = 1\right) = \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2^1\,, \\
\tilde{\mu}_2^0 &= \mathbb{E}\left(\widetilde{\boldsymbol{X}}_{\{2\}} | Y = 0\right) = \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2^0\,, \\
\tilde{\sigma}_1 &= \mathrm{Var}(\widetilde{\boldsymbol{X}}_{\{1\}} | (Y = 1)) = \mathrm{Var}(\widetilde{\boldsymbol{X}}_{\{1\}} | (Y = 0)) = \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)\,, \\
\tilde{\sigma}_2 &= \mathrm{Var}(\widetilde{\boldsymbol{X}}_{\{2\}} | (Y = 1)) = \mathrm{Var}(\widetilde{\boldsymbol{X}}_{\{2\}} | (Y = 0)) = \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)\,.
\end{aligned}$$

Note that

$$\tilde{\mu}_1^1 - \tilde{\mu}_1^0 = \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right)^\mathsf{T} \boldsymbol{\Sigma}_1^{-1} \left(\boldsymbol{\mu}_1^1 - \boldsymbol{\mu}_1^0\right) \geqslant 0\,,$$

$$\tilde{\mu}_2^1 - \tilde{\mu}_2^0 = \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right)^\mathsf{T} \boldsymbol{\Sigma}_2^{-1} \left(\boldsymbol{\mu}_2^1 - \boldsymbol{\mu}_2^0\right) \geqslant 0\,.$$

Apparently, when $\tilde{\mu}_1^0 \leqslant \tilde{\mu}_1^1$ and $\tilde{\mu}_2^0 \leqslant \tilde{\mu}_2^1$, $\mathbb{1}(\widetilde{\boldsymbol{X}}_{\{1\}} > c_1)$ and $\mathbb{1}(\widetilde{\boldsymbol{X}}_{\{2\}} > c_2)$ are the α-level NP oracle classifiers using respectively 1-dimensional features $\widetilde{\boldsymbol{X}}_{\{1\}}$ and $\widetilde{\boldsymbol{X}}_{\{2\}}$. Applying Lemma 2 to $\widetilde{\boldsymbol{X}}_{\{1\}}$ and $\widetilde{\boldsymbol{X}}_{\{2\}}$, we conclude that the given conditions in the Lemma guarantee invariance of importance ranking of the NP oracles regarding the level α.

Proof of Lemma 4

Let $K(\cdot)$ be a real-valued β-valid kernel function on $\mathbb{R}^d$ with the support $[-1,1]^d$. Let $\boldsymbol{u} = (v, \boldsymbol{w})$ where $v \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^{d-1}$. Define $K'(v, \boldsymbol{w}) := \int K(v, \boldsymbol{w}) dv$. Since $\int K'(\boldsymbol{w}) d\boldsymbol{w} = \int\int K((v, \boldsymbol{w})) dv d\boldsymbol{w} = \int K(\boldsymbol{u}) d\boldsymbol{u} = 1$, and $K'(\cdot)$ is clearly supported on $[-1,1]^{d-1}$, $K'(\cdot)$ is a real-valued kernel function on $\mathbb{R}^{d-1}$. For all $l \geqslant 1$, it follows from Jensen's inequality and the first property of β-valid kernel of K that

$$\int |K'(\boldsymbol{w})|^l d\boldsymbol{w} = \int \left| \int K((v, \boldsymbol{w})) dv \right|^l d\boldsymbol{w} \leqslant \int\int |K(\boldsymbol{u})|^l \, dv d\boldsymbol{w} = \int |K|^l < \infty \,. \tag{S2.15}$$

By the second property of β-valid kernel of K,

$$\begin{aligned} \int \|\boldsymbol{w}\|^\beta \, |K'(\boldsymbol{w})| \, d\boldsymbol{w} &= \int \|\boldsymbol{w}\|^\beta \left| \int K((v, \boldsymbol{w})) dv \right| d\boldsymbol{w} \leqslant \int\int \|\boldsymbol{w}\|^\beta \, |K((v, \boldsymbol{w}))| dv d\boldsymbol{w} \\ &\leqslant \int\int \|(v, \boldsymbol{w})\|^\beta \, |K((v, \boldsymbol{w}))| dv d\boldsymbol{w} \int = \int \|\boldsymbol{u}\|^\beta |K(\boldsymbol{u})| d\boldsymbol{u} < \infty \,. \end{aligned} \tag{S2.16}$$

By the third property of β-valid kernel of K, for all $\boldsymbol{t} \in \mathbb{N}^{d-1}$ such that $1 \leqslant |(0, \boldsymbol{t})| \leqslant \lfloor \beta \rfloor$, we have

$$\begin{aligned} \int \boldsymbol{w}^{\boldsymbol{t}} K'(\boldsymbol{w}) d\boldsymbol{w} &= \int \boldsymbol{w}^{\boldsymbol{t}} \int K((v, \boldsymbol{w})) dv d\boldsymbol{w} \\ &= \int\int \boldsymbol{w}^{\boldsymbol{t}} K((v, \boldsymbol{w})) dv d\boldsymbol{w} = \int (v, \boldsymbol{w})^{(0, \boldsymbol{t})} K(\boldsymbol{u}) d\boldsymbol{u} = 0 \,. \end{aligned} \tag{S2.17}$$

Inequalities (S2.15)-(S2.17) together show that $K'(\cdot)$ is a β-valid kernel on $\mathbb{R}^{d-1}$ with support $[-1,1]^{d-1}$.

Proof of Proposition 1

Given that the kernel K is β-valid, Lemma 4 implies that K_A is β-valid. Since K is L'-Lipschitz, for all $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^d$, we have

$$|K(\boldsymbol{u}_1) - K(\boldsymbol{u}_2)| \leqslant L' \|\boldsymbol{u}_1 - \boldsymbol{u}_2\| \,.$$

Now for any $\boldsymbol{u}_{1A}, \boldsymbol{u}_{2A} \in \mathbb{R}^l$,

$$\begin{aligned}
|K_A(\boldsymbol{u}_{1A}) - K_A(\boldsymbol{u}_{2A})| &= \left| \int K(\boldsymbol{u}_{1A}, \boldsymbol{u}_{A^c}) d\boldsymbol{u}_{A^c} - \int K(\boldsymbol{u}_{2A}, \boldsymbol{u}_{A^c}) d\boldsymbol{u}_{A^c} \right| \\
&= \int |K(\boldsymbol{u}_{1A}, \boldsymbol{u}_{A^c}) - K(\boldsymbol{u}_{2A}, \boldsymbol{u}_{A^c})| \, d\boldsymbol{u}_{A^c} \\
&= \int_{[-1,1]^{d-l}} |K(\boldsymbol{u}_{1A}, \boldsymbol{u}_{A^c}) - K(\boldsymbol{u}_{2A}, \boldsymbol{u}_{A^c})| \, d\boldsymbol{u}_{A^c} \\
&\leqslant \int_{[-1,1]^{d-l}} L' \|(\boldsymbol{u}_{1A}, \boldsymbol{u}_{A^c}) - (\boldsymbol{u}_{2A}, \boldsymbol{u}_{A^c})\| d\boldsymbol{u}_{A^c} \\
&= \int_{[-1,1]^{d-l}} L' \|\boldsymbol{u}_{1A} - \boldsymbol{u}_{2A}\| d\boldsymbol{u}_{A^c} \\
&= 2^{d-l} L' \|\boldsymbol{u}_{1A} - \boldsymbol{u}_{2A}\| \,,
\end{aligned}$$

where the second equality follows because K is assumed to have support $[-1,1]^d$. Therefore, for any $\boldsymbol{u}_{1A}, \boldsymbol{u}_{2A} \in \mathbb{R}^l$,

$$|K_A(\boldsymbol{u}_{1A}) - K_A(\boldsymbol{u}_{2A})| \leqslant L'_A \|\boldsymbol{u}_{1A} - \boldsymbol{u}_{2A}\| \,.$$

for some positive constant $L'_A (= 2^{d-l} L')$, i.e., K_A is L'_A-Lipshitz.

let $h_{m_1} = \left(\frac{\log m_1}{m_1} \right)^{\frac{1}{2\beta+\ell}}$. By Lemma 1 in [1], there exists some constant C_{0A} that does not depend on m_1 and δ_3, such that with probability at least $1 - \delta_3/2$,

$$\|\hat{p}_{0A}(\boldsymbol{x}_A) - p_{0A}(\boldsymbol{x}_A)\|_\infty \leqslant \varepsilon_{0A} \,,$$

where $\varepsilon_{0A} = C_{0A} \sqrt{\frac{\log(2m_1/\delta_3)}{m_1 h_{m_1}^\ell}}$, where $C_{0A} = \sqrt{48 c_{1A}} + 32 c_{2A} + 2L c_{3A} + L'_A + L + \tilde{C}_A \sum_{1 \leqslant |\boldsymbol{q}| \leqslant \lfloor \beta \rfloor} \frac{1}{\boldsymbol{q}!}$, in which $c_{1A} = \|p_{0A}\|_\infty \|K_A\|^2$, $c_{2A} = \|K_A\|_\infty + \|p_{0A}\|_\infty + \int |K_A| |\boldsymbol{t}|^\beta d\boldsymbol{t}$, $c_{3A} = \int |K_A| |\boldsymbol{t}|^\beta d\boldsymbol{t}$, $L'_A = 2^{d-l} L'$ and $\tilde{C}_A$ is such that $\tilde{C}_A \geqslant \sup_{1 \leqslant |\boldsymbol{q}| \leqslant \lfloor \beta \rfloor} \sup_{\boldsymbol{x}_A \in [-1,1]^l} |p_{0A}^{(\boldsymbol{q})}(\boldsymbol{x}_A)|$.

Since for fixed number of d, there are finite number of subsets, $C_0 = \max_A C_{0A}$ is finite. Therefore, we have with probability at least $1 - \delta_3/2$,

$$\|\hat{p}_{0A}(\boldsymbol{x}) - p_{0A}(\boldsymbol{x})\|_\infty \leqslant \varepsilon_0 \,,$$

where $\varepsilon_0 = C_0\sqrt{\frac{\log(2m_1/\delta_3)}{m_1 h_{m_1}^{\ell}}}$. Similarly let $h_{n_1} = \left(\frac{\log n_1}{n_1}\right)^{\frac{1}{2\beta+\ell}}$, there exists some constant C_1 that does not depend on n_1 and δ_3, such that with probability at least $1-\delta_3/2$,

$$\|\hat{p}_{1A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)\|_\infty \leqslant \varepsilon_1\,,$$

where $\varepsilon_1 = C_1\sqrt{\frac{\log(2n_1/\delta_3)}{n_1 h_{n_1}^{\ell}}}$. Also, because for all $A \in \{1,\cdots,d\}$, p_{1A} is Hölder class on compact set, and there is a bounded number of all p_{1A}'s, there is universal upper bound U of $\|p_{1A}\|_\infty$ for all $A \in \{1,\cdots,d\}$. Therefore, we have with probability at least $1-\delta_3$,

$$\begin{aligned}
&\left\|\frac{\hat{p}_{1A}(\boldsymbol{x}_A)}{\hat{p}_{0A}(\boldsymbol{x}_A)} - \frac{p_{1A}(\boldsymbol{x}_A)}{p_{0A}(\boldsymbol{x}_A)}\right\|_\infty \\
\leqslant\ &\left\|\frac{\hat{p}_{1A}(\boldsymbol{x}_A)}{\hat{p}_{0A}(\boldsymbol{x}_A)} - \frac{p_{1A}(\boldsymbol{x}_A)}{\hat{p}_{0A}(\boldsymbol{x}_A)}\right\|_\infty + \left\|\frac{p_{1A}(\boldsymbol{x}_A)}{\hat{p}_{0A}(\boldsymbol{x}_A)} - \frac{p_{1A}(\boldsymbol{x}_A)}{p_{0A}(\boldsymbol{x}_A)}\right\|_\infty \\
\leqslant\ &\left\|\frac{1}{\hat{p}_{0A}(\boldsymbol{x}_A)}\right\|_\infty \|\hat{p}_{1A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)\|_\infty + \left\|\frac{p_{1A}}{p_{0A}}\right\|_\infty \left\|\frac{p_{0A}}{\hat{p}_{0A}} - 1\right\|_\infty \\
\leqslant\ &\left\|\frac{1}{\hat{p}_{0A}(\boldsymbol{x}_A)}\right\|_\infty \|\hat{p}_{1A}(\boldsymbol{x}_A) - p_{1A}(\boldsymbol{x}_A)\|_\infty + \left\|\frac{p_{1A}}{p_{0A}}\right\|_\infty \left\|\frac{p_{0A} - \hat{p}_{0A}}{\hat{p}_{0A}}\right\|_\infty \\
\leqslant\ &\frac{\varepsilon_1 + \varepsilon_0 U/\mu_{\min}}{\mu_{\min} - \varepsilon_0} =: b_{m_1,n_1}\,.
\end{aligned}$$

When $n_1 \wedge m_1 \geqslant 2/\delta_3$,

$$\varepsilon_0 \leqslant \sqrt{2}C_0\left(\frac{\log m_1}{m_1}\right)^{\beta/(2\beta+\ell)}, \quad \varepsilon_1 \leqslant \sqrt{2}C_1\left(\frac{\log n_1}{n_1}\right)^{\beta/(2\beta+\ell)}.$$

These combined with $\sqrt{\frac{\log(2m_1/\delta_3)}{m_1 h_{m_1}^{\ell}}} < \frac{\mu_{\min}}{2C_0}$ imply that

$$b_{m_1,n_1} \leqslant \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\beta/(2\beta+\ell)} + \left(\frac{\log n_1}{n_1}\right)^{\beta/(2\beta+\ell)}\right], \tag{S2.18}$$

for some positive constant $\widetilde{C}$ that does not depend on the subset A.

Proof of Lemma 5

Given any feature set A, let $\{T_{iA} := \hat{s}_A(\boldsymbol{X}^0_{iA}), \boldsymbol{X}^0_i \in \mathcal{S}^0_{\text{lo}}\}$ be the scores by applying the scoring function $\hat{s}_A(\cdot) = \hat{p}_{0A}(\cdot)/\hat{p}_{1A}(\cdot)$ to $\mathcal{S}^0_{\text{lo}}$. Sort $\{T_{iA}\}$ in an increasing order such that $T_{(1)A} \leqslant, \ldots, T_{(m_2)A}$. Let $\widehat{C}'_{\alpha A} = T_{(k')A}$ be a score threshold using k'-th order statistic, where $k' = \lceil (m_2+1)d_{\alpha,\delta_1,m_2} \rceil$, in which

$$d_{\alpha,\delta_1,m_2} = \frac{1 + 2\delta_1(m_2+2)(1-\alpha) + \sqrt{1 + 4\delta_1(m_2+2)(1-\alpha)\alpha}}{2\left\{\delta_1(m_2+2)+1\right\}},$$

and $\lceil z \rceil$ denotes the smallest integer larger than or equal to z. Denote the corresponding NP classifier as

$$\hat{\phi}'_{\alpha A}(\boldsymbol{X}) = \mathbb{1}\left(\hat{s}_A(\boldsymbol{X}_A) > \widehat{C}'_{\alpha A}\right).$$

Because we use kernel density estimates and the kernels are β-valid, the scoring function $\hat{s}_A(\cdot)$ is continuous. Therefore, by Proposition 1 in [2], we have

$$\mathbb{P}\left(R_0\left(\hat{\phi}'_{\alpha A}\right) > \alpha\right) = \sum_{j=k'}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j},$$
$$\mathbb{P}\left(R_0\left(\hat{\phi}_{\alpha A}\right) > \alpha\right) = \sum_{j=k*}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j}.$$

Note that by the definition of k^*,

$$k^* = \min\left\{k : \sum_{j=k}^{m_2} \binom{m_2}{j} (1-\alpha)^j \alpha^{m_2-j} \leqslant \delta_1\right\}.$$

Proposition 2.2 in [3] implies $\mathbb{P}\left(R_0\left(\hat{\phi}'_{\alpha A}\right) > \alpha\right) \leqslant \delta_1$. So we also have $\sum_{j=k'}^{m_2} \binom{m_2}{j}(1-\alpha)^j \alpha^{m_2-j} \leqslant \delta_1$. This together with the definition of k^* implies that $k' \geqslant k^*$, and therefore $R_0(\hat{\phi}_{\alpha A}) \geqslant R_0(\hat{\phi}'_{\alpha A})$.

By Lemma 2.1 in [3], for any $\delta_2 \in (0,1)$, if $m_2 \geqslant \frac{4}{\alpha\delta_1}$,

$$\mathbb{P}\left(\left|R_0\left(\hat{\phi}'_{\alpha A}\right) - R_0(\varphi^*_{\alpha A})\right| > \xi\right) \leqslant \delta_2,$$

where ξ is defined by

$$\xi = \sqrt{\frac{\lceil d_{\alpha,\delta_1,m_2}\,(m_2+1)\rceil\,(m_2+1-\lceil d_{\alpha,\delta_1,m_2}\,(m_2+1)\rceil)}{(m_2+2)(m_2+1)^2\,\delta_2}} + d_{\alpha,\delta_1,m_2} + \frac{1}{m_2+1} - (1-\alpha)\,.$$

Let $\mathcal{E}_1 := \left\{R_0\left(\hat{\phi}_{\alpha A}\right) \leqslant \alpha\right\}$ and $\mathcal{E}_2 := \left\{\left|R_0\left(\hat{\phi}'_{\alpha A}\right) - R_0(\varphi^*_{\alpha A})\right| \leqslant \xi\right\}$. On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, which has probability at least $1-\delta_1-\delta_2$, we have

$$\alpha = R_0(\varphi^*_{\alpha A}) \geqslant R_0(\hat{\phi}_{\alpha A}) \geqslant R_0(\hat{\phi}'_{\alpha A}) \geqslant R_0(\varphi^*_{\alpha A}) - \xi\,,$$

which implies

$$\left|R_0(\hat{\phi}_{\alpha A}) - R_0(\varphi^*_{\alpha A})\right| \leqslant \xi\,.$$

If $m_2 \geqslant \max(\delta_1^{-2}, \delta_2^{-2})$, we have $\xi \leqslant (5/2)m_2^{-1/4}$, also by Lemma 2.1 of [3].

Proof of Theorem 1

Decompose $|\mathrm{NPC}_{\alpha A} - R_1\left(\varphi^*_{\alpha A}\right)|$ as follows:

$$|\mathrm{NPC}_{\alpha A} - R_1\left(\varphi^*_{\alpha A}\right)| \leqslant \left|\mathrm{NPC}_{\alpha A} - R_1\left(\hat{\phi}_{\alpha A}\right)\right| + \left|R_1\left(\hat{\phi}_{\alpha A}\right) - R_1\left(\varphi^*_{\alpha A}\right)\right|\,.$$

First we derive a bound for $|\mathrm{NPC}_{\alpha A} - R_1(\hat{\phi}_{\alpha A})|$. Let $D > 0$, then conditioning on $\hat{s}_A(\cdot)$ and $\widehat{C}_{\alpha A}$, by Hoeffding's inequality, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\mathbb{1}\left(\hat{s}_A(\boldsymbol{X}^1_{iA}) < \widehat{C}_{\alpha A}\right) - \mathbb{E}\left[\mathbb{1}\left(\hat{s}_A(\boldsymbol{X}^1_{A}) < \widehat{C}_{\alpha A}\right)\right]\right| > D \;\middle|\; \hat{s}_A(\cdot)\,, \widehat{C}_{\alpha A}\right)\\ \leqslant& 2e^{-2n_2D^2}\,.\end{aligned}$$

This implies the following unconditional result,

$$\mathbb{P}\left(\left|\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\mathbb{1}\left(\hat{s}_A(\boldsymbol{X}^1_{iA}) < \widehat{C}_{\alpha A}\right) - \mathbb{E}\left[\mathbb{1}\left(\hat{s}_A(\boldsymbol{X}^1_{A}) < \widehat{C}_{\alpha A}\right)\right]\right| \leqslant D\right) \geqslant 1 - 2e^{-2n_2D^2}\,.$$

Let $2e^{-2n_2D^2} = \delta_4$ and then $D = \sqrt{\frac{1}{2n_2}\log\frac{2}{\delta_4}}$. So we have with probability at least $1-\delta_4$,

$$\left|\mathrm{NPC}_{\alpha A} - R_1\left(\hat{\phi}_{\alpha A}\right)\right| \leqslant \sqrt{\frac{1}{2n_2}\log\frac{2}{\delta_4}}\,.$$

When $n_2 \geqslant (\log\frac{2}{\delta_4})^2$, $\left|\mathrm{NPC}_{\alpha A} - R_1\left(\hat{\phi}_{\alpha A}\right)\right| \leqslant \frac{1}{\sqrt{2}} n_2^{-\frac{1}{4}}$.

Propositions 1 and 2 imply that, it holds with probability at least $1-\delta_1-\delta_2-\delta_3$,

$$\begin{aligned}
&\left|R_1\left(\hat{\phi}_{\alpha A}\right) - R_1\left(\varphi^*_{\alpha A}\right)\right| \\
\leqslant& 2\bar{C}\left[\left(\frac{2}{5}m_2^{1/4}\underline{C}\right)^{-1/\underline{\gamma}} + \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\beta/(2\beta+\ell)} + \left(\frac{\log n_1}{n_1}\right)^{\beta/(2\beta+\ell)}\right]\right]^{1+\bar{\gamma}} + C^*_{\alpha A}\left(\frac{2}{5}m_2^{1/4}\right)^{-1} \\
\leqslant& \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + \left(\frac{\log n_1}{n_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + m_2^{-(\frac{1}{4}\wedge\frac{1+\bar{\gamma}}{\underline{\gamma}})}\right].
\end{aligned}$$

for some generic constant $\widetilde{C}$. Since we consider fixed d, there are only a finite number of constants $C^*_{\alpha A}$, and so they are bounded from above by a single constant that does not depend on A. Therefore, we have with probability at least $1-\delta_1-\delta_2-\delta_3-\delta_4$,

$$\left|\mathrm{NPC}_{\alpha A} - R_1\left(\hat{\phi}_{\alpha A}\right)\right| \leqslant \widetilde{C}\left[\left(\frac{\log m_1}{m_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + \left(\frac{\log n_1}{n_1}\right)^{\frac{\beta(1+\bar{\gamma})}{2\beta+\ell}} + m_2^{-(\frac{1}{4}\wedge\frac{1+\bar{\gamma}}{\underline{\gamma}})} + n_2^{-\frac{1}{4}}\right],$$

for some generic constant $\widetilde{C}$ that does not depend on A.

Proof of Theorem 2

By Theorem 1, the sample size requirement on m_1, m_2, n_1, n_2, and $|A_1| \leqslant d$, we have with probability at least $1-(\delta_1+\delta_2+\delta_3+\delta_4)$,

$$\mathrm{NPC}_{\alpha A_1} \leqslant R_1(\varphi^*_{\alpha A_1}) + |\mathrm{NPC}_{\alpha A_1} - R_1(\varphi^*_{\alpha A_1})| \leqslant R_1(\varphi^*_{\alpha A_1}) + \frac{g}{2}\,.$$

Similarly for each of $j = 2, \ldots, K$, we have with probability at least $1 - (\delta_1 + \delta_2 + \delta_3 + \delta_4)$,

$$\mathrm{NPC}_{\alpha A_j} \geqslant R_1(\varphi^*_{\alpha A_j}) - |\mathrm{NPC}_{\alpha A_j} - R_1(\varphi^*_{\alpha A_j})| \geqslant R_1(\varphi^*_{\alpha A_j}) - \frac{g}{2} > R_1(\varphi^*_{\alpha A_1}) + \frac{g}{2}\,,$$

where the last inequality follows from the assumption

$$\min_{A \in \{A_2, \ldots, A_K\}} R_1\left(\varphi^*_{\alpha A}\right) - R_1\left(\varphi^*_{\alpha A_1}\right) > g\,.$$

Therefore, with probability at least $1 - K(\delta_1 + \delta_2 + \delta_3 + \delta_4)$,

$$\mathrm{NPC}_{\alpha A_1} < \min_{j=2,\ldots,K} \mathrm{NPC}_{\alpha A_j}\,.$$

In other words, NPC selects the best model A_1 among $\{A_1, \ldots, A_K\}$.

CHAPTER 3

P-value free FDR control in independent multiple testing problems with small sample sizes: high-throughput enrichment and differential analyses

3.1 Introduction

High-throughput technologies are widely used to measure system-wide biological features, such as genes, genomic regions, and proteins ("high-throughput" means the number of features is large, at least in thousands). The most common goal of analyzing high-throughput data is to contrast two conditions so as to reliably screen "interesting features," where "interesting" means "enriched" or "differential." "Enriched features" are defined to have higher expected measurements (without measurement errors) under the experimental (i.e., treatment) condition than the background (i.e., the negative control) condition. The detection of enriched features is called "enrichment analysis." For example, typical enrichment analyses include calling protein-binding sites in a genome from chromatin immunoprecipitation sequencing (ChIP-seq) data [76, 77] and identifying peptides from mass spectrometry (MS) data [78]. In contrast, "differential features" are defined to have different expected measurements between two conditions, and their detection is called "differential analysis." For example, popular differential analyses include the identification of differentially expressed genes (DEGs) from genome-wide gene expression data (e.g., microarray and RNA sequencing (RNA-seq) data [79–85]) and differentially interacting chromatin regions (DIRs) from Hi-C data [86–88] (Fig. 3.1a). In most scientific research, the interesting features only constitute a small proportion of all features, and the remaining majority is referred to as "uninteresting

features."

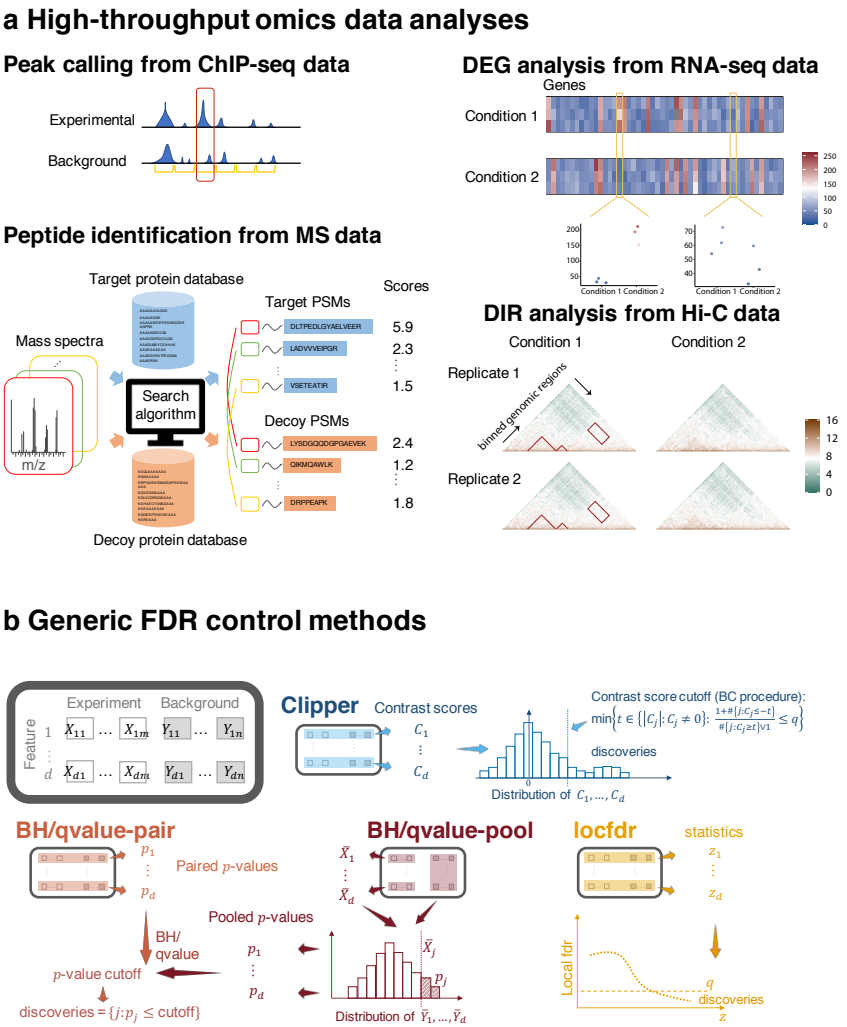

Figure 3.1: High-throughput omics data analyses and generic FDR control methods. (a) Illustration of four common high-throughput omics data analyses: peak calling from ChIP-seq data, peptide identification from MS data, DEG analysis from RNA-seq data, and DIR analysis from Hi-C data. In these four analyses, the corresponding features are genomic regions (yellow intervals), peptide-spectrum matches (PSMs; a pair of a mass spectrum and a peptide sequence), genes (columns in the heatmaps), and chromatin interacting regions (entries in the heatmaps). (b) Illustration of Clipper and five generic FDR control methods: BH-pair (and qvalue-pair), BH-pool (and qvalue-pool), and locfdr. The input data are d features with m and n repeated measurements under the experimental and background conditions, respectively. Clipper computes a contrast score for each feature based on the feature's m and n measurements, decides a contrast-score cutoff, and calls the features with contrast scores above the cutoff as discoveries. (This illustration is Clipper for enrichment analysis with $m = n$.) BH-pair or qvalue-pair computes a p-value for each feature based on the feature's m and n measurements, sets a p-value cutoff, and calls the features with p-values below the cutoff as discoveries. BH-pool or qvalue-pool constructs a null distribution from the d features' average (across the n replicates) measurements under the background condition, calculates a p-value for each feature based on the null distribution and the feature's average (across the m replicates) measurements under the experimental condition, sets a p-value cutoff, and calls the features with p-values below the cutoff as discoveries. The locfdr method computes a summary statistic for each feature based on the feature's m and n measurements, estimates the empirical null distribution and the empirical distribution of the statistic across features, computes a local fdr for each feature, sets a local fdr cutoff, and calls the features with local fdr below the cutoff as discoveries.

The identified features, also called the "discoveries" from enrichment or differential analysis, are subject to further investigation and validation. Hence, to reduce experimental

validation that is often laborious or expensive, researchers demand reliable discoveries that contain few false discoveries. Accordingly, the false discovery rate (FDR) [4] has been developed as a statistical criterion for ensuring discoveries' reliability. The FDR technically is defined as the expected proportion of uninteresting features among the discoveries under the frequentist statistical paradigm. In parallel, under the Bayesian paradigm, other criteria have been developed, including the Bayesian false discovery rate [89], the local false discovery rate (local fdr) [90], and the local false sign rate [91]. Among all these frequentist and Bayesian criteria, the FDR is the dominant criterion for setting thresholds in biological data analysis [76, 85, 92–98] and is thus the focus of this paper.

FDR control refers to the goal of finding discoveries such that the FDR is under a pre-specified threshold (e.g., 0.05). Existing computational methods for FDR control primarily rely on p-values, one per feature. Among the p-value-based methods, the most classic and popular ones are the Benjamini-Hochberg (BH) procedure [4] and the Storey's q-value [13]; later development introduced methods that incorporate feature weights [14] or covariates (e.g., independent hypothesis weighting (IHW) [15], adaptive p-value thresholding [16], and Boca and Leek's FDR regression [17]) to boost the detection power. All these methods set a p-value cutoff based on the pre-specified FDR threshold. However, the calculation of p-values requires either distributional assumptions, which are often questionable, or large numbers of replicates, which are often unachievable in biological studies (see Results). Due to these limitations of p-value-based methods in high-throughput biological data analysis, bioinformatics tools often output ill-posed p-values. This issue is evidenced by serious concerns about the widespread miscalculation and misuse of p-values in the scientific community [99]. As a result, bioinformatics tools using questionable p-values either cannot reliably control the FDR to a target level [97] or lack power to make discoveries [100]; see Results. Therefore, p-value-free control of FDR is desirable, as it would make data analysis more transparent and thus improve the reproducibility of scientific research.

Although p-value-free FDR control has been implemented in the MACS2 method for ChIP-seq peak calling [76] and the SAM method for microarray DEG identification [101], these two methods are restricted to specific applications and lack theoretical guarantee for

FDR control[1]. More recently, the Barber-Candès (BC) procedure has been proposed to achieve theoretical FDR control without using p-values [104], and it has been shown to perform comparably to the BH procedure with well-calibrated p-values [105]. The BC procedure is advantageous because it does not require well-calibrated p-values, so it holds tremendous potential in various high-throughput data analyses where p-value calibration is challenging [106]. For example, a recent paper has implemented the BC procedure to control the FDR in peptide identification from MS data [107].

Inspired by the BC procedure, we propose a general statistical framework Clipper to provide reliable FDR control for high-throughput biological data analysis, without using p-values or relying on specific data distributions. Clipper is a robust and flexible framework that applies to both enrichment and differential analyses and that works for high-throughput data with various characteristics, including data distributions, replicate numbers (from one to multiple), and outlier existence.

3.2 The Clipper methodology

Clipper is a flexible framework that reliably controls the FDR without using p-values in high-throughput data analysis with two conditions. Clipper has two functionalities: (I) enrichment analysis, which identifies the "interesting" features that have higher expected measurements (i.e., true signals) under the experimental condition than the background, a.k.a. negative control condition (if the goal is to identify the interesting features with smaller expected measurements under the experimental condition, enrichment analysis can be applied after the values are negated); (II) differential analysis, which identifies the interesting features that have different expected measurements between the two conditions. For both functionalities, uninteresting features are defined as those that have equal expected measurements under the two conditions.

Clipper only relies on two fundamental statistical assumptions of biological data analysis:

[1]Although later works have studied some theoretical properties of SAM, they are not about the exact control of the FDR [102, 103].

(1) measurement errors (i.e., differences between measurements and their expectations, with the expectations including biological signals and batch effects) are independent across all features and experiments; (2) every uninteresting feature has measurement errors identically distributed across all experiments. These two assumptions are used in almost all bioinformatics tools and commonly referred to as the "measurement model" in statistical genomics [108].

In the following subsections, we will first introduce notations and assumptions used in Clipper. Then we will detail how Clipper works and discuss its theoretical guarantee in three analysis tasks: the enrichment analysis with equal numbers of replicates under two conditions ($m = n$), the enrichment analysis with different numbers of replicates under two conditions ($m \neq n$), and the differential analysis (when $m + n > 2$)..

3.2.1 Notations and assumptions

To facilitate our discussion, we first introduce the following mathematical notations. For two random vectors $\boldsymbol{X} = (X_1, \ldots, X_m)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, or two sets of random variables $\mathcal{X} = \{X_1, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, \ldots, Y_n\}$, we write $\boldsymbol{X} \perp \boldsymbol{Y}$ or $\mathcal{X} \perp \mathcal{Y}$ if X_i is independent of Y_j for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$. To avoid confusion, we use $\mathrm{card}(A)$ to denote the cardinality of a set A and $|c|$ to denote the absolute value of a scalar c. We define $a \vee b := \max(a, b)$.

Clipper only requires two inputs: the target FDR threshold $q \in (0, 1)$ and the input data. Regarding the input data, we use d to denote the number of features with measurements under two conditions, and we use m and n to denote the numbers of replicates under the two conditions. For each feature $j = 1, \ldots, d$, we use $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top \in \mathbb{R}^m$ and $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top \in \mathbb{R}^n$ to denote its measurements under the two conditions, where $\mathbb{R}$ denotes the set of non-negative real numbers. We assume that all measurements are non-negative, as in the case of most high-throughput experiments. (If this assumption does not hold, transformations can be applied to make data satisfy this assumption.)

Clipper has the following assumptions on the joint distribution of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$.

For $j = 1, \ldots, d$, Clipper assumes that $X_{j1}, \ldots, X_{jm}$ are identically distributed, so are $Y_{j1}, \ldots, Y_{jn}$. Let $\mu_{Xj} = \mathbb{E}[X_{j1}]$ and $\mu_{Yj} = \mathbb{E}[Y_{j1}]$ denote the expected measurement of feature j under the two conditions, respectively. Then conditioning on $\{\mu_{Xj}\}_{j=1}^{d}$ and $\{\mu_{Yj}\}_{j=1}^{d}$,

$$\begin{aligned} &X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are mutually independent}\,; \\ &\boldsymbol{X}_j \perp \boldsymbol{X}_k, \boldsymbol{Y}_j \perp \boldsymbol{Y}_k \text{ and } \boldsymbol{X}_j \perp \boldsymbol{Y}_k\,,\ \forall j, k = 1, \ldots, d. \end{aligned} \tag{3.1}$$

An enrichment analysis aims to identify interesting features with $\mu_{Xj} > \mu_{Yj}$ (with $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ defined as the measurements under the experimental and background conditions, respectively), while a differential analysis aims to call interesting features with $\mu_{Xj} \neq \mu_{Yj}$. We define $\mathcal{N} := \{j : \mu_{Xj} = \mu_{Yj}\}$ as the set of uninteresting features and denote $N := \text{card}(\mathcal{N})$. In both analyses, Clipper further assumes that an uninteresting feature j satisfies

$$X_{j1}, \cdots, X_{jm}, Y_{j1}, \cdots, Y_{jn} \text{ are identically distributed}\,, \forall j \in \mathcal{N}\,. \tag{3.2}$$

Clipper consists of two main steps: construction and thresholding of contrast scores. First, Clipper computes contrast scores, one per feature, as summary statistics that reflect the extent to which features are interesting. Second, Clipper establishes a contrast-score cutoff and calls as discoveries the features whose contrast scores exceed the cutoff.

To construct contrast scores, Clipper uses two summary statistics $t(\cdot, \cdot) : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ to extract data information regarding whether a feature is interesting or not:

$$t^{\text{diff}}(\boldsymbol{x}, \boldsymbol{y}) := \bar{x} - \bar{y}\,; \tag{3.3}$$

$$t^{\max}(\boldsymbol{x}, \boldsymbol{y}) := \max(\bar{x}, \bar{y}) \cdot \text{sign}(\bar{x} - \bar{y})\,, \tag{3.4}$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^\top \in \mathbb{R}^m$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$, $\bar{x} = \sum_{i=1}^{m} x_i/m$, $\bar{y} = \sum_{i=1}^{n} y_i/n$, and $\text{sign}(\cdot) : \mathbb{R} \to \{-1, 0, 1\}$ with $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$, and $\text{sign}(x) = 0$ otherwise.

3.2.2 Enrichment analysis with equal numbers of replicates ($m = n$)

Under the enrichment analysis, we assume that $\boldsymbol{X}_j \in \mathbb{R}^m$ and $\boldsymbol{Y}_j \in \mathbb{R}^n$ are the measurements of feature j, $j = 1, \ldots, d$, under the experimental and background conditions with m and n replicates, respectively. We start with the simple case when $m = n$. Clipper defines a contrast score C_j of feature j in one of two ways:

$$C_j := t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{difference contrast score}\,, \tag{3.5}$$

or

$$C_j := t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \qquad \textbf{maximum contrast score}\,. \tag{3.6}$$

Accordingly, a large positive value of C_j bears evidence that $\mu_{Xj} > \mu_{Yj}$. Motivated by Barber and Candès [104] and Arias-Castro and Chen [105], Clipper proposes the following BC procedure to control the FDR under the target level $q \in (0, 1)$.

Definition 3.6 (Barber-Candès (BC) procedure for thresholding contrast scores [104])**.** *Given contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0\,;\; j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The BC procedure finds a contrast-score cutoff T^{BC} based on the target FDR threshold $q \in (0, 1)$ as*

$$T^{BC} := \min\left\{ t \in \mathcal{C} : \frac{\text{card}(\{j : C_j \leqslant -t\}) + 1}{\text{card}(\{j : C_j \geqslant t\}) \vee 1} \leqslant q \right\} \tag{3.7}$$

and outputs $\left\{j : C_j \geqslant T^{BC}\right\}$ as discoveries.

Theorem 3. *Suppose that the input data satisfy the Clipper assumptions (3.1)–(3.2) and $m = n$. Then for any $q \in (0, 1)$ and either definition of constrast scores in (3.5) or (3.6), the contrast-score cutoff T^{BC} found by the BC procedure guarantees that the discoveries have the FDR under q:*

$$FDR = \mathbb{E}\left[\frac{\text{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{BC}\right\}\right)}{\text{card}(\{j : C_j \geqslant T^{BC}\}) \vee 1} \right] \leqslant q\,,$$

where $\mathcal{N} = \{j : \mu_{Xj} = \mu_{Yj}\}$ denotes the set of uninteresting features.

The proof of Theorem 3 (Supp. Section S3.5.7) requires two key ingredients: Lemma 6,

which states important properties of contrast scores, and Lemma 7 from [109], which states a property of a Bernoulli process with independent but not necessarily identically distributed random variables. The cutoff T^{BC} can be viewed as a stopping time of a Bernoulli process.

Lemma 6. *Suppose that the input data that satisfy the Clipper assumptions (3.1)–(3.2) and $m = n$, and that Clipper constructs contrast scores $\{C_j\}_{j=1}^d$ based on (3.5) or (3.6). Denote $S_j = \mathrm{sign}\,(C_j) \in \{-1, 0, 1\}$. Then $\{S_j\}_{j=1}^d$ satisfy the following properties:*

(a) $S_1, \ldots, S_d$ are mutually independent ;

(b) $\mathbb{P}(S_j = 1) = \mathbb{P}(S_j = -1)$ for all $j \in \mathcal{N}$;

(c) $\{S_j\}_{j \in \mathcal{N}} \perp \mathcal{C}$.

Lemma 7. *Suppose that $Z_1, \ldots, Z_d$ are independent with $Z_j \sim \mathrm{Bernoulli}(\rho_j)$, and $\min_j \rho_j \geqslant \rho > 0$. Let J be a stopping time in reverse time with respect to the filtration $\{\mathcal{F}_j\}$, where*

$$\mathcal{F}_j = \sigma\left(\{(Z_1 + \cdots + Z_j),\, Z_{j+1},\, \cdots,\, Z_d\}\right), \tag{3.8}$$

with $\sigma(\cdot)$ denoting a σ-algebra. Then

$$\mathbb{E}\left[\frac{1+J}{1+Z_1+\cdots+Z_J}\right] \leqslant \rho^{-1}.$$

Here we give a brief intuition about how Lemma 7 bridges Lemma 6 and Theorem 3 for FDR control. First we note that the false discovery proportion (FDP), whose expectation is the FDR, satisfies

$$\mathrm{FDP} := \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j : C_j \geqslant T^{\mathrm{BC}}\}\right) \vee 1} \tag{3.9}$$

$$= \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}\right) + 1} \cdot \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\right\}\right) + 1}{\mathrm{card}\left(\{j : C_j \geqslant T^{\mathrm{BC}}\}\right) \vee 1} \tag{3.10}$$

$$\leqslant \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}\right) + 1} \cdot \frac{\mathrm{card}\left(\left\{j : C_j \leqslant -T^{\mathrm{BC}}\right\}\right) + 1}{\mathrm{card}\left(\{j : C_j \geqslant T^{\mathrm{BC}}\}\right) \vee 1} \tag{3.11}$$

$$\leqslant \frac{\mathrm{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\right\}\right)}{\mathrm{card}\left(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}\right) + 1} \cdot q\,, \tag{3.12}$$

where the last inequality follows from the definition of T^{BC} (3.7).

By its definition, if T^{BC} exists, it is positive. This implies that Clipper would never call the features with $C_j = 0$ as discoveries. Here we sketch the idea of proving Theorem 3 by considering a simplified case where $\mathcal{C}$ is fixed instead of being random; that is, we assume the features with non-zero contrast scores to be known. Then, without loss of generality, we assume $\mathcal{C} = \{1, \ldots, d\}$. Then we order the absolute values of uninteresting features' contrast scores, i.e., elements in $\{|C_j| : j \in \mathcal{N}\}$, from the largest to the smallest, denoted by $|C_{(1)}| \geqslant |C_{(2)}| \geqslant \cdots \geqslant |C_{(N)}|$. Let $J = \sum_{j \in \mathcal{N}} \mathbb{1}\left(|C_j| \geqslant T^{\mathrm{BC}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{BC}. When $J > 0$, $|C_{(1)}| \geqslant \cdots \geqslant |C_{(J)}| \geqslant T^{\mathrm{BC}}$. Define $Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, N$. Then for each order k, the following holds

$$C_{(k)} \geqslant T^{\mathrm{BC}} \Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{BC}} \text{ and } C_{(k)} > 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 0\,;$$

$$C_{(k)} \leqslant -T^{\mathrm{BC}} \Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{BC}} \text{ and } C_{(k)} < 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 1\,.$$

Then the upper bound of FDP becomes

$$\begin{aligned}\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} \cdot q &= \frac{\sum_{k=1}^{N} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{N} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{BC}}\right)} \cdot q \\ &= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{BC}}\right)} \cdot q \\ &= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \cdot q \\ &= \left(\frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\right) \cdot q\,.\end{aligned}$$

By Lemma 6(a)–(b), $Z_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(0.5)$, which together with Lemma 6(c) satisfy the condition of Lemma 7 and make $\rho = 0.5$. Then by Lemma 7, we have

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}] \leqslant \mathbb{E}\left[\frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\right] \cdot q \leqslant (\rho^{-1} - 1) \cdot q = q\,,$$

which is the statement of Theorem 3. The complete proof of Theorem 3 is in Supp. Section S3.5.7.

3.2.2.1 An optional, heuristic fix if the BC procedure makes no discoveries

Although the BC procedure has theoretical guarantee of FDR control, it lacks power when the number of replicates $m = n$, the target FDR threshold q, and the number of features d are all small (see Ge et al. [18] for evidence). As a result, the BC procedure may lead to no discoveries. In that case, Clipper implements a heuristic fix—an approximate p-value Benjamini-Hochberg (aBH) procedure—to increase the power. The aBH procedure constructs an empirical null distribution of contrast scores by additionally assuming that uninteresting features' contrast scores follow a symmetric distribution around zero; it then computes approximate p-values of features based on the empirical null distribution, and finally it uses the BH procedure [4] to threshold the approximate p-values.

Definition 3.7 (The aBH procedure)**.** *Given contrast scores* $\{C_j\}_{j=1}^{d}$*, an empirical null distribution is defined on* $\mathcal{E} := \{C_j : C_j < 0; j = 1, \ldots, d\} \cup \{-C_j : C_j < 0; j = 1, \ldots, d\}$.

The aBH procedure defines the approximate p-value of feature j as

$$p_j := \frac{\sum_{c \in \mathcal{E}} \mathbb{1}(c \geqslant C_j)}{\mathrm{card}(\mathcal{E}) \vee 1}.$$

Then it applies the BH procedure with the target FDR threshold q to $\{p_j\}_{j=1}^d$ to call discoveries.

3.2.3 Enrichment analysis with any numbers of replicates m and n

When $m \neq n$, the BC procedure cannot guarantee FDR control because Lemma 6 no longer holds. To control the FDR in a more general setting ($m = n$ or $m \neq n$), Clipper constructs contrast scores via permutation of replicates across conditions. The idea is that, after permutation, every feature becomes uninteresting and can serve as its own negative control.

Definition 3.8 (Permutation)**.** *We define σ as permutation, i.e., a bijection from the set $\{1, \cdots, m+n\}$ onto itself, and we rewrite the data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_d$ into a matrix $\mathbf{W}$:*

$$\mathbf{W} = \begin{bmatrix} W_{11} & \cdots & W_{1m} & W_{1(m+1)} & \cdots & W_{1(m+n)} \\ & \vdots & & & \vdots & \\ W_{d1} & \cdots & W_{dm} & W_{d(m+1)} & \cdots & W_{d(m+n)} \end{bmatrix} := \begin{bmatrix} X_{11} & \cdots & X_{1m} & Y_{11} & \cdots & Y_{1n} \\ & \vdots & & & \vdots & \\ X_{d1} & \cdots & X_{dm} & Y_{d1} & \cdots & Y_{dn} \end{bmatrix}.$$

We then apply σ to permute the columns of $\mathbf{W}$ and obtain

$$\mathbf{W}_\sigma := \begin{bmatrix} W_{1\sigma(1)} & \cdots & W_{1\sigma(m)} & W_{1\sigma(m+1)} & \cdots & W_{1\sigma(m+n)} \\ & \vdots & & & \vdots & \\ W_{d\sigma(1)} & \cdots & W_{d\sigma(m)} & W_{d\sigma(m+1)} & \cdots & W_{d\sigma(m+n)} \end{bmatrix},$$

from which we obtain the permuted measurements $\left\{(\boldsymbol{X}_j^\sigma, \boldsymbol{Y}_j^\sigma)\right\}_{j=1}^d$, where

$$\begin{aligned} \boldsymbol{X}_j^\sigma &:= \left(W_{j\sigma(1)}, \ldots, W_{j\sigma(m)}\right)^\top, \\ \boldsymbol{Y}_j^\sigma &:= \left(W_{j\sigma(m+1)}, \ldots, W_{j\sigma(m+n)}\right)^\top. \end{aligned} \tag{3.13}$$

In the enrichment analysis, if two permutations σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\},$$

then we define σ and σ' to be in one equivalence class. That is, permutations in the same equivalence class lead to the same division of $m+n$ replicates (from the two conditions) into two groups with sizes m and n. In total, there are $\binom{m+n}{m}$ equivalence classes of permutations.

We define σ_0 as the identity permutation such that $\sigma_0(i) = i$ for all $i \in \{1, \cdots, m+n\}$. In addition, Clipper randomly samples h equivalence classes $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the other $h_{\max} := \binom{m+n}{m} - 1$ equivalence classes (after excluding the equivalence class containing σ_0). Note that $h_{\max}$ is the maximum value h can take.

Clipper then obtains $\left\{(\boldsymbol{X}_j^{\sigma_0}, \boldsymbol{Y}_j^{\sigma_0}), (\boldsymbol{X}_j^{\sigma_1}, \boldsymbol{Y}_j^{\sigma_1}), \cdots, (\boldsymbol{X}_j^{\sigma_h}, \boldsymbol{Y}_j^{\sigma_h})\right\}_{j=1}^{d}$, where $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ are the permuted measurements based on σ_ℓ, $\ell = 0, \ldots, h$. Then Clipper computes $T_j^{\sigma_\ell} := t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$ to indicate the degree of "interestingness" of feature j reflected by $(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})$. Note that Clipper chooses t^{diff} instead of $t^{\max}$ because empirical evidence shows that t^{diff} leads to better power. Sorting $\{T_j^{\sigma_\ell}\}_{\ell=0}^{h}$ gives

$$T_j^{(0)} \geqslant T_j^{(1)} \geqslant \cdots \geqslant T_j^{(h)}.$$

Then Clipper defines the contrast score of feature j, $j = 1, \ldots, d$, in one of two ways:

$$C_j := \begin{cases} T_j^{(0)} - T_j^{(1)} & \text{if } T_j^{(0)} = T_j^{\sigma_0} \\ T_j^{(1)} - T_j^{(0)} & \text{otherwise} \end{cases} \qquad \textbf{difference contrast score}, \tag{3.14}$$

or

$$C_j := \begin{cases} \left|T_j^{(0)}\right| & \text{if } T_j^{(0)} = T_j^{\sigma_0} > T_j^{(1)} \\ 0 & \text{if } T_j^{(0)} = T_j^{(1)} \\ -\left|T_j^{(0)}\right| & \text{otherwise} \end{cases} \qquad \textbf{maximum contrast score}. \tag{3.15}$$

The intuition behind the contrast scores is that, if $C_j < 0$, then $\mathbb{1}(T_j^{(0)} = T_j^{\sigma_0}) = 0$,

which means that at least one of $T_j^{\sigma_1}, \ldots, T_j^{\sigma_h}$ (after random permutation) is greater than $T_j^{\sigma_0}$ calculated from the original data (identity permutation), suggesting that feature j is likely an uninteresting feature in enrichment analysis. Motivated by Gimenez and Zou [110], we propose the following procedure for Clipper to control the FDR under the target level $q \in (0, 1)$.

Definition 3.9 (Gimenez-Zou (GZ) procedure for thresholding contrast scores [110])**.** *Given $h \in \{1, \cdots, h_{\max}\}$ and contrast scores $\{C_j\}_{j=1}^d$, $\mathcal{C} = \{|C_j| : C_j \neq 0\ ;\ j = 1, \ldots, d\}$ is defined as the set of non-zero absolute values of C_j's. The GZ procedure finds a contrast-score cutoff T^{GZ} based on the target FDR threshold $q \in (0, 1)$ as:*

$$T^{GZ} := \min \left\{ t \in \mathcal{C} : \frac{\frac{1}{h} + \frac{1}{h}\text{card}\left(\{j : C_j \leqslant -t\}\right)}{\text{card}\left(\{j : C_j \geqslant t\}\right) \vee 1} \leqslant q \right\} \tag{3.16}$$

and outputs $\left\{j : C_j \geqslant T^{GZ}\right\}$ as discoveries.

Theorem 4. *Suppose that the input data that satisfy the Clipper assumptions (3.1)–(3.2). Then for any $q \in (0, 1)$ and either definition of contrast scores in (3.14) or (3.15), the contrast-score cutoff T^{GZ}found by the GZ procedure (3.16) guarantees that the discoveries have the FDR under q:*

$$FDR = \mathbb{E}\left[\frac{\text{card}\left(\left\{j \in \mathcal{N} : C_j \geqslant T^{GZ}\right\}\right)}{\text{card}\left(\{j : C_j \geqslant T^{GZ}\}\right) \vee 1}\right] \leqslant q\,,$$

where $\mathcal{N}$ denotes the set of uninteresting features.

The proof of Theorem 4 (Supp. Section S3.5.7) is similar to that of Theorem 3 and requires two key ingredients: Lemma 7, which is also used in the proof of Theorem 3, and Lemma 8, which is similar to Lemma 6 and is about the properties of signs of $\{C_j\}_{j=1}^d$. The cutoff T^{GZ} can also be viewed as a stopping time of a Bernoulli process.

Lemma 8. *For input data that satisfy the Clipper assumptions (3.1) and (3.2), Clipper constructs contrast scores $\{C_j\}_{j=1}^d$ based on (3.15) or (3.14). Denote $S_j = sign\,(C_j) \in \{-1, 0, 1\}$. Then $\{S_j\}_{j=1}^d$ and $\{C_j\}_{j=1}^d$ satisfy the following properties:*

(a) $S_1, \ldots, S_d$ *are mutually independent ;*

(b) $\mathbb{P}(S_j = 1) \leqslant \frac{1}{h+1}$ *for all* $j \in \mathcal{N}$*;*

(c) $\{S_j\}_{j\in\mathcal{N}} \perp \mathcal{C}$*.*

We note that the GZ procedure is also applicable to the enrichment analysis with equal numbers of replicates, i.e., $m = n$ (Section 3.2.2). We will compare the GZ procedure against the BC procedure in our results.

3.2.4 Differential analysis with $m + n > 2$

For differential analysis, Clipper also uses permutation to construct contrast scores. When $m \neq n$, the equivalence classes of permutations are defined the same as for the enrichment analysis with $m \neq n$. When $m = n$, there is a slight change in the definition of equivalence classes of permutations: if σ and σ' satisfy that

$$\{\sigma(1), \cdots, \sigma(m)\} = \{\sigma'(1), \cdots, \sigma'(m)\} \text{ or } \{\sigma'(m+1), \cdots, \sigma'(2m)\},$$

then we say that σ and σ' are in one equivalence class. In total, there are $h_{\text{total}} := \binom{m+n}{m}$ (when $m \neq n$) or $\binom{2m}{m}/2$ (when $m = n$) equivalence classes of permutations. Hence, to have more than one equivalence class, we cannot perform differential analysis with $m = n = 1$; in other words, the total number of replicates $m + n$ must be at least 3.

Then Clipper randomly samples $\sigma_1, \ldots, \sigma_h$ with equal probabilities without replacement from the $h_{\max} := h_{\text{total}} - 1$ equivalence classes that exclude the class containing σ_0, i.e., the identity permutation. Note that $h_{\max}$ is the maximum value h can take. Next, Clipper computes $T_j^{\sigma_\ell} := \left|t^{\text{diff}}(\boldsymbol{X}_j^{\sigma_\ell}, \boldsymbol{Y}_j^{\sigma_\ell})\right|$, where $\boldsymbol{X}_j^{\sigma_\ell}$ and $\boldsymbol{Y}_j^{\sigma_\ell}$ are the permuted data defined in (3.13), and it defines C_j as the contrast score of feature j, $j = 1, \ldots, d$, in the same ways as in (3.14) or (3.15).

Same as in the enrichment analysis with $m \neq n$, Clipper also uses the GZ procedure [110] to set a cutoff on contrast scores to control the FDR under the target level $q \in (0, 1)$, following Theorem 4.

3.2.5 Clipper variant algorithms

For nomenclature, we assign the following names to Clipper variant algorithms, each of which combines a contrast score definition with a thresholding procedure.

- **Clipper-diff-BC**: difference contrast score $C_j = t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ (3.5) and BC procedure (Definition 3.6);
- **Clipper-diff-aBH**: difference contrast score $C_j = t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and aBH procedure (Definition 3.7);
- **Clipper-diff-GZ**: difference contrast score $\tau_j = T_j^{(0)} - T_j^{(1)}$ (3.14) and GZ procedure (Definition 3.9);
- **Clipper-max-BC**: maximum contrast score $C_j = t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ (3.6) and BC procedure;
- **Clipper-max-aBH**: maximum contrast score $C_j = t^{\max}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and aBH procedure;
- **Clipper-max-GZ**: maximum contrast score $\tau_j = T_j^{(0)}$ (3.15) and GZ procedure.

3.2.6 `R` package "Clipper"

In the `R` package `Clipper`, the default implementation is as follows. Based on the power comparison results in our manuscripts Ge et al. [18], Clipper uses Clipper-diff-BC as the default algorithm for the enrichment analysis with equal numbers of replicates; when there are no discoveries, Clipper suggests users to increase the target FDR threshold q or to use the Clipper-diff-aBH algorithm with the current q. For the enrichment analysis with different numbers of replicates under two conditions or the differential analysis, Clipper uses the Clipper-max-GZ algorithm by default.

3.3 Clipper has broad applications in omics data analyses

We then demonstrate the use of Clipper in four omics data applications: peptide identification from MS data, DEG identification from RNA-seq data, and DIR identification from Hi-C data. The first two applications are enrichment analyses, and the last two are differential analyses. In each application, we compared Clipper with mainstream bioinformatics methods to demonstrate Clipper's superiority in FDR control and detection power.

Peptide identification from MS data (enrichment analysis I)

The state-of-the-art proteomics studies use MS experiments and database search algorithms to identify and quantify proteins in biological samples. In a typical proteomics experiment, a protein mixture sample is first digested into peptides and then measured by tandem MS technology as mass spectra, which encode peptide sequence information. "Peptide identification" is the process that decodes mass spectra and converts mass spectra into peptide sequences in a protein sequence database via search algorithms. The search process matches each mass spectrum to peptide sequences in the database and outputs the best match, called a "peptide-spectrum match" (PSM). The identified PSMs are used to infer and quantify proteins in a high-throughput manner.

False PSMs could occur when mass spectra are matched to wrong peptide sequences due to issues such as low-quality spectra, data-processing errors, and incomplete protein databases, causing problems in the downstream protein identification and quantification [111]. Therefore, a common goal of database search algorithms is to simultaneously control the FDR and maximize the number of identified PSMs, so as to maximize the number of proteins identified in a proteomics study [78, 112, 113]. A widely used FDR control strategy is the target-decoy search, where mass spectra of interest are matched to peptide sequences in both the original (target) database and a decoy database that contains artificial false protein sequences. The resulting PSMs are called the target PSMs and decoy PSMs, respectively. The decoy PSMs, i.e., matched mass spectrum and decoy peptide pairs, are known to be false and thus used by database search algorithms to control the FDR. Mainstream database

search algorithms output a q-value for each PSM, target or decoy. Discoveries are the target PSMs whose q-values are no greater than the target FDR threshold q.

We used the first comprehensive benchmark dataset from an archaea species *Pyrococcus furiosus* to examine the FDR control and power of a popular database search algorithm Mascot [78] (Supp. Section S3.5.4). Using this benchmark dataset (Supp. Section S3.5.5), we demonstrate that, as an add-on, Clipper improves the power of Mascot. Specifically, Clipper treats mass spectra as features. For each mass spectrum, Clipper considers its measurement under the experimental condition as the $-\log_{10}$-transformed q-value of the target PSM that includes it, and its measurement under the background condition as the $-\log_{10}$-transformed q-value of the decoy PSM that includes it. Then Clipper decides which mass spectra and their corresponding target PSMs are discoveries (Supp. Section S3.5.6). Based on the benchmark dataset, we examined the empirical FDR, i.e., the FDP calculated based on the true positives and negatives, and the power of Mascot with or without Clipper as an add-on, for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3.2a shows that although Mascot and Mascot+Clipper both control the FDR, Mascot+Clipper consistently improves the power, thus enhancing the peptide identification efficiency of proteomics experiments.

While preparing this manuscript, we found a recent work [107] that used a similar idea to identify PSMs without using p-values. Clipper differs from this work in two aspects: (1) Clipper is directly applicable as an add-on to any existing database search algorithms that output q-values; (2) Clipper is not restricted to the peptide identification application.

DEG identification from RNA-seq data (differential analysis I)

RNA-seq data measure genome-wide gene expression levels in biological samples. An important use of RNA-seq data is the DEG analysis, which aims to discover genes whose expression levels change between two conditions. The FDR is a widely used criterion in DEG analysis [79–84].

We compared Clipper with two popular DEG identification methods DESeq2 [80] and edgeR [79] (Supp. Section S3.5.4). Specifically, we applied Clipper to two gene expression

matrices, one per condition, to identify DEGs (Supp. Section S3.5.6). To verify the FDR control of Clipper, DESeq2, and edgeR, we generated realistic semi-synthetic data from real RNA-seq data of classical and non-classical human monocytes [114], i.e., two conditions (Supp. Section S3.5.5). With ground truths (true DEGs and non-DEGs), the semi-synthetic data allow evaluating each method's FDR and power for a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3.2b shows that Clipper consistently controls the FDR and achieves high power. In contrast, edgeR controls the FDR but has lower power than Clipper does, while DESeq2 fails to control the FDR. To explain this DESeq2 result, we examined the p-value distributions of 16 non-DEGs that were most frequently identified from 200 synthetic datasets by DESeq2 at the target FDR threshold $q = 0.05$. Our results in Fig. S3.3 show that the 16 non-DEGs' p-values are non-uniformly distributed with a mode close to 0. Such unusual enrichment of overly small p-values makes these non-DEGs mistakenly called discoveries by DESeq2.

IHW is a popular procedure for boosting the power of p-value-based FDR control methods by incorporating feature covariates [15]. We used IHW as an add-on for DESeq2 and edgeR by adding every gene's mean expression across replicates and conditions as that gene's covariate, as suggested in [115]. The result in Supp. Fig. 3.2c shows that DESeq2+IHW and edgeR+IHW are similar to DESeq2 and edgeR, respectively, in FDR control and power.

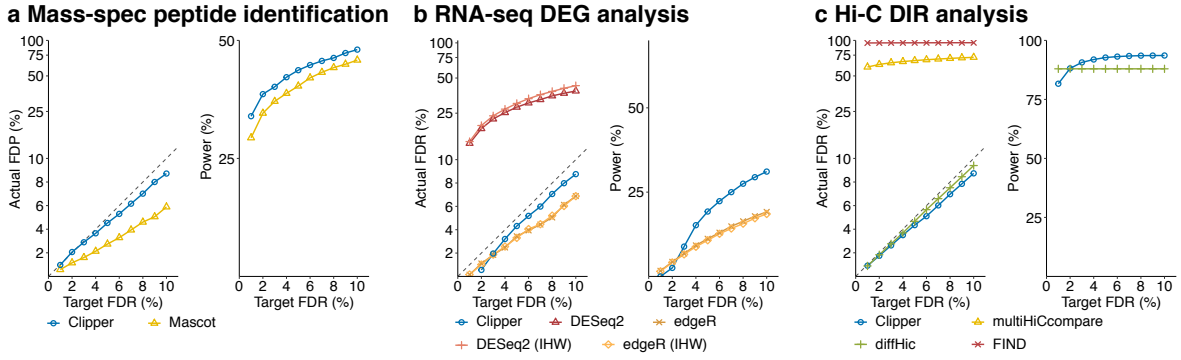

Figure 3.2: Comparison of Clipper and popular bioinformatics methods in terms of FDR control and power. (a) peptide identification on real proteomics data; (b) DEG analysis on synthetic bulk RNA-seq data; (c) DIR analysis on synthetic Hi-C data. In all four panels, the target FDR level q ranges from 1% to 10%. Points above the dashed line indicate failed FDR control; when this happens, the power of the corresponding methods is not shown, including HOMER in (a), MACS2 for target FDR less than 5% in (a), DESeq2 and DESeq2 (IHW) in (c), and multiHICcompare and FIND in (d). In all four applications, Clipper controls the FDR while maintaining high power, demonstrating Clipper's broad applicability in high-throughput data analyses.

DIR analysis of Hi-C data (differential analysis II)

Hi-C experiments are widely used to investigate spatial organizations of chromosomes and map chromatin interactions across the genome. A Hi-C dataset is often processed and summarized into an interaction matrix, whose rows and columns represent manually binned chromosomal regions and whose (i, j)-th entry represents the measured contact intensity between the i-th and j-th binned regions. The DIR analysis aims to identify pairs of genomic regions whose interaction intensities differ between conditions. Same as DEG analysis, DIR analysis also uses the FDR as a decision criterion [86–88].

We compared Clipper with three popular DIR identification methods: diffHic [88], FIND [87], and multiHiCcompare [86] (Supp. Section S3.5.4). Specifically, we applied Clipper to DIR identification by treating pairs of genomic regions as features and interaction intensities as measurements. To verify the FDR control of Clipper (Supp. Section S3.5.6), diffHiC, FIND, and multiHiCcompare, we generated realistic synthetic data from real interaction matrices of ENCODE cell line GM12878 [116] with true spiked-in DIRs to evaluate the FDR and power (Supp. Section S3.5.5). We examined the actual FDR and power in a range of target FDR thresholds: $q = 1\%, 2\%, \ldots, 10\%$. Fig. 3.2c shows that Clipper and diffHic are the only two methods that consistently control the FDR, while multiHiCcompare and FIND fail by a large margin. In terms of power, Clipper outperforms diffHic except for $q = 0.01$ and 0.02, even though Clipper has not been optimized for Hi-C data analysis. This result demonstrates Clipper's general applicability and strong potential for DIR analysis.

Discussion

In this paper, we proposed a new statistical framework, Clipper, for identifying interesting features with FDR control from high-throughput data. Clipper avoids the use of p-values and makes FDR control more reliable and flexible. We used comprehensive simulation studies to verify the FDR control by Clipper under various settings. We demonstrate that Clipper outperforms existing generic FDR control methods by having higher power and greater

robustness to model misspecification. We further applied Clipper to three popular bioinformatics analyses: peptide identification from MS data, DEG identification from RNA-seq data, and DIR identification from Hi-C data. Our results indicate that Clipper can provide a powerful add-on to existing bioinformatics tools to improve the reliability of FDR control and thus the reproducibility of scientific discoveries.

We validated the FDR control by Clipper using extensive and concrete simulations, including both model-based and real-data-based data generation with ground truths. In contrast, in most bioinformatics method papers, the FDR control was merely mentioned but rarely validated. Many of them assumed that using the BH procedure on p-values would lead to valid FDR control; however, the reality is often otherwise because p-values would be invalid when model assumptions were violated or the p-value calculation was problematic. Here we voice the importance of validating the FDR control in bioinformatics method development, and we use this work as a demonstration. We believe that Clipper provides a powerful booster to this movement. As a p-value-free alternative to the classic p-value-based BH procedure, Clipper relies less on model assumptions and is thus more robust to model misspecifications, making it an appealing choice for FDR control in diverse high-throughput biomedical data analyses.

Clipper is a flexible framework that is easily generalizable to identify a variety of interesting features. The core component of Clipper summarizes each feature's measurements under each condition into an informative statistic (e.g., the sample mean); then Clipper combines each feature's informative statistics under two conditions into a contrast score to enable FDR control. The current implementation of Clipper only uses the sample mean as the informative statistic to identify the interesting features that have distinct expected values under two conditions. However, by modifying the informative statistic, we can generalize Clipper to identify the features that are interesting in other aspects, e.g., having different variances between two conditions. Regarding the contrast score, Clipper makes careful choices between two contrast scores, difference and maximum, based on the number of replicates and the analysis task (enrichment vs. differential). Future studies are needed to explore other contrast scores and their power with respect to data characteristics and analysis tasks.

We have demonstrated the broad application potential of Clipper in various bioinformatics data analyses. Specifically, when used as an add-on to established, popular bioinformatics methods such as Mascot for peptide identification, Clipper guaranteed the desired FDR control and in some cases boosted the power. However, many more careful thoughts are needed to escalate Clipper into standalone bioinformatics methods for specific data analyses, for which data processing and characteristics (e.g., peak lengths, GC contents, proportions of zeros, and batch effects) must be appropriately accounted for before Clipper is used for the FDR control [117, 118]. We expect that the Clipper framework will propel future development of bioinformatics methods by providing a flexible p-value-free approach to control the FDR, thus improving the reliability of scientific discoveries.

3.4 Acknowledgments

This chapter is based on my joint work with Xinzhou Ge, Dongyuan Song, MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Dr. Wei Li, and my Ph.D. advisor Dr. Jingyi Jessica Li [18]. We would like to thank Dr. Yu-Cheng Yang for his suggestions on the figures and R package. The authors would also like to thank Mr. Nikos Ignatiadis, Dr. Lihua Lei, and Dr. Rina Barber for their insightful comments after we presented this work at the International Seminar on Selective Inference (https://www.selectiveinferenceseminar.com/past-talks). We also appreciate the comments and feedback from Mr. Tianyi Sun, Ms. Kexin Li, and other members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

3.5 Supplementary Material

S3.5.1 Review of generic FDR control methods

To facilitate our discussion, we introduce the notations for data. For feature $j = 1, \ldots, d$, we use $\boldsymbol{X}_j = (X_{j1}, \ldots, X_{jm})^\top \in \mathbb{R}^m$ and $\boldsymbol{Y}_j = (Y_{j1}, \ldots, Y_{jn})^\top \in \mathbb{R}^n$ to denote its measurements under the experimental and background conditions, respectively. We assume

that $X_{j1}, \ldots, X_{jm}$ are identically distributed, so are $Y_{j1}, \ldots, Y_{jn}$. Let $\mu_{Xj} = \mathbb{E}[X_{j1}]$ and $\mu_{Yj} = \mathbb{E}[Y_{j1}]$ denote the expected measurement of feature j under the two conditions, respectively. Then we denote by $\bar{X}_j$ the sample average of $X_{j1}, \cdots, X_{jm}$ and by $\bar{Y}_j$ the sample average of $Y_{j1}, \cdots, Y_{jn}$.

S3.5.2 P-value-based methods

Here we describe the details of p-value-based FDR control methods, including BH-pair, BH-pool, qvalue-pair, and qvalue-pool. Each of these four methods first computes p-values using either the pooled approach or the paired approach, and it then relies on the BH procedure [1] or Storey's qvalue procedure [2] for FDR control. In short, every p-value-based method is a combination of a p-value calculation approach and a p-value thresholding procedure. Below we introduce two p-value calculation approaches (paired and pooled) and two p-value thresholding procedures (BH and Storey's qvalue).

P-value calculation approaches

The paired approach. The paired approach examines one feature at a time and compares its measurements between two conditions. Besides the ideal implementation, i.e., the *correct paired approach* that uses the correct model to calculate p-values, we also include commonly-used flawed implementations that either misspecify the distribution, i.e., the *misspecified paired approach*, or misformulate the two-sample test as a one-sample test, i.e., the *2as1 paired approach*.

Here we use the negative binomial distribution as an example to demonstrate the ideas of the correct, misspecified, and 2as1 paried approaches. Suppose that for each feature j, its measurements under each condition follow a negative binomial distribution, and the two distributions under the two conditions have the same dispersion; that is, $X_{j1}, \cdots, X_{jm} \overset{\text{i.i.d.}}{\sim} \text{NB}(\mu_{Xj}, \theta_j)$; $Y_{j1}, \cdots, Y_{jn} \overset{\text{i.i.d.}}{\sim} \text{NB}(\mu_{Yj}, \theta_j)$, where θ_j is the dispersion parameter such that the variance $\text{Var}(X_{ji}) = \mu_{Xj} + \theta_j \mu_{Xj}^2$.

- The correct paired approach assumes that the two negative binomial distributions have

the same dispersion parameter θ_j, and it uses the two-sample test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \mu_{Yj}$ (differential analysis).

- The misspecified paired approach misspecifies the negative binomial distribution as Poisson, and it uses the two-sample test for the null hypothesis $H_0 : \mu_{Xj} = \mu_{Yj}$ against the alternative hypothesis $H_1 : \mu_{Xj} > \mu_{Yj}$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \mu_{Yj}$ (differential analysis).

- The 2as1 paired approach bluntly assumes $\mu_{Yj} = \bar{Y}_j$, and it performs the one-sample test based on $X_{j1}, \ldots, X_{jm}$ for the null hypotheses $H_0 : \mu_{Xj} = \bar{Y}_j$ against the alternative hypothesis $H_1 : \mu_{Xj} > \bar{Y}_j$ (enrichment analysis) or $H_1 : \mu_{Xj} \neq \bar{Y}_j$ (differential analysis).

The pooled approach. The pooled approach pools all features' average measurements under the background condition $\{\bar{Y}_j\}_{j=1}^d$ to form a null distribution, and it calculates a p-value for each feature j by comparing $\bar{X}_j$ to the null distribution. Specifically, in enrichment analysis, the p-value of feature j is computed as:

$$p_j = \frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \geqslant \bar{X}_j\right\}\right)}{d}.$$

In differential analysis, the p-value of feature j is computed as:

$$p_j = 2 \cdot \min\left(\frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \geqslant \bar{X}_j\right\}\right)}{d}, \frac{\operatorname{card}\left(\left\{k : \bar{Y}_k \leqslant \bar{X}_j\right\}\right)}{d}\right).$$

P-value thresholding procedures for FDR control

Definition S3.10 (BH procedure for thresholding p-values [1])**.** *The features' p-values $p_1, \ldots, p_d$ are sorted in an ascending order $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(d)}$. Given the target FDR threshold q, the Benjamini–Hochberg (BH) procedure finds a p-value cutoff T^{BH} as*

$$T^{BH} := p_{(k)}, \text{ where } k = \max\left\{j = 1, \ldots, d : p_{(j)} \leqslant \frac{j}{d}q\right\}. \tag{S3.17}$$

Then BH outputs $\left\{j : p_j \leqslant T^{BH}\right\}$ *as discoveries.*

Definition S3.11 (Storey's qvalue procedure for thresholding p-values [2])**.** *The features' p-values* $p_1, \ldots, p_d$ *are sorted in an ascending order* $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(d)}$*. Let* $\hat{\pi}_0$ *denote an estimate of the probability* P*(the i-th feature is uninteresting) (see Storey [2] for details). Storey's qvalue procedure defines the q-value for* $p_{(d)}$ *as*

$$\hat{q}(p_{(d)}) := \frac{\hat{\pi}_0 \cdot d \cdot p_{(d)}}{\text{card}\left(\left\{k : p_k \leqslant p_{(d)}\right\}\right)} = \hat{\pi}_0 \cdot p_{(d)}\,.$$

Then for $j = d-1, d-2, \ldots, 1$*, the q-value for* $p_{(j)}$ *is defined as:*

$$\hat{q}(p_{(j)}) := \min\left(\hat{q}(p_{(j+1)}),\ \frac{\hat{\pi}_0 \cdot d \cdot p_{(j)}}{\text{card}\left(\left\{k : p_k \leqslant p_{(j)}\right\}\right)}\right).$$

Then Storey's qvalue procedure outputs $\{j : \hat{q}(p_j) \leqslant q\}$ *as discoveries.*

We use function `qvalue` from `R` package `qvalue` (v 2.20.0; with default estimate $\hat{\pi}_0$) to calculate q-values.

Definition S3.12 (SeqStep+ procedure for thresholding p-values [3])**.** *Define* H_0^j *as the null hypothesis for feature* j *and* p_j *as the p-value for* H_0^j*,* $j = 1, \ldots, d$*. Order the null hypotheses* $H_0^1, \ldots, H_0^d$ *from the most to the least promising (here more promising means more likely to be interesting) and denote the resulting null hypotheses and p-values as* $H_0^{(1)}, \ldots, H_0^{(d)}$ *and* $p_{(1)}, \ldots, p_{(d)}$*. Given any target FDR threshold* q*, a pre-specified constant* $s \in (0,1)$*, and subset* $\mathcal{K} \subseteq \{1, \ldots, d\}$*, the SeqStep+ procedure finds a cutoff* $\hat{j}$ *as*

$$\hat{j} := \max\left\{j \in \mathcal{K} : \frac{1 + \text{card}\left(\left\{k \in \mathcal{K}, k \leqslant j : p_{(k)} > s\right\}\right)}{\text{card}\left(\left\{k \in \mathcal{K}, k \leqslant j : p_{(k)} \leqslant s\right\}\right) \vee 1} \leqslant \frac{1-s}{s} q\right\} \tag{S3.18}$$

Then SeqStep+ rejects $\left\{H_0^{(j)} : p_{(j)} \leqslant s,\ j \leqslant \hat{j},\ j \in \mathcal{K}\right\}$*. If the orders of the null hypotheses are independent of the p-values, the SeqStep+ procedure ensures FDR control.*

The GZ procedure (Definition 3.9) used in Clipper is a special case of the SeqStep+ procedure with $s = 1/(h+1)$. Recall that given the number of non-identical permutations $h \in$

$\{1, \cdots, h_{\max}\}$ and contrast scores $\{C_j\}_{j=1}^d$, the GZ procedure sorts $\{|C_j|\}_{j=1}^d$ in a decreasing order:

$$|C_{(1)}| \geqslant |C_{(2)}| \geqslant \cdots \geqslant |C_{(d)}| \,. \tag{S3.19}$$

To see the connection between the GZ procedure and SeqStep+, we consider the null hypothesis for the j-th ordered feature, $j = 1, \ldots, d$, as $H_0^{(j)} : \mu_{X(j)} = \mu_{Y(j)}$ and define the corresponding p-value $p_{(j)} := \frac{r(T_{(j)}^{\sigma_0})}{h+1}$, where $r(T_{(j)}^{\sigma_0})$ is the rank of $T_{(j)}^{\sigma_0}$ in $\{T_{(j)}^{\sigma_0}, \cdots, T_{(j)}^{\sigma_h}\}$ in a descending order. We also define $\mathcal{K} := \{j = 1, \ldots, d : C_j \neq 0\}$ as the subset of features with non-zero C_j's. Finally, we input the p-values, null hypothesis orders in (S3.19), $s = 1/(h+1)$, q and $\mathcal{K}$ into the SeqStep+ procedure, and we obtain the GZ procedure.

The BC procedure (Definition 3.6) is a further special case with $h = 1$, $p_{(j)} := \big(\mathbb{1}(C_{(j)} > 0) + 1\big)/2$, and $\mathcal{K} := \{j = 1, \ldots, d : C_j \neq 0\}$.

S3.5.3 Local-fdr-based methods

The FDR is statistical criterion that ensures the reliability of discoveries as a whole. In contrast, the local fdr focuses on the reliability of each discovery. The definition of the local fdr relies on some pre-computed summary statistics z_j for feature j, $j = 1, \ldots, d$. In the calculation of local fdr, $\{z_1, \ldots, z_d\}$ are assumed to be realizations of an abstract random variable Z that represents any feature. Let p_0 or p_1 denote the prior probability that any feature is uninteresting or interesting, with $p_0 + p_1 = 1$. Let $f_0(z) := \mathbb{P}(Z = z \,|\, \text{uninteresting})$ or $f_1(z) := \mathbb{P}(Z = z \,|\, \text{interesting})$ denote the conditional probability density of Z at z given that Z represents an uninteresting or interesting feature. Thus by Bayes' theorem, the posterior probability of any feature being uninteresting given its summary statistic $Z = z$ is

$$\mathbb{P}(\text{uninteresting} \mid Z = z) = p_0 f_0(z)/f(z) \,, \tag{S3.20}$$

where $f(z) := p_0 f_0(z) + p_1 f_1(z)$ is the marginal probability density of Z. Accordingly, the local fdr of feature j is defined as follows.

Definition S3.13 (Local fdr [4]). *Given notations defined above, the local fdr of feature j*

is defined as

$$\textit{local-fdr}_j := f_0(z_j)/f(z_j)\,.$$

Because $p_0 \leqslant 1$*,* $\textit{local-fdr}_j$ *is an upper bound of the posterior probability of feature* j *being uninteresting given its summary statistic* z_j*, defined in* (S3.20)*.*

Note that another definition of the local fdr is the posterior probability $\mathbb{P}(\text{uninteresting} \mid z)$ in (S3.20) [5]. Although this other definition is more reasonable, we do not use it but choose Definition S3.13 because the estimation of p_0 is ususally difficult. Another reason is that uninteresting features are the dominant majority in high-throughput biological data, so p_0 is often close to 1.

We define local-fdr-based methods as a type of FDR control methods by thresholding local fdrs of features under the target FDR threshold q. Although the local fdr is different from FDR, it has been shown that thresholding the local fdrs at q will approximately control the FDR under q [4]. This makes local-fdr-based methods competitors against Clipper and p-value-based methods.

Every local-fdr-based method is a combination of a local fdr calculation approach and a local fdr thresholding procedure. Below we introduce two local fdr calculation approaches (empirical null and swapping) and one local fdr thresholding procedure. After the combination, we have two local-fdr-based methods: locfdr-emp and locfdr-swap.

Local fdr calculation approaches

With $z_1, \ldots, z_d$, the calculation of local fdr defined in Definition S3.13 requires the estimation of f_0 and f, two probability densities. f is estimated by nonparametric density estimation, and f_0 is estimated by either the empirical null approach [4] or the swapping approach, which shuffles replicates between conditions [5]. With the estimated $\hat{f}$ and $\hat{f}_0$, the estimated local fdr of feature j is

$$\widehat{\text{local-fdr}}_j := \hat{f}_0(z_j)/\hat{f}(z_j)\,. \tag{S3.21}$$

The empirical null approach. This approach assumes a parametric distribution, typically

the Gaussian distribution, to estimate f_0. Then with the density estimate $\hat{f}$, the local fdr is estimated for each feature j. The implementation of this approach depends on the numbers of replicates.

- In 1vs1 enrichment and differential analyses, we define z_j as

$$z_j := \frac{D_j}{\sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(D_j - \bar{D}\right)^2}},$$

  where $D_j = X_{j1} - Y_{j1}$ and $\bar{D} = \sum_{j=1}^{d} D_j/d$.

- In 2vs1 enrichment and differential analyses, we define z_j as

$$z_j := \frac{\bar{X}_j - Y_{j1}}{\sqrt{\frac{s_{Xj}^2}{2}}},$$

  where $s_{Xj}^2 = \sum_{i=1}^{2}(X_{ji} - \bar{X}_j)^2$.

- In $m$vs$n$ enrichment and differential analyses with $m, n \geqslant 2$, we define z_j as the two-sample t-statistic with unequal variances:

$$z_j := \frac{\bar{X}_j - \bar{Y}_j}{\sqrt{\frac{s_{Xj}^2}{m} + \frac{s_{Yj}^2}{n}}},$$

  where $s_{Xj}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{ji} - \bar{X}_j)^2$ and $s_{Yj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ji} - \bar{Y}_j)^2$ are the sample variances of feature j under the experimental and background conditions.

Then $\{\widehat{\text{locfdr}}_j\}_{j=1}^{d}$ are estimated from $\{z_j\}_{j=1}^{d}$ by function `locfdr` in `R` package `locfdr` (v 1.1-8; with default arguments).

The swapping approach. This approach swaps $\lceil m/2 \rceil$ replicates under the experimental condition with $\lceil n/2 \rceil$ replicates under the background condition. Then it calculates the summary statistic for each feature on the swapped data, obtaining $z'_1, \ldots, z'_d$. Finally, it estimates f_0 and f by applying kernel density estimation to $z'_1, \ldots, z'_d$ and $z_1, \ldots, z_d$, respectively (by function `kde` in `R` package `ks`). With $\hat{f}_0$ and $\hat{f}$, $\{\widehat{\text{locfdr}}_j\}_{j=1}^{d}$ are calculated by

Definition S3.13.

The implementation of this approach depends on the numbers of replicates. Below are three special cases included in this work.

- In 1vs1 enrichment and differential analyses, the swapping approach is inapplicable because interesting features would not become uninteresting after the swapping.
- In 2vs1 enrichment and differential analyses, we define z_j and z'_j as

$$z_j = \frac{X_{j1} + X_{j2}}{2} - Y_{j1}\,,$$
$$z'_j = \frac{X_{j1} + Y_{j1}}{2} - X_{j2}\,.$$

- In 3vs3 enrichment and differential analyses with, we define z_j and z'_j as

$$z_j = \frac{X_{j1} + X_{j2}}{2} - \frac{Y_{j1} + Y_{j2}}{2}\,,$$
$$z'_j = \frac{X_{j1} + Y_{j1}}{2} - \frac{X_{j2} + Y_{j2}}{2}\,.$$

Then we apply kernel density estimation to $\{z_j\}_{j=1}^d$ and $\{z'_j\}_{j=1}^d$ to obtain $\hat{f}$ and $\hat{f}_0$, respectively. By (S3.21), we calculate $\{\widehat{\text{locfdr}}_j\}_{j=1}^d$.

The local fdr thresholding procedure

Definition S3.14 (locfdr procedure)**.** *Given the local fdr estimates* $\{\widehat{\textit{local-fdr}}_j\}_{j=1}^d$ *and the target FDR threshold* q*, the locfdr procedure outputs* $\{j = 1, \ldots, d : \widehat{\textit{local-fdr}}_j \leqslant q\}$ *as discoveries.*

S3.5.4 Bioinformatic methods with FDR control functionality

Mascot for peptide identification from MS data

Mascot uses probability-based scoring to identify PSMs from mass-spectrometry data. We ran Mascot in Proteome Discoverer 2.3.0.523 (ThermoScientific) with the following settings: 10 ppm precursor tolerance; 0.6 Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M). We ran Percolator [6] in conjunction with Mascot with the target/decoy selection mode set to "separate." For Mascot, for a range of target FDR thresholds ($q \in \{1\%, 2\%, \ldots, 10\%\}$), we identified the target PSMs with Mascot q-values no greater than q as discoveries. To prepare the input for Clipper, we set peptide and protein FDRs to 100% to obtain the entire lists of target PSMs and decoy PSMs with their Mascot q-values.

Differentially expressed gene (DEG) methods for bulk RNA-seq data

edgeR edgeR models each gene's read counts by using a negative binomial regression, where the condition is incorporated as an indicator covariate, and the condition's coefficient represents the gene-wise differential expression effect [7]. We used `R` package `edgeR` version 3.30.0.

DESeq2 DESeq2 uses a similar negative binomial regression as edgeR to model each gene's read counts under two conditions. DESeq2 differs from edgeR mainly in their estimation of the dispersion parameter in the negative binomial distribution [8]. We used the `R` package `DESeq2` version 1.28.1.

Differentially interacting chromatin regions (DIR) methods for Hi-C data

MultiHiCcompare MultiHiCcompare relies on a non-parametric method to jointly normalize multiple Hi-C interaction matrices [9]. It uses a generalized linear model to detect DIRs. MultiHiCcompare is an extension of the `HiCcompare` package [10]. We used `R` package

`multiHiCcompare` version 1.6.0.

diffHic diffHic uses the statistical framework of the `edgeR` package to model biological variability and to test for significant differences between conditions [11]. We used `R` package `diffHic` version 1.20.0.

FIND FIND uses a spatial Poisson process to detect chromosomal regions that display a significant difference between two regions' interaction intensity and their neighbouring interaction intensities [12]. We used `R` package `FIND` version 0.99.

S3.5.5 Benchmark data generation in omics data applications

Real MS standard data

The data generation information will be published in a future manuscript. Interested readers should contact Dr. Leo Wang at lewang@coh.org.

Bulk RNA-seq data with synthetic spike-in DEGs

We used the human monocyte RNA-Seq dataset including 17 samples of classical monocytes and 17 samples of nonclassical monocytes [13]. Each sample contains expression levels of $d = 52{,}376$ genes.

(i) We first performed normalization on all 34 samples using the edgeR normalization method Trimmed Mean of M-values (TMM) [14]. We denote the resulting normalized read count matrix of classical and non-classical monocytes by $\mathbf{X}^{\text{cl}}$ and $\mathbf{X}^{\text{ncl}}$, respectively. Following the convention in bioinformatics, the columns and rows of $\mathbf{X}^{\text{cl}}$ and $\mathbf{X}^{\text{ncl}}$ represent biological samples and genes, respectively.

(ii) To define true DEGs, we first computed the log fold change of gene j by $\text{logfc}_j = \log_2\left[(\bar{\mathbf{X}}_j^{\text{ncl}} + 1)/(\bar{\mathbf{X}}_j^{\text{cl}} + 1)\right]$ for $j = 1, \ldots, d$, where $\mathbf{X}_i^{\text{cl}}$ and $\mathbf{X}_i^{\text{ncl}}$ denote the i-th row vector of $\mathbf{X}^{\text{cl}}$ and $\mathbf{X}^{\text{ncl}}$ respectively and $\bar{\cdot}$ denotes the average of elements in a vector.

We added the pseudo-count of 1 to avoid division by 0. We defined true DEGs as those with $\text{logfc}_j \geqslant 4$ or $\text{logfc}_j \leqslant -4$.

(iii) We generated synthetic data with 3 samples under both the experimental and background conditions, a typical design in bulk RNA-seq experiments. Specifically, if gene j is a DEG, we randomly sampled without replacement 3 values from $\mathbf{X}_j^{\text{cl}}$ as counts under the experimental condition, and another 3 values from $\mathbf{X}_j^{\text{ncl}}$ as counts under the background condition. If gene j is not a DEG, we randomly sampled 6 values without replacement from $(\mathbf{X}_j^{\text{cl}}, \mathbf{X}_j^{\text{ncl}})$ and randomly split them into 3 and 3 counts under two conditions.

(iv) We repeated Step (iii) for 200 times to generate 200 synthetic datasets.

The human monocyte RNA-Seq data set is available in the NCBI Sequence Read Archive (SRA) under accession number SRP082682 (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=srp082

Hi-C data with synthetic spike-in DIRs

The real Hi-C interaction matrix contains the pairwise interaction intensities of 250 binned genomic regions in Chromosome 1. It is from the cell line GM12878 and available in the NCBI Gene Expression Omnibus(GEO) under accession number GSE63525. We denote the real interaction matrix as $\mathbf{X}^{\text{real}}$. Because $\mathbf{X}^{\text{real}}$ is symmetric, we only focus on its upper triangular part.

(i) Among the $(250 \times 250 - 250)/2 = 31{,}125$ upper triangular entries (i.e., region pairs), we selected 404 entries as true up-regulated DIRs, and 550 entries as true down-regulated DIRs (Fig. S3.5).

(ii) Next, for the (i, j)-th entry, we generated a log fold change, denoted by f_{ij}, between the two conditions as follows. We simulated f_{ij} from truncated Normal$(100/|i-j|, 0.5^2)$ with support $[0.05, \infty)$ if the (i, j)-th entry is up-regulated, or from truncated Normal$(-100/|i-j|, 0.5^2)$ with support $(-\infty, -0.05]$ if the (i, j)-th entry is down-regulated; if the (i, j)-th entry is not differential, we set $f_{ij} = 0$.

(iii) Then we specify the mean measurement of the (i, j)-th entry under the two conditions as $\mu_{Xij} = [\mathbf{X}^{\text{real}}]_{ij}$ and $\mu_{Yij} = [\mathbf{X}^{\text{real}}]_{ij} \cdot e^{f_{ij}}$, respectively.

(iv) We generated synthetic read counts of the (i, j)-th entry from $\text{NB}(\mu_{Xij}, 1000^{-1})$ and $\text{NB}(\mu_{Yij}, 1000^{-1})$ respectively under the two conditions.

(v) We repeated Step (iv) for 200 times to generate 200 synthetic datasets.

S3.5.6 Implementation of Clipper in omics data applications

Below we briefly introduce the implementation of Clipper in the four omics data applications. All the results were obtained by running using `R` package `Clipper` (see package vignette for details: https://github.com/JSB-UCLA/Clipper/blob/master/vignettes/Clipper.pdf).

Peptide identification from mass spectrometry data

(i) We consider each mass spectrum as a feature and its target/decoy PSM as a replicate under the experimental/background condition respectively. Then we consider $-\log_{10}(\text{q-value} + 0.01)$ as the measurement of each PSM, where the q-value is output by Mascot. Doing so, we summarized the Mascot output into a $d \times (m + n)$ matrix, where d is the number of mass spectra, and m and n are the numbers of experimental and control samples, respectively. We then applied Clipper to perform an enrichment analysis to obtain a contrast score C_j for each mass spectrum j. If the mass spectrum has no decoy or background measurement, we set $C_j = 0$. In our study, $m = n = 1$, so the default Clipper implementation is Clipper-diff-BC.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The target PSMs whose mass spectra have contrast scores greater than or equal to T_q are called discoveries.

DEG identification from bulk RNA-seq data

(i) We consider each gene as a feature and the class label—classical and non-classical human monocytes—as the two conditions. Then we consider $\log_2$-transformed read counts with a pseudocount 1 as measurements. Doing so, we summarized the gene expression matrix into a $d \times (m + n)$ matrix, where d is the number of genes, and m and n are the numbers of samples under the two conditions, respectively. We then applied Clipper to perform a differential analysis to obtain a contrast score C_j for each gene. In our study, $m = n = 3$, so the default Clipper implementation is Clipper-max-GZ with $h = 1$.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The genes with contrast scores greater than or equal to T_q are called discoveries.

DIR identification from Hi-C data

(i) We consider each pair of genomic regions as a feature and manually created two conditions. Then we consider log-transformed read counts as measurements. Doing so, we summarized the gene expression matrix into a $d \times (m + n)$ matrix, where d is the total pairs of genomic regions, and m and n are the numbers of samples under the two conditions, respectively. We then applied Clipper to perform a differential analysis to obtain a contrast score C_j for each pair of genomic regions. In our study, $m = n = 2$, so the default Clipper implementation is Clipper-max-GZ with $h = 1$.

(ii) For any target FDR threshold q, Clipper gives a cutoff T_q on contrast scores.

(iii) The pairs of genomic regions with contrast scores greater than or equal to T_q are called discoveries.

S3.5.7 Proofs

Proof of Theorem 1

We first prove Theorem 3, which relies on Lemmas 6 and 7. Here we only include the proof of Lemma 6 and defer the proof of Lemma 7 to Section S3.5.7.

Proof of Lemma 6. Here we prove that Lemma 6 holds when C_j is constructed using (3.5); the proof is similar when C_j is constructed using (3.6).

When input data satisfy (3.1) and (3.2) and $m = n$, properties (a) and (b) can be derived directly. To prove property (c), it suffices to prove that for any $j \in \mathcal{N}$ with $C_j \neq 0$, S_j is independent of $|C_j|$.

Note that $\bar{X}_j$ and $\bar{Y}_j$ are i.i.d for $j \in \mathcal{N}$ when $m = n$. Hence for any measurable set $\mathcal{A} \subset [0, +\infty)$,

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) = \mathbb{P}\left(t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) \in \mathcal{A}\right) = \mathbb{P}\left(t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j) \in \mathcal{A}\right) = \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A}\right).$$

The first equality holds because $t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j) = C_j = |C_j|$ when $S_j = 1$. The second equality holds because $t^{\text{diff}}(\boldsymbol{X}_j, \boldsymbol{Y}_j)$ and $t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j)$ are identically distributed when $j \in \mathcal{N}$. The third equality holds because $t^{\text{diff}}(\boldsymbol{Y}_j, \boldsymbol{X}_j) = -C_j$; if $-C_j \in \mathcal{A}$, then $S_j = -1$.

Because $\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) + \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A}\right) = \mathbb{P}\left(|C_j| \in \mathcal{A}\right)$, it follows that

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A}\right) = \frac{1}{2}\mathbb{P}\left(|C_j| \in \mathcal{A}\right) = \mathbb{P}(S_j = 1)\mathbb{P}\left(|C_j| \in \mathcal{A}\right),$$

where the last equality holds because $\mathbb{P}(S_j = 1) = 1/2$ by property (b).

Hence, S_j and $|C_j|$ are independent $\forall j \in \mathcal{N}$. □

Proof of Theorem 3. Define a random subset of $\mathcal{N}$ as $\mathcal{M} := \mathcal{N} \backslash \{j \in \mathcal{N} : C_j = 0\} = \{j \in \mathcal{N} : S_j \neq 0\}$.

First note that by Lemma 6(b), $\mathbb{P}(S_j = -1) = \mathbb{P}(C_j < 0) = 1/2$ for all $j \in \mathcal{M} \subset \mathcal{N}$. Assume without loss of generality that $\mathcal{M} = \{1, \ldots, d'\}$. We order $\{|C_j| : j \in \mathcal{M}\}$,

from the largest to the smallest, denoted by $|C_{(1)}| \geqslant |C_{(2)}| \geqslant \cdots \geqslant |C_{(d')}|$. Let $J = \sum_{j\in\mathcal{N}} \mathbb{1}\left(|C_j| \geqslant T^{\mathrm{BC}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{BC}. When $J > 0$, $|C_{(1)}| \geqslant \cdots \geqslant |C_{(J)}| \geqslant T^{\mathrm{BC}}$. Define $Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, d'$. Then for each order k, the following holds

$$
\begin{aligned}
C_{(k)} \geqslant T^{\mathrm{BC}} &\Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{BC}} \text{ and } C_{(k)} > 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 0\,; \\
C_{(k)} \leqslant -T^{\mathrm{BC}} &\Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{BC}} \text{ and } C_{(k)} < 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 1\,.
\end{aligned}
$$

Then

$$
\begin{aligned}
\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} &= \frac{\sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{BC}}\right)} \\
&= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{BC}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{BC}}\right)} \\
&= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \\
&= \frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\,.
\end{aligned}
$$

Because $\{S_j\}_{j\in\mathcal{N}}$ is independent of $\mathcal{C}$ (Lemma 6(c)), Lemma 6(a)-(b) still holds after $C_1, \ldots, C_{d'}$ are reordered as $C_{(1)}, \ldots, C_{(d')}$. Thus $Z_1, \ldots, Z_{d'}$ are i.i.d. from Bernoulli(1/2). To summarize, it holds that

$$
\{Z_j\}_{j\in\mathcal{M}} \;\Big|\; \mathcal{M} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(1/2)\,.
$$

Then by applying Lemma 7 and making $\rho = 0.5$, we have:

$$
\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} \;\Big|\; \mathcal{M}\right] \leqslant 1 \tag{S3.22}
$$

Then

$$
\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{BC}}\}) \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} \cdot \frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{BC}}\}) \vee 1}\right] \\
&\leqslant \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} \cdot \frac{\mathrm{card}(\{j : C_j \leqslant -T^{\mathrm{BC}}\}) + 1}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{BC}}\}) \vee 1}\right] \\
&\leqslant q \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1}\right] \\
&\leqslant q \cdot \mathbb{E}\left[\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{BC}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{BC}}\}) + 1} \;\middle|\; \mathcal{M}\right]\right] \\
&\leqslant q,
\end{aligned}
$$

where $\mathcal{M}$ is random subset of $\mathcal{N}$ such that for each $j \in \mathcal{M}$, $|C_j| > 0$. The last inequality follows from (S3.22).

□

Proof of Theorem 2

We then prove Theorem 4, which relies on Lemmas 7 and 8. Here we introduce the proof of Lemma 8 and defer the proof of Lemma 7 to Section S3.5.7.

Proof of Lemma 8. With input data satisfying (3.1) and (3.2), C_j constructed from (3.14) or (3.15), property (a) can be derived directly.

To show property (b), note that for each uninteresting feature $j \in \mathcal{N}$, $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ are from the same distribution; thus $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ are identically distributed. Define an event $\mathcal{E}_j := \left\{\sum_{\ell=0}^h \mathbb{1}(T_j^{\sigma_\ell} = T_j^{(0)}) = 1\right\}$, which indicates that $T_j^{(0)}$, the maximizer of $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$, is unique. Then conditional on $\mathcal{E}_j$, the maximizer is equally likely to be any of $\{0, \ldots, h\}$, and it follows that $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j) = \mathbb{P}(T_j^{\sigma_0} = T_j^{(0)} \mid \mathcal{E}_j) = 1/(h+1)$. Conditioning on that $\mathcal{E}_j$ does not happen, $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j^c) = 0$. Thus $\mathbb{P}(S_j = 1) = \mathbb{P}(S_j = 1 \mid \mathcal{E}_j)\mathbb{P}(\mathcal{E}_j) + \mathbb{P}(S_j =$

$1 \mid \mathcal{E}_j^c)\mathbb{P}(\mathcal{E}_j^c) \leqslant 1/(h+1)$.

The proof of property (c) is similar to the Proof of Lemma 6(c). It suffices to show that for any $j \in \mathcal{N}$ with $C_j \neq 0$ (that is, $\mathcal{E}_j$ occurs), S_j is independent of $|C_j|$. As $\boldsymbol{X}_j$ and $\boldsymbol{Y}_j$ are from the same distribution, $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ are identically distributed. Hence for any measurable set $\mathcal{A} \subset [0, +\infty)$,

$$\begin{aligned}\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) &= \mathbb{P}\left(T_j^{\sigma_0} = T_j^{(0)}, |C_j| \in \mathcal{A} \;\middle|\; \mathcal{E}_j\right) \\ &= \frac{1}{h}\mathbb{P}\left(T_j^{\sigma_0} \neq T_j^{(0)}, |C_j| \in \mathcal{A} \;\middle|\; \mathcal{E}_j\right) \\ &= \frac{1}{h}\mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right).\end{aligned}$$

The first equality holds because $T_j^{\sigma_0} = T_j^{(0)}$ when $S_j = 1$. The second equality holds because $\{T_j^{\sigma_\ell}\}_{\ell=0}^h$ are identically distributed when $j \in \mathcal{N}$. The third equality holds because $T_j^{\sigma_0} \neq T_j^{(0)}$ when $S_j = -1$.

Because $\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) + \mathbb{P}\left(S_j = -1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right)$, it follows that

$$\mathbb{P}\left(S_j = 1, |C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \frac{1}{h+1}\mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right) = \mathbb{P}(S_j = 1 \mid \mathcal{E}_j)\mathbb{P}\left(|C_j| \in \mathcal{A} \mid \mathcal{E}_j\right),$$

where the last equality holds because $\mathbb{P}(S_j = 1 \mid \mathcal{E}_j) = 1/(h+1)$.

Hence, S_j and $|C_j|$ are independent $\forall j \in \mathcal{N}$ with $C_j \neq 0$. □

Proof of Theorem 4. Define a random subset of $\mathcal{N}$ as $\mathcal{M} := \mathcal{N} \backslash \{j \in \mathcal{N} : C_j = 0\} = \{j \in \mathcal{N} : S_j \neq 0\}$. Assume without loss of generality that $\mathcal{M} = \{1, \ldots, d'\}$. We order $\{|C_j| : j \in \mathcal{M}\}$, from the largest to the smallest, denoted by $|C_{(1)}| \geqslant |C_{(2)}| \geqslant \cdots \geqslant |C_{(d')}|$. Let $J = \sum_{j \in \mathcal{N}} \mathbb{1}\left(|C_j| \geqslant T^{\mathrm{GZ}}\right)$, the number of uninteresting features whose contrast scores have absolute values no less than T^{GZ}. When $J > 0$, $|C_{(1)}| \geqslant \cdots \geqslant |C_{(J)}| \geqslant T^{\mathrm{GZ}}$. Define

$Z_k = \mathbb{1}\left(C_{(k)} < 0\right)$, $k = 1, \ldots, d'$. Then for each order k, the following holds:

$$C_{(k)} \geqslant T^{\mathrm{GZ}} \Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{GZ}} \text{ and } C_{(k)} > 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 0\,;$$
$$C_{(k)} \leqslant -T^{\mathrm{GZ}} \Longleftrightarrow \left|C_{(k)}\right| \geqslant T^{\mathrm{GZ}} \text{ and } C_{(k)} < 0 \Longleftrightarrow k \leqslant J \text{ and } Z_k = 1\,.$$

Then it follows that

$$\begin{aligned}
\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1} &= \frac{\sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{GZ}}\right)}{1 + \sum_{k=1}^{d'} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{GZ}}\right)} \\
&= \frac{\sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \geqslant T^{\mathrm{GZ}}\right)}{1 + \sum_{k=1}^{J} \mathbb{1}\left(C_{(k)} \leqslant -T^{\mathrm{GZ}}\right)} \\
&= \frac{(1 - Z_1) + \cdots + (1 - Z_J)}{1 + Z_1 + \cdots + Z_J} \\
&= \frac{1 + J}{1 + Z_1 + \cdots + Z_J} - 1\,.
\end{aligned}$$

Because $\{S_j\}_{j \in \mathcal{N}}$ is independent of $\mathcal{C}$ (Lemma 6(c)), Lemma 6(a)-(b) still holds after $C_1, \ldots, C_{d'}$ are reordered as $C_{(1)}, \ldots, C_{(d')}$. Thus $Z_1, \ldots, Z_{d'}$ are i.i.d. from Bernoulli(ρ_k). To summarize, it holds that

$$\{Z_j\}_{j \in \mathcal{M}} \;\Big|\; \mathcal{M} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\rho_k)\,.$$

Then by applying Lemma 7 and making $\rho = h/(h+1)$, we have:

$$\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1}\right] \leqslant 1/h\,. \tag{S3.23}$$

Then

$$
\begin{aligned}
\mathrm{FDR} &= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&= \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1} \cdot \frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&\leqslant h \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1} \cdot \frac{\frac{1}{h}\mathrm{card}(\{j : C_j \leqslant -T^{\mathrm{GZ}}\}) + \frac{1}{h}}{\mathrm{card}(\{j : C_j \geqslant T^{\mathrm{GZ}}\}) \vee 1}\right] \\
&\leqslant hq \cdot \mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{N} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{N} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1}\right] \\
&\leqslant hq \cdot \mathbb{E}\left[\mathbb{E}\left[\frac{\mathrm{card}(\{j \in \mathcal{M} : C_j \geqslant T^{\mathrm{GZ}}\})}{\mathrm{card}(\{j \in \mathcal{M} : C_j \leqslant -T^{\mathrm{GZ}}\}) + 1} \;\middle|\; \mathcal{M}\right]\right] \\
&\leqslant q\,,
\end{aligned}
$$

where the second inequality follows from the definition of T_{GZ} (3.16) and the last inequality follows from (S3.23).

□

Proof of Lemma 2

Finally, we derive Lemma 7 by following the same proof same as in [15], which relies on Lemma 9 and Corollary 1.

Lemma 9. *Suppose that* $Z_1, \ldots, Z_d \overset{i.i.d.}{\sim}$ *Bernoulli*(ρ). *Let* J *be a stopping time in reverse time with respect to the filtration* $\{\mathcal{F}_j\}$, *where* $\mathcal{F}_j = \sigma\left(\{(Z_1 + \cdots + Z_j),\, Z_{j+1},\, \cdots,\, Z_d\}\right)$ *with* $\sigma(\cdot)$ *denoting a* σ*-algebra, and the variables* $Z_1, \ldots, Z_j$ *are exchangeable with respect to* $\{\mathcal{F}_j\}$. *Then*

$$
\mathbb{E}\left[\frac{1+J}{1+Z_1+\cdots+Z_J}\right] \leqslant \rho^{-1}.
$$

Proof of Lemma 9. Define

$$
Y_j = Z_1 + \cdots + Z_j \in \mathcal{F}_j
$$

and define the process

$$M_j = \frac{1+j}{1+Z_1+\cdots+Z_j} = \frac{1+j}{1+Y_j} \in \mathcal{F}_j.$$

In [3], it is shown that $\mathbb{E}[M_d] \leqslant \rho^{-1}$. Therefore, by the optional stopping time theorem it suffices to show that $\{M_j\}$ is a supermartingale with respect to $\{\mathcal{F}_j\}$. As $\{Z_1, \ldots, Z_{j+1}\}$ are exchangeable with respect to $\mathcal{F}_{j+1}$, we have

$$\mathbb{P}\left(Z_{j+1} = 1 \mid \mathcal{F}_{j+1}\right) = \frac{Y_j + 1}{1+j}.$$

Therefore, if $Y_{j+1} > 0$,

$$\begin{aligned}
\mathbb{E}\left[M_j \mid \mathcal{F}_{j+1}\right] &= \frac{1+j}{1+Y_{j+1}} \cdot \mathbb{P}\left(Z_{j+1} = 0 \mid \mathcal{F}_{j+1}\right) + \frac{1+j}{1+Y_{j+1}-1} \cdot \mathbb{P}\left(Z_{j+1} = 1 \mid \mathcal{F}_{j+1}\right) \\
&= \frac{1+j}{1+Y_{j+1}} \cdot \frac{1+j-Y_{j+1}}{1+j} + \frac{1+j}{1+Y_{j+1}-1} \cdot \frac{Y_{j+1}}{1+j} \\
&= \frac{1+j-Y_{j+1}}{1+Y_{j+1}} + 1 \\
&= \frac{1+(j+1)}{1+Y_{j+1}} \\
&= M_{j+1}.
\end{aligned}$$

If instead $Y_{j+1} = 0$, then trivially $Y_j = 0$, and $M_j = 1 + j < 2 + j = M_{j+1}$. This proves that $\{M_j\}$ is a supermartingale with respect to $\{\mathcal{F}_j\}$ as desired.

□

Corollary 1. *Suppose that $\mathcal{A} \subseteq \{1, \ldots, d\}$ is fixed, while $Z_1, \ldots, Z_d \overset{i.i.d.}{\sim}$ Bernoulli(ρ). Let J be a stopping time in reverse time with respect to the filtration $\{\mathcal{F}_j\}$, where $\mathcal{F}_j = \sigma\left(\{\sum_{k \leqslant j, k \in \mathcal{A}} Z_k\} \cup \{Z_k : j < k < d, k \in \mathcal{A}\}\right)$ with $\sigma(\cdot)$ denoting a σ-algebra, and the variables $\{Z_k : k \leqslant j, k \in \mathcal{A}\}$ are exchangeable with respect to $\mathcal{F}_j$. Then*

$$\mathbb{E}\left[\frac{1 + \operatorname{card}\left(\{k : k \leqslant J, k \in \mathcal{A}\}\right)}{1 + \sum_{k \leqslant J, k \in \mathcal{A}} Z_k}\right] \leqslant \rho^{-1}.$$

Proof of Corollary 1. Let $\mathcal{A} = \{j_1, \ldots, j_m\}$ where $1 \leqslant j_1 < \cdots < j_m \leqslant d$. Then by considering the i.i.d. sequence

$$Z_{j_1}, \ldots, Z_{j_m}$$

in place of $Z_1, \ldots, Z_d$, we see that this result is equivalent to Lemma 9. □

Proof of Lemma 7. [From [3]] We may assume $\rho < 1$ to avoid the trivial case. We first introduce a different definition for $\{Z_j\}_{j=1}^d$ by defining a random set $\mathcal{A} \subseteq \{1, \ldots, d\}$ where for each j, independently,

$$\mathbb{P}(j \in \mathcal{A}) = \frac{1-\rho_j}{1-\rho}.$$

We then define random variables $Q_1, \ldots, Q_d \overset{i.i.d.}{\sim} \text{Bernoulli}(\rho)$, which are generated independently of the random set $\mathcal{A}$. Finally, we define

$$Z_j = Q_j \cdot \mathbb{1}(j \in \mathcal{A}) + \mathbb{1}(j \notin \mathcal{A}). \tag{S3.24}$$

Then $\{Z_j\}_{k=1}^d$ are mutually independent and $\mathbb{P}(Z_j = 1) = 1 - \mathbb{P}(j \in \mathcal{A}) \cdot \mathbb{P}(Q_j = 0) = \rho_j$, that is, $Z_j \sim \text{Bernoulli}(\rho_j)$. This new definition of $\{Z_j\}_{j=1}^d$ meet all the conditions required by Lemma 7, so that we can apply this new definition in the following proof.

As $Z_j = Q_j \cdot \mathbb{1}(j \in \mathcal{A}) + \mathbb{1}(j \notin \mathcal{A})$ for all j, we have

$$\frac{1+J}{1+Z_1+\cdots+Z_J} = \frac{1+\text{card}(\{j \leqslant J : j \in \mathcal{A}\}) + \text{card}(\{j \leqslant J : j \notin \mathcal{A}\})}{1+\sum_{j \leqslant J, j \in \mathcal{A}} Q_j + \text{card}(\{j \leqslant J : j \notin \mathcal{A}\})} \leqslant \frac{1+\text{card}(\{j \leqslant J : j \in \mathcal{A}\})}{1+\sum_{j \leqslant J, j \in \mathcal{A}} Q_j}, \tag{S3.25}$$

where the last step uses the identify $\frac{a+c}{b+c} \leqslant \frac{a}{b}$ whenever $0 < b \leqslant a$ and $c \geqslant 0$. Therefore, it will be sufficient to prove that

$$\mathbb{E}\left[\frac{1+\text{card}(\{j \leqslant J : j \in \mathcal{A}\})}{1+\sum_{j \leqslant J, j \in \mathcal{A}} Q_j} \,\middle|\, \mathcal{A}\right] \leqslant \rho^{-1}, \tag{S3.26}$$

To prove (S3.26), first let $\widetilde{Q}_j = Q_j \cdot \mathbb{1}(j \in \mathcal{A})$, and define a filtration $\{\mathcal{F}'_j\}$ where $\mathcal{F}'_j$ is the

σ-algebra generated as

$$\mathcal{F}'_j = \sigma\left(\left\{\widetilde{Q}_1 + \cdots + \widetilde{Q}_j, \widetilde{Q}_{j+1}, \cdots, \widetilde{Q}_d, \mathcal{A}\right\}\right).$$

Next for any j, by (S3.24) we see that

$$Z_1 + \cdots + Z_j, Z_{j+1}, \ldots, Z_d \in \mathcal{F}'_j \Rightarrow \mathcal{F}_j \subseteq \mathcal{F}'_j,$$

so J is a stopping time (in reverse time) with respect to $\mathcal{F}'_j$. Finally, since the Q_j's are independent of $\mathcal{A}$, (S3.26) follows from Corollary 1 after conditioning on $\mathcal{A}$. □

S3.5.8 Supplementary figures

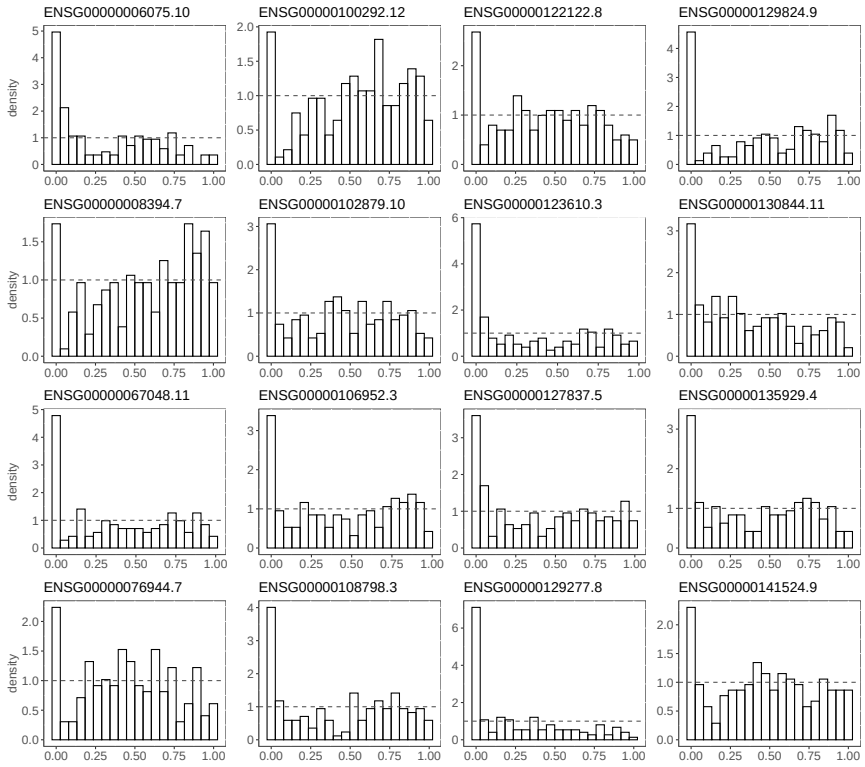

Figure S3.3: The p-value distributions of 16 non-DEGs that are most frequently identified by DESeq2 at $q = 5\%$ from 200 synthetic datasets. The p-values of these 16 genes tend to be overly small, and their distributions are non-uniform with a mode close to 0.

a

GO terms enriched in Clipper−specific DEGs in Clipper vs. DESeq2 comparison

GO term (ID)	qvalue (Clipper)
neutrophil activation (GO:0042119)	3.104557e−10
granulocyte activation (GO:0036230)	3.104557e−10
neutrophil degranulation (GO:0043312)	8.587750e−10
neutrophil activation involved in immune response (GO:0002283)	8.591455e−10
neutrophil mediated immunity (GO:0002446)	3.104557e−10

b

GO terms enriched in Clipper−specific DEGs in Clipper vs. edgeR comparison

GO term (ID)	qvalue (Clipper)
neutrophil degranulation (GO:0043312)	8.587750e−10
neutrophil activation involved in immune response (GO:0002283)	8.591455e−10
neutrophil activation (GO:0042119)	3.104557e−10
neutrophil mediated immunity (GO:0002446)	3.104557e−10
granulocyte activation (GO:0036230)	3.104557e−10
cellular response to chemical stress (GO:0062197)	2.157116e−03
response to oxidative stress (GO:0006979)	3.141033e−03
cellular response to oxidative stress (GO:0034599)	2.902893e−03

Figure S3.4: Enrichment q-values of GO terms that are found enriched in the DEGs that are uniquely identified by Clipper in pairwise comparison of (a) Clipper vs. edgeR and (b) Clipper vs. DESeq2. These GO terms are all related to immune response and thus biologically meaningful.

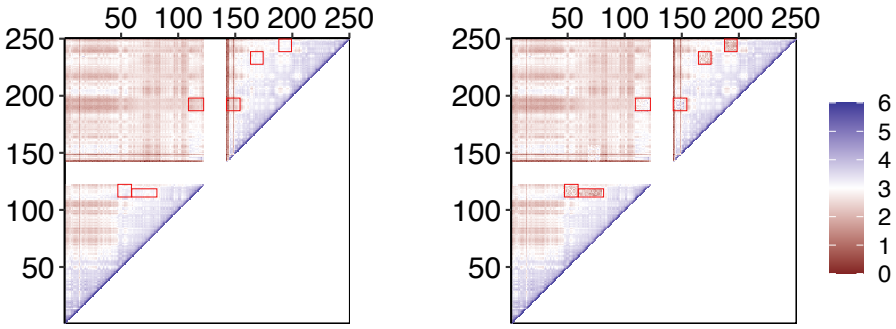

Figure S3.5: $\log_{10}$-transformed mean Hi-C interaction matrices (μ_X and μ_Y in Section S3.5.5) under the two conditions. DIR regions are highlighted in red squares.

CHAPTER 4

FDR control in aggregating peptides identified by multiple database search algorithms from mass spectrometry data

4.1 Introduction

Proteomics studies have discovered essential roles of proteins in complex disease such as neurodegenerative disease [119] and cancer [120–122]. These studies have demonstrated the potential of using proteomics to identify clinical biomarkers for disease diagnosis and therapeutic targets for disease treatment. In recent years, proteomics analytical technologies, particularly tandem mass spectrometry (MS)-based shotgun proteomics, have advanced immensely, thus enabling high-throughput identification and quantification of proteins in biological samples. Compared to prior technologies, shotgun proteomics has simplified sample preparation and protein separation, reduced time and cost, and saved procedures that may result in sample degradation and loss [123]. In a typical shotgun proteomics experiment, a protein mixture is first enzymatically digested into peptides, i.e., short amino acid chains up to approximately 40-residue long; the resulting peptide mixture is then separated and measured by tandem MS into tens of thousands of mass spectra. Each mass spectrum encodes the chemical composition of a peptide; thus, the spectrum can be used to identify the peptide's amino acid sequence and post-translational modifications, as well as to quantify the peptide's abundance with additional weight information (Fig. 4.1a).

Since the development of shotgun proteomics, numerous database search algorithms have been developed to automatically covert mass spectra into peptide sequences. Pop-

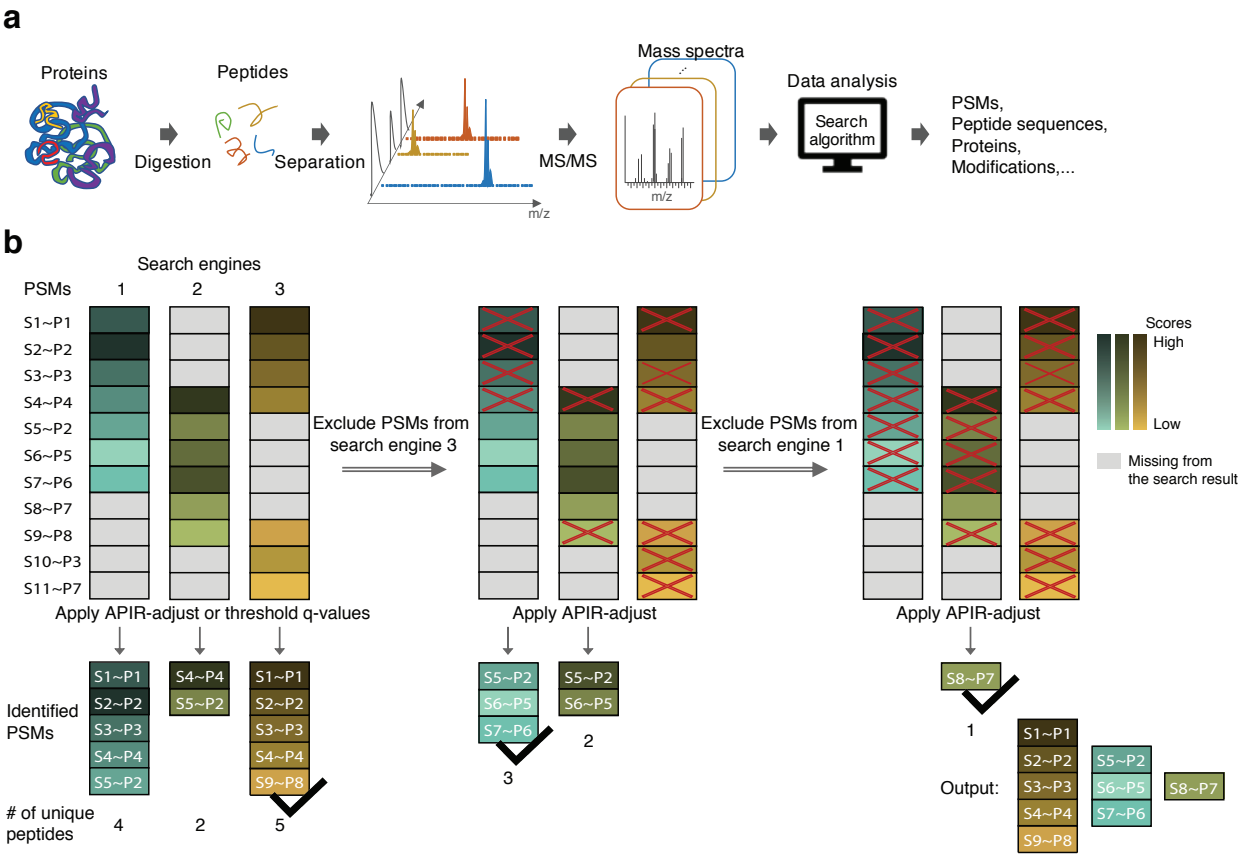

Figure 4.1: (a) The workflow of a typical shotgun proteomics experiment. The protein mixture is first digested into peptides, short amino acid chains. The resulting peptide mixture is separated and measured by tandem mass spectrometry (MS) as mass spectra, which encode the chemical composition of peptides. Then database search algorithms are used to decode these mass spectra by identifying PSMs, peptides, proteins, modifications and etc. (b) Illustration of APIR in aggregating three database search algorithms. We use S1∼P1 to denote a PSM of mass spectrum S1 and peptide sequence P1 and etc. In the output of a database search algorithm, a PSM with a higher score is marked by a darker color. Gray PSMs are missing from the output. APIR adopts a sequential approach to aggregate database search algorithms 1, 2, and 3. In the first round, APIR applies APIR-adjust or q-value/PEP thresholding to identify a set of identified target PSMs from the output of each database search algorithm. APIR then selects the algorithm whose identified PSMs by APIR-adjust contain the highest number of unique peptides and treats the corresponding identified PSMs as identified by APIR. In this example, APIR identified equal numbers of PSMs from algorithms 1 and 3 but more unique peptides from algorithm 3; therefore, APIR selects algorithm 3 in the first round. In the second round, APIR excludes all PSMs, both identified and unidentified by the selected database search algorithm in the first round (algorithm 3 in this example), from the output of the remaining database search algorithms. Then it applies APIR-adjust again to find the algorithm whose identified PSMs by APIR-adjust contain the highest number of unique peptides (algorithm 1 in this example). APIR repeats this procedure in the subsequent rounds until all database search algorithms are exhausted and outputs the union of PSMs identified in each round.

ular database search algorithms include SEQUEST [112], Mascot [78], MaxQuant [124], Byonic [113], and MS-GF+ [125], among many others. A database search algorithm takes as input the mass spectra from a shotgun proteomics experiment and a protein database that contains all known protein sequences. For each mass spectrum, the algorithm identifies the best matching peptide sequence, a subsequence of a protein sequence, from the database; we call this process "peptide identification," whose result is a "peptide-spectrum match"

(PSM). However, due to data imperfection (such as low-quality mass spectra, mistakes in data processing, and incomplete protein database), the resulting PSMs often consist of many false PSMs, causing issues in the downstream, system-wide identification and quantification of proteins [126].

To ensure the accuracy of PSMs, the false discovery rate (FDR) has been used as the most popular statistical criterion [127–136]. Technically, the FDR is defined as the expected proportion of false PSMs among the identified PSMs; in other words, a small FDR indicates good accuracy of PSMs. However, controlling the FDR is only one side of the story. Because shotgun proteomics experiments are costly, a common goal of database search algorithms is to identify as many true PSMs as possible to maximize the experimental output, in other words, to maximize the identification power given a target, user-specified FDR threshold (e.g., 1% or 5%). To achieve this goal, existing database search algorithms have predominantly relied on the target-decoy search strategy [126] to estimate the FDR.

The key idea of the target-decoy search strategy is to generate a negative control of PSMs by matching mass spectra against artificially created, false protein sequences, called "decoy" sequences. Decoy sequences can be created in multiple ways, and a typical way is to reverse each protein sequence to obtain a corresponding decoy sequence. Given the decoy sequences, the target-decoy search strategy can be implemented as the concatenated search or parallel search. In the concatenated search, a concatenated protein database is created by pooling original protein sequences, called "target" sequences, with the decoy sequences; then a database search algorithm uses the concatenated protein database to find PSMs; consequently, each mass spectra is mapped to either a target sequence or a decoy sequence with only one matching score (Fig. 4.2a). In the parallel search, a database search algorithm conducts two parallel searches: a target search where each mass spectrum is matched to target sequences and a decoy search where the mass spectrum is matched to decoy sequences; consequently, each mass spectrum receives two matching scores from the two searches (Fig. 4.2b). In both implementations, a PSM is called a target PSM or simply a PSM if it contains a target sequence; otherwise, it is called a decoy PSM. Finally, a database search algorithm uses the decoy PSMs, i.e., the PSMs known to be false, to estimate the

FDR [126, 131]. In technical terms, each target PSM receives a q-value from an algorithm such as Byonic, Mascot, SEQUEST, and MS-GF+ [78, 112, 113, 125] or a posterior error probability (PEP) from an algorithm such as MaxQuant [124] (see Online Methods). Both q-value and PEP are related to the FDR so that users can control the FDR under a threshold q if they keep only the target PSMs with q-values or PEPs not exceeding q; however, the FDR control is only guaranteed when the q-values and PEPs are valid [130].

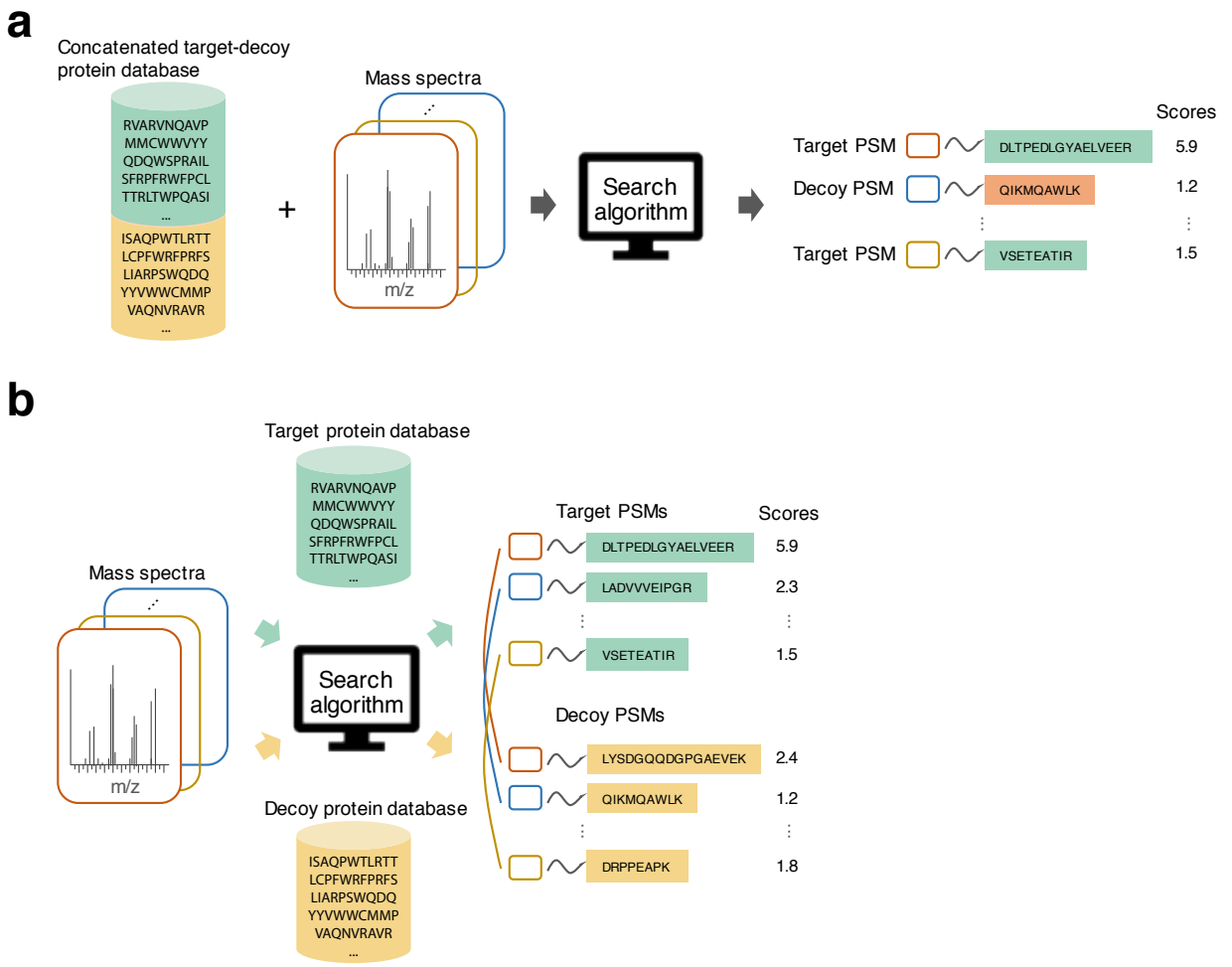

Figure 4.2: Two implementations of the target-decoy search strategy: concatenated (a) and parallel (b). In the concatenated search, a concatenated protein database is created by pooling original protein sequences, called "target" sequences, with the decoy sequences; then a database search algorithm uses the concatenated protein database to find PSMs; consequently, each mass spectra is mapped to either a target sequence or a decoy sequence with only one matching score. In the parallel search, a database search algorithm conducts two parallel searches: a target search where each mass spectrum is matched to target sequences and a decoy search where the mass spectrum is matched to decoy sequences; consequently, each mass spectrum receives two matching scores from the two searches. In both implementations, a PSM is called a target PSM or simply a PSM if it contains a target sequence; otherwise, it is called a decoy PSM.

It has been observed that, with the same input mass spectra and FDR threshold, different database search algorithms may find largely distinct sets of target PSMs [19–23]. There are

two possible reaons of this phenomenon. One is that different algorithms find different sets of true PSMs by design. The other is that some algorithms have identified excessive false PSMs due to failed FDR control. It is important to disentangle these two reasons because, if the former is true, we may aggregate the distinct sets of target PSMs to increase the peptide identification power; otherwise, we must refine the output of the algorithms that have failed to control the FDR before performing the aggregation.

To leverage database search algorithms' distinct advantages, statistical methods have been developed to aggregate search results from multiple database search algorithms. We refer to these methods as aggregation methods. Existing aggregation methods include Scaffold [22], MSblender [136], FDRAnalysis [137], iProphet [135], ConsensusID [134], and PepArML [127]. Among these six methods, except FDRAnalysis that has been shown infeasible for high-throughput proteomics [19], the rest have two major drawbacks: (1) limited compatibility with database search algorithms and (2) lack of guarantee to identify more peptides. About the first drawback, these aggregation methods unanimously limit the choices of database search algorithms. In particular, only Scaffold supports Byonic, which demonstrates superior performance in both FDR control and power on our newly generated proteomics standard dataset (see Results and Fig. 4.3a and b for details). Moreover, none of these aggregation methods support some recently published database search algorithms such as TagGraph [138] or Bolt [139] (see Online Methods for a list of database search algorithms compatible with each aggregation method). As for the second drawback, although empirical evidence shows that, on some datasets, these aggregation methods may identify more target PSMs than those identified by individual database search algorithms, none of these aggregation methods is guanranteed to do so.

In addition to the above aggregation methods developed for proteomics data, generic statistical methods developed for aggregating rank lists are in theory applicable to aggregating the target PSM lists output by different database search algorithms. However, none of these methods have been used for proteomics data, nor are they guaranteed to increase the identified target PSMs given an FDR threshold. Therefore, the field calls for a robust, powerful, and flexible aggregation method that allows researchers to reap the benefits of the

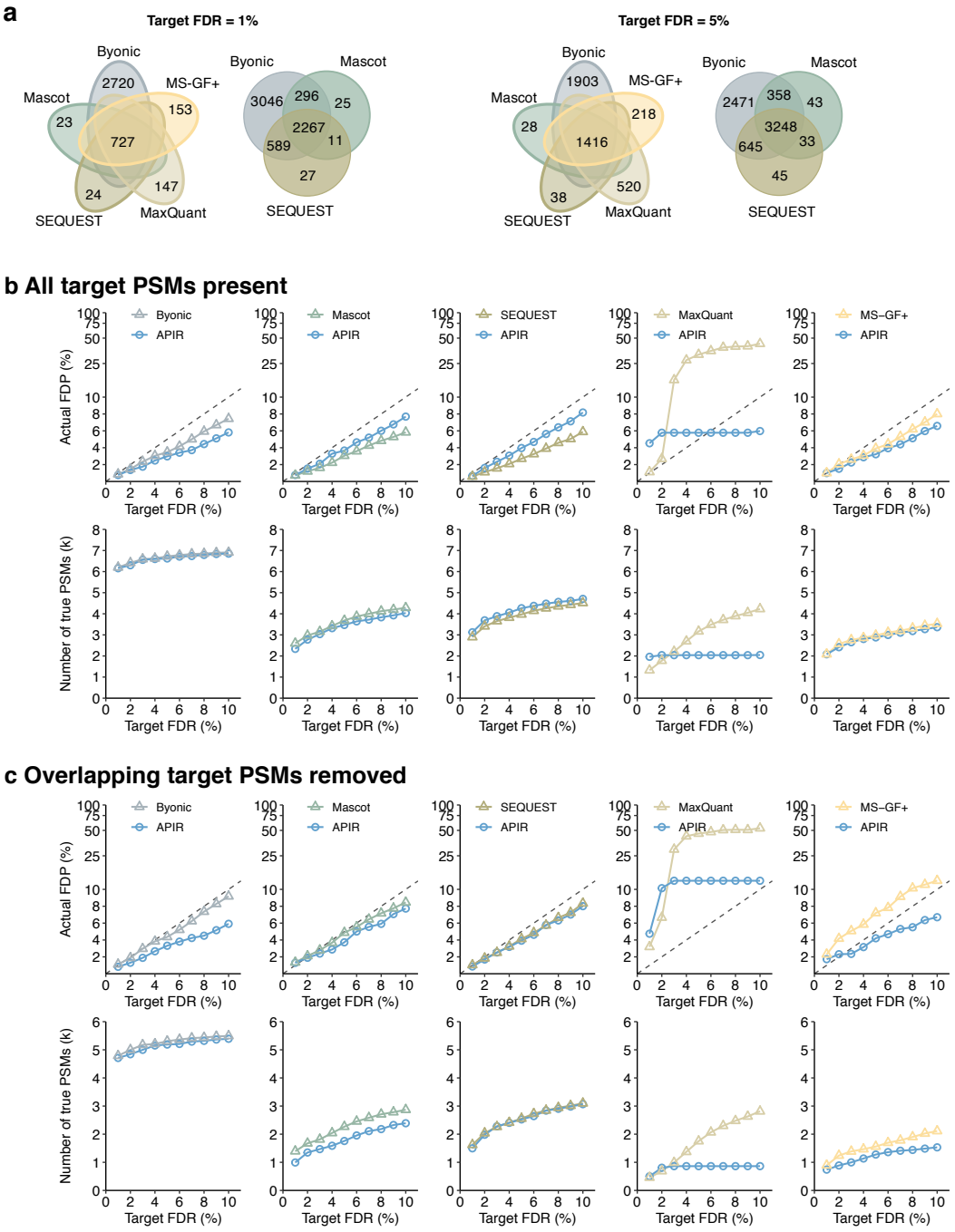

Figure 4.3: Benchmarking APIR-adjust and the five popular database search algorithms—Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+—on the complex proteomics standard dataset in terms of FDR control and power. (a) Venn diagrams of the true target PSMs identified by Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ at the FDR threshold $q = 1\%$ (left) and $q = 5\%$ (right). (b)-(c) At the FDR threshold $q \in \{1\%, 2\%, \cdots, 10\%\}$, FDPs and power of each of the five database search algorithms when all target PSMs are present (b) or when the 1416 target PSMs identified by all five database search algorithms at the FDR threshold $q = 5\%$ are removed from each database search algorithm (c).

diverse and ever-growing database search algorithms.

Here we develop Aggregate Peptide Identification Results (APIR), a statistical framework

that aggregates peptide identification results from multiple database search algorithms with FDR control. APIR is the first statistical framework that is universally adaptive to database search algorithms outputting PSMs with scores (e.g., q-values or PEPs) and is guaranteed to identify at least the same number of, if not more, peptides than individual database search algorithm. APIR is a robust, flexible, and powerful framework that enhances the power while controlling the FDR of peptide identification from shotgun proteomics data.

4.2 APIR methodology

APIR aims to combine the PSMs identified from multiple database search algorithms with valid FDR control. Aside from an FDR threshold q (e.g., 5%), from the output of each database search algorithm, APIR inputs a list of target PSMs with scores and a list of decoy PSMs with scores. APIR is a sequential FDR control framework that relies on APIR-adjust, a core component of APIR, to control FDR in each step. Below we introduce the details of APIR by first introducing APIR-adjust and then the general framework based on APIR-adjust for aggregating search results.

4.2.1 APIR-adjust: FDR control on the target PSMs identified by individual search algorithms

The core component of APIR is APIR-adjust, an FDR-control method that re-identifies target PSMs from a single database search algorithm. APIR-adjust takes as input an FDR threshold q, a list of target PSMs with scores, and a list of decoy PSMs with scores. APIR-adjust then outputs identified target PSMs.

We first define the target coverage proportion as the proportion of target PSMs whose mass spectra also appear among the decoy PSMs. Depending on the database search algorithms and the implementation of target-decoy search strategy (concatenated or parallel), the target coverage proportion could vary from 0 to 1. When the target coverage proportion is high, most of the target PSMs could be one-to-one paired with decoy PSMs by their mass spectra so that in each pair, the decoy PSM score serves as a negative control for the target

PSM score. When the proportion is low, we cannot form many pair-decoy score pairs but to use decoy PSM scores collectively as a negative control. We thus design two approaches, tailored specifically for these two scenarios, into APIR-adjust.

Here we introduce notations to facilitate our discussion. Suppose a database search algorithm outputs m target PSMs with scores $T_1, \ldots, T_m$ and n decoy PSMs with scores $D_1, \ldots, D_n$. Also, suppose that among the m target PSMs, the first $s \leqslant \min(m, n)$ target PSMs can be paired one-to-one with decoy PSMs; accordingly, the target coverage proportion is s/m. Without loss of generality, we rearrange decoy PSM indices such that the i-th decoy PSM shares the same mass spectrum with the i-th target PSM for $1 \leqslant i \leqslant s$.

When the target coverage proportion is relatively high ($s/m \geqslant 40\%$), APIR-adjust identifies target PSMs using Clipper in Chapter 3, a p-value-free statistical framework for FDR control on high-throughput data by contrasting two conditions. Specifically, Clipper constructs a contrast score $C_i = T_i - D_i$ if $i = 1, \ldots, s$ and $C_i = 0$ if $i = s+1, \ldots, m$; then it finds a cutoff $C_{\text{thre}} = \min\left\{t \in \{|C_i| : C_i \neq 0\} : \frac{|i:C_i \leqslant -t|+1}{\max(|i:C_i \geqslant t|, 1)} \leqslant q\right\}$, and outputs $\{i : C_i \geqslant C_{\text{thre}}\}$ as the indices of identified target PSMs. Based on Clipper, APIR-adjust requires two assumptions to control the FDR: first, $T_1, \ldots, T_m, D_1, \ldots, D_n$ are mutually independent, and second, T_i and D_i are identically distributed if the i-th target PSM is false. See the original paper for detailed proofs that guarantee FDR control [18].

When the target coverage proportion is relatively low ($s/m < 40\%$), APIR-adjust uses the pooled approach, a p-value-based approach described in Section S3.5.2, to identify target PSMs. By assuming that the scores of decoy PSMs and false target PSMs are independently and identically distributed, the p-value-based approach constructs a null distribution by pooling $D_j, j = 1, \ldots, n$. Then APIR-adjust computes a p-value for the i-th target PSM as the tail probability right of T_i, i.e., $p_i = |\{j : D_j \geqslant T_i\}|/n, i = 1, \ldots, m$, and controls FDR using the Benjamini-Horchberg procedure [4].

4.2.2 APIR: a sequential framework for aggregating multiple search algorithms' identified target PSMs with FDR control

Suppose we are interested in aggregating K algorithms. Let W_k denote the set of target PSMs output by the k-th algorithm, $k = 1, \ldots, K$. APIR adopts a sequential approach that consists of K rounds.

- In the first round, APIR applies APIR-adjust or q-value/PEP thresholding to each algorithm's output with the FDR threshold q. Denote the identified target PSMs from the k-th algorithm by $U_{1k} \subset W_k$. Define $J_1 \in \{1, \cdots, K\}$ to be the algorithm such that U_{1J_1} contains the highest number of unique peptides among $U_{11}, \ldots, U_{1K}$. We use the number of unique peptides rather than the number of PSMs because peptides are more biologically relevant than PSMs.

- In the second round, APIR first excludes all target PSMs output by the J_1-th algorithm, identified or unidentified in the first round, i.e., W_{J_1}, from the outputs of the remaining database search algorithms, resulting in reduced sets of candidate target PSMs $W_1 \backslash W_{J_1}$, ..., $W_K \backslash W_{J_1}$. Then APIR applies APIR-adjust with FDR threshold q to these reduced sets except $W_{J_1} \backslash W_{J_1} = \varnothing$. Denote the resulting sets of identified target PSMs by $U_{2k} \subset (W_k \backslash W_{J_1})$, $k \in \{1, \ldots, K\} \backslash \{J_1\}$. Again APIR finds the J_2-th algorithm such that U_{2J_2} contains the most unique peptides.

- APIR repeats this in the subsequent rounds. In Round ℓ with $\ell \geqslant 2$, APIR first excludes all target PSMs output by the selected $\ell - 1$ database search algorithms from the outputs of remaining database search algorithms and applies APIR-adjust. That is, APIR applies APIR-adjust with FDR threshold q to identify a set of identified PSMs $U_{\ell k}$ from $W_k \backslash (W_{J_1} \cup \cdots \cup W_{J_{\ell-1}})$, the reduced candidate pool of algorithm k after the previous $\ell - 1$ rounds, for algorithms $k \in \{1, \cdots, K\} \backslash \{J_1, \cdots, J_{\ell-1}\}$. Then APIR finds the algorithm, which we denote by J_ℓ, such that $U_{\ell J_\ell}$ contains the most unique peptides.

- Finally, APIR outputs $U_{1J_1} \cup \cdots \cup U_{KJ_K}$ as the identified target PSMs.

By adopting this sequential approach, APIR is guaranteed to identify at least as many, if not more, unique peptides as those identified by a single database search algorithm; under reasonable assumptions, APIR controls the FDR of the identified target PSMs under q. See Fig. 4.1b for graphical illustration and Section S4.6.10 for the the theoretical guarantee of FDR control by APIR.

4.3 Results

To benchmark existing database search algorithms and aggregation methods including APIR, we generated the first publicly available complex proteomics standard dataset that approaches the dynamic range of a typical proteomics experiment from *Pyrococcus Furiosus (Pfu)*. We also designed simulation studies to benchmark APIR against naive aggregation approaches: intersection and union. To demonstrate the power of APIR, we applied it to five real datasets, including the proteomics standard dataset, three acute myeloid leukemia (AML) datasets, and a triple-negative breast cancer (TNBC) dataset. Notably, out of the three AML datasets, we generated two from bone marrow samples of acute myeloid leukemia (AML) patients with either enriched or depleted leukemia-stem-cells (LSC) for studying the disease mechanisms of AML.

4.3.1 Byonic, Mascot, SEQUEST, MaxQuant and MS-GF+ capture unique true PSMs on the proteomics standard dataset, but MaxQuant fails to control the FDR

We first benchmarked five popular database search algorithms: Byonic [113], Mascot [78], SEQUEST [112], MaxQuant [124], and MS-GF+ [125] on the proteomics standard dataset. Specifically, we ran tandem MS analysis to generate $49,303$ mass spectra from *Pfu*. We then generated a reference database by concatenating the *Pfu* database, the Uniprot Human database, and two contaminant databases: the contaminant repository for affinity purification (the CRAPome) [140] and the contaminant database from MaxQuant. During the process, we performed *in silico* digestion to remove the overlapping peptides between *Pfu*

and human from the human database. Finally, we input the resulting database and the *Pfu* mass spectra into a database search algorithm. We consider a target PSM, a peptide, or a protein as true if the database search algorithm reports its master protein as from either *Pfu* or the two contaminants and as false otherwise. See Methods for experimental details about the generation of this benchmark data and how we benchmark these five database search algorithms.

Our results in Fig. 4.3a show that individual database search algorithms indeed capture unique PSMs. At both $q = 1\%$ and 5%, all five database search algorithms identify unique true target PSMs. Notably, at $q = 1\%$, the number of true target PSMs identified by Byonic alone $(2,720)$ is nearly four times the number of overlapping PSMs among the five database search algorithms (727). At $q = 5\%$, Byonic again identifies more unique true target PSMs $(1,903)$ than the overlap among the five database search algorithms $(1,416)$. Additionally, MaxQuant and MS-GF+ also demonstrate their distinctive advantages: MaxQuant identifies 147 and 520 unique true PSMs while MS-GF+ identifies 153 and 218 at $q = 1\%$ and 5% respectively. In contrast, SEQUEST and Mascot show little advantage in the presence of Byonic: Byonic nearly covers the identified true PSMs from SEQUENST and Mascot (Fig. 4.3a). Our results confirm some database search algorithms' distinctive advantages in identifying unique PSMs, which aligns well with existing literature [19–23, 141].

In terms of FDR control, the four database search algorithms—Byonic, Mascot, SEQUEST, and MS-GF+—demonstrate robust FDR control as they keep FDP on the benchmark data under the FDR thresholds $q \in \{1\%, \ldots, 10\%\}$. In contrast, except at small values of q such as 1% or 2%, MaxQuant fails FDR control by a large margin (Fig. 4.3b).

4.3.2 For individual database search algorithms, APIR-adjust shows robust FDR control and power advantage on the proteomics standard dataset

To demonstrate the use of APIR-adjust, we applied it as an add-on to the five database search algorithms for adjusting their identified target PSMs on the proteomics standard dataset. We examined the FDP and power for a range of FDR thresholds: $q \in \{1\%, 2\%, \ldots, 10\%\}$.

Our results in Fig. 4.3a show that APIR-adjust controls FDR without sacrificing power when applied to Byonic, Mascot, SEQUEST, or MS-GF+. As for MaxQuant, even though APIR-adjust fails to control the FDR when q is small ($q \leqslant 5\%$), the FDPs of adjusted MaxQuant results are much closer to the target level than the FDPs of the original results. When q is above 5%, APIR-adjust enables MaxQuant to achieve a good FDR control.

Even with valid q-values or PEPs, q-value/PEP thresholding could only work when all target PSMs with q-values smaller than or equal to q are present in the output of database search algorithms. In other words, q-values are no longer guaranteed to control the FDR after a subset of target PSMs are removed. To verify this, we apply q-value/PEP thresholding after excluding from each database search algorithm the 1416 shared true PSMs that are identified at the FDR threshold $q = 5\%$ (Fig. 4.3a). Our results in Fig. 4.3 show that thresholding the q-values of MS-GF+ could no longer control the FDR. In contrast, because APIR-adjust ignores the FDR-control property of q-values and treats them as scores, APIR-adjust demonstrates a robust FDR control even with missing target PSMs.

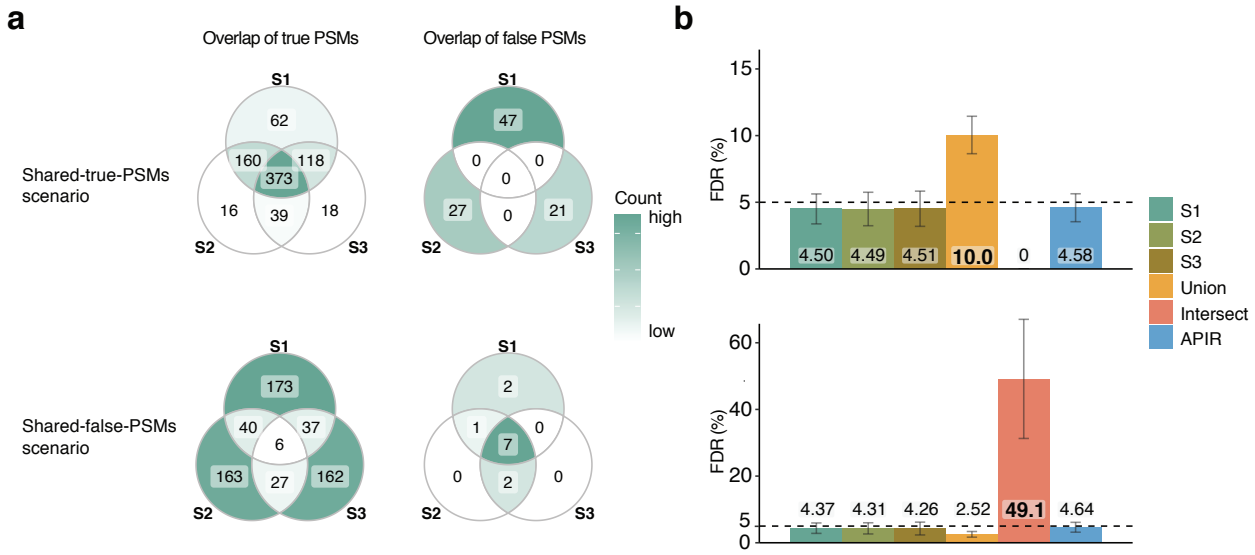

Figure 4.4: Comparison of APIR, intersection, and union in the FDR control of aggregating three database search algorithms. At the FDR threshold $q = 5\%$, each database search algorithm's and each aggregation method's actual FDRs are evaluated on 200 simulated datasets under two scenarios: the shared-true-PSMs scenario (top) and the shared-false-PSMs scenario (bottom). (a) Venn diagrams of true PSMs and false PSMs from one simulated dataset under either scenario. In the shared-true-PSMs scenario, the three database search algorithms tend to identify overlapping true PSMs but non-overlapping false PSMs. In the shared-false-PSMs scenario where the database search algorithms tend to identify overlapping false PSMs but non-overlapping true PSMs. (b) The FDR of each database search algorithm and each aggregation method. Union fails to control the FDR in the shared-true-PSMs scenario, while intersection fails in the shared-false-PSMs scenario. APIR controls FDR in either scenario.

4.3.3 For aggregating multiple database search algorithms, APIR has verified FDR control and power advantage in simulation and on the proteomics standard dataset

Next, we validated the FDR control of APIR using simulation studies. We compared APIR with two naive aggregation approaches: intersection or union. A popular misconception exists that if each database search algorithm controls FDR well, taking their intersection guarantees FDR control while taking their union does not [127]. To clarify this misconception, we generated target PSMs with scores from three toy database search algorithms under two simulation scenarios: shared-true-PSMs and shared-false-PSMs. Under the shared-true-PSMs scenario, the three toy database search algorithms tend to identify overlapping true PSMs but non-overlapping false PSMs. In comparison, under the shared-false-PSMs scenario, the toy database search algorithms tend to identify overlapping false PSMs but non-overlapping true PSMs (Fig. 4.4a). Under both scenarios, we first applied APIR-adjust to each toy database search algorithm. Then we aggregated their results using intersection, union, or APIR and compared their FDR-control performances.

Our results in Fig. 4.4b confirm that APIR-adjust controls FDR of individual database search algorithms and APIR controls FDR in aggregating them. In contrast, intersection fails to control the FDR under the shared-false-PSMs scenario, and union fails under the shared-true-PSMs scenario. It is not hard to see mathematically that the FDR of union is approximately upper bounded by q times the number of database search algorithms. In contrast, the maximum FDR of intersection could potentially approach 1. See Online Methods for details of the simulation.

We further demonstrate that APIR controls FDR and improves power on the complex protein standard. Because running database search algorithms is time-consuming, we expect users to aggregate typically no more than three database search algorithms. Therefore, we examined 20 combinations in total, with 10 combinations of two database search algorithms out of five and 10 combinations of three database search algorithms out of five at two FDR threshold $q \in \{1\%, 5\%\}$. We compared APIR with Scaffold because it is the only

existing aggregation software that is compatible with Byonic. Notably, Scaffold does not aim to control the FDR; instead, it relies on two sets of scores to ensure the quality of identified PSMs: a peptide identification probability, which is the probability of the peptide present in the sample and is thresholded by a peptide threshold, and a protein identification probability, which is the probability of the protein present in the sample and is thresholded by a protein threshold. Moreover, Scaffold requires both thresholds to be specified before it outputs search results. Thus Scaffold is not directly comparable with APIR in terms of FDR control. Accordingly, we conducted the following two comparisons: in the first comparison, we implemented Scaffold by setting both the peptide threshold and the protein threshold to be q FDR; in the second comparison, we set the peptide threshold to be q and varies the protein threshold to maximize the number of peptides. See Online Methods for details.

For each combination under either comparison, we examined the FDP in identified PSMs and compared power increase by computing the percentage increase in identified true PSMs, true peptides or true proteins. The percentage increase in true PSMs, peptides, or proteins is computed by treating as the baseline the maximal number of true identified PSMs, peptides, or proteins by individual database search algorithms in the first round of APIR. For example, if we aim to aggregate Byonic and MaxQuant on the proteomics standard dataset, based on our benchmarking results in Fig. 4.3b we would choose to apply q-value thresholding to Byonic and apply APIR-adjust to MaxQuant to identify PSMs in the first round. Then when we calculate percentage increases in identified true PSMs, the baseline would be the larger value between the number of correctly identified PSMs by thresholding the PEPs of Byonic and the number of correctly identified PSMs by applying APIR-adjust to MaxQuant. Because of the trade-off between FDR and power, it is reasonable to compare power only when FDR is controlled. Therefore, it is unfair to compare APIR with the original MaxQuant in terms of power since the latter fails to control the FDR on the proteomics standard dataset. When calculating the percentage increase in true proteins, we identify proteins from the identified PSMs using the majority rule (See Methods for details on protein identification and quantification).

In our first comparison, we set both the peptide threshold and the protein threshold of

Scaffold to be q. Our results in Fig. 4.5 and Fig. 4.6 show that at either FDR threshold $q = 5\%$ or 1%, APIR controls FDR of PSM identification; it also improves power of peptide identification compared to individual database search algorithms in nearly all combinations. The exceptions occur when APIR aggregates Mascot or SEQUEST with Byonic and results in zero power improvement. The reason, as shown in Fig. 4.3a, is that Byonic nearly covers all true PSMs output by either Mascot or SEQUEST. Consequently, APIR calls the PSMs identified by Byonic as discoveries in the first round and fails to identify any more PSMs from either Mascot or SEQUEST in the subsequent rounds. In contrast to APIR's stable FDR control and power improvement, Scaffold demonstrates good FDR control but highly unstable power improvement. Specifically, Scaffold fails to identify more PSMs, peptides or proteins than the most powerful database search algorithm in all combinations but one at $q = 5\%$ and in more than half of the combinations at $q = 1\%$.

In our second comparison, we set the peptide threshold to be q and varied the protein threshold to maximize the number of unique peptides. Our results in Fig. 4.7 and Fig. 4.8 show that at both FDR thresholds $q = 5\%$ and 1%, Scaffold demonstrates a slightly inflated FDP in many combinations. In terms of power, although our implementation favors Scaffold, it still fails to outperform the most powerful individual database search algorithm in more combinations than APIR. Our results confirm the stable performance of APIR. See Online Methods for details on how we implemented Scaffold.

4.3.4 APIR empowers peptide identification by aggregating the search results from Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ on four real datasets

We next applied APIR to four real datasets: two phospho-proteomics (explained below) datasets of AML that we generated (phospho AML1 and phospho AML2) for studying the properties of leukemia stem cells in AML pateints; a published phospho-proteomics dataset of triple-negative breast cancer (TNBC) from Fang et al. [142] that studies the drug effect of genistein on breast cancer; and a published nonphospho-proteomics dataset of AML from

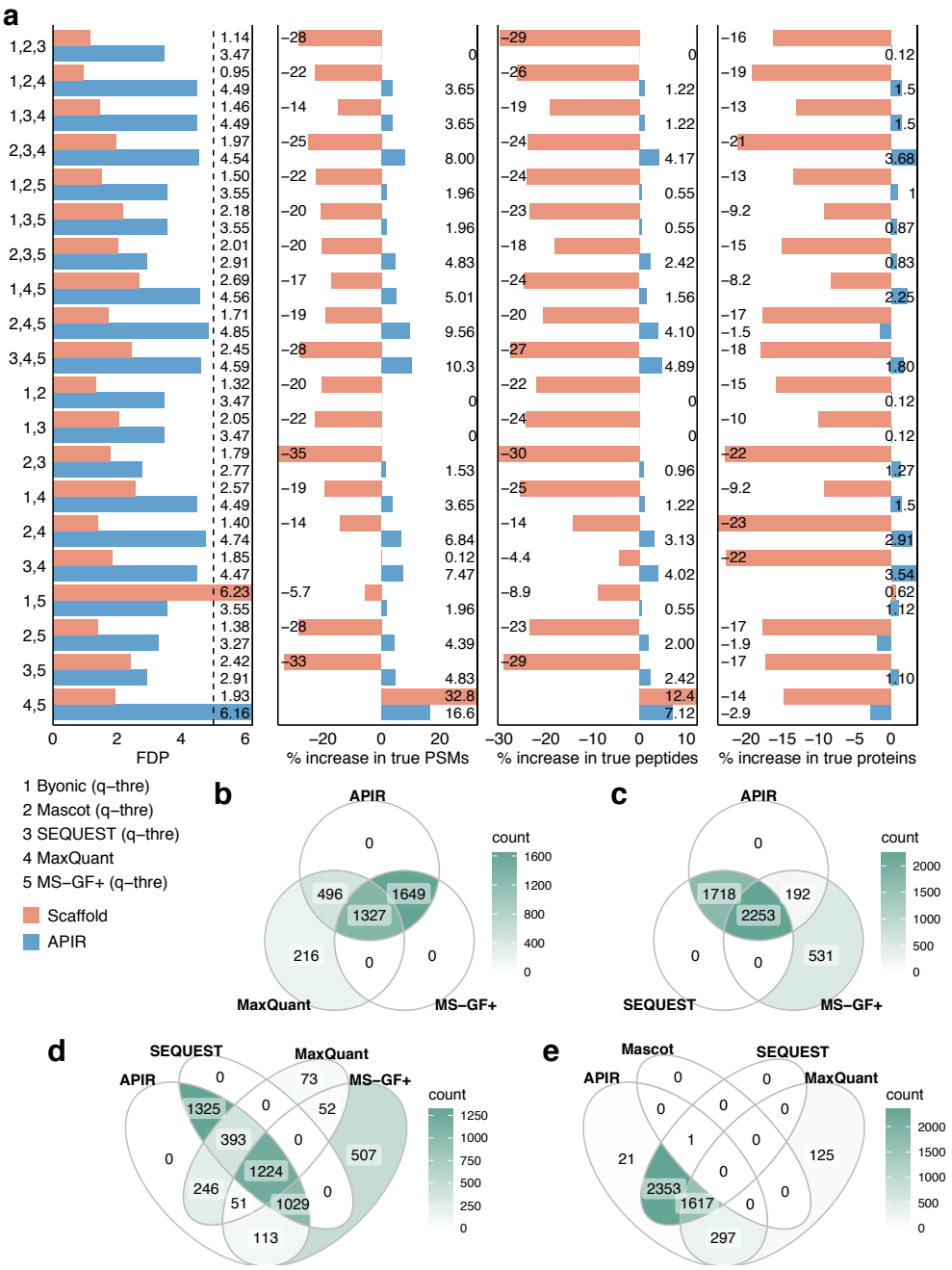

Figure 4.5: On the proteomics standard, comparison of APIR and Scaffold at the FDR threshold $q = 5\%$ in terms of FDR control and power. We set both the peptide threshold and the protein threshold of Scaffold to be 95%. (a) FDPs (first column), the percentage increase in true PSMs (second column), the percentage increase in true peptides (third column), and the percentage increase in true proteins (fourth column) in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). Based on the benchmarking results in Fig. 4.3b, we applied q-value thresholding to Byonic, Mascot, SEQUEST, and MS-GF+, and applied APIR-adjust to MaxQuant in the first round of APIR. The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by individual database search algorithms in the first round of APIR. (b)-(e) Venn diagrams of true PSMs by APIR and individual database search algorithms from four example combinations in (a). Venn diagrams comparing APIR with (b) MaxQuant (adjusted by APIR-adjust) and MS-GF+; with (c) SEQUEST, MaxQuant (adjusted by APIR-adjust), and MS-GF+; with (d) SEQUEST and MS-GF+; with (e) Mascot, SEQUEST, and MaxQuant (adjusted by APIR-adjust) demonstrate that APIR identifies almost all true PSMs by individual database search algorithms at the same FDR threshold $q = 5\%$.

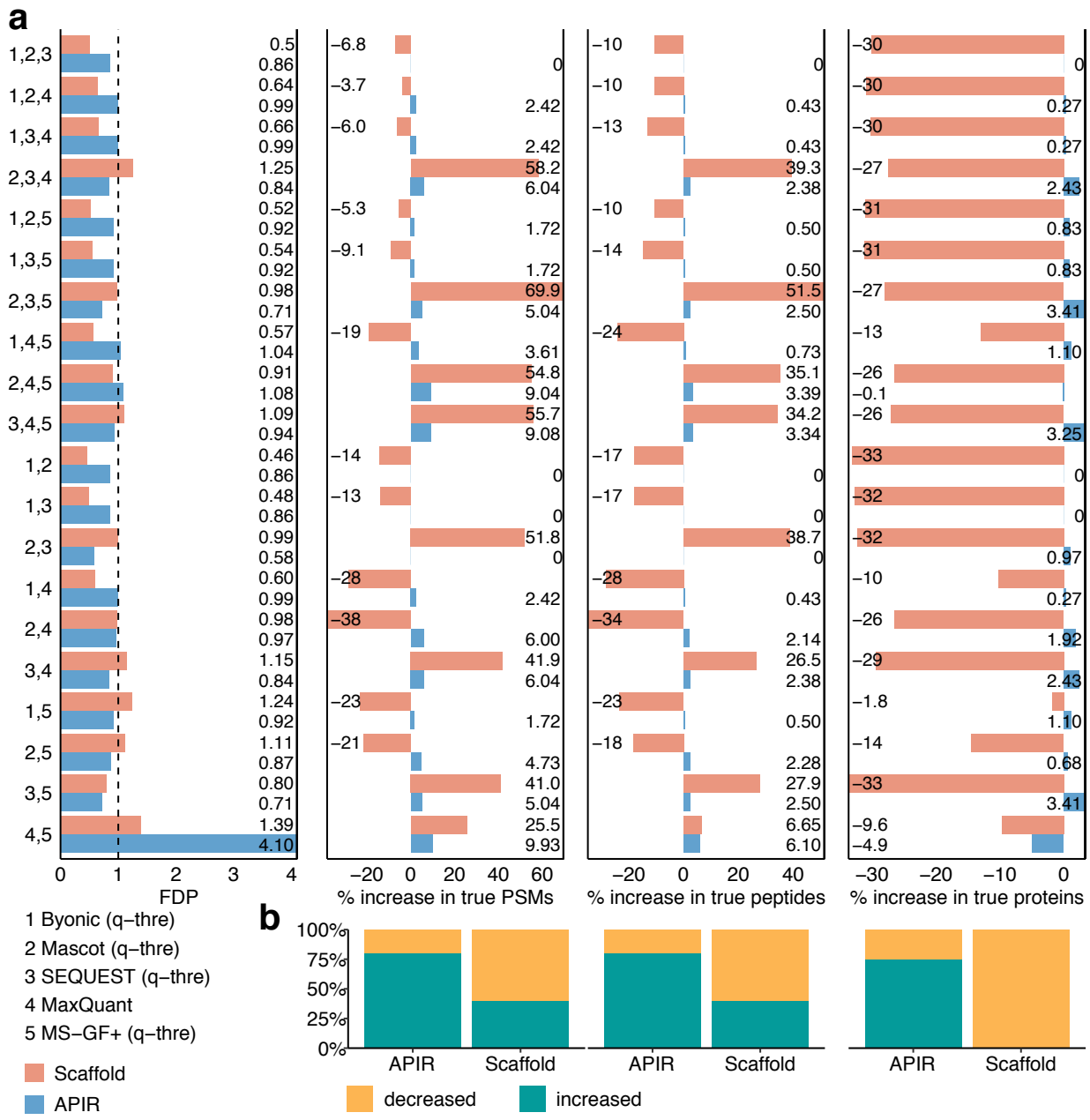

Figure 4.6: On the proteomics standard, comparison of APIR and Scaffold at the FDR threshold $q = 1\%$ in terms of FDR control and power. We set both the peptide threshold and the protein threshold of Scaffold to be 99%. (a) FDPs (first column), the percentage increase in true PSMs (second column), the percentage increase in true peptides (third column), and the percentage increase in true proteins (fourth column) in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). Based on the benchmarking results in Fig. 4.3c, we applied q-value thresholding to Byonic, Mascot, SEQUEST, and MS-GF+, and applied APIR-adjust to MaxQuant in the first round of APIR. The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by individual database search algorithms in the first round of APIR. (b) Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column).

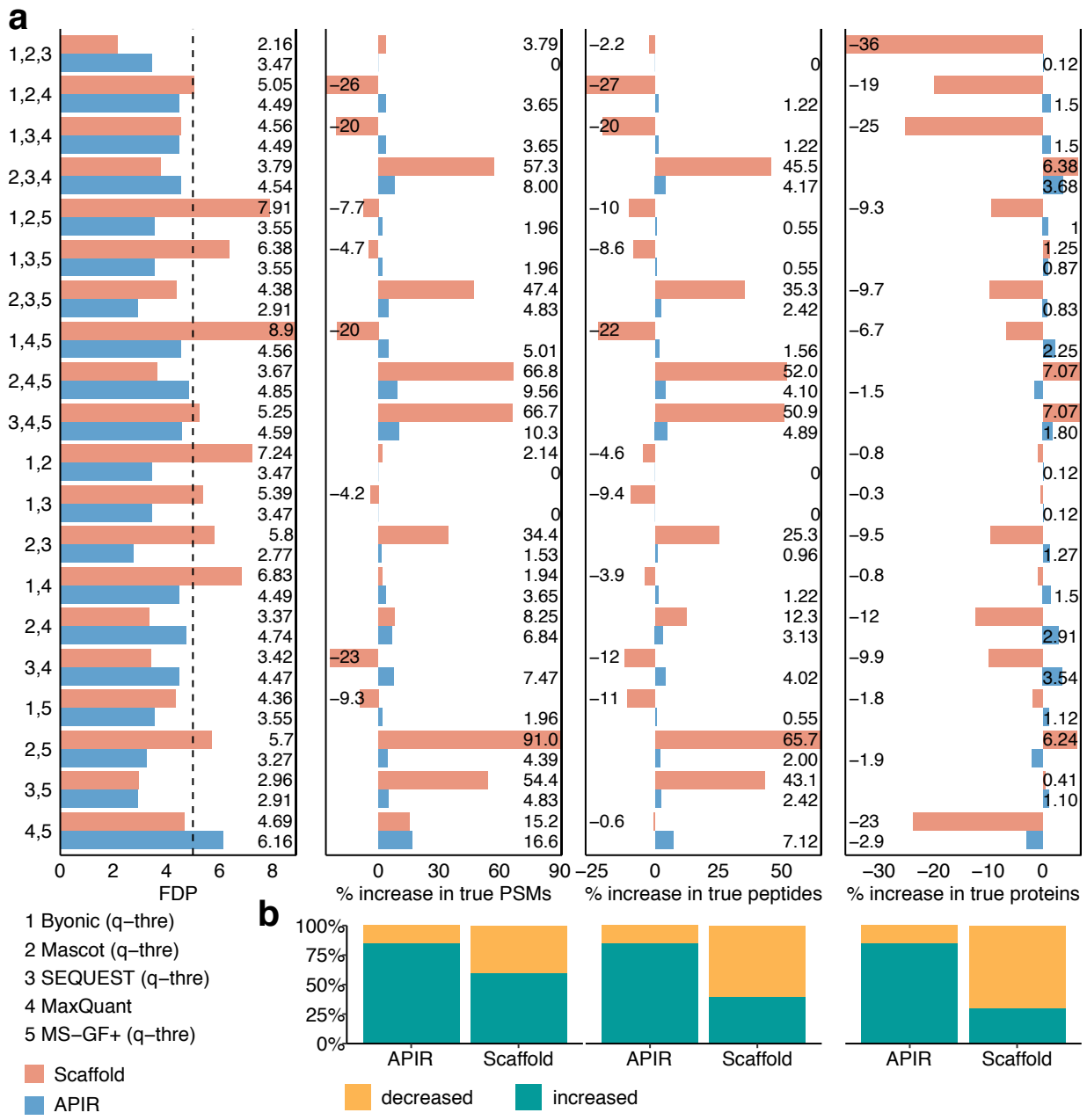

Figure 4.7: On the proteomics standard, comparison of APIR and Scaffold at the FDR threshold $q = 5\%$ in terms of FDR control and power. We set the peptide threshold to be 95% and varied the protein threshold to find the maximal number of identified peptides. (a) FDPs (first column), the percentage increase in true PSMs (second column), the percentage increase in true peptides (third column), and the percentage increase in true proteins (fourth column) in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). Based on the benchmarking results in Fig. 4.3b, we applied q-value thresholding to Byonic, Mascot, SEQUEST, and MS-GF+, and applied APIR-adjust to MaxQuant in the first round of APIR. The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by individual database search algorithms in the first round of APIR. (b) Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column).

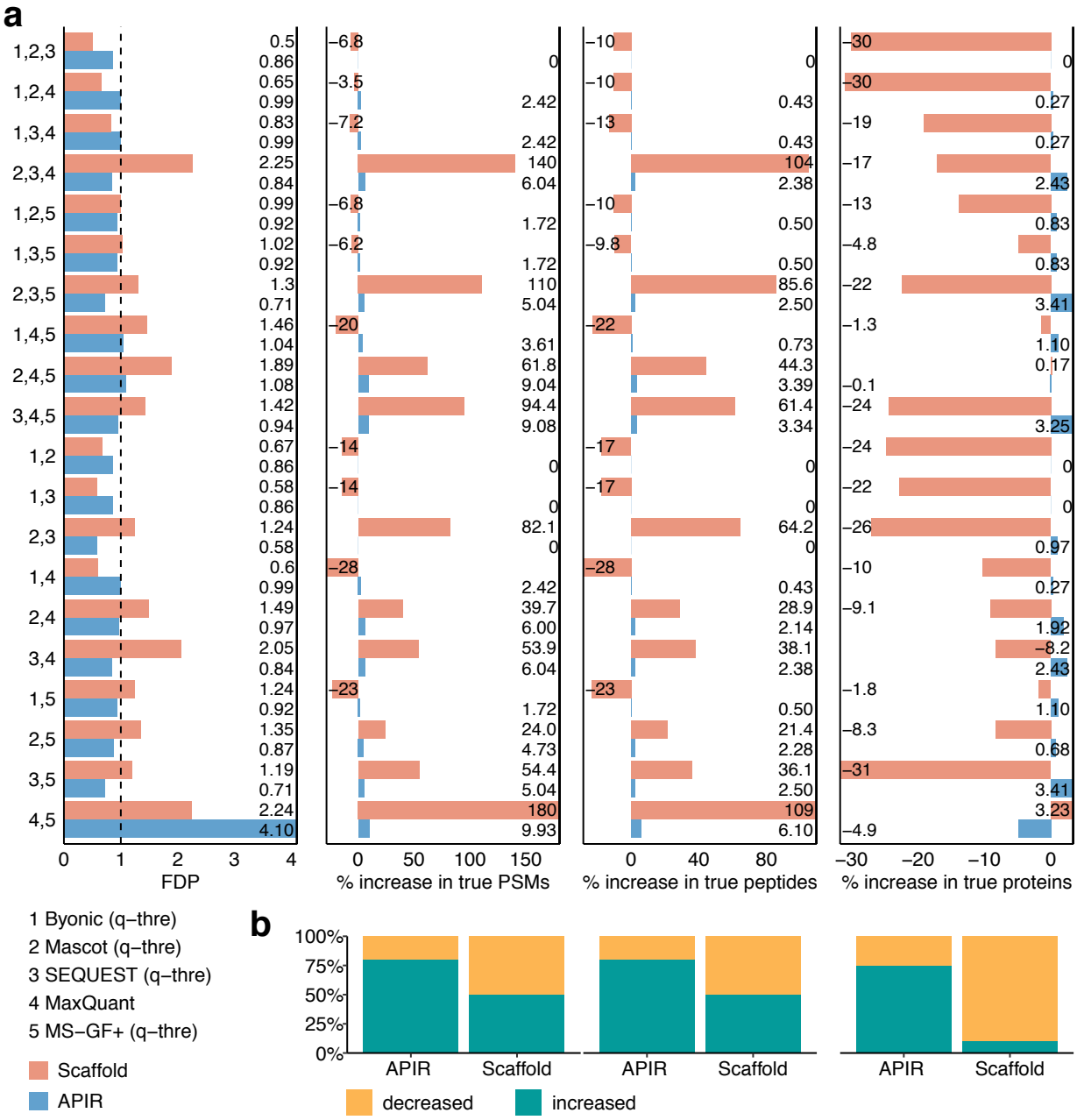

Figure 4.8: On the proteomics standard, comparison of APIR and Scaffold at the FDR threshold $q = 1\%$ in terms of FDR control and power. We set the peptide threshold to be 99% and varied the protein threshold to find the maximal number of identified peptides. (a) FDPs (first column), the percentage increase in true PSMs (second column), the percentage increase in true peptides (third column), and the percentage increase in true proteins (fourth column) in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+). Based on the benchmarking results in Fig. 4.3b, we applied q-value thresholding to Byonic, Mascot, SEQUEST, and MS-GF+, and applied APIR-adjust to MaxQuant in the first round of APIR. The percentage increase in true PSMs/peptides/proteins is computed by treating as the baseline the maximal number of correctly identified PSMs/peptides/proteins by individual database search algorithms in the first round of APIR. (b) Proportion of combinations that show a non-negative percentage increase (green bars) in true PSMs (first column), true peptides (second column), and true proteins (third column).

Raffel et al. [143] (nonphospho AML) that also compares the stem cells with non-stem cells in AML patients. Phospho-proteomics is a branch of proteomics; while traditional proteomics aims to capture all peptides in a sample, phospho-proteomics focuses on phosphorylated ones, also called phosphopeptides, because phosphorylation regulates essentially all cellular processes [144]. On each dataset, we applied APIR and examined its performance at two FDR thresholds $q \in \{1\%, 5\%\}$ in four aspects: the percentage increase in PSMs, peptides, peptides with modifications, and proteins, which we calculated in a similar fashion to what we did on the proteomics standard dataset.

Our results in Fig. 4.9 and 4.10 show that APIR improved power on all four levels across all four datasets at both FDR thresholds $q \in \{1\%, 5\%\}$. Specifically, at both FDR thresholds, APIR consistently improved power on the peptide level on all four datasets, a result that aligned with our expectation because APIR is guaranteed to do so. Interestingly, on both the peptide level and the peptide-with-modification level, APIR also achieved improved power across 20 combinations on all four datasets at both FDR thresholds, with only one exception: APIR fell short by a negligible 0.1% when aggregating the search results from Byonic, Mascot and SEQUEST on the TNBC dataset at the FDR threshold $q = 5\%$. On the protein level, APIR still managed to outperform single database search algorithms across all combinations on both phospho-proteomics AML datasets and in more than half of the combinations on either the TNBC dataset or the nonphospho-proteomics AML dataset. Our results demonstrate that APIR could boost the power of mass spectrometry data analysis.

4.3.5 APIR identifies biologically meaningful proteins from a phospho AML datasets and a TNBC dataset

We investigated the biological functions of additional proteins that APIR found on the phospho AML datasets and the TNBC dataset.

On both phospho AML1 and AML2 datasets, APIR identified biologically meaningful proteins that were missed by unadjusted individual database search algorithms. On phospho AML1, APIR identified across the 20 combinations 80 additional proteins at the FDR thresh-

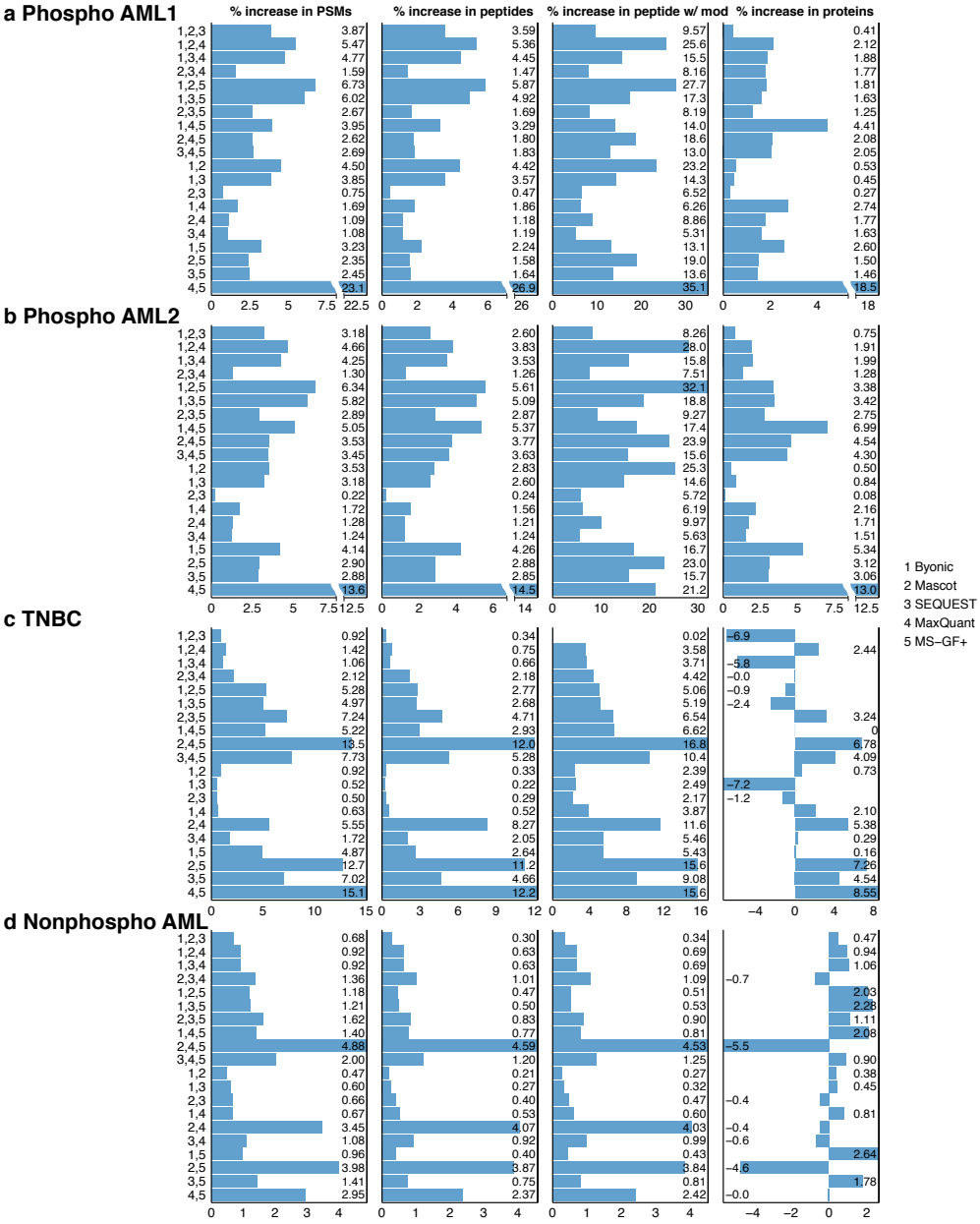

Figure 4.9: Power improvement of APIR over individual database search algorithms at the FDR threshold $q = 5\%$. The percentage increase in PSMs (first column), the percentage increase in peptides (second column), the percentage increase in peptides with modifications (third column), and the percentage increase in true proteins (fourth column) of APIR in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+) at the FDR threshold $q = 5\%$ on the phospho-proteomics AML datasets (a)-(b), the TNBC dataset (c) and the nonphospho-proteomics AML dataset (d). The percentage increase in PSMs/peptides/peptides with modifications/proteins is computed by treating as the baseline the maximal number of PSMs/peptides/peptides and modifications/proteins by individual database search algorithms in the first round of APIR.

old $q = 1\%$ and 121 additional proteins at the FDR threshold $q = 5\%$, including transcription intermediary factor 1-alpha (TIF1α), phosphatidylinositol 4,5-bisphosphate 5-phosphatase A (PIB5PA), sterile alpha motif domain containing protein 3 (SAMD3), homeobox protein

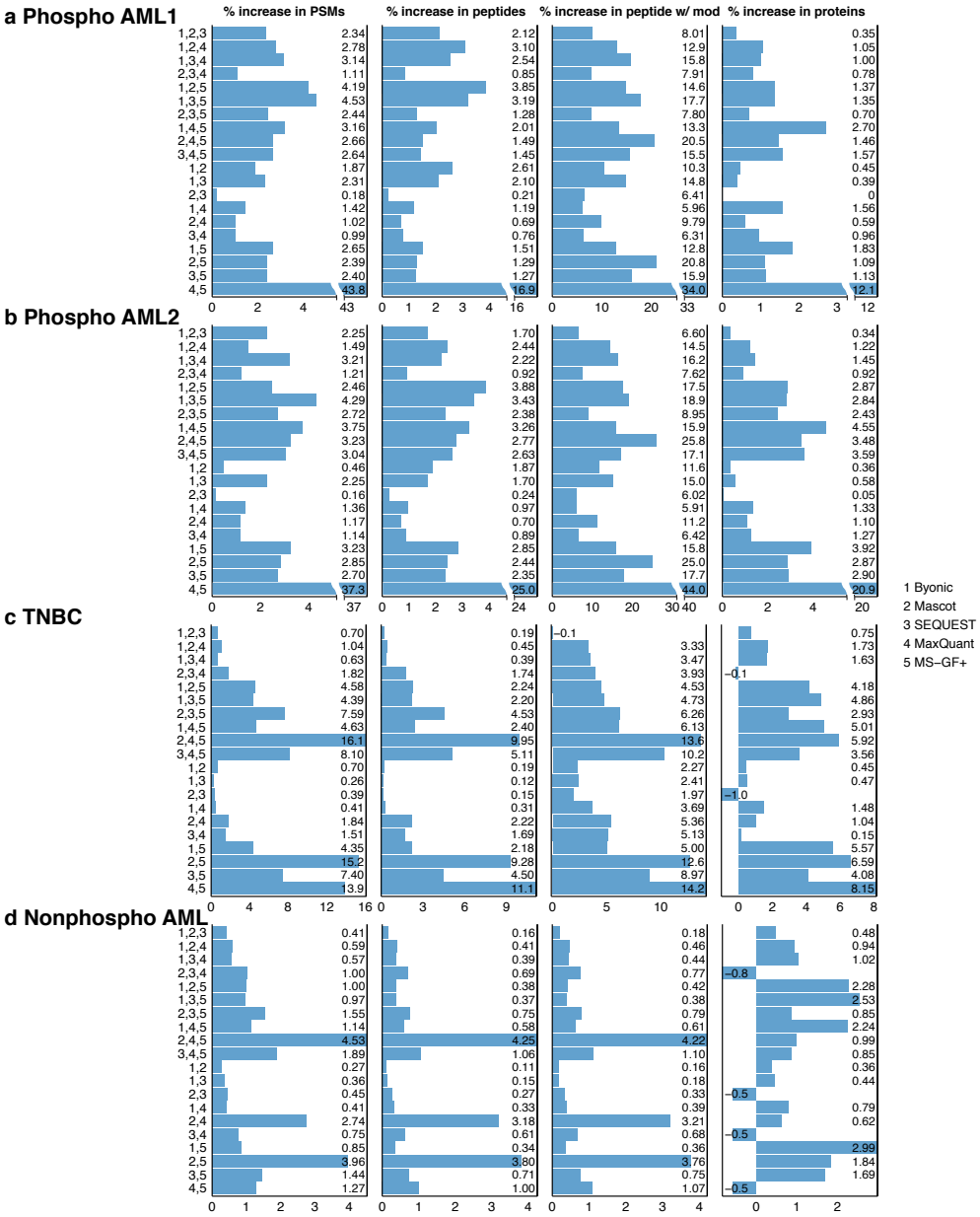

Figure 4.10: Power improvement of APIR over individual database search algorithms at the FDR threshold $q = 1\%$. The percentage increase in PSMs (first column), the percentage increase in peptides (second column), the percentage increase in peptides with modifications (third column), and the percentage increase in true proteins (fourth column) of APIR in aggregating two or three database search algorithms out of the five (Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+) at the FDR threshold $q = 1\%$ on the phospho-proteomics AML datasets (a)-(b), the TNBC dataset (c) and the nonphospho-proteomics AML dataset (d). The percentage increase in PSMs/peptides/peptides with modifications/proteins is computed by treating as the baseline the maximal number of PSMs/peptides/peptides and modifications/proteins by individual database search algorithms in the first round of APIR.

Hox-B5 (HOXB5), small ubiquitin-related modifier 2 (SUMO-2), transcription factor jun-D (JUND), glypican-2 (GPC2), dnaJ homolog subfamily C member 21 (DNAJC21), mRNA decay activator protein ZFP36L2 (ZFP36L2), leucine-rich repeats and immunoglobulin-like domains protein 1 (LRIG-1), and mitochondrial intermembrane space import and assembly

protein 40 (CHCHD4). High levels of TIF1α are associated with oncogenesis and disease progression in a variety of cancer lineages such as AML [145–151]. PIB5PA has been shown to have a tumor-suppressive role in human melanoma [152]. Its high expression has been correlated with limited tumor progression and better prognosis in breast cancer patients [153]. SMAD3 is known to play key roles in the development and progression of various types of tumor [154–159]. HOXB5 is among the most affected transcription factors by the genetic mutations that initiate AML [160–162]. SUMO-2 has been found to play a key role in regulating CBX2, which is overexpressed in several human tumors, including leukemia and whose expression is correlated with lower overall survival [163]. JUND has been shown to play a central role in the oncogenic process leading to adult T-cell leukemia [164]. GPC2 has been identified as an oncoprotein and a candidate immunotherapeutic target in high-risk neuroblastoma [165]. DNAJC21 mutations have been linked to cancer-prone bone marrow failure syndrome [166]. ZFP36L2 has been found to induce AML cell apoptosis and inhibits cell proliferation [167]; its mutation has been associated with the pathogenesis of acute leukemia [168]. LRIG-1 has been found to regulate the self-renewing ability of leukemia stem cells in AML [169]. CHCHD4 plays key roles in regulating tumor proliferation [170]. On phospho AML2, APIR has identified 62 additional proteins at FDR 1% and 19 additional proteins at FDR 5%, including JUND and myeloperoxidase (MPO). MPO is expressed in hematopoietic progenitor cells in prenatal bone marrow, which are considered initial targets for the development of leukemia [171–173].

On the TNBC dataset, APIR identified 92 proteins that were not found by unadjusted single database search algorithms at the FDR threshold $q = 1\%$ and 69 such proteins at FDR $q = 5\%$. In particular, at the FDR threshold $q = 1\%$, APIR has uniquely identified BRCA2, DNA repair associated (BRCA2), and Fanconi anemia complementation group E (FANCE). BRCA2 is a well-known breast cancer susceptibility gene. An inherited genetic mutation inactivating the BRCA2 gene can be found in people with TNBC [174–179]. FANC-BRCA pathway, including FANCE and BRCA2, is known for its roles in DNA damage response. Inactivation of the FANC–BRCA pathway has been identified in ovarian cancer cell lines and sporadic primary tumor tissues [180, 181]. Additionally, at both FDR thresholds, we have

identified JUND and roundabout guidance receptor 4 (ROBO4); the latter regulates tumor growth and metastasis in multiple types of cancer, including breast cancer [182–185]. Our results demonstrate APIR's strong potential in identifying novel disease-related proteins.

4.3.6 APIR identifies differentially expressed peptides that are biologically meaningful from a phospho AML datasets

An important use of proteomics data is the differential expression (DE) analysis, which aims to identify proteins whose expression levels change between two conditions. The ideal unit of measurements is proteins; however, due to the difficulties in quantifying protein levels from tandem MS data, an alternative approach has been proposed and used, which identifies differentially expressed peptides first and then investigates their corresponding proteins along with their modifications. Because it is less error-prone to quantify peptides than quantify proteins, doing so could dramatically reduce errors in the DE analysis.

Here we compared APIR with MaxQuant and MS-GF+ by performing DE analysis on the phospho AML1 dataset. We focused on this dataset instead of the TNBC dataset or the non-phospho AML dataset because the phosphos AML datasets are unpublished. The phospho AML1 dataset contains six bone marrow samples: three enriched with leukemia stem cells (LSCs), two depleted of LSCs, and one control. To simplify our DE analysis, we selected two enriched and two depleted samples as shown in Fig. 4.11a. Specifically, we first applied APIR to aggregate the search results by MaxQuant and MS-GF+ on the phospho AML1 dataset using all six samples. Then we applied DESeq2 to identify DE peptides from the peptide level results of APIR, APIR-adjusted MaxQuant, and APIR-adjusted MS-GF+ using the four selected samples. Our results in Fig. 4.11 show that at the FDR threshold 5%, we identified 318 DE peptides from 224 proteins based on APIR, 251 DE peptides from 180 proteins based on MaxQuant, and 242 DE peptides from 190 proteins based on MS-GF+ respectively. In particular, APIR has identified 6 leukemia related proteins: the promyelocytic leukemia zinc finger (PLZF), Serine/threonine-protein kinase B-raf (B-raf), Signal transducer and activator of transcription 5B (STAT5B), Promyelocytic Leukemia Protein (PML), cyclin-dependent

kinase inhibitor 1B (CDKN1B), and retinoblastoma-associated protein (RB1), all of which belong to the AML KEGG pathway or the chronic myeloid leukemia KEGG pathway [186–188]. In particular, PLZF and CDKN1B were uniquely identified from the APIR aggregated results but not by either APIR-adjusted MaxQuant or APIR-adjusted MS-GF+.

We next investigated the phosphorylation on the identified DE peptides of PLZF or CDKN1B. With regard to PLZF, APIR has identified phosphorylation at Threonine 282, which is known to activate cyclin-A2 [189], a core cell cycle regulator of which the deregulation seems to be closely related to chromosomal instability and tumor proliferation [190–192]. As for CDKN1B, APIR has identified phosphorylation at Serine 140. Previous studies have revealed that ATM phosphorylation of CDKN1B at Serine 140 is important for stabilization and enforcement of the CDKN1B-mediated G1 checkpoint in response to DNA damage [193]. A recent study shows that inability to phosphorylate CDKN1B at Serine 140 is associated with enhanced cellular proliferation and colony formation [194]. Our results illustrate that APIR could assist in discovering interesting proteins and relevant post-translational modifications.

a

Channels	Sample description
1	P5337 LSC depleted
2	P5337 LSC enriched
3	P5340 LSC depleted
4	P5340 LSC enriched
5	~~P5340 LSC enriched~~
6	~~Control~~

b

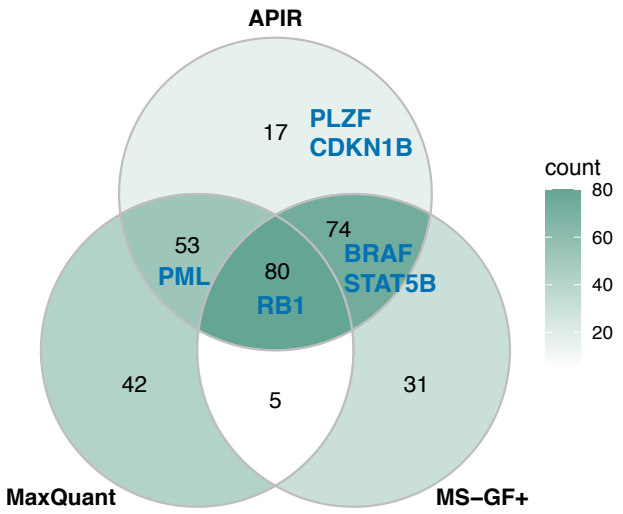

Figure 4.11: Comparison of APIR with MaxQuant and MS-GF+ by DE analysis on the phospho AML1 dataset. (a) Sample description of the phospho AML1 dataset. This dataset contains six bone marrow samples from two patients: P5337 and P5340. From P5337, one LSC enriched sample and one LSC depleted sample were taken. From P5340, two LSC enriched samples and one LSC depleted sample were taken. We ignored one LSC enriched sample from P5340 and the control sample while conducting DE analysis (crossed out). (b) Venn diagrams of proteins from the identified DE peptides based on APIR aggregating MaxQuant and MS-GF+, APIR-adjusted MaxQuant and APIR-adjusted MS-GF+. APIR has identified 6 leukemia-related proteins: PLZF, B-raf, STAT5B, PML, CDKN1B, and RB1, all of which belong to the AML KEGG pathway or the chronic myeloid leukemia KEGG pathway. Note that PLZF and CDKN1B were uniquely identified from the APIR aggregated results.

4.4 Discussion

We developed a statistical framework APIR to identify PSMs with FDR control from multiple database search algorithms. The core component of APIR is APIR-adjust, an FDR-control method that re-identifies PSMs from a single database search algorithm without restrictive distribution assumptions. Based on APIR-adjust, APIR aggregates target PSMs from multiple database search algorithms with FDR control. APIR offers a great advantage of flexibility: APIR is compatible with any database search algorithm that outputs scores. The reason lies in that APIR is a sequential approach based on a simple idea: given multiple disjoint sets of discoveries with each FDP smaller than or equal to q, their union also has FDP smaller than or equal to q. This sequential approach not only allows APIR to circumvent the need to impose restrictive distribution assumptions on each database search algorithm's scores, but also ensures that APIR would identify at least as many, if not more, unique peptides as a single database search algorithm does. By assessing APIR on the first publicly available complex proteomics standard dataset, we verify that APIR consistently improves the sensitivity of peptide identification analysis with FDR control of PSMs. Our extensive studies on leukemia and TNBC data suggest that APIR could lead to discoveries of additional disease-relevant peptides and proteins that are otherwise missed by individual database search algorithms.

The current implementation of APIR controls FDR on the PSM levels. However, in shotgun proteomics experiments, PSMs serve merely as an intermediate to identify peptides and proteins, the real molecules of biological interest; thus, an ideal FDR control should occur on the peptide or protein level. Besides, FDR control on the PSM level does not entail FDR control on the peptide or protein level because the same peptide sequences could be coded into multiple PSMs, and a protein consists of multiple peptides. To realize FDR control of peptides or proteins, APIR-adjust needs to be carefully modified. Take the FDR control on the peptide level as an example. One possible modification would be to engineer a score for each peptide from the scores of PSMs that contain this peptide. Future studies are needed to explore possible ways of engineering a peptide score. Once we modify APIR-

adjust to control the FDR on the peptide level or the protein level, the current sequential approach of APIR still applies: in each round, we apply the modified APIR-adjust to identify peptides or proteins and exclude the peptides or proteins output by the selected database search algorithm from the previous round.

4.5 Acknowledgments

This chapter is based on my joint work with MeiLu McDermott, Kyla Woyshner, Antigoni Manousopoulou, Xinzhou Ge, Dr. Leo David Wang, and my Ph.D. advisor Dr. Jingyi Jessica Li.

4.6 Supplementary Material

S4.6.1 Complex proteomics standard dataset generation

The complex proteomics standard (CPS) (part number 400510) was purchased by Agilent (Agilent, Santa Clara, CA, USA). CPS contains soluble proteins extracted from the archaeon *Pyrococcus furiosus* (*Pfu*). All other chemicals were purchased from Sigma Aldrich (Sigma Aldrich, St. Louis, MO, USA). The fully sequenced genome of *Pfu* encodes for approximately 2000 proteins that cover a wide range of size, pI, concentration levels, hydrophobic/hydrophilic character, etc. CPS (500ug total protein) was dissolved in 100uL of 0.5 M tri-etrhylammonium bicarbonate (TEAB) and 0.05% sodium dodecyl sulfate (SDS) solution. Proteins were reduced using tris(2-carboxyethyl)phosphine hydrochloride (TCEP) (4 uL of 50mM solution added in the protein mixture and sample incubated at 60 0 C for 1hour) and alkylated using methyl methyl methanethiosulfonate (MMTS) (2 uL of 50mM solution added in the protein mixture and sample incubated at room temperature for 15 minutes). To enzymatically digest the proteins, 20ug trypsin dissolved 1:1 in ultrapure water was added in the sample and this was incubated overnight (16 hours) in dark at 37 0 C. The tryptic peptides were cleaned with C-18 tips (part number 87784) from Thermo Fisher Scientific (Thermo Fisher Scientific, Waltham, MA, USA) following the manufacturer's instructions.

Peptides were LC-MS analysed using the Ultimate 3000 uPLC system (EASY-Spray column, part number ES803A, Thermo Fisher Scientific) hyphenated with the Orbitrap Fusion Lumos mass spectrometry instrument (Thermo Fisher Scientific). Peptides were fragmented using low energy CID and detected with the linear ion trap detector.

On this complex proteomics standard dataset, we benchmarked the five database search algorithms—SEQUEST [1], Mascot [2], MaxQuant [3], Byonic [4], and MS-GF+ [5]—in terms of peptide identification. Specifically, we first generated a reference database by concatenating the Uniprot *Pyrococcus furiosus* (*Pfu*) database, the Uniprot Human database, and two contaminant databases: the CRAPome [6] and the contaminant databases from MaxQuant. During the process, we performed *in silico* digestion of *Pfu* proteins and removed human proteins that contained *Pfu* peptides from the reference database. We then input the *Pfu* mass spectra and the resulting database into a database search algorithm. We consider a target PSM as true if the database search algorithm reports its master protein as from *Pfu* or the two contaminants and false if from the human. The *in silico* digestion was performed in Python using the `pyteomics.parser` function from `pyteomics` with the following settings: Trypsin digestion, two allowed missed cleavages, minimum peptide length of six [7, 8].

S4.6.2 TNBC data and non-phospho AML data availability

- The raw MS data files of the TNBC dataset is available at the PRoteomics IDEntifications Database (PRIDE) with the dataset identifier PXD002735 [9].
- The raw MS data files of the non-phospho dataset is available at the (PRIDE) with the dataset identifier PXD008307 [9].

S4.6.3 Existing aggregation methods

Scaffold Scaffold (Proteome Software, Portland, Oregan, USA) adopts a Bayesian approach to aggregate probabilities of the individual database search algorithm results into a single probability for each PSM. One of its key step is to generate for each database

search algorithm a peptide probability model that estimates the probability of an individual spectrum being correctly assigned to a peptide based on that database search algorithm's score. To realize this, Scaffold designs a different statistical model for the internal scores from each database search algorithm [10], making it difficult to generalize its approach to other database search algorithms. Scaffold supports Byonic (Protein Metrics), Mascot (Matrix Science), Mascot Distiller (Matrix Science), MaxQuant/Andromeda (Max Planck Institute), Peaks (Bioinformatics Solutions), and Proteome Discoverer (Thermo Fisher Scientific) database search algorithms including Byonic, SEQUEST, and Mascot.

MSblender MSblender is an open-source software that uses a probability mixture model to model the scores of correct and incorrect PSMs. In particular, the correct PSM scores across database search algorithms are assumed to follow a two-component (by default) multivariate Gaussian [11]. Search engines that are compatible with MSblender include SEQUEST (Thermo Fisher Scientific), X!Tandem [12], OMSSA [13], InsPecT [14], MyriMatch [15], MSGFDB [16] (`http://www.marcottelab.org/index.php/MSblender#Prerequisites`).

ConsensusID ConsensusID is part of the OpenMS Proteomics Pipeline [17]. It adopts a probabilistic approach to aggregate the top-scoring PSM results from several database search algorithms. A key feature of this tool is its sequence similarity scoring mechanism, which is a method to estimate the scores for PSMs in cases when the peptide is missing from the high-ranking results of a database search algorithm. It involves fitting the scores from each database search algorithm as a two-component mixture model. The two components are a Gumbel distribution for the incorrect PSMs and a normal distribution for the correct PSMs [18]. Although the paper Nahnsen et al. [18] claims that ConsensusID supports all database search algorithms, the OpenMS pipeline only supports the following search algorithms: Comet [19], CompNovo [20], Crux [21], Mascot (Thermo Fisher Scientific), MS-GF+ [5], MyriMatch [15], OMSSA [13], PepNovo [22], X!Tandem [12].

PepArML PepArML is an unsupervised, model-free, machine-learning-based method to aggregate search results. It is compatible with Mascot [2], Tandem [12] with native, K-score, and s-score [23] scoring, OMSSA [13], MyriMatch [15], InSpecT [14], and MS-GF [24] spectral probability scores.

iProphet The iProphet software is an open-source software within the Trans Proteomic Pipeline (TPP) suite. It is used between PeptideProphet [25] and ProteinProphet [26]. It calculates peptide level probabilities via mixture models. The TPP suite is compatible with COMET [19], X!Tandem [12], SEQUEST (Thermo Fisher Scientific), MS-GF+ [5], InSpecT [14], OMSSA [13], MyriMatch [15], ProbID [27], Mascot (Matrix Science), Phenyx [28].

S4.6.4 Implementation of database search algorithms on the proteomics standard

Byonic, SEQUEST, and Mascot Byonic, SEQEUST, and Mascot were each run in Proteome Discoverer 2.3.0.523 (ThermoScientific). The following settings were used for all 5 database search algorithms: 10 ppm precursor tolerance; 0.6 Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M). Percolator was used in conjunction with both SEQUEST and Mascot, and the target decoy mode was set to separate. To acquire the total list of identified PSMs, peptides, and proteins, internal FDRs for all database search algorithms were set to 100%.

MaxQuant MaxQuant was implemented with the following settings: 10 ppm precursor tolerance; 0.6 Da fragment tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ), oxidation (M); second peptide search: True. To acquire the total list of identified PSMs, peptides, and proteins, the internal FDR was set to 100%. MaxQuant outputs a posterior error probability (PEP) for each target PSM and decoy PSM.

MS-GF+ MS-GF+ was implemented with the following settings: 10 ppm precursor tolerance; static modifications: methylthio (C); dynamic modifications: deamination (NQ),

oxidation (M). To acquire the total list of identified PSMs, peptides, and proteins, the internal FDR was set to 100%.

S4.6.5 Implementation of database search algorithms on the phospho AML datasets

Byonic, SEQUEST, and Mascot The phospho AML spectra were searched with the following settings: 10ppm precursor tolerance; 0.02 Da fragment tolerance; static modifications: TMT6plex (N-term, K), Carbamindomethyl (C); dynamic modifications: Oxidation (M), Phopho (STY).

S4.6.6 Implementation of database search algorithms on the TNBC dataset

Byonic, SEQUEST, and Mascot The Genistein spectra were searched with the following settings: 20ppm precursor tolerance; 0.02 Da fragment tolerance; static modifications: TMT6plex (N-term, K), Carbamindomethyl (C); dynamic modifications: Oxidation (M), Phopho (STY).

S4.6.7 Implementation of database search algorithms on the non-phospho AML dataset

Byonic, SEQUEST, and Mascot The non-phospho spectra were searched with the following settings: 10ppm precursor tolerance; 0.6 Da fragment tolerance; (digestion enzyme Lys-c - do I need to state if this is reflected in experimental condition) static modifications: TMT6plex (N-term, K), Carbamindomethyl (C); dynamic modifications: Oxidation (M).

S4.6.8 Implementation of Scaffold

We used Scaffold to combine the search results of Byonic, Mascot, SEQUEST, MaxQuant, and MS-GF+ on the proteomics standard. For each combination of database search algorithms, the result files were inputted into Scaffold Q+ (version 4.10.0, Proteome Software

Inc., Portland, OR) to generate peptide and protein identification probabilities. Peptide probabilities were assigned by the Scaffold Local FDR algorithm, protein groups were generated using standard experiment wide protein grouping, and protein probabilities were assigned by the Protein Prophet algorithm [26]. To compare Scaffold with APIR which aims to control FDR at the PSM level, we implemented Scaffold in two ways. In the first implementation, we set both the peptide threshold and the protein threshold to be q FDR, where q is the FDR threshold of APIR. In the second comparison, we set the peptide threshold to be q FDR and varied the protein threshold among all default thresholds: $20\%, 50\%, 80\%, 90\%, 95\%, 99\%, 99.9\%$, 1% FDR, 2% FDR, 3% FDR, 5% FDR and 10% FDR to maximize the number of identified peptides.

S4.6.9 DE peptides analysis of the phospho AML1 dataset

Here we describe how we performed DE analysis on the phospho AML1 dataset. This dataset contains six bone marrow samples: one LSC enriched sample and one LSC depleted sample from patient P5337, two LSC enriched samples and one LSC depleted sample from patient P5340, and one control.

Using all six samples from the phospho AML1 dataset, we first applied APIR to combine the search results by MaxQuant and MS-GF+. Then we applied APIR to adjust the search results of MaxQuant and MS-GF+ separately. Next, we selected fours sample: the two samples from P5337 and the LSC depleted sample from P5340 and one of the two LSC enriched samples from patient P5340, as shown in Fig. 4.11a. We treated the LSC enriched samples and the LSC depleted samples as from two conditions and applied DESeq2 with FDR threshold 5% for DE analysis [29]. We use package `DESeq2` version 1.28.1.

S4.6.10 Theoretical results of APIR

To facilitate our discussion, we start with notations for the mathematical abstraction of APIR, followed by assumptions and proofs.

Let Ω denote the set of all possible PSMs from a tandem MS experiment and $W_k \subset \Omega$

denote the set of target PSMs output by the k-th database search algorithm, $k = 1, \ldots, K$. Let S_k denote the set containing scores of the PSMs in W_k. The exact definition of S_k depends on the implementation of APIR-adjust: Clipper or the pooled approach. Specifically, if APIR-adjust adopts Clipper, $S_k = \{C_i : i \in W_k\} \subset \mathbb{R}$, where C_i is the contrast score of Clipper (See Section 4.2.1). If APIR-adjust adopts the pooled approach, $S_k = \{p_i : i \in W_k\} \subset \mathbb{R}$, where p_i is the p-value calculated using the pooled approach (See Section 4.2.2). We define $\mathcal{W} := \{w : w \subset \Omega\}$ to be the power set of Ω and $\mathcal{S} := \{s : s \subset \mathbb{R}\}$ to be the power set of $\mathbb{R}$.

Here we introduce the mathematical abstraction of APIR-adjust. Given an FDR threshold $q \in (0, 1)$ and a set of target PSMs W with their scores S from a single database search algorithm, we define $\mathcal{P}_q : \mathcal{W} \times \mathcal{S} \to \mathcal{W}$ as an *identification procedure* that takes W and S as input and outputs a subset of W with FDR controlled under q.

Next, we introduce a *selection procedure*, denoted by $\mathcal{Q}$, that finds the index of the "best" set among multiple sets of identified PSMs, where "best" in default APIR means having the most unique peptides. Specifically,

$$\mathcal{Q} : \underbrace{\mathcal{W} \times \cdots \times \mathcal{W}}_{\text{any finite number}} \to \{1, \ldots, K\}$$

takes as input multiple sets of identified PSMs by APIR-adjust, each from a distinct database search algorithm, and outputs the index of the database search algorithm whose set is selected as the best. In case the "best" set is not unique, $\mathcal{Q}$ randomly selects one of the "best" sets and outputs its index.

Then APIR consists of K rounds:

$$\text{Round } 1: \left.\begin{array}{rl} U_{11} := & \mathcal{P}_q\left(W_1, S_1\right); \\ & \vdots \\ U_{1K} := & \mathcal{P}_q\left(W_K, S_K\right); \end{array}\right\} K \text{ sets of identified PSMs}$$

$$J_1 := \mathcal{Q}\left(\{U_{1k} : k = 1, \cdots, K\}\right); \text{ the index of the selected algorithm}$$

$$\vdots$$

$$\text{Round } \ell: \left.\begin{array}{rl} U_{\ell 1} := & \mathcal{P}_q\left(W_1 \backslash \left(\cup_{k'=1}^{\ell-1} W_{J_{k'}}\right), S_1\right); \\ & \vdots \\ U_{\ell K} := & \mathcal{P}_q\left(W_K \backslash \left(\cup_{k'=1}^{\ell-1} W_{J_{k'}}\right), S_K\right); \end{array}\right\} K \text{ sets of identified PSMs}$$

$$J_\ell = \mathcal{Q}\left(\{U_{\ell k} : k \neq J_1, \cdots, J_{\ell-1}\}\right); \text{the index of the selected algorithm}$$

$$\vdots$$

$$\text{Round } K: \left.\begin{array}{rl} U_{K1} := & \mathcal{P}_q\left(W_1 \backslash \left(\cup_{k'=1}^{K-1} W_{J_{k'}}\right), S_1\right); \\ & \vdots \\ U_{KK} := & \mathcal{P}_q\left(W_K \backslash \left(\cup_{k'=1}^{K-1} W_{J_{k'}}\right), S_K\right); \end{array}\right\} K \text{ sets of identified PSMs}$$

$$J_K = \mathcal{Q}\left(\{U_{Kk} : k \neq J_1, \cdots, J_{K-1}\}\right); \text{ the index of the selected algorithm}$$

and outputs $\cup_{\ell=1}^{K} U_{\ell J_\ell}$ as the final set of identified PSMs. Let $V_{\ell k}$ denote the number of false PSMs in $U_{\ell k}$, $\ell, k = 1, \ldots, K$. Because $U_{1J_1}, \ldots, U_{KJ_K}$ are mutually disjoint, to show FDR control we only need to show that

$$\mathbb{E}\left[\frac{\sum_{\ell=1}^{K} V_{\ell J_\ell}}{\left(\sum_{\ell=1}^{K} |U_{\ell J_\ell}|\right) \vee 1}\right] \leqslant q\,, \tag{S4.1}$$

where $a \vee b$ means $\max(a, b)$.

To facilitate our theoretical discussion, we would like to emphasize the source of randomness and how we represent them in our notations. First, both $\{W_k\}_{k=1}^K$ and $\{S_k\}_{k=1}^K$ are random because the shotgun proteomics technology is innately random. Consequently,

the mass spectra from tandem MS experiments could vary in terms of numbers and quality, leading to random lists of target/decoy PSMs and random scores output by database search algorithms. By convention, we use capital letters "W" and "S" to represent a random set of PSMs and a random set of scores respectively. Second, $\mathcal{P}_q$ and $\mathcal{Q}$ could be random or deterministic functions. Third, $\{U_{\ell k} : \ell, k = 1, \cdots, K\}$ are random due to the random input of $\mathcal{P}_q$; therefore, they are also represented by capital U. For similar reasons, $\{J_1\}_{k=1}^K$ are also random. Lastly, as a result, notations such as W_{J_ℓ} and S_{J_ℓ} have two layers of randomness: random PSM sets and scores represented by W and S and random database search algorithm index J_ℓ. Notably, although a capital letter, K is deterministic because it represents the number of database search algorithms we want to aggregate.

Here we introduce assumptions of APIR for FDR control. Conditioning on W_k, let $\mu_k := \mathbb{E}[S_k]$ denote the set of expected scores output by algorithm k. We impose three sets of assumptions respectively on $\{W_k, \mu_k, S_k\}_{k=1}^K$, $\mathcal{P}_q$ and $\mathcal{Q}$.

As for $\{W_k, \mu_k, S_k\}_{k=1}^K$, we require

(A.1) conditioning on $\{W_k, \mu_k\}_{k=1}^K$, $S_1, \ldots, S_K$ are mutually independent. That is, given the set of target PSMs and their expected scores, the observed scores are mutually independent subject to independent sources of randomness.

As for $\mathcal{P}_q$, we assume the following.

(A.2) Conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and given any subset $\widetilde{W}_k \subset W_k$ for any $k = 1, \ldots, K$, we obtain $\mathcal{P}_q(\widetilde{W}_k, S_k)$. Let $\widetilde{U}_k := \mathcal{P}_q(\widetilde{W}_k, S_k)$ and $\widetilde{V}_k$ denote the number of false PSMs in $\widetilde{U}_k$. Then $\mathbb{E}[\widetilde{V}_k/(|\widetilde{U}_k| \vee 1) \mid \{W_k, \mu_k\}_{k=1}^K] \leqslant q$ for all $k = 1, \ldots, K$. That is, $\mathcal{P}_q$ controls FDR when applied to any subset of target PSMs from each of the K database search algorithms. Notably, this assumption is guaranteed for APIR-adjust if the assumptions in Section 4.2.1 hold.

(A.3) Following (A.2), if we assume that $\widetilde{V}_k/(|\widetilde{U}_k| \vee 1)$ is independent of $|\widetilde{U}_k|/(\sum_{k=1}^K |\widetilde{U}_k| \vee 1)$ for $k = 1, \ldots, K$. That is, the FDP of the discoveries, i.e., identified PSMs from a

subset of the target PSMs, from algorithm k is independent of the proportion of the discoveries from algorithm k among all discoveries.

Finally, we assume the following about $\mathcal{Q}$.

(A.4) Conditioning on $\{W_k, \mu_k\}_{k=1}^K$, $\{J_\ell\}_{\ell=1}^K$ is independent of $\{S_k\}_{k=1}^K$. That is, the output of procedure $\mathcal{Q}$ is conditionally independent of the randomness of the scores output by the K algorithms.

We start our proof by first showing that conditioning on $\{W_k = w_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$ and $\{J_k = j_k\}_{k=1}^\ell$,

$$\{J_{\ell+1}, \cdots, J_K\} \perp \frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}. \tag{S4.2}$$

Because $V_{\ell j_\ell}/(|U_{\ell j_\ell}| \vee 1)$ is the FDP of $\mathcal{P}_q(w_{j_\ell} \backslash (\cup_{k'=1}^{\ell-1} w_{j_{k'}}), S_{j_\ell})$, the randomness of $V_{\ell j_\ell}/(|U_{\ell j_\ell}| \vee 1)$ results solely from the randomness of scores in S_{j_ℓ}. By (A.4), S_{j_ℓ} is independent of $\{J_{\ell+1}, \cdots, J_K\}$ conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and $J_1, \ldots, J_\ell$. Equation (S4.2) follows accordingly.

We can then show FDR control in the ℓ-th round conditioning on $\{W_k, \mu_k\}_{k=1}^K$ and $\{J_\ell\}_{\ell=1}^K$:

$$\begin{aligned}
&\mathbb{E}\left[\frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_K = j_K\right] \\
=&\mathbb{E}\left[\frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_K = j_K\right] \\
=&\mathbb{E}\left[\frac{V_{\ell j_\ell}}{|U_{\ell j_\ell}| \vee 1}\middle| W_1 = w_1, \cdots, W_K = w_K, \{\mu_k\}_{k=1}^K, J_1 = j_1, \cdots, J_\ell = j_\ell\right] \\
\leqslant& q\,,
\end{aligned}$$

where the last equality results from (S4.2). The last inequality holds by (A.2).

Finally, we prove (S4.1):

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\sum_{\ell=1}^{K} V_{\ell J_\ell}}{\left(\sum_{\ell=1}^{K}|U_{\ell J_\ell}|\right)\vee 1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{\ell=1}^{K}\frac{|U_{\ell J_\ell}|}{(|U_{1J_1}|+\cdots+|U_{KJ_K}|)\vee 1}\frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}|\vee 1}\middle|\{W_k,\mu_k,J_k\}_{k=1}^{K}\right]\right] \\
&= \mathbb{E}\left[\sum_{\ell=1}^{K}\mathbb{E}\left[\frac{|U_{\ell J_\ell}|}{(|U_{1J_1}|+\cdots+|U_{KJ_K}|)\vee 1}\middle|\{W_k,\mu_k,J_k\}_{k=1}^{K}\right]\mathbb{E}\left[\frac{V_{\ell J_\ell}}{|U_{\ell J_\ell}|\vee 1}\middle|\{W_k,\mu_k,J_k\}_{k=1}^{K}\right]\right] \qquad \text{(S4.3)} \\
&\leqslant \sum_{\ell=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\frac{|U_{\ell J_\ell}|}{(|U_{1J_1}|+\cdots+|U_{KJ_K}|)\vee 1}\middle|\{W_k,\mu_k,J_k\}_{k=1}^{K}\right]\right]\cdot q \\
&\leqslant q\,,
\end{aligned}
$$

where (S4.3) holds as a result of (A.3).

S4.6.11 Post-processing

Master protein recommendation For a given PSM, database search algorithms may disagree on its master protein, causing difficulties in downstream analysis. APIR tackles this issue using a majority vote. Specifically, APIR selects the most frequently reported master protein across database search algorithms for the given PSM. If there is a tie, APIR outputs all tied master proteins.

Post-translational Modification recommendation For a given PSM, how APIR aggregates its modifications across database search algorithms depends on the type of modifications: static or variable. Static modifications occur universally at every instance of a specified amino acid residue or terminus. For example, tandem mass tags occur at every N-termimal. Since static modifications are known and could be specified in the database search process, different database search algorithms will agree in terms of the locations and types of static modifications. Therefore, for any PSM, APIR simply outputs its static modifications by any database search algorithm based on user specification. The default static

modification used by APIR includes cysteine carbamidomethylation and tandem mass tags at N-terminal and lysine.

Unlike static modifications, variable modifications do not apply to all instances of an amino acid residue. For example, phosphorylation typically occurs at only one or few serines in a peptide with many serines. Because variable modifications are hard to detect, database search algorithms may disagree in the types (such as phosphorylation versus oxidation) and/or sites of modifications; however, they always agree on the number of modifications. Suppose database search algorithms report M modifications for the given PSM. To handle these potential disagreements, APIR uses one of the two strategies to recommend variable modifications for a given PSM: PhosphoSitePlus (PSP)-free or PSP-based. In a PSP-free modification recommendation, APIR first counts the number of database search algorithms that report each modification—a combination of modification type and site. Then APIR reports the top M most frequently reported variable modifications. A PSP-based modification strategy is similar to PSP-free except for the handling of tied phosphorylation sites. When there is a tie among phosphorylation sites, APIR reports the most frequently studied phosphorylation sites by searching the literature hits on PhosphoSitePlus(https://www.phosphosite.org/) (PSP), a manually curated and interactive resource for studying protein modifications. In particular, PSP has cataloged and counted existing literature and experiments by phosphorylation. Based on PSP, APIR reports the modification with the highest number of high-throughput literature hits if there is a tie between phosphorylations. If doing so fails to identify a unique modification, APIR compares their numbers of Cell Signaling Technology mass spectrometry studies that found the given phosphorylation and report the highest-numbered phosphorylation. If this fails to provide a unique modification, APIR will report the ties.

Abundance aggregation At the PSM level, APIR first averages a PSM's abundance across database search algorithms. Then APIR performs normalization by scaling a_{ij}, which denotes the averaged abundance of PSM i in channel j, by $10^6/(\sum_i a_{ij})$ so that resulting normalized samples will have total abundance 10^6.

To obtain the abundance at the peptide level, APIR averages the abundance of PSMs containing the same peptide and then performs a scaling across channels such that its cross-channel average equals 100. Specifically, let b_{ij} denotes the averaged abundance of peptide i in sample j. The normalized abundance would be $100b_{ij}/(\sum_j b_{ij})$.

To obtain the abundance at the protein level, APIR averages the abundance of PSMs with the same recommended master protein and then performs the same row normalization as it does at the peptide level.

S4.6.12 Simulation studies

Here we describe how we conducted the simulation studies. Suppose that we have a total of 10^4 mass spectra and that target PSMs and decoy PSMs are ordered in such a way that the i-th target PSM shares the same mass spectrum as the i-th decoy PSM. Among the 10^4 target PSMs, 1500 are true PSMs, and the rest are false. Let T_{i1}, T_{i2}, T_{i3} denote the scores of the i-th target PSM by toy database search algorithm $1, 2, 3$ respectively and D_{i1}, D_{i2}, D_{i3} denote the scores of the i-th decoy PSM. In addition, we generate $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3 \subset \{1, 2, \ldots, 10^4\}$ by randomly sampling without replacement 1000, 2000, and 3000 indices from $\{1, 2, \ldots, 10^4\}$. We generate 200 simulated datasets under either the shared-true-PSMs scenario or the shared-false-PSMs scenario using the following procedures.

Under the shared-true-PSMs scenario, if the i-th target PSM is true, we generate X_i from the exponential distribution with mean 8, Y_i from the exponential distribution with mean 1 and set $T_{i1} = T_{i2} = T_{i3} = X_i$ and $D_{i1} = D_{i2} = D_{i3} = Y_i$; if the i-th target PSM is false, we generate $T_{i1}, T_{i2}, T_{i3}, D_{i1}, D_{i2}, D_{i3}$ independently from exponential with mean 1. Under the shared-false-PSMs scenario, if the i-th target PSM is true, we generate T_{i1}, T_{i2}, T_{i3} independently from exponential with mean 4 and D_{i1}, D_{i2}, D_{i3} independently from exponential with mean 1; if the i-th target PSM is false, we first generate X_i and Y_i independently from the exponential distribution with mean 1 and then set $T_{i1} = T_{i2} = T_{i3} = X_i$ and $D_{i1} = D_{i2} = D_{i3} = Y_i$. Under either scenario, we set T_{ij} to be a missing value if $i \in \mathcal{M}_j$ so that each algorithm captures unique target PSMs.

We examine the actual FDRs of APIR-adjust on each toy database search algorithm and of aggregation methods: union, intersection, and APIR at the FDR threshold $q = 5\%$. For each FDR-control method, we calculate an FDP—the proportion of identified PSMs that are false—on each simulated data and average those 200 FDPs to compute the FDR. To obtain the FDP of APIR-adjust, we apply APIR-adjust (in this case, Clipper because the coverage target proportion is 100%) with the FDR threshold $q = 5\%$ to each toy database search algorithm. To obtain the FDP of union/intersection, we take the union/intersection of the three sets of identified target PSMs by APIR-adjust, one per each toy database search algorithm. To obtain the FDP of APIR, we apply the default APIR to aggregate the three toy database search algorithms with the FDR threshold $q = 5\%$.

Bibliography

[1] Eric D Green et al. “Strategic vision for improving human health at The Forefront of Genomics”. In: *Nature* 586.7831 (2020), pp. 683–692.

[2] Stanislaw Supplitt et al. “Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine”. In: *International Journal of Molecular Sciences* 22.3 (2021), p. 1422.

[3] Ali Khodadadian et al. “Genomics and transcriptomics: the powerful technologies in precision medicine”. In: *International Journal of General Medicine* 13 (2020), p. 627.

[4] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[5] Hirotugu Akaike. “A new look at the statistical model identification”. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.

[6] Gideon Schwarz et al. “Estimating the dimension of a model”. In: *The annals of statistics* 6.2 (1978), pp. 461–464.

[7] Colin L Mallows. “Some comments on C p”. In: *Technometrics* 15.4 (1973), pp. 661–675.

[8] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[9] Jianqing Fan and Runze Li. “Variable selection via nonconcave penalized likelihood and its oracle properties”. In: *Journal of the American statistical Association* 96.456 (2001), pp. 1348–1360.

[10] Cun-Hui Zhang et al. “Nearly unbiased variable selection under minimax concave penalty”. In: *The Annals of statistics* 38.2 (2010), pp. 894–942.

[11] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 301–320.

[12] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1 (2006), pp. 49–67.

[13] John D Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[14] Yoav Benjamini and Yosef Hochberg. "Multiple hypotheses testing with weights". In: *Scandinavian Journal of Statistics* 24.3 (1997), pp. 407–418.

[15] Nikolaos Ignatiadis et al. "Data-driven hypothesis weighting increases detection power in genome-scale multiple testing". In: *Nature methods* 13.7 (2016), pp. 577–580.

[16] Lihua Lei and William Fithian. "Adapt: an interactive procedure for multiple testing with side information". In: *arXiv preprint arXiv:1609.06035* (2016).

[17] Simina M Boca and Jeffrey T Leek. "A direct approach to estimating false discovery rates conditional on covariates". In: *PeerJ* 6 (2018), e6035.

[18] Xinzhou Ge et al. "Clipper: p-value-free FDR control on high-throughput data from two conditions". In: *bioRxiv* (2020).

[19] David Shteynberg et al. "Combining results of multiple search engines in proteomics". In: *Molecular & Cellular Proteomics* 12.9 (2013), pp. 2383–2393.

[20] Tommi Välikangas, Tomi Suomi, and Laura L Elo. "A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1344–1355.

[21] Ruben K Dagda, Tamanna Sultana, and James Lyons-Weiler. "Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics". In: *Journal of proteomics & bioinformatics* 3 (2010), p. 39.

[22] Brian C Searle, Mark Turner, and Alexey I Nesvizhskii. "Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies". In: *The Journal of Proteome Research* 7.1 (2008), pp. 245–253.

[23] Dominique Tessier et al. "Origin of disagreements in tandem mass spectra interpretation by search engines". In: *Journal of proteome research* 15.10 (2016), pp. 3481–3488.

[24] Wei Vivian Li, Yiling Chen, and Jingyi Jessica Li. "TROM: A testing-based method for finding transcriptomic similarity of biological samples". In: *Statistics in biosciences* 9.1 (2017), pp. 105–136.

[25] Jie Lyu et al. "DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features". In: *Science advances* 6.46 (2020), eaba6784.

[26] Lei Li et al. "An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability". In: *Nature Genetics* (2021), pp. 1–12.

[27] Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag Inc, 2009. ISBN: 0-387-95284-5.

[28] Gareth James et al. *An introduction to statistical learning*. Springer, 2013.

[29] Ramón Díaz-Uriarte and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest". In: *BMC bioinformatics* 7.1 (2006), p. 3.

[30] Alexander Statnikov, Lily Wang, and Constantin F Aliferis. "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification". In: *BMC bioinformatics* 9.1 (2008), p. 319.

[31] Javad Salimi Sartakhti, Mohammad Hossein Zangooei, and Kourosh Mozafari. "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)". In: *Computer methods and programs in biomedicine* 108.2 (2012), pp. 570–579.

[32] Bo Xin et al. “Efficient Generalized Fused Lasso and its Application to the Diagnosis of Alzheimer’s Disease.” In: *AAAI*. 2014, pp. 2163–2169.

[33] T Maruthi Padmaja et al. “Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection”. In: *Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on*. IEEE. 2007, pp. 511–516.

[34] Chandree L Beaulieu et al. “FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project”. In: *The American Journal of Human Genetics* 94.6 (2014), pp. 809–817.

[35] Kym M Boycott et al. “Rare-disease genetics in the era of next-generation sequencing: discovery to translation”. In: *Nature Reviews Genetics* 14.10 (2013), p. 681.

[36] Adam Cannon et al. “Learning with the Neyman-Pearson and min-max criteria”. In: *Los Alamos National Laboratory, Tech. Rep. LA-UR* (2002), pp. 02–2951.

[37] Clayton Scott and Robert Nowak. “A Neyman-Pearson approach to statistical learning”. In: *IEEE Transactions on Information Theory* 51.11 (2005), pp. 3806–3819.

[38] Philippe Rigollet and Xin Tong. “Neyman-pearson classification, convexity and stochastic constraints”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2831–2855.

[39] Xin Tong. “A plug-in approach to Neyman-Pearson classification”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3011–3040.

[40] Anqi Zhao et al. “Neyman-Pearson classification under high-dimensional settings”. In: *Journal of Machine Learning Research* 17.213 (2016), pp. 1–39.

[41] Xin Tong, Yang Feng, and Jingyi Jessica Li. “Neyman-Pearson (NP) classification algorithms and NP receiver operating characteristic (NP-ROC) curves”. In: *arXiv preprint arXiv:1608.03109* (2016).

[42] Jianqing Fan, Rui Song, et al. "Sure independence screening in generalized linear models with NP-dimensionality". In: *The Annals of Statistics* 38.6 (2010), pp. 3567–3604.

[43] Jianqing Fan and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5 (2008), pp. 849–911.

[44] Jianqing Fan, Yang Feng, and Rui Song. "Nonparametric independence screening in sparse ultra-high-dimensional additive models". In: *Journal of the American Statistical Association* 106.494 (2011), pp. 544–557.

[45] Runze Li, Wei Zhong, and Liping Zhu. "Feature screening via distance correlation learning". In: *Journal of the American Statistical Association* 107.499 (2012), pp. 1129–1139.

[46] Sihai Dave Zhao and Yi Li. "Principled sure independence screening for Cox models with ultra-high-dimensional covariates". In: *Journal of multivariate analysis* 105.1 (2012), pp. 397–411.

[47] Qin. Mai and Hui Zou. "The Kolmogorov filter for variable screening in high- dimensional binary classification". In: *Biometrika* 100 (2013), pp. 229–234.

[48] Jinyuan Chang, Tang Cheng-Yong, and Yichao Wu. "Local Independence Feature Screening for Nonparametric and Semiparametric Models by Marginal Empirical Likelihood". In: *Annals of Statistics* 44 (2016), pp. 515–539.

[49] Xu Han. "TNonparametric Screening under Conditional Strictly Convex Loss for Ultrahigh Dimensional Sparse Data". In: *Annals of Statistics* (2018+).

[50] Ning Hao and Hao Helen Zhang. "Interaction screening for ultrahigh-dimensional data". In: *Journal of the American Statistical Association* 109.507 (2014), pp. 1285–1301.

[51] Yingying Fan et al. "Innovated interaction screening for high-dimensional nonlinear classification". In: *The Annals of Statistics* 43.3 (2015), pp. 1243–1272.

[52] Jacob Bien, Noah Simon, Robert Tibshirani, et al. “Convex hierarchical testing of interactions”. In: *The Annals of Applied Statistics* 9.1 (2015), pp. 27–42.

[53] Ning Hao, Yang Feng, and Hao Helen Zhang. “Model Selection for High Dimensional Quadratic Regression via Regularization”. In: *Journal of the American Statistical Association* (2017), to appear. URL: http://dx.doi.org/10.1080/01621459.2016.1264956.

[54] Min Zhou et al. “A Statistical Approach to Screening Interaction Effects for Ultra-High Dimensional Data”. In: *arXiv:1902.03525* (2019).

[55] Vladimir Koltchinskii. “Introduction”. In: *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011, pp. 1–16.

[56] J Neyman and ES Pearson. “On the problem of the most efficient tests of statistical inference”. In: *Biometrika A* 20 (1933), pp. 175–240.

[57] Xin Tong et al. “Neyman-Pearson classification: parametrics and power enhancement”. In: *https://arxiv.org/abs/1802.02557* (2018).

[58] J. Audibert and A. Tsybakov. “Fast learning rates for plug-in classifiers under the margin condition”. In: *The Annals of Statistics* 35 (2007), pp. 608–633.

[59] P. Rigollet and R. Vert. “Optimal rates for plug-in estimators of density level sets”. In: *Bernoulli* 15.4 (2009), pp. 1154–1178.

[60] Wolfgang Polonik. “Measuring mass concentrations and estimating density contour clusters-an excess mass approach”. In: *The Annals of Statistics* (1995), pp. 855–881.

[61] E. Mammen and A.B. Tsybakov. “Smooth discrimination analysis”. In: *Annals of Statistics* 27 (1999), pp. 1808–1829.

[62] Thomas Fleischer et al. “Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis”. In: *Genome biology* 15.8 (2014), p. 435.

[63] Ron Edgar, Michael Domrachev, and Alex E Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". In: *Nucleic acids research* 30.1 (2002), pp. 207–210.

[64] HT Lynch, C Snyder, and MJ Casey. "Hereditary ovarian and breast cancer: what have we learned?" In: *Annals of oncology* 24.suppl_8 (2013), pp. viii83–viii95.

[65] Fang Cao et al. "Clinicopathological significance of reduced SPARCL1 expression in human breast cancer". In: *Asian Pacific Journal of Cancer Prevention* 14.1 (2013), pp. 195–200.

[66] Karin Milde-Langosch et al. "Prognostic relevance of glycosylation-associated genes in breast cancer". In: *Breast cancer research and treatment* 145.2 (2014), pp. 295–305.

[67] Graeme I Murray et al. "Profiling the expression of cytochrome P450 in breast cancer". In: *Histopathology* 57.2 (2010), pp. 202–211.

[68] Boon Shing Tan et al. "CYP2S1 and CYP2W1 mediate 2-(3, 4-dimethoxyphenyl)-5-fluorobenzothiazole (GW 610, NSC 721648) sensitivity in breast and colorectal cancer cells". In: *Molecular cancer therapeutics* (2011), molcanther–0391.

[69] Yan Li et al. "Tumoral expression of drug and xenobiotic metabolizing enzymes in breast cancer patients of different ethnicities with implications to personalized medicine". In: *Scientific reports* 7.1 (2017), p. 4747.

[70] Richa Singh, Vikas Yadav, Neeru Saini, et al. "MicroRNA-195 inhibits proliferation, invasion and metastasis in breast cancer cells by targeting FASN, HMGCR, ACACA and CYP27B1". In: *Scientific reports* 5 (2015), p. 17454.

[71] Franz Zehentmayr et al. "Hsa-miR-375 is a predictor of local control in early stage breast cancer". In: *Clinical epigenetics* 8.1 (2016), p. 28.

[72] Li-Xu Yan et al. "MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis". In: *Rna* (2008).

[73] Li Xu Yan et al. "Knockdown of miR-21 in human breast cancer cell lines inhibits proliferation, in vitro migration and in vivo tumor growth". In: *Breast cancer research* 13.1 (2011), R2.

[74] Xuan Pan, Rui Wang, and Zhao-Xia Wang. "The potential role of miR-451 in cancer diagnosis, prognosis, and therapy". In: *Molecular cancer therapeutics* 12.7 (2013), pp. 1153–1162.

[75] Jingyi Jessica Li, Yiling Elaine Chen, and Xin Tong. "A flexible model-free prediction-based framework for feature ranking". In: *Journal of Machine Learning Research* 22.124 (2021), pp. 1–54.

[76] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome biology* 9.9 (2008), pp. 1–9.

[77] Sven Heinz et al. "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". In: *Molecular cell* 38.4 (2010), pp. 576–589.

[78] David N Perkins et al. "Probability-based protein identification by searching sequence databases using mass spectrometry data". In: *ELECTROPHORESIS: An International Journal* 20.18 (1999), pp. 3551–3567.

[79] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[80] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15.12 (2014), p. 550.

[81] Cole Trapnell et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". In: *Nature biotechnology* 31.1 (2013), pp. 46–53.

[82] Jun Li et al. "Normalization, testing, and false discovery rate estimation for RNA-sequencing data". In: *Biostatistics* 13.3 (2012), pp. 523–538.

[83] Thomas J Hardcastle and Krystyna A Kelly. "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data". In: *BMC bioinformatics* 11.1 (2010), pp. 1–14.

[84] Gordon K Smyth. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". In: *Statistical applications in genetics and molecular biology* 3.1 (2004), pp. 1–25.

[85] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[86] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. "multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments". In: *Bioinformatics* 35.17 (2019), pp. 2916–2923.

[87] Mohamed Nadhir Djekidel, Yang Chen, and Michael Q Zhang. "FIND: difFerential chromatin INteractions Detection using a spatial Poisson process". In: *Genome research* 28.3 (2018), pp. 412–422.

[88] Aaron TL Lun and Gordon K Smyth. "diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data". In: *BMC bioinformatics* 16.1 (2015), pp. 1–11.

[89] Bradley Efron and Robert Tibshirani. "Empirical Bayes methods and false discovery rates for microarrays". In: *Genetic epidemiology* 23.1 (2002), pp. 70–86.

[90] Bradley Efron et al. "Empirical Bayes analysis of a microarray experiment". In: *Journal of the American statistical association* 96.456 (2001), pp. 1151–1160.

[91] Matthew Stephens. "False discovery rates: a new deal". In: *Biostatistics* 18.2 (2017), pp. 275–294.

[92] John D Storey and Robert Tibshirani. "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445.

[93] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. "Identifying differentially expressed genes using false discovery rate controlling procedures". In: *Bioinformatics* 19.3 (2003), pp. 368–375.

[94] Bing Yang et al. "Identification of cross-linked peptides from complex samples". In: *Nature methods* 9.9 (2012), pp. 904–906.

[95] James Robert White, Niranjan Nagarajan, and Mihai Pop. "Statistical methods for detecting differentially abundant features in clinical metagenomic samples". In: *PLoS Comput Biol* 5.4 (2009), e1000352.

[96] Andrey A Shabalin. "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". In: *Bioinformatics* 28.10 (2012), pp. 1353–1358.

[97] Stijn Hawinkel et al. "A broken promise: microbiome differential abundance methods do not control the false discovery rate". In: *Briefings in bioinformatics* 20.1 (2019), pp. 210–221.

[98] Ye Zheng and Sündüz Keleş. "FreeHi-C simulates high-fidelity Hi-C data for benchmarking and data augmentation". In: *Nature Methods* 17.1 (2020), pp. 37–40.

[99] Joses Ho et al. "Moving beyond P values: data analysis with estimation graphics". In: *Nature methods* 16.7 (2019), pp. 565–566.

[100] Dongyuan Song and Jingyi Jessica Li. "PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data". In: *bioRxiv* (2020).

[101] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. "Significance analysis of microarrays applied to the ionizing radiation response". In: *Proceedings of the National Academy of Sciences* 98.9 (2001), pp. 5116–5121.

[102] Jesse Hemerik and Jelle J Goeman. "False discovery proportion estimation by permutations: confidence for significance analysis of microarrays". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.1 (2018), pp. 137–155.

[103] Jesse Hemerik, Aldo Solari, and Jelle J Goeman. "Permutation-based simultaneous confidence bounds for the false discovery proportion". In: *Biometrika* 106.3 (2019), pp. 635–649.

[104] Rina Foygel Barber and Emmanuel J Candès. "Controlling the false discovery rate via knockoffs". In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[105] Ery Arias-Castro and Shiyun Chen. "Distribution-free multiple testing". In: *Electronic Journal of Statistics* 11.1 (2017), pp. 1983–2001.

[106] Yoav Benjamini. "Selective inference: The silent killer of replicability". In: *Issue 2.4* 2.4 (2020).

[107] Kristen Emery et al. "Multiple Competition-Based FDR Control and Its Application to Peptide Detection". In: *International Conference on Research in Computational Molecular Biology*. Springer. 2020, pp. 54–71.

[108] Abhishek K Sarkar and Matthew Stephens. "Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis". In: *BioRxiv* (2020).

[109] Rina Foygel Barber, Emmanuel J Candès, et al. "A knockoff filter for high-dimensional selective inference". In: *The Annals of Statistics* 47.5 (2019), pp. 2504–2537.

[110] Jaime Roquero Gimenez and James Zou. "Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization". In: *arXiv preprint arXiv:1810.11378* (2018).

[111] Boris Bogdanow, Henrik Zauber, and Matthias Selbach. "Systematic errors in peptide and protein identification and quantification by modified peptides". In: *Molecular & Cellular Proteomics* 15.8 (2016), pp. 2791–2801.

[112] Michael P Washburn, Dirk Wolters, and John R Yates. "Large-scale analysis of the yeast proteome by multidimensional protein identification technology". In: *Nature biotechnology* 19.3 (2001), pp. 242–247.

[113] Marshall Bern, Yong J Kil, and Christopher Becker. "Byonic: advanced peptide and protein identification software". In: *Current protocols in bioinformatics* 40.1 (2012), pp. 13–20.

[114] Claire R Williams et al. "Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq". In: *BMC bioinformatics* 18.1 (2017), p. 38.

[115] Keegan Korthauer et al. "A practical guide to methods controlling false discoveries in computational biology". In: *Genome biology* 20.1 (2019), pp. 1–21.

[116] ENCODE Project Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489.7414 (2012), pp. 57–74.

[117] Jonathan Thorsen et al. "Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies". In: *Microbiome* 4.1 (2016), p. 62.

[118] Charlotte Soneson and Mark D Robinson. "Bias, robustness and scalability in single-cell differential expression analysis". In: *Nature methods* 15.4 (2018), p. 255.

[119] Oscar Alzate. *Neuroproteomics*. CRC Press, 2009.

[120] John M Koomen et al. "Proteomic contributions to personalized cancer care". In: *Molecular & Cellular Proteomics* 7.10 (2008), pp. 1780–1794.

[121] Mark A Eckert et al. "Proteomics reveals NNMT as a master metabolic regulator of cancer-associated fibroblasts". In: *Nature* 569.7758 (2019), pp. 723–728.

[122] Gali Yanovich et al. "Clinical proteomics of breast cancer reveals a novel layer of breast cancer classification". In: *Cancer research* 78.20 (2018), pp. 6001–6010.

[123] Marjorie L Fournier et al. "Multidimensional separations-based shotgun proteomics". In: *Chemical reviews* 107.8 (2007), pp. 3654–3686.

[124] Jürgen Cox and Matthias Mann. "MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification". In: *Nature biotechnology* 26.12 (2008), pp. 1367–1372.

[125] Sangtae Kim and Pavel A Pevzner. "MS-GF+ makes progress towards a universal database search tool for proteomics". In: *Nature communications* 5 (2014), p. 5277.

[126] Joshua E Elias and Steven P Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". In: *Nature methods* 4.3 (2007), pp. 207–214.

[127] Nathan Edwards, Xue Wu, and Chau-Wen Tseng. "An unsupervised, model-free, machine-learning combiner for peptide identifications from tandem mass spectra". In: *Clinical Proteomics* 5.1 (2009), pp. 23–36.

[128] Kyowon Jeong, Sangtae Kim, and Nuno Bandeira. "False discovery rates in spectral identification". In: *BMC bioinformatics* 13.16 (2012), pp. 1–15.

[129] Kristen Emery et al. "Multiple competition-based FDR control for peptide detection". In: *arXiv preprint arXiv:1907.01458* (2019).

[130] Lukas Käll et al. "Posterior error probabilities and false discovery rates: two sides of the same coin". In: *Journal of proteome research* 7.01 (2008), pp. 40–44.

[131] Lukas Käll et al. "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases". In: *Journal of proteome research* 7.01 (2008), pp. 29–34.

[132] Oliver Serang and William Noble. "A review of statistical methods for protein identification using tandem mass spectrometry". In: *Statistics and its interface* 5.1 (2012), p. 3.

[133] Alexey I Nesvizhskii. "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics". In: *Journal of proteomics* 73.11 (2010), pp. 2092–2123.

[134] Sven Nahnsen et al. "Probabilistic consensus scoring improves tandem mass spectrometry peptide identification". In: *Journal of proteome research* 10.8 (2011), pp. 3332–3343.

[135] David Shteynberg et al. “iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates”. In: *Molecular & cellular proteomics* 10.12 (2011).

[136] Taejoon Kwon et al. “MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines”. In: *Journal of proteome research* 10.7 (2011), pp. 2949–2958.

[137] David C Wedge et al. “FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines”. In: *Journal of proteome research* 10.4 (2011), pp. 2088–2094.

[138] Arun Devabhaktuni et al. “TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets”. In: *Nature biotechnology* 37.4 (2019), pp. 469–479.

[139] Amol Prakash et al. “Bolt: A new age peptide search engine for comprehensive MS/MS sequencing through vast protein databases in minutes”. In: *Journal of The American Society for Mass Spectrometry* 30.11 (2019), pp. 2408–2418.

[140] Dattatreya Mellacheruvu et al. “The CRAPome: a contaminant repository for affinity purification–mass spectrometry data”. In: *Nature methods* 10.8 (2013), pp. 730–736.

[141] Joao A Paulo. “Practical and efficient searching in proteomics: a cross engine comparison”. In: *Webmedcentral* 4.10 (2013).

[142] Yi Fang et al. “Quantitative phosphoproteomics reveals genistein as a modulator of cell cycle and DNA damage response pathways in triple-negative breast cancer cells”. In: *International journal of oncology* 48.3 (2016), pp. 1016–1028.

[143] Simon Raffel et al. “BCAT1 restricts αKG levels in AML stem cells leading to IDH mut-like DNA hypermethylation”. In: *Nature* 551.7680 (2017), pp. 384–388.

[144] Sean J Humphrey, David E James, and Matthias Mann. “Protein phosphorylation: a major switch mechanism for metabolic regulation”. In: *Trends in Endocrinology & Metabolism* 26.12 (2015), pp. 676–687.

[145] Wen-Wei Tsai et al. “TRIM24 links a non-canonical histone signature to breast cancer”. In: *Nature* 468.7326 (2010), pp. 927–932.

[146] Zhibin Cui et al. “TRIM24 overexpression is common in locally advanced head and neck squamous cell carcinoma and correlates with aggressive malignant phenotypes”. In: *PloS one* 8.5 (2013), e63887.

[147] Anna C Groner et al. “TRIM24 is an oncogenic transcriptional activator in prostate cancer”. In: *Cancer cell* 29.6 (2016), pp. 846–858.

[148] Haiying Li et al. “Overexpression of TRIM24 correlates with tumor progression in non-small cell lung cancer”. In: *PloS one* 7.5 (2012), e37657.

[149] Xiao Liu et al. “Overexpression of TRIM24 is associated with the onset and progress of human hepatocellular carcinoma”. In: *PloS one* 9.1 (2014), e85462.

[150] Jianwei Wang et al. “Knockdown of tripartite motif containing 24 by lentivirus suppresses cell growth and induces apoptosis in human colorectal cancer cells”. In: *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* 22.1 (2014), pp. 39–45.

[151] C Li et al. “Knockdown of TRIM24 suppresses growth and induces apoptosis in acute myeloid leukemia through downregulation of Wnt/GSK-3β/β-catenin signaling”. In: *Human & Experimental Toxicology* 39.12 (2020), pp. 1725–1736.

[152] Yan Ye et al. “PI (4, 5) P2 5-phosphatase A regulates PI3K/Akt signalling and has a tumour suppressive role in human melanoma”. In: *Nature communications* 4.1 (2013), pp. 1–15.

[153] Laura J Van't Veer et al. “Gene expression profiling predicts clinical outcome of breast cancer”. In: *nature* 415.6871 (2002), pp. 530–536.

[154] Sang-Uk Han et al. “Loss of the Smad3 expression increases susceptibility to tumorigenicity in human gastric cancer”. In: *Oncogene* 23.7 (2004), pp. 1333–1341.

[155] Patrick Ming-Kuen Tang et al. "Smad3 promotes cancer progression by inhibiting E4BP4-mediated NK cell development". In: *Nature communications* 8.1 (2017), pp. 1–15.

[156] C Liu et al. "MicroRNA-34b inhibits pancreatic cancer metastasis through repressing Smad3". In: *Current molecular medicine* 13.4 (2013), pp. 467–478.

[157] Maj Petersen et al. "Smad2 and Smad3 have opposing roles in breast cancer bone metastasis by differentially affecting tumor angiogenesis". In: *Oncogene* 29.9 (2010), pp. 1351–1361.

[158] Nicholas I Fleming et al. "SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer". In: *Cancer research* 73.2 (2013), pp. 725–735.

[159] Jianfei Xue et al. "Sustained activation of SMAD3/SMAD4 by FOXM1 promotes TGF-β–dependent cancer metastasis". In: *The Journal of clinical investigation* 124.2 (2014), pp. 564–579.

[160] Konstanze Döhner and Hartmut Döhner. "Molecular characterization of acute myeloid leukemia". In: *Haematologica* 93.7 (2008), pp. 976–982.

[161] Raed A Alharbi et al. "The role of HOX genes in normal hematopoiesis and acute leukemia". In: *Leukemia* 27.5 (2013), pp. 1000–1008.

[162] A Renneville et al. "Cooperating gene mutations in acute myeloid leukemia: a review of the literature". In: *leukemia* 22.5 (2008), pp. 915–931.

[163] Antonella Di Costanzo et al. "The HDAC inhibitor SAHA regulates CBX2 stability via a SUMO-triggered ubiquitin-mediated pathway in leukemia". In: *Oncogene* 37.19 (2018), pp. 2559–2572.

[164] M Terol et al. "HBZ-mediated shift of JunD from growth suppressor to tumor promoter in leukemic cells by inhibition of ribosomal protein S25 expression". In: *Leukemia* 31.10 (2017), pp. 2235–2243.

[165] Kristopher R Bosse et al. "Identification of GPC2 as an oncoprotein and candidate immunotherapeutic target in high-risk neuroblastoma". In: *Cancer cell* 32.3 (2017), pp. 295–309.

[166] Hemanth Tummala et al. "DNAJC21 mutations link a cancer-prone bone marrow failure syndrome to corruption in 60S ribosome subunit maturation". In: *The American Journal of Human Genetics* 99.1 (2016), pp. 115–124.

[167] Jia Liu et al. "ZFP36L2, a novel AML1 target gene, induces AML cells apoptosis and inhibits cell proliferation". In: *Leukemia research* 68 (2018), pp. 15–21.

[168] Eisaku Iwanaga et al. "Mutation in the RNA binding protein TIS11D/ZFP36L2 is associated with the pathogenesis of acute leukemia". In: *International journal of oncology* 38.1 (2011), pp. 25–31.

[169] Lijuan Chen et al. "LncRNA MAGI2-AS3 inhibits the self-renewal of leukaemic stem cells by promoting TET2-dependent DNA demethylation of the LRIG1 promoter in acute myeloid leukaemia". In: *RNA biology* 17.6 (2020), pp. 784–793.

[170] Luke W Thomas et al. "CHCHD4 regulates tumour proliferation and EMT-related phenotypes, through respiratory chain-mediated metabolism". In: *Cancer & metabolism* 7.1 (2019), pp. 1–17.

[171] David Ross et al. "Cell-specific activation and detoxification of benzene metabolites in mouse and human bone marrow: identification of target cells and a potential role for modulation of apoptosis in benzene toxicity." In: *Environmental health perspectives* 104.suppl 6 (1996), pp. 1177–1182.

[172] William B Slayton et al. "The first-appearance of neutrophils in the human fetal bone marrow cavity". In: *Early human development* 53.2 (1998), pp. 129–144.

[173] Diane G Schattenberg et al. "Peroxidase activity in murine and human hematopoietic progenitor cells: potential relevance to benzene-induced toxicity." In: *Molecular pharmacology* 46.2 (1994), pp. 346–351.

[174] Michelle W Wong-Brown et al. "Prevalence of BRCA1 and BRCA2 germline mutations in patients with triple-negative breast cancer". In: *Breast cancer research and treatment* 150.1 (2015), pp. 71–80.

[175] DG Evans et al. "Prevalence of BRCA1 and BRCA2 mutations in triple negative breast cancer". In: *Journal of medical genetics* 48.8 (2011), pp. 520–522.

[176] E Comen et al. "Relative contributions of BRCA1 and BRCA2 mutations to "triple-negative" breast cancer in Ashkenazi Women". In: *Breast cancer research and treatment* 129.1 (2011), pp. 185–190.

[177] C Villarreal-Garza et al. "The prevalence of BRCA1 and BRCA2 mutations among young Mexican women with triple-negative breast cancer". In: *Breast cancer research and treatment* 150.2 (2015), pp. 389–394.

[178] Rachel Greenup et al. "Prevalence of BRCA mutations among women with triple-negative breast cancer (TNBC) in a genetic counseling cohort". In: *Annals of surgical oncology* 20.10 (2013), pp. 3254–3258.

[179] Johanna Tommiska et al. "The DNA damage signalling kinase ATM is aberrantly reduced or lost in BRCA1/BRCA2-deficient and ER/PR/ERBB2-triple-negative breast cancer". In: *Oncogene* 27.17 (2008), pp. 2501–2506.

[180] Toshiyasu Taniguchi et al. "Disruption of the Fanconi anemia–BRCA pathway in cisplatin-sensitive ovarian tumors". In: *Nature medicine* 9.5 (2003), pp. 568–574.

[181] Carmen J Marsit et al. "Inactivation of the Fanconi anemia/BRCA pathway in lung and oral cancers: implications for treatment and survival". In: *Oncogene* 23.4 (2004), pp. 1000–1004.

[182] Helong Zhao et al. "Endothelial Robo4 suppresses breast cancer growth and metastasis through regulation of tumor angiogenesis". In: *Molecular oncology* 10.2 (2016), pp. 272–281.

[183] Rebecca Marlow et al. “Vascular Robo4 restricts proangiogenic VEGF signaling in breast”. In: *Proceedings of the National Academy of Sciences* 107.23 (2010), pp. 10520–10525.

[184] Steven Suchting et al. “Soluble Robo4 receptor inhibits in vivo angiogenesis and endothelial cell migration”. In: *The FASEB Journal* 19.1 (2005), pp. 121–123.

[185] Xiaodong Zhuang et al. “Robo4 vaccines induce antibodies that retard tumor growth”. In: *Angiogenesis* 18.1 (2015), pp. 83–95.

[186] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[187] Minoru Kanehisa. “Toward understanding the origin and evolution of cellular organisms”. In: *Protein Science* 28.11 (2019), pp. 1947–1951.

[188] Minoru Kanehisa et al. “KEGG: integrating viruses and cellular organisms”. In: *Nucleic Acids Research* 49.D1 (2021), pp. D545–D551.

[189] JA Costoya, RM Hobbs, and PP Pandolfi. “Cyclin-dependent kinase antagonizes promyelocytic leukemia zinc-finger through phosphorylation”. In: *Oncogene* 27.27 (2008), pp. 3789–3796.

[190] CH Yam, TK Fung, and RYC Poon. “Cyclin A in cell cycle control and cancer”. In: *Cellular and Molecular Life Sciences CMLS* 59.8 (2002), pp. 1317–1326.

[191] Ida RK Bukholm, Geir Bukholm, and Jahn M Nesland. “Over-expression of cyclin A is highly associated with early relapse and reduced survival in patients with primary breast carcinomas”. In: *International journal of cancer* 93.2 (2001), pp. 283–287.

[192] Marcos Malumbres and Mariano Barbacid. “Cell cycle, CDKs and cancer: a changing paradigm”. In: *Nature reviews cancer* 9.3 (2009), pp. 153–166.

[193] Erica K Cassimere, Claire Mauvais, and Catherine Denicourt. “p27Kip1 is required to mediate a G1 cell cycle arrest downstream of ATM following genotoxic stress”. In: *PLoS One* 11.9 (2016), e0162806.

[194] Byung-Kwon Choi et al. “WIP1 dephosphorylation of p27Kip1 Serine 140 destabilizes p27Kip1 and reverses anti-proliferative effects of ATM phosphorylation”. In: *Cell Cycle* 19.4 (2020), pp. 479–491.