Outcome-Guided Disease Subtyping and Power Calculation for High-Dimensional Omics Studies

by

Peng Liu

M.S. in Biostatistics, University of Pittsburgh, 2016

B.M. in Preventive Medicine, Sun Yat-sen University, China 2012

Submitted to the Graduate Faculty of

the Graduate school of Public health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Peng Liu

It was defended on

June 17, 2021

and approved by

George C. Tseng, ScD, Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Lu Tang, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Danial E. Weeks, PhD, Professor, Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh

Yongseok Park, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Copyright © by Peng Liu
2021

Outcome-Guided Disease Subtyping and Power Calculation for High-Dimensional Omics Studies

Peng Liu, PhD

University of Pittsburgh, 2021

With the rapid advancement of high-throughput technologies, a large amount of high-dimensional data has been generated in the public domain, which gives rise to various statistical and computational challenges in the design and analysis of omics experiments. This proposal focuses on addressing disease subtyping (Chapters 2&3) and power calculation issues (Chapter 4) in the analysis of high-dimensional omics studies.

In Chapter 2, we proposed an outcome-guided disease subgrouping framework called ogClust. Disease subtyping by omics data usually applies conventional clustering methods, which primarily concerns identifying subpopulations with similar patterns in gene features. Since outcome information is not considered in clustering, the identified disease subtypes are often not associated with the outcome. ogClust uses a continuous or survival clinical outcome to guide disease subtypes, which identifies disease subtypes with their driving genes, and guarantees that the resulting subtypes are associated with disease of interest.

In Chapter 3, we extended the ogClust model by integrating multi-omics data and incorporating biological information via the sparse overlapping group lasso to improve the accuracy and interpretability of feature selection and disease subtyping. An EM algorithm with alternating direction method of multiplier (ADMM) approach is applied for fast optimization.

In Chapter 4, we proposed a power calculation and study design method "MethylSeqDesign" for bisulfite DNA methylation sequencing (Methyl-Seq) studies. A three sequential steps power calculation method is designed to perform genome-wide power calculation and simultaneously consider sample size and sequencing depth. The performance of the method was evaluated with extensive simulations. Two real examples are analyzed to illustrate our approach.

Contribution to public health:

The methods proposed in Chapters 2 & 3 are useful for identifying outcome-associated clusters that are more likely to have distinct biological mechanisms or clinical significance, which is an essential first step towards precision medicine. The proposed method in Chapter 4 provides a useful tool to perform genome-wide power calculation and study design for Methyl-Seq studies.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation is made as a completion of the Doctoral degree in Biostatistics at the University of Pittsburgh. My journey as a PhD student has come to an end, while the journey of my academic career has just begun. When I look back upon my life, I see a shy and smiling boy, thanks to my parents who gave me a childhood free of worries and pressure. I was not the most talented nor popular student in the class. I spent my teenage years studying and dreaming, graduated, worked, married, and study abroad like many others.

I never thought I could pursue a career in academia until I met my advisor, Professor George C. Tseng, with whom I spent very challenging and fruitful five years. I would sincerely thank my advisor for your tremendous help and support along the way. There were a couple of times that are extremely difficult, especially in the years when my first kid was born and I had no publication. You patiently guided me to transform from a student to a dedicated researcher and mentored me on parenting and taking care of my family. Considering I have a family to support, you have generously provided extra personal and financial help. You are a fantastic advisor in my academic career and my life.

I want to extend my thanks to co-advisor Dr. Lu Tang for your valuable advice and timely support. It is such an enjoyable experience discussing academic questions with you. Additionally, I would like to thank Dr. Yongseok Park, Dr. Zhao Ren, and my previous and current labmates for their constructive comments and valuable contribution to this dissertation. Besides, I would like to thank Dr. Liza Konnikova, Dr. Pamjeet Randhawa, Dr. Kunhong Xiao, and Dr. Jianhua Luo for our productive collaborations. Furthermore, I appreciate all my committee members Prof. George C. Tseng, Dr. Lu Tang, Dr. Yongseok Park, and Prof. Daniel E. Weeks, for your time and valuable input.

Finally, I am so grateful for having my beloved family: my wife Shan Wu, and my two adorable kids Alex and Shana. Thanks my dear wife, for always being so supportive and holding half sky of our family. Thanks Alex, for loving mom and dad and being a good big brother. And thanks Shana as your birth has brought us so much joy and happiness. How blessed I am to have a place of comfort and being genuinely loved.

1.0 Introduction

1.1 Overview of High-Throughput Omics Data and Technologies

The rapid advancement of high-throughput technology has revolutionized the ability to parallelly measure a large number of molecules at an unprecedented resolution. The technologies have been widely used in various omics experiments, including genomics, transcriptomics, and epigenomics, which correspond to the global assessment of DNA, RNA, or epigenomic changes. These omics experiments are essential in creating a comprehensive understanding of biological mechanisms, revealing molecular subtypes, identifying biomarkers and developing targeted therapies in clinical practice. With the increasing number of omics studies that have been performed and large amount of data that have been accumulated in the public domain, there are emerging statistical and computational challenges in the design and analysis of omics studies.

1.1.1 Different types of omics data

Transcriptomics is the study of all types of RNA transcripts in an organism (transcriptome), including messenger RNA (mRNA), micro RNA(miRNA), and other non-coding RNAs. mRNA carries the protein-coding information transcribed from DNA. miRNA functions in post-transcriptional gene expression regulation. Other non-coding RNAs such as Long non-coding RNA(lncRNA) regulate the transcription of other RNAs. The study of transcriptome enables to characterize the transcriptional activity and study the gene expression profiles.

Microarray and RNA seq/scRNAseq are the two primary high throughput techniques to measure mRNA expression levels. Transcriptomics data has been used for disease diagnosis and prognosis, molecular subtypes identification and biomarkers detection (Wang et al., 2009; Raghavachari and Garcia-Reyero, 2018).

Epigenomics is the global study of the complete set of epigenetic modifications of DNA

or histones bind to DNA without change of nucleic acid sequence. DNA methylation and histone modification are the two most characterized epigenomic processes. Both of them act to regulate gene expression. Chapter 4 of this dissertation focuses on DNA methylation, which is a reversible epigenetic modification of DNA nucleotides where methyl groups are added to DNA molecules. DNA methylation is one of the best characterized and most studied epigenetic markers, which controls gene expression in normal cell development and abnormal biological processes such as cancer. Particularly in gene promoter regions, hyper-methylation is shown closely related to silencing gene expressions. In mammals, such as human, DNA methylation happens almost exclusively at cytosine site that follows with guanine known as CpG site. There are tens of thousands of regions with a high frequency of CpG sites in the whole genome that are classified as CpG islands, which typically exist at or near the transcription starting sites of genes. DNA methylation process has been found to link to many important biological processes, such as genomic imprinting, X-chromosome inactivation, repression of repetitive elements, aging and carcinogenesis (Li et al., 1993; Paulsen and Ferguson-Smith, 2001; Robertson, 2005). In cancer studies, aberrant DNA methylation changes are considered as one of the leading factors in developing tumors (Esteller, 2005; Baylin, 2005; Delpu et al., 2013; Licht, 2015).

Other omics studies that are not the focus of this dissertation include genomics, proteomics, and metabolomics. Genomics studies the complete genetic complement of an organism, focusing on the analysis of structure, function, and variation of the genome. Proteomics and metabolomics study the complete set of proteins and metabolites in a cell or organism, respectively.

1.1.2 Microarray and next generation sequencing (NGS)

Microarray is a technology where a large set of oligonucleotide or cDNA probes are attached on a solid support (often referred to as a "chip"), and hybridize with the unknown fragment of sequences. Then fluorescence intensity is measured to quantify the abundance of target molecules. Microarray has a broad range of applications, such as gene expression analysis and genetic variation discovery (Brown and Botstein, 1999). However, it has two

major limitations: the microarray design depends on the prior knowledge of the genome or epigenome features, which hinders the discovery of novel genetic variation. In addition, the cross-hybridization across similar sequences gives high level of background noise and complicates the analysis of the related features (Hurd and Nelson, 2009).

The NGS technology offers solutions to the above problems by directly sequencing molecules at single-nucleotide resolution without probe hybridization, allowing detection of novel genetic variations and removing the cross-hybridization issue. NGS is capable of sequencing hundreds of millions of molecules in parallel. With the cost of sequencing drops quickly to $\$1,000$ or less for an entire human genome, NGS quickly gains its popularity in the last decades, overtaking micro-array. It has been widely used in genomics(Koboldt et al., 2013), transcriptomic(Wang et al., 2009), methylation alterations(Ku et al., 2011), and chromatin immunoprecipitation (Mardis, 2008).

1.1.3 Statistical challenges for analyzing high throughput omics data

The availability of omics data brings statistical and computational challenges and opportunities to analyze, integrate and interprate the data. This dissertation will focus on the following statistical issues: 1) The statistical community has been seeking to study complex diseases by identifying disease subtypes with heterogeneous molecular profiles and disease mechanisms, such that treatment can be tailored to improve disease prognosis. 2) Integrate multi-omics data to provide a holistic molecular view of the biological problem, improving interpretability and power. 3) Additionally, genome-wide power calculation for high-throughput sequencing experiments is crucial for an adequate study design and successful data analysis. In the following sections of the introduction, we will briefly introduce the existing statistical methods of disease subtyping, integrative analysis, and power/sample size calculation for high-dimensional omics data.

1.2 Overview of Disease Subtyping Methods

The increasing number of omics data generated has provided new opportunities for unveiling underlying disease subtypes with heterogeneous molecular patterns and disease mechanisms for many complex diseases. Such disease subtyping by omics data has become a practical approach in identifying clinically relevant subtypes with tailored therapy towards precision medicine. This section will give an overview of the disease subtyping methods.

1.2.1 Clustering methods

In the literature, disease subtyping by omics data usually applies conventional clustering methods, which primarily concerns the identification of subpopulations with similar patterns in gene features. Popular methods, including sparse K means (Witten and Tibshirani, 2010), penalized model-based clustering (Pan and Shen, 2007), can effectively select gene features and perform sample clustering simultaneously. Since outcome information is not considered in clustering, the identified disease subtypes are often not associated with the outcome. In the literature, Bair and Tibshirani (2004) and Koestler et al. (2010) have developed a two-stage semi-supervised method, where K-means or other standard clustering methods is applied to the top M selected features with the highest marginal outcome association. The two-stage approach is suboptimal in that not all outcome-associated features are good descriptors of the desired subtypes, and this type of approach is inherently ad hoc in selecting the top features. Ahmad and Fröhlich (2017) proposed a Bayesian method to cluster omics data with survival outcomes and molecular features. Wang et al. (2020b) proposed a supervised convex clustering algorithm. However, these two methods are computationally intensive and only affordable up to ~ 100 genes. In Chapter 2&3 of this dissertation, we proposed outcome guided clustering methods for high dimensional omics data.

1.2.2 Latent class models

The latent class models in the literature (see Vermunt and Magidson (2003); Dean and Raftery (2010) for review) has the following two categories: one popular category of latent

class analysis (Lanza and Rhoades, 2013) is equivalent to unsupervised model-based clustering and such clustering does not meet our purpose as we previously discussed. Another set of latent class model (Houseman et al., 2006; DeSantis et al., 2008; Desantis et al., 2012) links outcome with latent class variable via a finite mixture model. Set Y as the outcome variable and $\boldsymbol{X}$ as a vector of covariates that are associated with outcome Y. $k(1 \leq k \leq K)$ denotes the latent classes and $i(1 \leq i \leq n)$ denotes the subjects. A typical probability density function is:

$$f\left(y_i; \boldsymbol{x}_i\right) = \sum_{k=1}^{K} \pi_k f_k\left(y_i; \boldsymbol{x}_i\right)$$

where π_k is the mixing probability for subgroup k, and $\sum_{k=1}^{K} \pi_k = 1$. While without variables (signature) to characterize the cluster membership, this type of model is incapable of classifying future patients into the subtypes. On the other hand, the generative latent class model have been developed Dayton and Macready (1988); Bandeen-Roche et al. (1997); Guo et al. (2006) and are widely applied in social sciences, psychology and public health. These models are in the following form:

$$f\left(y_i; \boldsymbol{x}_i\right) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_i) f_k\left(y_i; \boldsymbol{x}_i\right)$$

where $\pi_k(\boldsymbol{x}_i) = \frac{f(z_i=k|\boldsymbol{x}_i)}{\sum_{l=1}^{K} f(z_i=l|\boldsymbol{x}_i)}$, and z_i is the class label for subject i. It has been extended for various applications (Larsen, 2004; Lin et al., 2002) and for joint latent class modelling for survival and longitudinal data (Lin et al., 2002; Proust-Lima and Taylor, 2009; Proust-Lima et al., 2014; Furgal et al., 2019; Sun et al., 2019a). A shortcoming of these methods is the low dimensionality of clinical covariates used to characterize subtypes (normally less than a dozen even if variable selection is applied) and the lack of extensibility for various biological scenarios needed.

1.2.3 Precision medicine in clinical trials

There are three main categories of subgrouping methods in clinical trials for precision medicine. The first category of the methods identify patients with beneficial treatment effect

by maximizing the overall clinical benefit. A popular type of methods is individualized treatment regime (ITR) (Cai et al., 2011; Zhao et al., 2012, 2013; Shi et al., 2018), which aims to identify the optimal treatment regime given a patient by maximizing the value function of clinical benefit. Zhao et al. (2012) proposed the outcome weighted learning method, which transforms the value function maximization problem into a weighted classification problem, and many publications follow this framework (Xu et al., 2015; Zhu et al., 2017). The second category of subgrouping methods in clinical trials is conditional outcome modeling (Imai et al., 2013; Foster et al., 2011; Sugasawa and Noma, 2019), which estimates individualized treatment effect by modeling potential outcomes under either treatment, then stratify the overall population based on the individualized treatment effect. Lastly, the third type of subgrouping method in clinical trials directly estimate the covariate and treatment interaction, bypassing estimating the main effect of treatment and baseline characteristics. Su et al. (2008) proposed a tree-based method to identify subgroups by exploring interaction structure in survival analysis. Tian et al. (2014) proposed a method to directly model the interactions between treatment and covariates, without modeling the main effect. These methods are limited to randomized clinical trials and not applicable to the general disease subtyping. In addition, they identify subgroups with different treatment effect using clinical variables, and often can not handle high-dimensional omics data.

1.3 Data Integration and Meta-Analysis

A single omics data set is limited by its restricted biological information it contains and sample size it has. Meta-analysis methods integrating multi-studies and multi-omics data are attractive to improve the model's power, discovery, and interpretability. The integrative methods for omics data can be categorized into horizontal (across studies) and vertical (across omics modalities) integration analysis methods (Tseng et al., 2015).

1.3.1 Horizontal integration

Horizontal integration unites information from different sources or studies of the same type of data to improve power and reproducibility. This approach has been used in GWAS studies to strengthen the power to identify susceptible loci (Fritsche et al., 2013; Liu et al., 2014), and in the analysis of biological networks to study gene regulation in complex disease (Li et al., 2015). In transcriptomics studies, it has been used for biomarker detection, pathway analysis, subtyping, and differential expression analysis (Ma et al., 2019). For disease subtyping, Planey and Gevaert (2016) proposed a method to identify patient subtypes across multiple studies by measuring similarity between the study-specific clusters. Huo and Tseng (2017) has proposed the meta-analytic framework for sparse K-means to identify disease subtypes integrating multiple studies.

1.3.2 Vertical integration

Vertical integration unites information across multiple omics data types. The statistical methods for the vertical integration can be characterized as unsupervised (e.g. multi-omics clustering) and supervised(regression based) approaches. Shen et al. (2009) proposed a penalized latent variable model-based clustering method, iCluster, for joint modeling of multiple types of omics data. It assumes a consistent clustering across multi-level omics data, which may not hold in some cases. JIVE(Lock et al., 2013) performs decomposition of variation into joint and individual components, allowing common and omic-specific molecular profiling structures. Lock and Dunson (2013) fitted a finite Dirichlet mixture model to perform Bayesian consensus clustering (BCC), which extends the JIVE modeling strategy within a Bayesian framework. Kim et al. (2017) improves feature selection of iCluster by incorporating prior knowledge of inter-omics regularization. It also allows scattered samples to achieve tight clustering. Huo and Tseng (2017) proposed integrative sparse K-means approach to integratively cluster multi-omics data with feature selection and incorporating biological information.

Mankoo et al. (2011), Zhao et al. (2015) and Jiang et al. (2016) has conducted supervised integration of TCGA multi-omics data using additive multivariate cox model to predict

cancer prognosis, results show improved prediction of survival response. Wang et al. (2013) proposed integrative Bayesian method(iBAG) to associate the gene expression and DNA methylation with patient's survival outcome. These supervised integrative methods focus on prediction without considering subgroups of patients. Chapter 3 of the dissertation cover a new vertical integrative method for subgrouping as well as prediction.

1.4 Power Calculation Methods for High Dimensional Omics Data

Due to rapid development of NGS techniques and dropping prices, an increasing number of omics experiments have been performed. As the cost for sequencing is still substantial in terms of budget, power calculation is essential for a well-designed study and analysis. It is desirable to develop easy-to-use study design and power calculation tools for high-dimensional omics experiments.

The conventional methods of power calculation focus on a single gene, meaning that the statistical power is estimated for one gene given type-I error, sample size and effect size. However, in high-throughput studies, statistical power should be considered simultaneously for thousands of genes, where genome-wide power and genome-wide type I error should be defined and calculated. As a consequence, in the literature, the conservative family-wise error rate (FWER) and the scientifically more applicable false discovery rate (FDR;**?**) were suggested to replace the type-I error α. Lee and Whitmore (2002) first discussed the significance of measuring the power and sample size for microarray data and offered a method for regulating FWER based on ANOVA. Since then, several other methods for FWER control of microarray power calculation have been suggested (Jung, 2005; Dobbin and Simon, 2005; Jung and Young, 2012). In addition, Ferreira and Zwinderman (2006), Liu and Hwang (2007) and van Iterson et al. (2009) incorporated the FDR principle and used pilot data to account for the more practical estimation of power in the genome-wide scenario. Gadbury et al. (2004a) developed the concept of the predicted discovery rate (EDR) for genome-wide detection power in replace of $1 - \beta$ in univariate case.

Unlike earlier fluorescence-based technologies such as microarray, modeling of NGS data

should consider count data. In addition, both sequencing depth and sample size play important role in power calculation and study cost. For DNA sequencing, power calculaton methods have been developed to detect association of common and rare variants in GWAS studies, somatic mutation, and heterozygous variant (Li et al., 2018). For RNA-seq analysis, several power calculation methods have been proposed for differential expression analysis between two groups. Wu et al. (2015) have proposed the simulation based method for power calculation, stratified by sequencing depth. Lin et al. (2019) has proposed the RNASeqDesign method, where sequencing depth is one dimension in power calculation. In Methyl-Seq studies, however, the complexity and large scale of methylation data brings statistical challenges for power calculation. To the best of our knowledge, there is no well-designed power calculation methods for differential methylation analysis. We will focus on power calculation for Methyl-Seq studies in Chapter 4.

1.5 Motivation and Overview of this Dissertation

My dissertation contains five chapters. Chapter 1 includes a general overview of omics data and experimental technologies, the motivation and methods for disease subtyping, integrative analysis and power calculation methods for high-dimensional omics data. These contents serve as the background knowledge for the methodology development for Chapter 2, 3 and 4. Figure 1 gives a graphical overview of this disserataion.

In Chapter 2, we proposed a unified latent generative model to perform disease subtyping constructed from omics data with outcome guidance, which improves the resulting subtypes concerning the disease of interest. Feature selection is embedded in a regularization regression. A modified EM algorithm is applied for numerical computation and parameter estimation. The proposed method performs feature selection, latent subtype characterization and outcome prediction simultaneously. To account for possible outliers or violation of mixture Gaussian assumption, we incorporate robust estimation using adaptive Huber or median-truncated loss function. Extensive simulations and an application to complex lung diseases with transcriptomic and clinical data demonstrate the ability of the proposed

method to identify clinically relevant disease subtypes and signature genes suitable to explore toward precision medicine.

In Chapter 3, we further extended the model to jointly incorporate pathway information and integrate multi-omics data via sparse overlapping group lasso, such that the prior biological knowledge can be used to guide the disease subtyping. An algorithm using an alternating direction method of multipliers(ADMM) is applied for optimization. Simulations and real data applications will be applied performed to compare the performance with existing methods such as IS-Kmeans.

In Chapter 4, we proposed a power calculation and study design method MethylSeqDesign for DNA methylation studies, which is inspired by our previous publication Lin et al. (2019). The proposed method utilizes pilot data for power calculation and experimental

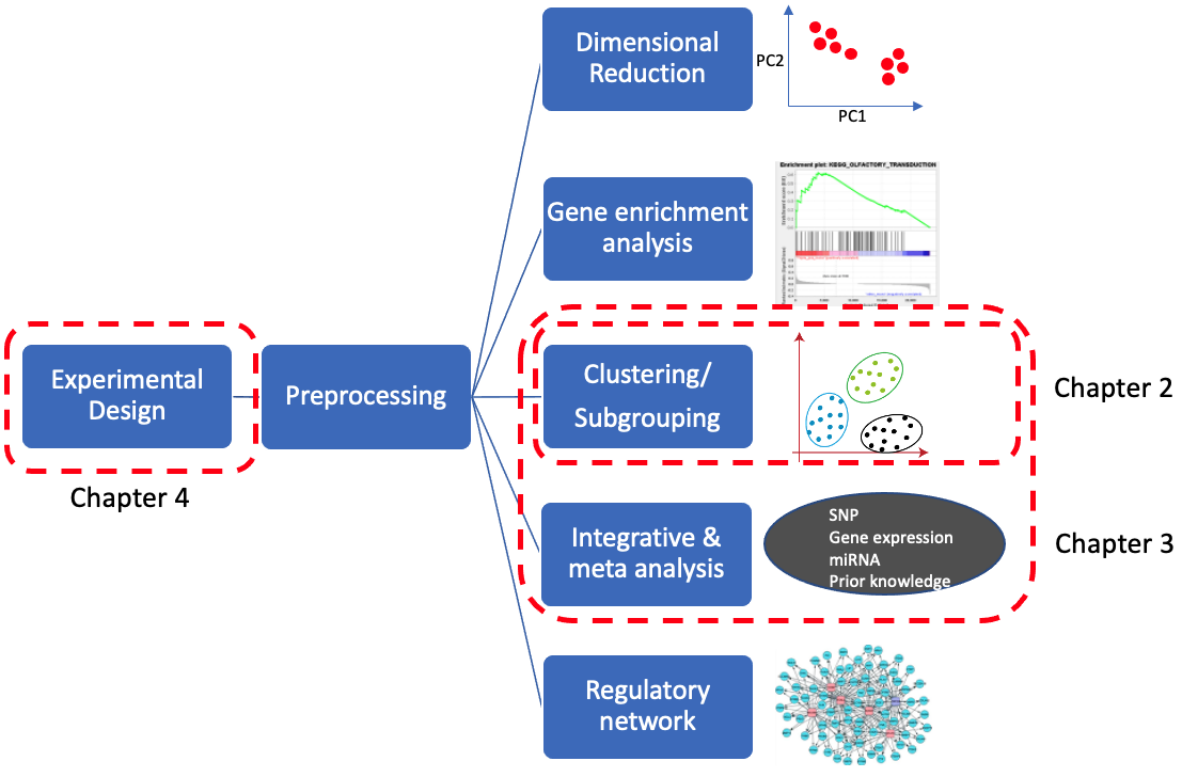


Figure 1: A graphical illustration of the overall structure of this dissertation.

design for Methyl-Seq experiments. The approach is based on a mixture model fitting of p-value distribution from pilot data and a parametric bootstrap procedure based on approximated Wald test statistics to infer genome-wide power for optimal sample size and sequencing depth. The performance of the method was evaluated with simulations. Two real examples are analyzed to illustrate our method.

Chapter 5 includes discussion and future work. We are interested in extending our current ogClust framework to use the guide of multivariate or even multi-types of outcomes(e.g., continuous, survival, and categorical). We are also interested in performing joint modeling to adjust the level of outcome guidance in subtyping. In addition, it is attractive to capture subtype-specific network dynamics via a sparse Gaussian graphical model.

2.0 Outcome-Guided Disease Subtyping for High-Dimensional Omics Data

2.1 Introduction

[1]Many complex diseases were once considered a single disorder, within which all patients receive a uniform screening, diagnosis and treatment strategy. With better understanding of the underlying disease mechanisms, evidences have emerged to define novel subtypes of many complex diseases using clinical variables, selected biomarkers, imaging measurements, molecular profiling or genetic alterations, where the therapeutic plan can be tailored to each subtype to improve disease prognosis. In breast cancer, for example, four intrinsic subtypes (Lumina A, Lumina B, HER2-enriched and Basal-like) and a Normal Breast-like group were first identified in Perou et al. (2000) by cluster analysis of 42 patients based on microarray expression profile of 8102 genes and the result has been validated in many follow-up studies. Of the subtypes, Lumina A and Lumina B patients tend to have longer survival and lower recurrence rate, which require less aggressive treatment to reduce side effects. Basal-like (triple negative) tumors are often more malignant and have a poorer prognosis but can be successfully treated with certain combinations of surgery, radiotherapy and chemotherapy. HER2-enriched patients can be treated with HER2-targeted therapy such as trastuzumab, which is surprisingly harmful to those in the Lumina subtypes. Subsequent tailored screening/prevention programs and novel treatment strategies from successful disease subtyping have decreased breast cancer mortality over the years (Jemal et al., 2009a). Cluster analysis in high-dimensional omics data to characterize novel disease subtypes is an essential first step towards precision medicine and is the focus of this chapter.

Classical clustering methods, such as hierarchical clustering, K-means clustering and Gaussian mixture model, have been widely used in the literature for disease subtyping. These methods are effective when the dimension of features is low and the clusters are well separated. The clustering task, however, becomes more challenging in high-dimensional omics data (e.g., thousands of genes in transcriptomic data) and the classical methods often

[1]This chapter has been submitted to Annuals of Applied Statistics.

fail to identify clinically meaningful clusters since they naively treat all features as equally important. Similar to most small-n-large-p problems, it is generally believed that only a small portion of features are relevant in the cluster characterization. A large amount of work has been devoted to dimension reduction and feature selection in cluster analysis, such as sparse principal component analysis or sparse factor analysis coupled with standard clustering (Zou et al., 2006; Bair et al., 2006), model-based clustering with variable selection (Tadesse et al., 2005; Pan and Shen, 2007) and sparse K-means (Witten and Tibshirani, 2010). Interested readers may refer to Bouveyron and Brunet-Saumard (2014) for further references.

Although the aforementioned methods are powerful to simultaneously identify clusters and relevant features, the resulting clusters of patients may not guarantee biological meaning or clinical impact. A common practice is to perform post-hoc analyses to assess association between the identified clusters and disease relevant measures or clinical outcomes, such as survival. Such association justifies potential clinical relevance of the novel disease subtypes and supports further investigation. However, if no association is observed, the cluster analysis is considered a failed effort to bring clinical impact. In the clustering of high-dimensional omics data, the latter situation happens frequently since decision of final clusters largely depends on the selected features. The data may contain multi-faceted cluster structures that can be defined by different sets of gene features. In Figure 2, we demonstrate this phenomenon using a lung disease transcriptomic dataset. When we select the top 50 X/Y chromosome genes (annotated in the GeneCards database; www.genecards.org) that are most associated with the gender variable and perform simple K-means, Figure 2A identifies two clear male/female clusters. Similarly, if the top 50 genes associated with the age variable are selected from age-related genes annotated in the HAGR database (Tacutu et al., 2018), Figure 2B finds three clusters of young, middle-aged and old patients through K-means clustering. Although heatmaps in Figures 2A and 2B show well-separated clusters, they are not novel for the clinical purpose of disease subtyping. Figure 2C shows result of the proposed outcome-guided clustering method to be introduced. With guidance from the clinical outcome FEV1 (measuring the volume of air a person can exhale during the first second of forced expiration), three clusters of patients are identified with distinct clinical behavior and molecular mechanisms (see Chapter 2.4 for detailed results). When gene signals

are largely driven by potentially disease-irrelevant factors (e.g., as in Figures 2A and 2B), genes that are directly relevant to the disease with greater clinical potential (e.g. Figure 2C) are less likely to be uncovered. In the literature, constraints in the forms of prior knowledge in samples (Wagstaff et al., 2001) or pathway structure in features (Huo and Tseng, 2017) have been used to restrict the free parameters in high-dimensional space during clustering. The approaches improve biological relevance of the finding, but still cannot prevent the true outcome-associated disease subtypes from being masked by disease-irrelevant clusters.

This practical example raises a fundamental question in clustering of high-dimensional omics data for disease subtyping: can we simultaneously identify disease subtypes and the

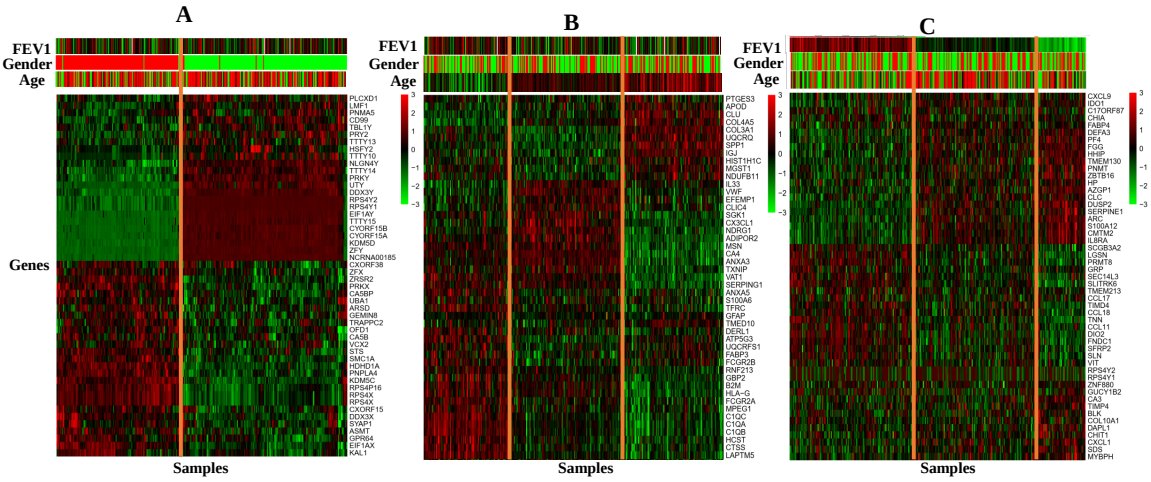


Figure 2: A real example illustrates (A) two gender-associated clusters are found by the top 50 X/Y chromosome genes and K-means; (B) three age-associated clusters are detected by the top 50 age-related genes and K-means; (C) three clusters are identified from our algorithm, which are associated with clinical outcome FEV1 but neither associated with gender nor age.

driving gene signatures, where the detection of disease subtypes is guided by outcome association? This question is unique as it touches both supervised and unsupervised components in the context of machine learning. In the process of detecting novel disease subtypes, we focus on identifying disease-related subtypes and hope to disentangle and reduce impact of factors driven by clinically irrelevant variables (e.g. demographic variables, such as gender, age and race). In the literature, little has been done in this proposed direction. Bair and Tibshirani (2004), Koestler et al. (2010) and Gaynor and Bair (2017) have developed a two-stage semi-supervised method, where K-means or other conventional clustering methods are applied to the pre-selected top features with the highest marginal outcome association. These two-stage approach is, however, ad hoc in selecting the number of top features and has difficulty in incorporating confounding variables in the outcome association. Ahmad and Fröhlich (2017) proposed a Bayesian method to cluster omics data with surivial outcomes and molecular features. Wang et al. (2020b) proposed a supervised convex clustering algorithm. however, these two methods are computationally intensive and only feasible with up to 100 genes. Approaches of these five papers also cannot extend to extensive biological scenarios. In this Chapter, we propose an outcome-guided clustering (ogClust) model to provide a unified solution high-dimensional omics data.

Throughout this Chapter, we avoid the term "semi-supervised" adopted by Bair and Tibshirani (2004) and Koestler et al. (2010). Instead, we name by "outcome-guided disease subtyping" or "outcome-guided clustering" since the term "semi-supervised learning" has been used in at least two other machine learning scenarios: (1) A small set of labeled data and a larger set of unlabeled data are jointly analyzed for machine learning; (2) Cluster analysis is pursued with known constraints (e.g. pairs of observation must or must not be clustered together). Interested readers may refer to Bair (2013) for a review of semi-supervised clustering methods. One should also note that the outcome-guided clustering discussed in this chapter substantially differs from latent class analysis methods in regression setting by Houseman et al. (2006), DeSantis et al. (2007) and Desantis et al. (2012). In this case, patients in latent classes are identified to have heterogeneous intercepts or regression slopes. The latent classes, in a sense, represent patient clusters (or disease subtypes), but there is lack of a gene signature and prediction model to classify future patients into the

disease subtypes (latent classes), presenting a major obstacle towards precision medicine.

The Chapter is structured as follows. Section 2.2 introduces the ogClust model (Section 2.2.1), an EM algorithm for parameter estimation (Section 2.2.2), extensions to robust estimation procedures in outcome association (Section 2.2.3), and its extension to survival outcome (Section 2.2.4). We perform extensive simulations to evaluate ogClust and compare it with existing methods in Section 2.3, and evaluate its robust estimation in Section 2.3.2. A disease subtyping application using a lung disease transcriptomic dataset is presented in Section 2.4. We include final conclusion and discussion in Section 2.5.

2.2 Proposed Method

2.2.1 Model and notations

We consider the problem of disease subtyping (clustering) of n observations from high-dimensional data $\mathbb{G} = \{g_{ij}, 1 \leq i \leq n, 1 \leq j \leq q\}$, where $\mathbb{G}$ can be mRNA expression, miRNA expression, methylation or phenomic data and q can be at the scale of hundreds to thousands. Our ultimate goal is to cluster n observations into K clinically meaningful clusters represented by latent group label $\mathbb{Z} = \{z_i, 1 \leq i \leq n\}$, $z_i \in \{1, \ldots, K\}$, and $z_i = k$ means that observation i is assigned to cluster k ($1 \leq k \leq K$). Since clustering result purely from $\mathbb{G}$ may not necessarily be clinically useful as discussed in Section 2.1, we assume that a clinical outcome $\mathbb{Y} = \{y_i, 1 \leq i \leq n\}$ is given to guide the clustering (e.g. survival outcome or FEV1 in the lung disease example in Section 2.4). We also assume a set of pre-specified covariates $\mathbb{X} = \{x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p\}$, where the p covariates (e.g. age, gender, etc.) are potentially associated with the outcome and may confound with the association between Z and Y. Denote by $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{iq})^T$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$. We assume observed data $(y_i, \boldsymbol{x}_i, \boldsymbol{g}_i)$ for subject i ($1 \leq i \leq n$) are independent realizations of the model for $(Y, \boldsymbol{X}, \boldsymbol{G})$.

As shown in Figure 3, the proposed ogClust framework consists of two components: *disease subtyping model* and *outcome association model*. The disease subtyping model is a conventional high-dimensional discriminant analysis where we train to characterize $\pi_k =$

$Pr(Z = k|\boldsymbol{G})$ (or $\pi_{ik} = Pr(Z_i = k|\boldsymbol{g}_i)$ for observation i). In this chapter, we apply a multinomial logistic regression $\pi_{ik}|\boldsymbol{\gamma} = \frac{\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_k)}{\sum_{l=1}^{K}\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_l)}$, where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k, 1 \leq k \leq K\}$ and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{qk})^T$. Since q is usually large, we assume only a small subset $\mathcal{A} \subset \{1, \ldots, q\}$ of features effective in characterizing the clusters that affect the outcome, where its cardinality $\text{card}(\mathcal{A}) < \min(n, q)$. In other words, $\boldsymbol{\gamma}_{[j]} \neq \mathbf{0}$ if $j \in \mathcal{A}$ and $\boldsymbol{\gamma}_{[j]} = \mathbf{0}$ if $j \in \mathcal{A}^c$, where $\boldsymbol{\gamma}_{[j]} = \{\gamma_{j1}, \ldots, \gamma_{jK}\}$. We apply LASSO regularization, or group LASSO regularization (Tibshirani et al., 2012) with parameters in $\boldsymbol{\gamma}_{[j]}$ as a group to the multinomial logistic regression to select subtyping features.

In the outcome association model, we assume the following mixture model:

$$f(y_i; \boldsymbol{x_i}) = \sum_{k=1}^{K} \pi_{ik} f_k(y_i; \boldsymbol{x_i}), \tag{1}$$

where $f_k(y; \boldsymbol{x})$ is density function of cluster k. We assume a continuous response Y where the k-th mixture density $f_k(y; \boldsymbol{x}, \beta_{0k}, \boldsymbol{\beta}, \sigma)$ is parameterized by cluster specific intercept β_{0k}, com-

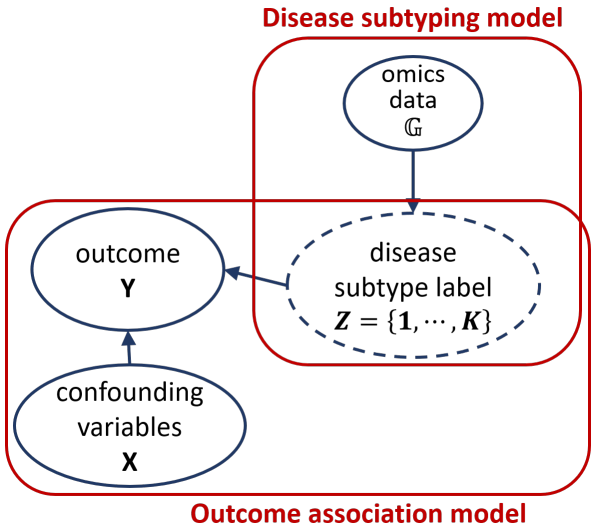


Figure 3: A graphical illustration of the unified regression model. Y is the outcome to guide clustering, X are the baseline covariates that are believed to have effects on Y. G are the variables (e.g. gene expression) that defines the outcome associated subgroups. Z is the unobserved latent subgroup index to define final clustering.

mon covariate effect $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and a homogeneous error σ. In this chapter, we specifically assume $y_i|z_i = k \sim N(\beta_{0k} + \boldsymbol{\beta}^T \boldsymbol{x_i}, \sigma^2)$ with mixture probability $\pi_{ik} = \frac{\exp\left(\boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{g}_i^T \boldsymbol{\gamma}_l\right)}$, $k = 1, \ldots, K$. Denote by $\boldsymbol{\theta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma\}$ the collection of all parameters from the two models in ogClust ($\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0K})^T$), given $\mathbb{Y}$, $\mathbb{X}$ and $\mathbb{G}$, $\boldsymbol{\theta}$ can be estimated by maximizing the following sample likelihood of the basic model:

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{g_i}, \boldsymbol{\gamma}) f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right). \tag{2}$$

Remarks:

1. Generalization from continuous outcome Y to other types of outcome Y is relatively straightforward. Section 2.2.4 discusses the extension to survival outcome.
2. In the current model, we assume only several important and pre-selected covariates for $\boldsymbol{X}$ and no variable selection is implemented in the outcome association model. Including $\boldsymbol{X}$ (e.g. age or gender) in the outcome association model has two main advantages: (i) it corrects for potential confounding effects between the association of outcome Y and subtype Z, (ii) if a covariate, say, gender, is indeed predictive of Y and there exist many strong gender-associated genes in $\boldsymbol{G}$, the model will avoid identification of gender-related clusters in Z. In this case, although gender-associated subtypes are predictive of the outcome, their information has been captured by observable covariate and thus can be avoided in subtyping.
3. The current model assumes a simplified common covariate effect $\boldsymbol{\beta}$ across all clusters. It is straightforward to extend for cluster-specific interaction term $\boldsymbol{\beta_k}$, meaning cluster-specific age or gender effects.
4. We apply multinomial logistic regression in this chapter but other high-dimensional discriminant analysis methods, such as sparse linear discriminant analysis, can also be used.
5. The conditional probability $\hat{\pi}_{ik}|\hat{\boldsymbol{\gamma}} = \frac{\exp\left(\boldsymbol{g}_i^T \hat{\boldsymbol{\gamma}}_k\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{g}_i^T \hat{\boldsymbol{\gamma}}_l\right)}$ can be used to predict the cluster label of new observations.

2.2.2 Numerical estimation by EM algorithm

A numerical method using EM algorithm is proposed for ogClust parameter estimation in Equation (2). By introducing $z_{ik}, k = 1, \ldots, K$, as missing indicator variables, following the seminal idea in Dempster et al. (1977), the complete log likelihood function can be written as

$$l_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \{z_{ik} \log \pi_{ik} + z_{ik} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right)\}, \tag{3}$$

where $z_{ik} = 1$ if subject i belongs to subgroup k, and $z_{ik} = 0$ otherwise.

Since gene expression is usually high dimensional, including genes in $\mathcal{A}^c$ with non-predictive effect will introduce extra noise to the disease subtyping model and may produce irrelevant subtypes that are not necessarily related to the disease outcome of interest. In the following, we will illustrate with a LASSO penalty or an alternative group LASSO regularization framework for gene selection. We define the penalized log-likelihood function as

$$\tilde{l}_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \{z_{ik} \log \pi_{ik} + z_{ik} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right)\} - \lambda R(\boldsymbol{\gamma}), \tag{4}$$

where λ is the regularization tuning parameter and $R(\boldsymbol{\gamma}) = \sum_{j=1}^{q} \sum_{k=1}^{K} |\gamma_{jk}|$ for LASSO penalty. Alternatively, we can use group LASSO penalty plus ℓ_2 regularization $R(\boldsymbol{\gamma}) = \sum_{j=1}^{q} \|\boldsymbol{\gamma}_{[j]}\|_2 + \alpha \sum_{j=1}^{q} \sum_{k=1}^{K} \gamma_{jk}^2$, where $\|\boldsymbol{\gamma}_{[j]}\|_2 = \sqrt{\sum_{k=1}^{K} \gamma_{jk}^2}$. The first term is a group LASSO penalty to select or deselect $\boldsymbol{\gamma}_{[j]}$ for gene j. The second term encourages joint selection of predictive genes with high collinearity. Detecting multiple genes with high collinearity offers better molecular insight to the subtype mechanism and provides more stable cluster prediction for future patients. The irrelevant features are removed by shrinking corresponding elements of $\boldsymbol{\gamma}_{[j]}$ to zero, thus a sub-model is automatically selected. This procedure performs feature selection and numerical estimation of parameters simultaneously.

Maximization of $\tilde{l}_n^c(\boldsymbol{\theta})$ can be achieved by sequentially and iteratively updating $\boldsymbol{\beta_0}$, $\boldsymbol{\beta}$, σ and $\boldsymbol{\gamma}$ in an EM algorithm, which takes the following steps:

- The E step computes the conditional expectation of the function $\tilde{l}_n^c(\boldsymbol{\theta})$ with respect to z_{ik}, given the observed data y_i, $\boldsymbol{x}_i$ and the current parameter estimates $\boldsymbol{\theta}^{(m)}$,

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log \pi_{ik} + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) - \lambda \sum_{j=1}^{q} R(\boldsymbol{\gamma}_j),$$

where the posterior weights

$$w_{ik}^{(m)} = E\left(Z_{ik}|y_i, \boldsymbol{x}_i, \boldsymbol{\theta}^{(m)}\right) = \frac{\pi_{ik}^{(m)} f\left(y_i; \boldsymbol{x}_i, \beta_{0k}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\right)}{\sum_{l=1}^{K} \pi_{il}^{(m)} f\left(y_i; \boldsymbol{x}_i, \beta_{0l}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\right)}. \tag{5}$$

- The M step on the $(m+1)$-th iteration maximizes the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right)$ with respect to $\boldsymbol{\theta}$. By taking partial derivatives, it is easy to show that $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ and σ^2 are updated by the following updating equations:

$$\beta_{0k}^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(m)} \left(y_i - (\boldsymbol{\beta}^{(m)})^T \boldsymbol{x}_i\right)}{\sum_{i=1}^{n} w_{ik}^{(m)}}, \quad k = 1, \ldots, K, \tag{6}$$

$$\beta_{\ell}^{(m+1)} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} x_{i\ell} \left(y_i - \beta_{0k}^{(m+1)} - \sum_{h \neq \ell} \beta_h^{(m)} x_{ih}\right)}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} x_{i\ell}^2}, \quad \ell = 1, \ldots, p, \tag{7}$$

$$(\sigma^{(m+1)})^2 = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \left(y_i - \beta_{0k}^{(m+1)} - (\boldsymbol{\beta}^{(m+1)})^T \boldsymbol{x}_i\right)^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)}}. \tag{8}$$

The updated estimates $\boldsymbol{\gamma}^{(m+1)}$ is obtained following an approximation procedure of Friedman et al. (2010). For lasso penalty $R(\boldsymbol{\gamma}) = \sum_{k=1}^{K} R_k(\boldsymbol{\gamma}_k) = \sum_{k=1}^{K} \sum_{j=1}^{q} |\gamma_{jk}|$, the likelihood for estimating $\boldsymbol{\gamma}^{(m+1)}$ given $w^{(m)}$ is

$$\begin{aligned}
\tilde{l}_p\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right) &= \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log \pi_{ik} - \lambda R(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log \frac{\exp\left(\boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)}{\sum_{l=1}^{K} \exp\left(\boldsymbol{g}_i^T \boldsymbol{\gamma}_l\right)} - \lambda R(\boldsymbol{\gamma}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \left\{ \boldsymbol{g}_i^T \boldsymbol{\gamma}_k - \log\left(\sum_{l=1}^{K} \exp\left(\boldsymbol{g}_i^T \boldsymbol{\gamma}_l\right)\right)\right\} - \lambda R(\boldsymbol{\gamma})
\end{aligned}$$

We approximate the partial log likelihood $\tilde{l}_p(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ by quadratic approximation. The resulting partial likelihood $\tilde{l}_{Qk}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ for subgroup k is in the form of a weighted least square:

$$\tilde{l}_{Qk}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right) = -\frac{1}{2} \sum_{i=1}^{n} W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)^2 - \lambda R_k(\boldsymbol{\gamma}_k) + C,$$

where $h_{ik} = \boldsymbol{g}_i^T \boldsymbol{\gamma}_k^{(m)} + \frac{w_{ik}^{(m)} - \pi_{ik}^{(m)}}{W_{ik}}$, $W_{ik} = \pi_{ik}^{(m)}(1 - \pi_{ik}^{(m)})$, and C is independent of $\boldsymbol{\gamma}_k$. Thus the solution to $\boldsymbol{\gamma}^{(m+1)}$ can be obtained by coordinate descent, i.e., individually solving

Algorithm 1 Pseudo code for ogClust model estimation.

input: $\mathbb{Y}$, $\mathbb{X}$, $\mathbb{G}$ and K

Initialize $\boldsymbol{\theta}^{(0)}$ and set $m = 0$;

repeat

E-step: compute the posterior weights $w_{ik}^{(m)}$ by Equation (5);

M-step:

1. Update $\{\boldsymbol{\beta}_0^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\}$ to $\{\boldsymbol{\beta}_0^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \sigma^{(m+1)}\}$ by Equations (6)-(8);

2. Update $\boldsymbol{\gamma}^{(m)}$ to $\boldsymbol{\gamma}^{(m+1)}$ by coordinate descent:

Set $\tilde{\boldsymbol{\gamma}}^{old} = \boldsymbol{\gamma}^{(m)}$;

repeat

Update $\tilde{\gamma}_{kj}^{old}$ to $\tilde{\gamma}_{kj}^{new}$ by Equation (9), for $k = 1, \ldots, K$ and $j = 1, \ldots, q$;

until $||\tilde{\boldsymbol{\gamma}}^{old} - \tilde{\boldsymbol{\gamma}}^{new}|| < 10^{-7}$;

Set $\boldsymbol{\gamma}^{(m+1)} = \tilde{\gamma}_{kj}^{new}$, $\boldsymbol{\theta}^{(m+1)} = \{\boldsymbol{\beta}_0^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, \sigma^{(m+1)}\}$, $m = m + 1$;

until $||\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}|| < 10^{-7}$;

output: Parameter estimates $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(m)}$

$\max_{\boldsymbol{\gamma}_k \in R^q} \tilde{l}_{Qk}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ for each k. By some algebraic manipulation, we obtain the estimate

$$\tilde{\gamma}_{kj} = \frac{S\left(\sum_{i=1}^{N} g_{ij} W_{ik} \left(h_{ik} - (\boldsymbol{g}_i^{(j)})^T \tilde{\boldsymbol{\gamma}}_k^{(j)}\right), \lambda\right)}{\sum_{i=1}^{N} W_{ik} g_{ij}^2}, \tag{9}$$

where $S(z, \lambda) = \text{sign}(z)(|z| - \lambda)_+$ is a soft thresholding operator, $(a)_+ = \max(0, a)$, $\tilde{\boldsymbol{\gamma}}_k^{(j)}$ is the parameter vector $\tilde{\boldsymbol{\gamma}}_k$ omitting $\tilde{\gamma}_{kj}$, and $\boldsymbol{g}_i^{(j)}$ is the gene vector $\boldsymbol{g}_i$ omitting g_{ij}. The coordinate descent procedure iteratively updates the current estimate $\tilde{\boldsymbol{\gamma}}$ until convergence. For the group LASSO + ℓ_2 regularization, we apply the glmnet function in R package glmnet, setting multinomial family, grouped type and α equals 0.5.

The pseudo code for fitting the unified ogClust model is given in Algorithm 1. Multiple initializations could be used to avoid convergence to local minimums and increase the numerical stability of parameter estimates. We use Bayesian information criterion (BIC) to determine the tuning parameter λ and the number of subgroups K in simulation. BIC is defined as $\ln(n)\text{df}(\hat{\boldsymbol{\theta}}) - 2\ln(\text{L}(\hat{\boldsymbol{\theta}}))$, where $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}\}$ and $\text{df}(\hat{\boldsymbol{\theta}})$ is the number of

non-zero estimated parameters. In the real application, because of potential data noises and violation of Gaussian assumption, BIC may fail to choose the correct K. To address this issue, we plot the trend of root mean square error (RMSE) and R^2 as a function of K and identify the elbow point as the optimal number of clusters K as shown in Figure 4.

2.2.3 Robust estimation procedures

The ogClust model is based on and could be sensitive to the Gaussian mixture assumption in outcome Y. There are three common types of model misspecification: (A) heavy-tailed or skewed error term in the outcome association model, (B) outliers in outcome Y in the outcome association model, and (C) scattered observations who do not fit into any of the K subtypes in the disease subtyping model. Our model is relatively robust to type C misspecifications because of the soft assignment using multinomial logistic probability function. One may iteratively remove a small number of samples with unconfident cluster assignments. To guard against the first two types of model misspecification, we propose 1)

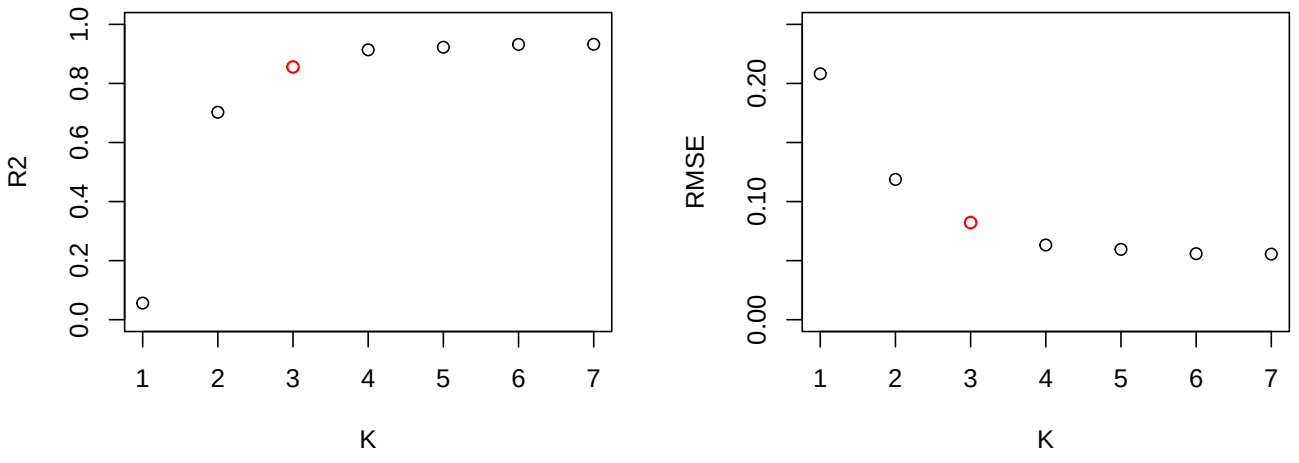


Figure 4: Plot of (A) R^2 and (B) RMSE against the number of clusters.

ogClust with median-truncated loss (ogClust-median-truncation) 2) ogClust with Huber loss (ogClust-Huber) 3) ogClust with adaptive-Huber loss (ogClust-adHuber) to replace the original ogClust with quadratic loss. Intuitively, median-truncation and Huber loss functions are effective in dealing with potential outliers. As we will introduce later, the adaptive-Huber loss is particularly useful for heavy-tailed and skewed error terms. Hence, the penalized log-likelihood function is defined as

$$l_n^c(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\{z_{ik}\log\pi_{ik} + z_{ik}\ell_\tau(e_{ik})\} - \lambda\sum_{j=1}^{p}R(\boldsymbol{\gamma}_j).$$

where $\ell_\tau(e_{ik})$ denotes the robust loss function to replace $\log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right)$. We follow the same EM procedure with modified loss functions to compute numerical solutions.

2.2.3.1 Median-truncated loss

The median-truncated loss (Chi et al., 2019) describes the loss function for subject i in subgroup k as:

$$\ell_\tau(e_{ik}) = \begin{cases} e_{ik}^2/2 & \text{if } |e_{ik}| \le \tau_k \\ 0 & \text{if } |e_{ik}| > \tau_k \end{cases},$$

where $e_{ik} = y_i - \hat{\beta}_{0k} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}_i$, and $\tau_k = \text{median}\left\{|e_{ik}|\right\}_{i=1}^n$. The loss function remains the same for e_{ik} smaller or equal to median τ_k, and the loss function equals to 0 for e_{ik} larger than the median τ_k. The cutoff τ_k is chosen as the median of $e_{1k}, ..., e_{nk}$. By taking partial derivatives, the estimates in the $(m+1)$th iteration for $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ and σ are obtained by the following equations:

$$\begin{aligned}
\beta_{0k}^{(m+1)} &= \frac{\sum_{i=1}^n w_{ik}^{(m)}\left(y_i - (\boldsymbol{\beta}^{(m)})^T\boldsymbol{x}_i\right) I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}{\sum_{i=1}^n w_{ik}^{(m)} I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}, \\
\beta_\ell^{(m+1)} &= \frac{\sum_{i=1}^n\sum_{k=1}^K w_{ik}^{(m)} x_{i\ell}\left(y_i - \beta_{0k}^{(m+1)} - \sum_{h\neq\ell}\beta_h^{(m)}x_{ih}\right) I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}{\sum_{i=1}^n\sum_{k=1}^K w_{ik}^{(m)} x_{i\ell}^2 I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}, \\
(\sigma^{(m+1)})^2 &= \frac{\sum_{i=1}^n\sum_{k=1}^K w_{ik}^{(m)}\left(y_i - \beta_{0k}^{(m+1)} - (\boldsymbol{\beta}^{(m+1)})^T\boldsymbol{x}_i\right)^2 I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}{\sum_{i=1}^n\sum_{k=1}^K w_{ik}^{(m)} I\left(\left|e_{ik}^{(m)}\right| \le \tau_k\right)}.
\end{aligned}$$

2.2.3.2 Huber loss

The Huber loss alternatively describes the loss function for subject i in subgroup k as:

$$\ell_\tau(e_{ik}) = \begin{cases} e_{ik}^2/2 & \text{if } |e_{ik}| \le \tau \\ \tau|e_{ik}| - \tau^2/2 & \text{if } |e_{ik}| > \tau \end{cases}.$$

This loss function is quadratic for small values of e, and linear for large values of e. The cutoff τ is suggested as a fixed constant ($\tau = 1.345$) which gives 95% efficiency under Gaussian assumption in regression setting (Huber, 2004). By EM algorithm, the estimates in the $(m+1)$th iteration for $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ and σ^2 are obtained by the following equations:

$$\begin{aligned}
\beta_{0k}^{(m+1)} &= \frac{\sum_{i=1}^n w_{ik}^{(m)} \left(y_i - (\boldsymbol{\beta}^{(m)})^T \boldsymbol{x}_i\right) I\left(\left|e_{ik}^{(m)}\right| \le \tau\right) + \sum_{i=1}^n w_{ik}^{(m)} \cdot \tau \cdot \text{sign}\left(e_{ik}^{(m)}\right) \cdot I\left(\left|e_{ik}^{(m)}\right| > \tau\right)}{\sum_{i=1}^n w_{ik}^{(m)} I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)}, \\
\beta_{\ell}^{(m+1)} &= \frac{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} x_{i\ell} \left(\left(y_i - \beta_{0k}^{(m+1)} - \sum_{h\neq\ell} \beta_h^{(m)} x_{ih}\right) I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)\right)}{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} x_{i\ell}^2 I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)} + \\
&\quad \frac{\tau \,\text{sign}\left(e_{ik}^{(m)}\right) I\left(\left|e_{ik}^{(m)}\right| > \tau\right)}{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} x_{i\ell}^2 I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)}, \\
(\sigma^{(m+1)})^2 &= \frac{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} \left(y_i - \beta_{0k}^{(m+1)} - (\boldsymbol{\beta}^{(m+1)})^T \boldsymbol{x}_i\right)^2 I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)}{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)} + \\
&\quad \frac{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} \left(2\tau \left|y_i - \beta_{0k}^{(m+1)} - \boldsymbol{X}_i^T \boldsymbol{\beta}^{(m+1)}\right| - \tau^2\right) I\left(\left|e_{ik}^{(m)}\right| > \tau\right)}{\sum_{i=1}^n \sum_{k=1}^K w_{ik}^{(m)} I\left(\left|e_{ik}^{(m)}\right| \le \tau\right)}.
\end{aligned}$$

2.2.3.3 Adaptive Huber loss

When there are no outliers but the error term is heavy-tailed asymmetric, median-truncated loss or Huber loss using constant τ would introduces bias (Sun et al., 2019b). To mitigate this bias, we use an adaptive Huber loss in the EM algorithm by adopting the method of Wang et al. (2020a). In this method, the cutoff τ is data-driven and estimated

adaptively, taking into account sample size, n, dimension of $\boldsymbol{\beta}$, p, by iteratively solving the following equations:

$$\begin{cases} g_1(\boldsymbol{\theta}, \tau) := \sum_{i=1}^{n} w_{ik}^{(m)} \sum_{k=1}^{K} \ell_\tau'(e_{ik}) \boldsymbol{X}_i = \mathbf{0} \\ g_2(\boldsymbol{\theta}, \tau) := (n-p)^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} \min\{e_{ik}^2, \tau^2\} / \tau^2 - n^{-1}(p+z) = 0 \end{cases}$$

where $z = log(n)$ by default. This method is implemented in R package tfHuber. We adapt it into the M-step of our EM algorithm to update $\{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \sigma\}$. At a high level, by allowing increasing value of cutoff τ as n increases, there is a trade-off between the robustness and bias. By picking an optimal τ, the bias becomes negligible while the result is still robust to outliers caused by heavy-tailed noise.

2.2.4 ogClust model with survival outcome

ogClust model can be extended to use survival outcomes. To facilitate model fitting, we choose accelerated failure time (AFT) model with log-logistic distribution to model time-to-event data as $(log(Y)|Z = k) = \beta_{0k} + X\beta + W\sigma$, where W $\sim$ standard logistic distribution and σ is the standard deviation. Therefore, the likelihood of mixture model can be written as

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik} L_{ik}\left(y_i | \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma_k, \right).$$

Denote δ as a binary indicator of event, $\delta = 1$ means event and 0 means right-censored. The likelihood function $L_{ik}\left(Y_i | \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right)$ is defined as

$$L_{ik}\left(y_i | \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) = \left\{\frac{1}{\sigma} f_W\left(w_i\right)\right\}^{\delta_i} \left\{S_W\left(w_i\right)\right\}^{1-\delta_i}$$

where

$$w_i = \frac{z_i - \beta_{0k} - X_i\beta}{\sigma}$$

$$S_W(w_i) = 1/\left(1 + e^{w_i}\right)$$

$$f_W(w_i) = e^{w_i}/\left(1 + e^{w_i}\right)^2.$$

Therefore, the penalized log-likelihood function is defined as:

$$\tilde{l}_n^c(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{z_{ik} \log \pi_{ik} + z_{ik} log L_{ik}\left(y_i | \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right)\right\} - \lambda R(\boldsymbol{\gamma}).$$

We follow the same EM algorithm in the original model, except that the likelihood of the ATF model is maximized by using the function "survreg" in R package"survival".

2.3 Simulations

In this section, we conduct three simulations to evaluate the performance of clustering, feature selection, and outcome prediction for ogClust, robust estimation procedures of ogClust, and its extension for survival outcome respectively. In Section 2.3 we assume that the continuous outcome Y follows mixture of Gaussian distribution and compare the performance of ogClust with three other methods. In Section 2.3.2 we introduce outliers or skewed and heavy-tailed errors to outcome Y, and compare the performance of three robust estimation procedures with the non-robust ogClust method. In Section 2.3.3 we show the advantage of ogClust over three other methods with survival outcome Y to guide the clustering.

2.3.1 Simulations to evaluate ogClust

Simulation scheme

1. Simulate $q = 1000$ genes ($\boldsymbol{G} = \{G_1, ..., G_{1000}\}$), among which G_1 to G_{30} are differentially expressed (DE) across clusters while the rest of the genes are not differentially expressed and their expression values are randomly drawn from the standard normal distribution (Figure 5B). Expression levels of the 30 DE genes are randomly drawn from N(1,1) and N(0,1) to form 3×3 clusters as specified in Figure 5A: gene set $\boldsymbol{G}_{\mathcal{A}_1}$, $\mathcal{A}_1 = \{1, \ldots, 15\}$, defines three clusters associated with the outcome Y; gene set $\boldsymbol{G}_{\mathcal{A}_2}$, $\mathcal{A}_2 = \{16, \ldots, 30\}$, defines three "clinically irrelevant clusters" that are independent of Y.
2. Use parameters corresponding to $\mathcal{A}_1$, $\boldsymbol{\gamma}_{\mathcal{A}_1} = (\boldsymbol{\gamma}_{1\mathcal{A}_1}, \boldsymbol{\gamma}_{2\mathcal{A}_1}, \boldsymbol{\gamma}_{3\mathcal{A}_1})^T$, to represent the effect of gene expression on subtyping. For identifiability, we set $\boldsymbol{\gamma}_{3\mathcal{A}_1} = \mathbf{0}$. $\boldsymbol{\gamma}_{1\mathcal{A}_1}$ and $\boldsymbol{\gamma}_{2\mathcal{A}_1}$ vary in different models. The active set for outcome-guided subtypes is restricted to $\mathcal{A}_1$, in other words, $\boldsymbol{\gamma}_{\mathcal{A}_1^c} = \mathbf{0}$.
3. Given gene expression of $\boldsymbol{G}_{\mathcal{A}_1}$ and $\boldsymbol{\gamma}_{\mathcal{A}_1}$, we obtain $\pi_{ik} = \frac{\exp(\boldsymbol{g}_{i\mathcal{A}_1}^T \boldsymbol{\gamma}_{k\mathcal{A}_1})}{\sum_{l=1}^{3} \exp(\boldsymbol{g}_{i\mathcal{A}_1}^T \boldsymbol{\gamma}_{l\mathcal{A}_1})}$, $k \in \{1, 2, 3\}$, which represent the probability of subject i belonging to the kth subgroup. Therefore, subgroup indicator Z_i for subject i is randomly drawn from a multinomial distribution with probability $\boldsymbol{p}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$

4. Sample independent covariates X_1 and X_2 are sampled from normal distributions $N(1,1)$ and $N(2,1)$ respectively. Recall that $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ is the set of regression coefficients of the two covariates and $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \beta_{03})^T$ represents the baseline mean of the three subgroups. We set $\boldsymbol{\beta} = (1,1)^T$, and $\boldsymbol{\beta}_0$ varies according to different models.
5. Given the latent subgroup index Z_i, the outcome for subject i can be simulated by $(Y_i|Z_i = k) = \beta_{0k} + \boldsymbol{X_i^T}\boldsymbol{\beta} + e_i$, where $e_i \sim N(0, \sigma^2)$ and we set $\sigma^2 = 1$.

The simulation scheme is illustrated in detail in Figure 5. Let $\boldsymbol{\beta}_0 = (1, 1+\delta, 1+2\delta)^T$ and $\boldsymbol{\gamma}_{\mathcal{A}_1} = ((\text{rep}(\gamma, 5), \text{rep}(0,5), \text{rep}(\text{-}\gamma, 5))^T, (\text{rep}(\text{-}\gamma, 5), \text{rep}(0,5), \text{rep}(\gamma, 5))^T, (\text{rep}(0, 15))^T)$, where $\text{rep}(a, b) = (a, \ldots, a)_{(1\times b)}$. We consider four models with different choices of $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_{\mathcal{A}_1}$ specified below:

- Model 1: $\gamma = 1$ and $\delta = 2$
- Model 2: $\gamma = 1$ and $\delta = 3$
- Model 3: $\gamma = 1$ and $\delta = 5$
- Model 4: $\gamma = 3$ and $\delta = 3$

Essentially, γ controls the level of cluster separation in the omics space and δ represents the difference of subgroup effect on the value of outcome Y. (refer to Appendix A for detailed explanation). We first evaluate Models 1-3 with lower level of cluster separation $\gamma = 1$ and varying outcome association $\delta = 2, 3, 5$. Model 4 evaluates $\gamma = \delta = 3$.

We compare the performance of the proposed ogClust using group LASSO + ℓ_2 penalty with three other competing clustering methods: 1) SKM: sparse K-means clustering (Witten and Tibshirani, 2010), a modified K-means algorithm with variable selection; 2) PMBC: penalized model based clustering (Pan and Shen, 2007), an unsupervised method based on Gaussian mixture model; 3) SC: supervised clustering (Bair and Tibshirani, 2004), a post-screening clustering method. SKM and PMBC are not outcome-guided and could be sensitive to any "clinically irrelevant" clusters, while SC has a variable pre-screening by outcome association. To evaluate the performance of these methods, we simulate 100 datasets with sample size $n = 600$, where there are 1000 genes and three subgroups with equal size. To implement SKM and PMBC and compare with ogClust, we assign observations to the cluster with closest center (SKM) or with the highest posterior probability (PMBC), then fit

linear regression with covariate X and outcome Y in each resulting cluster to make outcome prediction. For SC, we apply a pre-screen step to pre-select M outcome associated genes

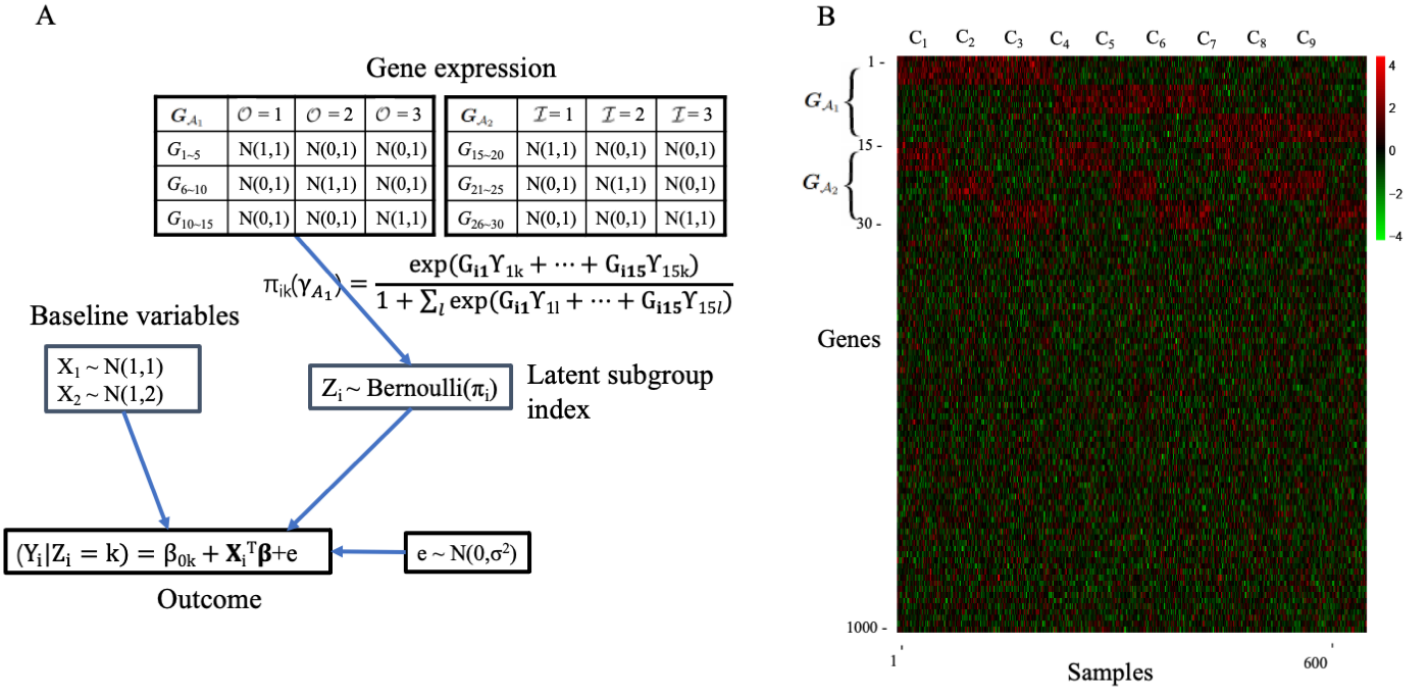


Figure 5: (A) Data generation scheme. $\mathcal{O} = \{1, 2, 3\}$ denotes three clusters defined by genes set $G_{\mathcal{A}_1}$, $\mathcal{A}_1 = \{1, \ldots, 15\}$, and $\mathcal{I} = \{1, 2, 3\}$ denotes another three independent clusters defined by $G_{\mathcal{A}_2}$, $\mathcal{A}_2 = \{16, \ldots, 30\}$. Expression of genes in $G_{\mathcal{A}_1}$ and $G_{\mathcal{A}_2}$ are generated from the distributions listed on the above table. For subject i, only $G_{\mathcal{A}_1}$ have real signals effecting Z_i, which is drawn from a Multinomial distribution with probability $\boldsymbol{\pi}_i = \{\pi_{i1}, \pi_{i2}, 1 - \pi_{i1} - \pi_{i2}\}$. Baseline variables X_1 and X_2 are generated from $N(1, 1)$ and $N(1, 2)$ respectively. Given X_i, G_i and Z_i, the outcome Y_i is generated finally. (B) Heatmap of the expression of 1000 genes across samples. A total of nine subgroups C_1, ..., C_9 are jointly defined by genes sets $G_{\mathcal{A}_1}$ and $G_{\mathcal{A}_2}$.

before we perform K-means clustering and fit linear regression in each resulting cluster, the value of M is determined by cross-validation.

The performance of these methods is evaluated by their clustering accuracy, gene selection and outcome prediction by 10-fold cross-validation. Within each fold of training/testing split, we fit each of the methods using the training set and then predict both latent subgroup label and outcome value for testing set. We compute $RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}$ and $R^2 = 1 - SS_{residual}/SS_{total} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2}$ from the 10-fold cross validation and average the results to measure the prediction error of outcome (Table 1). We also compute the average number of false positives (FPs) and false negatives (FNs) for evaluating the accuracy of feature selection (Table 1). For clustering accuracy, we compute the adjusted Rand index (ARI) (Hubert and Arabie, 1985), which has 0 expectation when clustering is random and bounded by 1 with perfect partition, to measure the consistency of predicted subgroup label with true latent subgroup index (Table 1).

Table 1 shows results of ogClust compared to SKM, PMBC and SC under the four simulation settings. To measure clustering performance, SKM and PMBC identifies $K = 2$ clusters in 100 simulations and 74-83 of the 100 simulations respectively, but since the algorithm has no outcome guidance, they mostly obtain clinically irrelevant clusters and have ARI= 0.04-0.16 when compared with the three true outcome-associated clusters. SC pre-selects outcome-associated gene features to perform clustering and generates improved ARI=0.35-0.41, but the method identifies $K = 2$ clusters for all simulations. In contrast, for Models 2-4, ogClust identifies $K = 3$ clusters for 98-99 out of 100 simulations and produces ARI=0.86-0.91. For the weak signal Model 1, ogClust identifies $K = 3$ clusters for 37 of the 100 simulations and the ARI reduces to 0.45. When evaluating gene selection, PMBC misses majority of the first 15 true clustering genes (8.7-11.1 FNs) and both SKM and PMBC add many false positives (776.1-813.8 FPs for SKM and 87.0-100.3 FPs for PMBC). SC contains outcome association gene selection but still misses 4.8-8.7 FNs and adds 17.5-88.0 FPs. In contrast, ogClust almost does not miss true clustering genes (FN=0 for Model 2-4 and FN=3 for Model 1) and only adds $\sim$14 false positives. For outcome prediction result, ogClust generates the lowest RMSE and the highest R^2, showing better clinical relevance of produced disease subtypes. In summary, SKM and PMBC are vulnerable to missing clinically relevant

clusters and related predictive genes. SC only modestly improves in detecting outcome-associated genes and clusters, and the two-stage approach reduces performance and rigor of inference. ogClust outperforms the three methods in clustering accuracy, gene selection and clinical outcome prediction.

Table A.1.1 and A.1.2 in Appendix A show the simulation results when there is a stronger and weaker signal in $\boldsymbol{G}_{\mathcal{A}_2}$ compared with $\boldsymbol{G}_{\mathcal{A}_1}$ respectively. When the signal in $\boldsymbol{G}_{\mathcal{A}_2}$is stronger, SKM and PMBC are dominated by $\boldsymbol{G}_{\mathcal{A}_2}$ and returns clinically irrelevant clusters with ARI=0. When the signal in $\boldsymbol{G}_{\mathcal{A}_2}$is weaker, SKM and PMBC performs slightly better in identifying the three outcome associated clusters and outcome prediction with higher ARI and R^2. However, the expression of $\boldsymbol{G}_{\mathcal{A}_2}$ has little influence on the performance of SC and ogClust. Overall, ogClust performs consistently the best among all the simulation settings.

2.3.2 Robust estimation under outliers or heavy-tailed errors

To compare the performance of robust methods in guarding against outliers or violation of Gaussian mixture assumption, we perform simulation using the following settings:

- Setting A: The error term in the outcome association model is randomly drawn from standard normal distribution; normal assumption is not violated.
- Setting B: 10% of the observations are outliers and the error term is randomly drawn from unif(min-10,max+10).
- Setting C: The error term is randomly drawn from heavy-tailed lognormal distribution with log-mean 0 and log-standard deviation 1.

The simulation scheme follows Model 2 in Section 2.3, except that in step 5, the generation of outcome Y varies according to the different settings above. Under each setting, we compare the performance of ogClust, ogClust-Huber, ogClust-adHuber, and ogClust-median-truncation. Models are fit in the training data and tested in the testing data where four measures, i.e. RMSE, R^2, ARI and FNs, are calculated. We tune the number of selected genes by altering the parameter λ. The analysis above is performed on 100 sets of training

Table 1: Comparison of sparse K-means (SKM), penalized model based clustering (PMBC), supervised clustering (SC) and outcome-guided clustering (ogClust) under four simulation model settings with 600 observations and 2 baseline covariates, 1000 genes and 100 repetitions.

Methods	Estimated K			ARI	Selected Genes		Outcome	
	2	3	> 3		FPs	FNs	RMSE	R^2
Model 1: $\gamma = 1; \delta = 2$								
SKM	100	**0**	0	0.04	776.1	1.9	1.93	0.25
PMBC	82	**6**	12	0.08	88.0	11.1	1.93	0.24
SC	100	**0**	0	0.35	41.3	4.8	1.58	0.48
ogClust	62	**37**	1	0.45	5.9	3.0	1.55	0.51
Model 2: $\gamma = 1; \delta = 3$								
SKM	100	**0**	0	0.04	776.1	1.9	2.65	0.15
PMBC	82	**11**	7	0.10	87.0	10.2	2.67	0.14
SC	100	**0**	0	0.36	33.4	4.9	2.08	0.47
ogClust	2	**98**	0	0.86	14.6	0.0	1.90	0.56
Model 3: $\gamma = 1; \delta = 5$								
SKM	100	**0**	0	0.04	776.1	1.9	4.20	0.05
PMBC	74	**11**	15	0.09	100.3	10.2	4.22	0.05
SC	100	**0**	0	0.36	37.9	4.9	3.20	0.46
ogClust	0	**99**	1	0.91	14.5	0.0	2.70	0.61
Model 4: $\gamma = 3; \delta = 3$								
SKM	100	**0**	0	0.05	813.8	1.6	2.61	0.15
PMBC	83	**5**	12	0.16	96.7	8.7	2.64	0.15
SC	100	**0**	0	0.41	17.5	5.0	2.01	0.48
ogClust	1	**99**	0	0.88	12.0	0.0	1.75	0.63

and testing data such that we can obtained smooth curves capturing the trend of the four measures against the varying number of selected genes.

As shown in Figure 6, in the first column when the normal assumption is satisfied, the non-robust ogClust model performs the best and the three robust methods have only very slightly worse performance. This shows that robust estimation methods only minimally reduce efficiency when the Gaussian mixture assumption is true. On the other hand, when the Gaussian assumption is violated in the second and the third columns, the three robust methods greatly outperform the original model. ogClust-adHuber consistently outperforms ogClust-Huber with fixed cutoff. Compared to median truncation, ogClust-adHuber performs better for heavy-tailed error term but slightly worse with existence of outliers. ogClust-median-truncation can quickly capture the outcome associated DE genes with relatively low number of selected genes, but it performs worse than ogClust-adHuber in setting C because of the bias in parameter estimates. Since ogClust-adHuber outperforms ogClust-huber overall and performs well in most settings, it is recommended for general applications and will be evaluated in real data in Section 2.4.

2.3.3 Simulation to evaluate ogClust for survival outcome

The simulation scheme is the same as in Section 2.3, except that in step 5, a survival outcome is generated as follows: given subgroup index Z, survival time Y follows AFT model with log-logistic distribution, i.e. $(log(Y)|Z = k) = \beta_{0k} + X\beta + W\sigma$, where W $\sim$ standard logistic distribution and $\sigma = 0.5$. We set the end of follow-up time to be 100, any time that is greater than 100 is right-censored.

We evaluate the performance under four settings: (A) $\gamma = 1$ and $\delta = 1$, (B) $\gamma = 3$ and $\delta = 1$, (C) $\gamma = 1$ and $\delta = 2$, and (D) $\gamma = 3$ and $\delta = 2$, representing varying level of cluster separation (reflected by γ) and outcome association (δ). Similar to Section 2.3.2, we compare the performance of SKM, PMBC, SC and ogClust in terms of RMSE, R^2, ARI and FNs under each setting. Models are evaluated in 100 sets of simulated training and testing data. We vary the number of selected genes by tuning the penalty parameter λ and obtain smooth curves representing the trend of the four measures against the varying number of

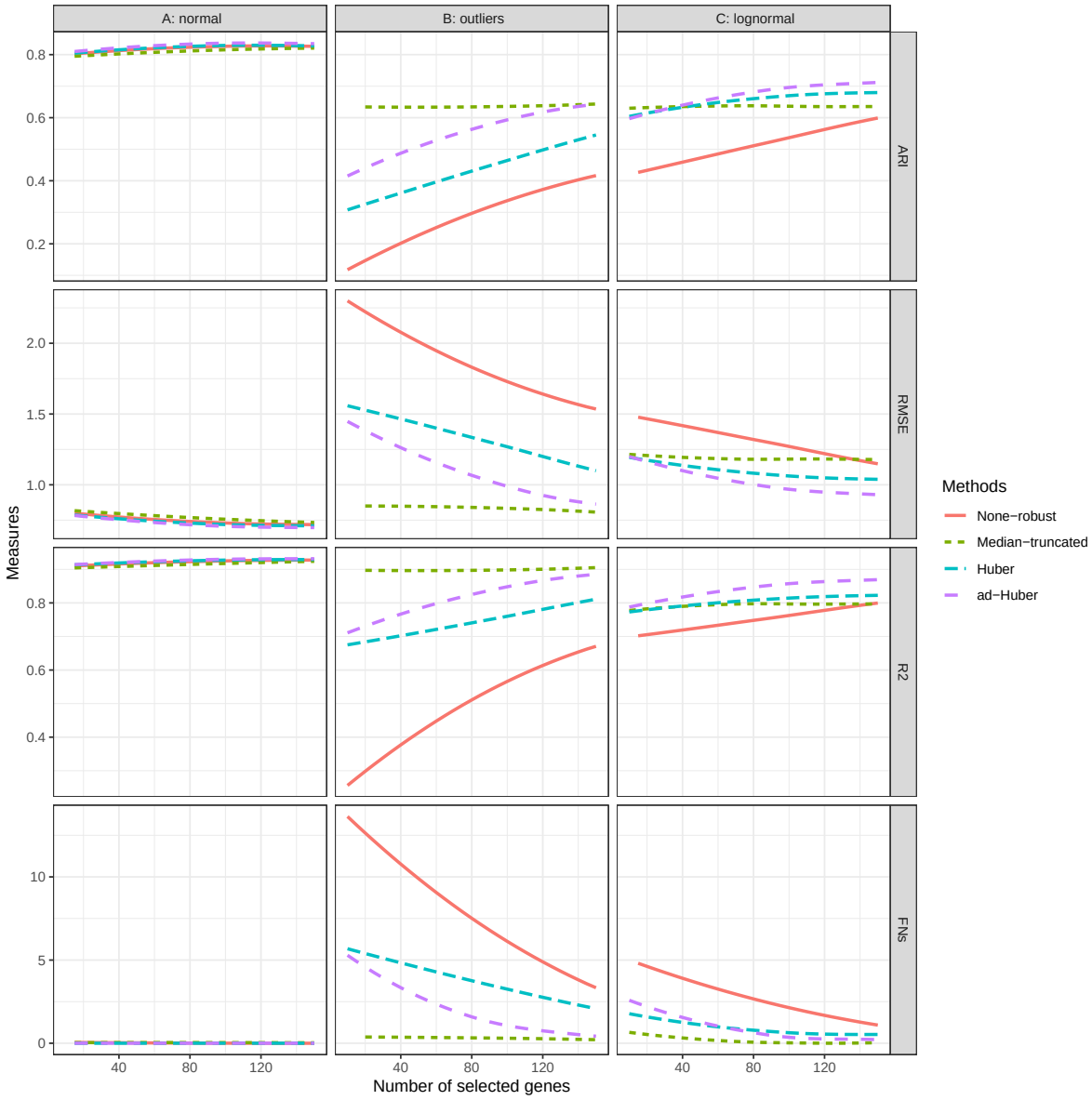


Figure 6: Comparison of ogClust and three robust ogClust methods under settings A: error term is randomly drawn from standard normal distribution, setting B: 10% of the observations are outliers, and setting C: error term is randomly drawn from heavy-tailed lognormal distribution. We compare RMSE, R^2, ARI and FNs (y-axis) vs number of genes selected in each setting (x-axis).

selected genes.

As the result shown in Figure 7, SKM and PMBC have the lowest ARI, highest RMSE, lowest R^2 and highest FNs among all four settings because they lack outcome guidance. SC has improved the four measures when compared with SKM and PMBC, and ogClust consistently outperforms the other three methods in all simulation settings.

2.4 Real Data Application

2.4.1 Apply to LGRC dataset

We apply the ogClust model to a lung disease transcriptomic dataset with $n = 319$ patients. Gene expression data are collected from Gene Expression Omnibus (GEO) GSE47460 and clinical information obtained from Lung Genomics Research Consortium (https://ltrcpublic.com/). The majority of patients were diagnosed with one of the two most representative lung disease subtypes: chronic obstructive pulmonary disease (COPD) and interstitial lung disease (ILD). COPD is a progressive lung disease caused by the repeated exposure to a noxious agent and is classified by symptoms, airflow obstruction and exacerbation history. ILD is a loosely defined group of patients characterized by changes in the interstitium of the lung, causing pulmonary restriction and impaired gas exchange. Current clinical classification criteria of the subtypes evolve over time and are debatable. They often fail to accommodate patients with atypical features, who are left unclassified. The current criteria also fail to reflect advances of high-throughput mRNA expression techniques to improve understanding and interpretation of the disease subtypes. In this section, we utilize the standardized form of a patient's forced expiratory score (FEV1%prd), a person's measured FEV1 normalized by the predicted FEV1 with healthy lung, as the clinical outcome Y to guide the disease subtyping. Age, gender and BMI are included as covariates X in the ogClust model.

Similar to simulations, we apply ogClust and compare with two existing methods, sparse K-means and supervised clustering. Data are first preprocessed by conventional procedures following an earlier publication (Kim et al., 2015). Non-expressed genes (mean expression

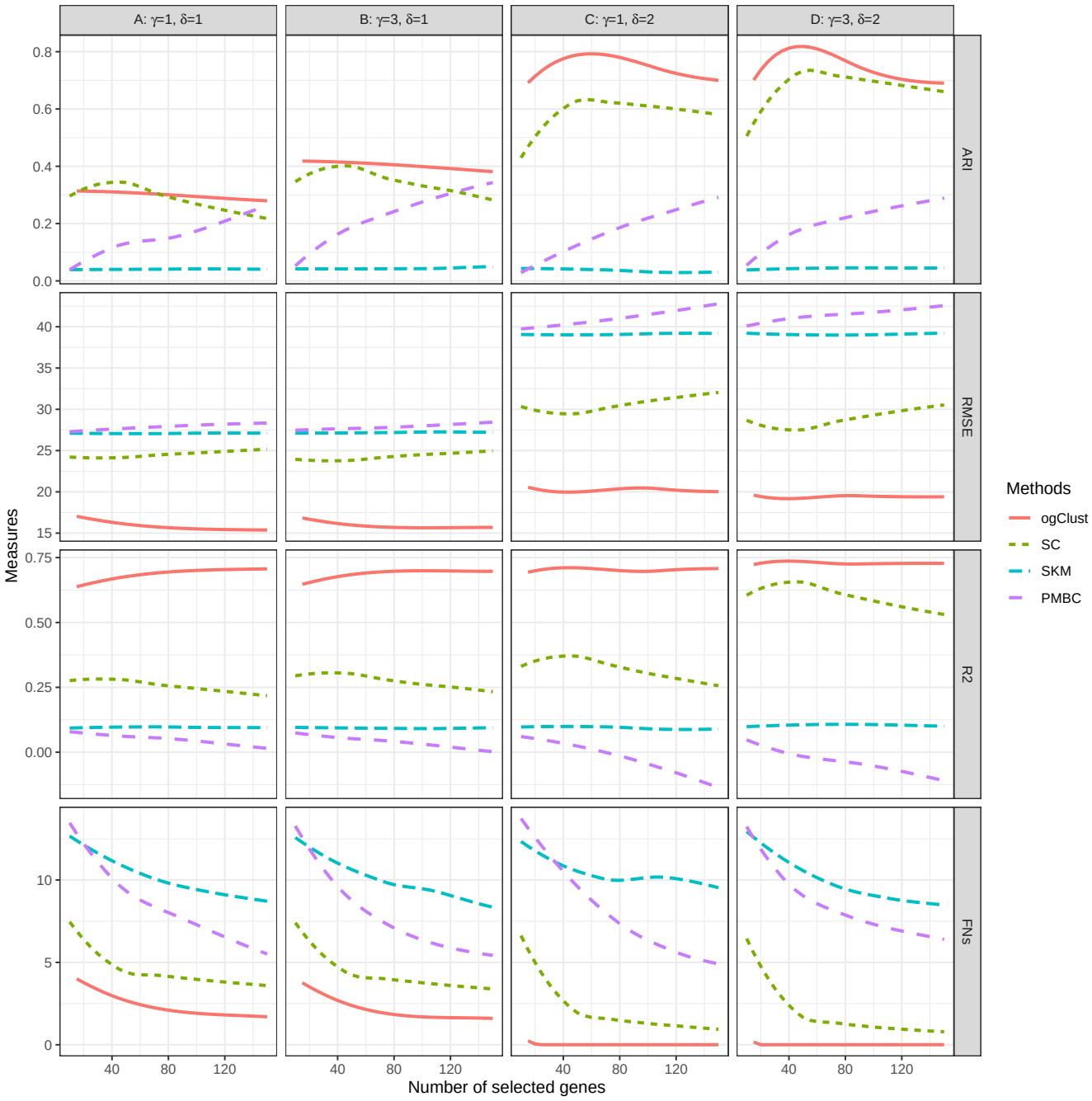


Figure 7: Comparison of ogClust and SC,SKM and PMBC under four simulation settings with survival outcome. We compare RMSE, R^2, ARI and FNs (y-axis) vs number of genes selected in each setting (x-axis).

in the lower 50 percentile) are filtered and top informative genes (genes with the largest variance) are selected for analysis. Table 4 shows the result when setting the number of subgroups $K = 3$ (see Figure 4 for analysis of justifying selection of K), and using the top 500, 1000 and 2000 pre-filtered genes (by the largest variance) in the comparison. Since, unlike in simulations, the underlying true class labels Z are unknown, we benchmark the clustering performance in several measures. We compare the outcome prediction error using RMSE and R^2 and evaluate p-value of the association between subgroups and the FEV1%prd outcome by Kruskal-Wallis test. We also show the number of selected genes, which has at least one non-zero $\hat{\gamma}_{jk}$ $(1 \leq k \leq K)$, used to characterize the disease subtypes. The result in Table 4 shows that ogClust identifies disease subtypes with better association with clinical outcome with smaller number of genes compared to sparse K-means and supervised clustering. For example, when the top 2000 pre-filtered genes are used, ogClust selects 22 genes to define three disease subtypes that explain FEV1%prd outcome with $R^2 = 0.350$ and association $p = 1.84 \times 10^{-57}$. In contrast, sparse K-means needs 253 genes to reach $R^2 = 0.055$ and $p = 5.11 \times 10^{-7}$. Although supervised clustering also aims to detect subtypes associated with outcome, it only improves slightly from sparse K-means with R^2=0.058 and $p = 2.11 \times 10^{-8}$. Compared with ogClust, ogClust-adHuber better explains outcome with $R^2 = 0.455$, and has relatively lower association with $p = 9.49 \times 10^{-24}$. Figure 8A shows the clinical diagnosis (piechart above), expression of the selected genes (heatmap in the middle), distribution of outcome (boxplot below) for each method. For the three clusters identified by ogClust, one cluster is almost purely COPD (blue bar), one cluster is almost purely ILD (red bar) and one cluster in between with mixed COPD and ILD (green bar). The result indicates existence of a COPD/ILD intermediate subtype of patients that have distinct molecular expression pattern and FEV1%prd clinical outcome. SKM and SC, however, identify three clusters with more mixed diagnosis of COPD and ILD and are dominated by non-outcome-related genes. We next evaluate the enriched pathways and canonical functions using Ingenuity Pathway Analysis (IPA) tool. To account for the randomness of gene selection, we repeat the analysis in 500 bootstrapped datasets and select the top 200 most frequently selected genes as our final input gene list for IPA. As shown in Figure 8B, the genes selected by ogClust are more significantly enriched in pathways associated with immune responses and

organismal injury, while other methods select genes largely irrelevant to lung disease (e.g. cancer and dermatological diseases).

Table 2: Comparison of sparse K-means (SKM), supervised clustering (SC), outcome-guided clustering (ogClust), and ogClust with adaptive-Huber loss (ogClust-adHuber) when applied to the lung disease transcriptomic dataset. We set the number of subgroups K equals 3, top 500, 1000, and 2000 genes are used. RMSE and R^2 measure outcome prediction performance. Kruskal-Wallis test measures whether outcome is associated with the clusters. Fisher's exact test measures whether subgroup label is consistent with the clinical diagnosis.

K	Total number of genes	Methods	RMSE	R^2	Kruskal-Wallis test	Genes selected	Fisher's exact test
3	500	SKM	0.208	0.060	7.28×10^{-5}	218	1.34×10^{-9}
		SC	0.203	0.101	1.36×10^{-7}	70	3.65×10^{-24}
		ogClust	0.189	0.226	7.21×10^{-56}	33	2.81×10^{-41}
		ogClust-adHuber	0.168	0.386	2.24×10^{-47}	11	1.12×10^{-18}
3	1000	SKM	0.209	0.052	1.79×10^{-6}	172	1.27×10^{-7}
		SC	0.204	0.086	2.31×10^{-5}	60	4.43×10^{-21}
		ogClust	0.186	0.249	7.62×10^{-56}	40	8.05×10^{-41}
		ogClust-adHuber	0.161	0.432	1.00×10^{-57}	25	1.87×10^{-31}
3	2000	SKM	0.208	0.055	5.11×10^{-7}	253	4.16×10^{-23}
		SC	0.207	0.058	2.11×10^{-8}	45	8.52×10^{-14}
		ogClust	0.173	0.350	1.84×10^{-57}	22	8.56×10^{-34}
		ogClust-adHuber	0.158	0.455	9.49×10^{-24}	24	5.51×10^{-43}

2.5 Discussion and Conclusion

In this chapter, we propose a unified outcome-guided clustering (ogClust) framework for disease subtyping from omics data. ogClust links the disease subtyping model and the

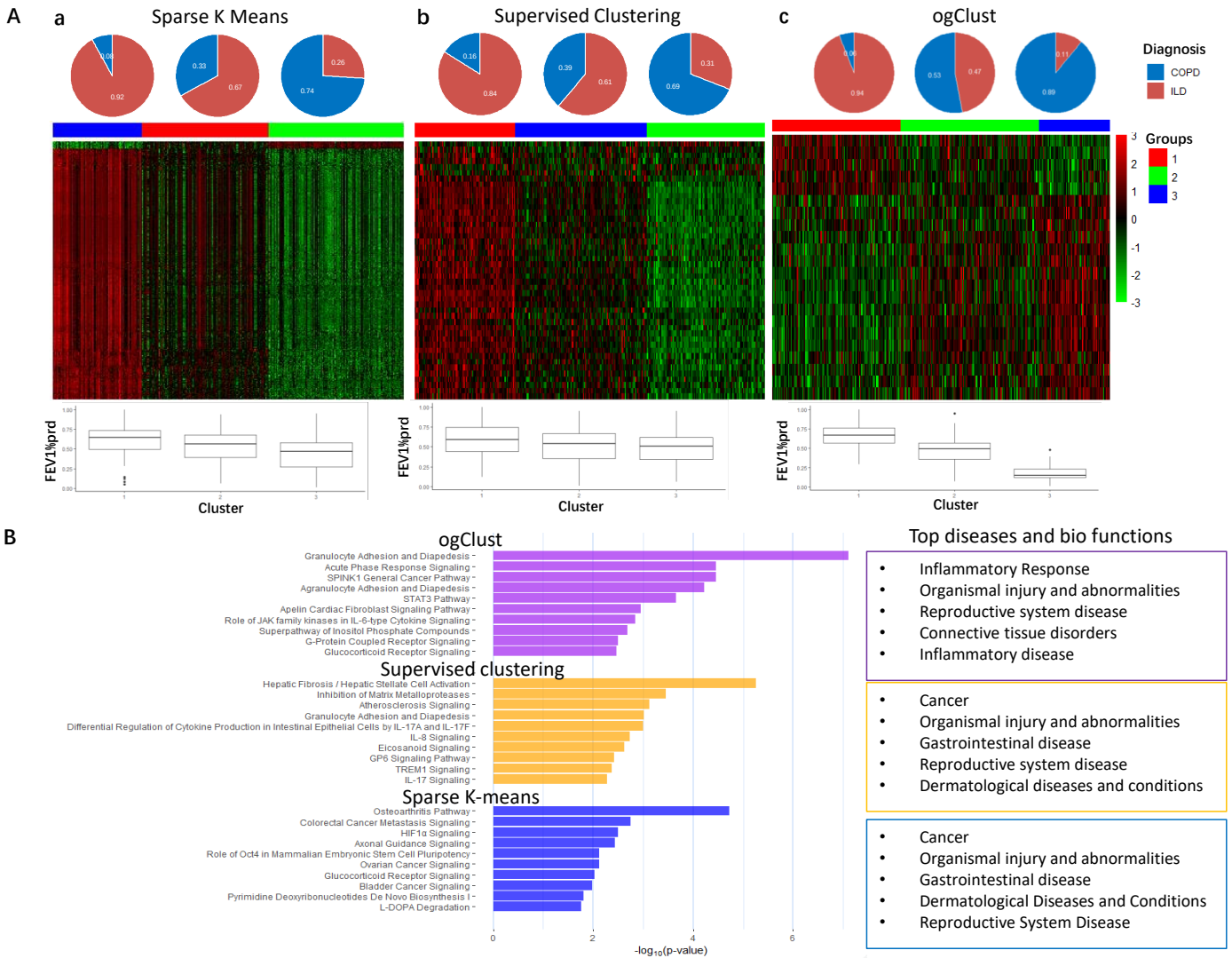


Figure 8: (A) Pie chart of clinical diagnosis (top), heatmap of expression of selected genes (middle), and boxplot of outcome FEV1%prd (bottom) in each cluster for (a) SKM ,(b) SC, and (c) ogClust. (B) Enriched pathways and top disease annotations of the selected genes for SKM, SC and ogClust.

outcome association model through a latent cluster label Z. From extensive simulations and a real data application on lung disease transcriptomic data, we demonstrate the ability of ogClust to identify outcome associated clusters (disease subtypes) that are otherwise easily masked by other facets of clinically irrelevant cluster structure. Additionally, ogClust is immediately applicable to future patients to predict their disease subtypes. Unlike hard (deterministic) assignment in hierarchical clustering or K-means, the prediction is a soft assignment with classification probability, reflecting the confidence of the subtyping prediction of each patient.

As mentioned in the Introduction section, the concept of outcome-guided clustering is novel in the field. It involves both supervised and unsupervised components in the framework but differs from classical clustering or classification problems. It should not be confused with two types of semi-supervised machine learning, where mixing of labeled and unlabeled data are trained or constrained prior knowledge is imposed in clustering. To some extent, it is similar to latent class models in outcome association, but the latter model cannot provide latent class assignments for future observations, while the ogClust model can predict disease subtypes for precision medicine purpose.

In the current ogClust model, omics data $\boldsymbol{G}$ from a single source are used to characterize the subtype Z and covariates $\boldsymbol{X}$ do not contribute to clustering. Integration of multi-source of data (e.g. multiple transcriptomic studies or a single study with multi-omics data) requires more careful modeling for each problem setting and will be a future direction.

ogClust parameter estimation is implemented via a modified EM algorithm and thus provides fast computing for high-dimensional data. In the lung disease example, the model fitting can be finished in 2.17 minutes using 1 core (Intel Xeon 6130) for $n = 319$ patients, $q = 2000$ genes and $p = 3$ covariates. To select tuning parameters K and λ by BIC, multiple runs are necessary. An R package is freely available on https://github.com/liupeng2117/ogClust, along with all data and code to reproduce results in this chapter.

3.0 Outcome-Guided Disease Subtyping Integrating Prior Biological Information and Multiomics Datasets

3.1 Introduction

[1]Many complex diseases were once thought of as a single entity within which all patients receive uniform diagnosis and treatment. Modern omics studies, however, have revealed numerous molecular subtypes with differential disease mechanisms, therapeutic targets and survival outcomes. For example, breast cancer subtypes (Luminal A, Luminal B, Basal and Her2) were identified in 2000 (Perou et al., 2000) and repeatedly validated afterwards. Importantly, these subtypes have clinical relevance since they show different prognostic survival and respond to different treatments (Masuda et al., 2013; Burstein et al., 2015). Moving towards clinical practice, disease subtyping in breast cancer has developed tailored screening/prevention programs and novel treatment strategies to decrease mortality (Jemal et al., 2009b). Other notable examples include six molecular subtypes identified in triple negative breast cancer (TNBC) (Lehmann et al., 2011; Chen et al., 2012), four subtypes in colorectal cancer (Guinney et al., 2015), and many different cancers (Linnekamp et al., 2015; Rojas et al., 2016) and psychiatric disorders (Stessman et al., 2016; Demkow and Wolańczyk, 2017; Bowen et al., 2019). Such disease subtyping by omics data has become an effective approach to dissect the heterogeneous patient population into homogeneous subgroups towards precision medicine.

Disease subtyping using single cohort/omics analysis suffers from sample size limitation and reproducibility issues. Over the years large amount of omics data are accumulated in public databases and depositories, for example, The Cancer Genome Atlas (TCGA) http://cancergenome.nih.gov, Gene Expression Omnibus (GEO) http://www.ncbi.nlm.nih.gov/geo/, Sequence Read Archive (SRA) http://www.ncbi.nlm.nih.gov/sra, just to name a few. The ever increasing number of omics data provide unprecedented opportunities for unveiling the disease subtypes and mechanisms via omics data integrative analysis. On the other hand,

[1]This chapter will be submitted to Bioinformatics.

there are a tremendous amount of biological information (e.g. pathway information) that can be incorporated to guide the omics data integrative analysis.

In the literature, integrating analysis for omics data can be categorized in two categories: 1) horizontal integration and 2) vertical integration (Tseng et al., 2015). In horizontal meta-analysis, multiple studies of the same type of omics data (e.g., gene expression) from different cohorts are combined to increase sample size and statistical power, which is widedly used in differential expression analysis (Ramasamy et al., 2008), pathway analysis (Shen and Tseng, 2010) and subtype discovery (Huo et al., 2016).

In contrast, vertical integrative analysis aims to integrate multi-level omics data from the same patient cohort (e.g., gene expression data, genome-wide profiling of somatic mutation, DNA copy number, DNA methylation or miRNA expression from the same set of biological samples (Shen et al., 2009; Richardson et al., 2016; Kim et al., 2017). In this chapter, we focus on vertical omics integrative analysis for disease subtype discovery.

Existing vertical integrative methods on disease subtyping usually apply unsupervised (multi-omics clustering) approaches. Shen et al. (2009) proposed a penalized latent variable model-based clustering method, iCluster, for joint modeling of multiple types of omics data. It assumes a consistent clustering across multi-level omics data, which may not hold in some cases. JIVE (Lock et al., 2013) performed decomposition of variation into joint and individual components, allowing common and omic-specific molecular profiling structures. Lock and Dunson (2013) fitted a finite Dirichlet mixture model to perform Bayesian consensus clustering (BCC), which extends the JIVE modeling strategy within a Bayesian framework. Huo and Tseng (2017) proposed integrative sparse K-means approach to integratively cluster multi-omics data with feature selection and incorporating biological information.

Since outcome information is not considered in the aforementioned clustering methods, the identified disease subtypes and subtype-specific molecular profiles are often not associated with the outcome. In the literature, Bair and Tibshirani (2004), Koestler et al. (2010) and Gaynor and Bair (2013) have developed a two-stage semi-supervised method, where K-means or other standard clustering methods, are applied to the top M features with the highest marginal outcome association. The two-stage approach can only handle single omics data and is suboptimal in the ad hoc selection of the top features. Ahmad and Fröhlich (2017) proposed

a Bayesian method to cluster omics data with survival outcomes and molecular features, however, it is computationally intensive and only affordable for up to ∼100 features. Little effort has been done to develop integrative disease subtyping methods with the guidance of outcome.

In this chapter, we propose an extended integrative outcome guided clustering model (ogClust)to solve the followings issues: 1) integrative analysis of multi-level omics data and prior biological information 2) associating outcome with identified disease subtypes 3) feature selection 4) prediction of subgroup label and outcome for future patients. This method is a multi-omics extension of our previous ogClust method in Chapter 2, where only single omics data was used in outcome guided disease subtyping.

The chapter is structured as follows. Section 3.2 introduces the integrative ogClust model for multi-omics data and benchmarking criteria for evaluation. Section 3.3 evaluates its performance by extensive simulations. An application to lung disease transcriptome (LGRC) dataset is illustrated in Section 3.4. Discussion and conclusion is in 3.5

3.2 Proposed Method

3.2.1 Model and notations

We consider the problem of disease subtyping (clustering) of n subjects from high-dimensional datasets $\boldsymbol{G} = \{\mathbb{G}^{(1)}, ..., \mathbb{G}^{(S)}\}$, where $\mathbb{G}^{(s)} = \{g_{ij}^{(s)}, 1 \leq i \leq n, 1 \leq j \leq q^{(s)}, 1 \leq s \leq S\}$. The multi-omics datasets $\boldsymbol{G}$ could be a combination of mRNA expression, miRNA expression, methylation or phenomic data. Dimension $q^{(s)}$ of the s^{th} omics dataset can be at the scale of hundreds to thousands and let $q = \sum_{s=1}^{S} q^{(s)}$. Our goal is to cluster n observations into K clinically meaningful clusters represented by latent group label $\mathbb{Z} = \{z_i, 1 \leq i \leq n\}$, $z_i \in \{1, \ldots, K\}$, and $z_i = k$ means that observation i is assigned to cluster k $(1 \leq k \leq K)$. Clustering using single omics dataset fails to consider the regulatory mechanisms across omics and it is not uncommon that different omics dataset produce distinct clustering results. It is desirable to combines multiple datasets and taking into account prior knowledge $\mathbb{P}$

such as pathways or miRNA targeting database, to improve reproducibility of subtyping and interpretation of feature selection. Furthermore, clustering result purely from $\boldsymbol{G}$ may not necessarily be clinically useful as discussed in Section 2.1, we assume that a clinical outcome $\mathbb{Y} = \{y_i, 1 \leq i \leq n\}$ is given to guide the clustering. We also assume a set of pre-specified covariates $\mathbb{X} = \{x_{ij}, 1 \leq i \leq n, 1 \leq j \leq p\}$, where the p covariates (e.g. age, gender, etc.) are potentially associated with the outcome and may confound with the association between Z and Y. Let $\boldsymbol{g}_i = (\boldsymbol{g}_i^{(1)}, \ldots, \boldsymbol{g}_i^{(S)})^T$ where $\boldsymbol{g}_i^{(s)} = (g_{i1}^{(s)}, \ldots, g_{iq}^{(s)})$ and let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, we assume observed data $(y_i, \boldsymbol{x}_i, \boldsymbol{g}_i)$ for subject i $(1 \leq i \leq n)$ are independent realizations of the model for $(Y, \boldsymbol{X}, \boldsymbol{G})$.

As shown in Figure 9, the proposed integrative ogClust framework consists of two components: *disease subtyping model* and *outcome association model*. The disease subtyping

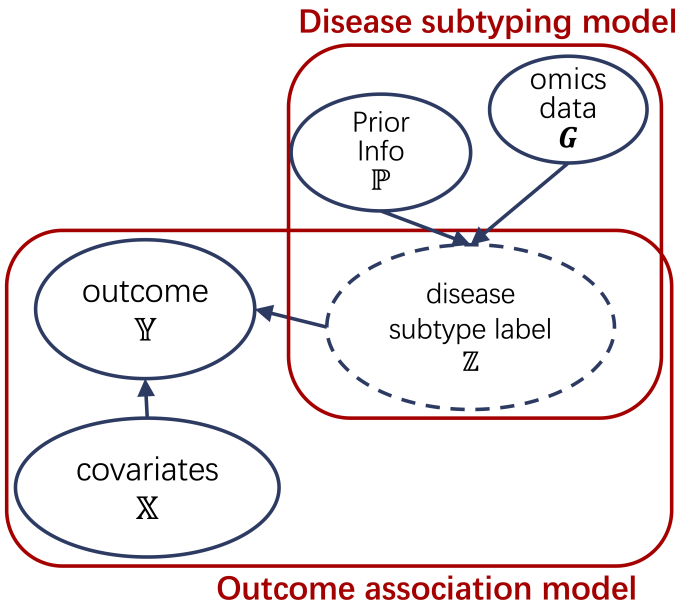


Figure 9: A graphical illustration of the unified regression model. Y is the outcome to guide clustering, X are the baseline covariates that are believed to have effects on Y. G are the combined datasets that defines the outcome associated subgroups. P is prior knowledge to incoporate into the model, and Z is the unobserved latent subgroup index.

model is specified in the same manner as in Section 2.2.1. Specifically, it is a conventional high-dimensional discriminant analysis where we train to characterize $\pi_k = Pr(Z = k|\boldsymbol{G})$ (or $\pi_{ik} = Pr(Z_i = k|\boldsymbol{g}_i)$ for observation i). In this chapter, we apply a multinomial logistic regression $\pi_{ik}|\boldsymbol{\gamma} = \frac{\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_k)}{\sum_{l=1}^{K}\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_l)}$, where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k, 1 \leq k \leq K\}$ and $\boldsymbol{\gamma}_k = (\gamma_{1k}, \ldots, \gamma_{qk})^T$. Since q is usually large, we assume only a small subset $\mathcal{A} \subset \{1, \ldots, q\}$ of features effective in characterizing the clusters that affect the outcome, where its cardinality $\text{card}(\mathcal{A}) < \min(n, q)$. In other words, $\boldsymbol{\gamma}_{[j]} \neq \mathbf{0}$ if $j \in \mathcal{A}$ and $\boldsymbol{\gamma}_{[j]} = \mathbf{0}$ if $j \in (\mathcal{A})^c$, where $\boldsymbol{\gamma}_{[j]} = \{\gamma_{j1}, \ldots, \gamma_{jK}\}$. We designed a sparse overlapping group lasso regularization (Huo and Tseng, 2017) to incorporate prior knowledge of feature groups (e.g. pathways) to select features for subgrouping.

In the outcome association model, we assume the same mixture model as Equation (1) in Section 2.2.1.

$$f(y_i; \boldsymbol{x_i}) = \sum_{k=1}^{K} \pi_{ik} f_k(y_i; \boldsymbol{x_i}),$$

We assume a continuous response Y where the k-th mixture density $f_k(y; \boldsymbol{x}, \beta_{0k}, \boldsymbol{\beta}, \sigma)$ is parameterized by cluster specific intercept β_{0k}, common covariate effect $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ and a homogeneous error σ. In this chapter, we specifically assume $y_i|z_i = k \sim N(\beta_{0k} + \boldsymbol{\beta}^T\boldsymbol{x_i}, \sigma^2)$ with mixture probability $\pi_{ik} = \frac{\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_k)}{\sum_{l=1}^{K}\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_l)}$, $k = 1, \ldots, K$. Denote by $\boldsymbol{\theta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma\}$ the collection of all parameters from the two models in ogClust ($\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0K})^T$), given $\mathbb{Y}$, $\mathbb{X}$ and $\boldsymbol{G}$, $\boldsymbol{\theta}$ can be estimated by maximizing the following sample likelihood of the basic model, which is the same as Equation (2):

$$\mathrm{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_{ik}(\boldsymbol{g_i}, \boldsymbol{\gamma}) f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right).$$

3.2.2 Design of overlapping group lasso penalty

Since $\boldsymbol{G}$ is usually high dimensional, including features with non-predictive effect will introduce extra noise to the disease subtyping model and may produce irrelevant subtypes that are not necessarily related to the disease outcome of interest. Therefore we need to add a penalty term to the likelihood objective function for selecting the informative features. Considering the multi-omics structure of the datasets and the need of incorporating prior

information, a sparse overlapping group structure which allows for overlapping group features and sparse informative features within groups is desirable. In this section, we consider three motivating scenarios as illustrated in Figure 10. Figure 10 (A) demonstrate a multi-omics non-overlapping scenario where all omics features of the same gene are in a group for integrative analysis. In 10 (B) where there are gene and miRNA expression datasets, a miRNA and its regulating genes are in a group, and 10 (C) shows the transcriptomic application where pathways are overlapping groups.

In the following of this section, we will illustrate with a sparse overlapping group LASSO penalty. The penalized likelihood can be specified as

$$l_n(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{g_i}, \boldsymbol{\gamma}) f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) - R(\boldsymbol{\gamma})$$

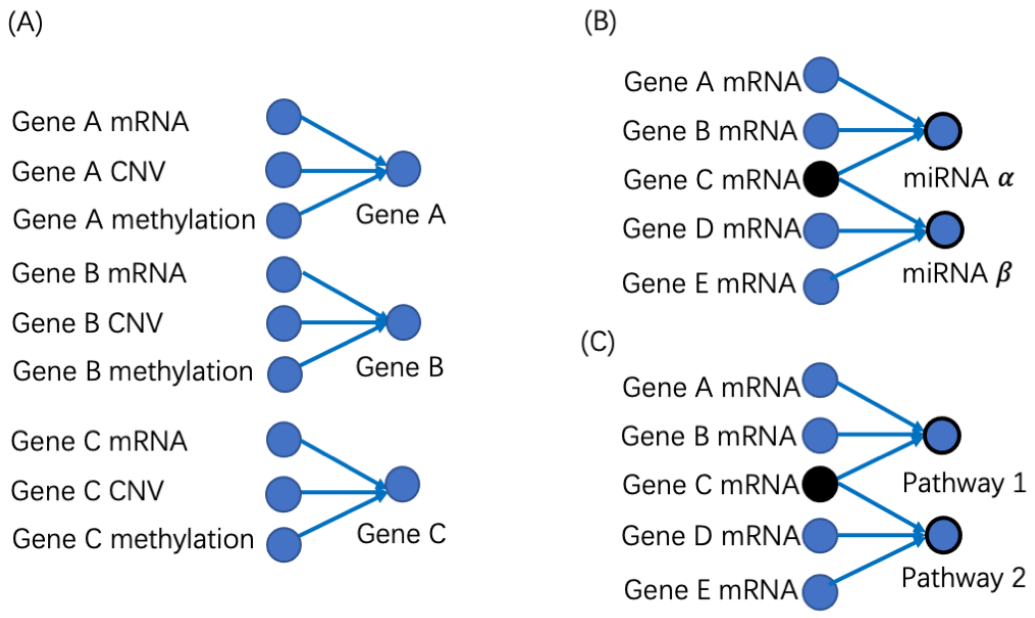


Figure 10: Motivating examples of using sparse overlapping group LASSO structure for feature selection in multi-omics datasets. (A) Combine all omics features of the same gene as a groups. (B) Use each miRNA and its targeted genes as a group. (C) For transcriptomic application, use pathways as the overlapping groups.

The sparse overlapping group lasso penalty term $R(\gamma)$ can be specified as

$$R(\gamma) = \lambda\alpha \sum_{j=1}^{q} ||\boldsymbol{\gamma}_j||_2 + \lambda(1-\alpha) \sum_{g=1}^{G} w_g \left\|\boldsymbol{m}_g \circ \boldsymbol{\gamma}\right\|_2 \tag{10}$$

where λ is the penalty tuning parameter controlling the number of nonzero features/pathways. α is a term controlling the balance between the individual feature and group feature penalties. If $\alpha = 1$ there is no group feature penalty and only the individual feature penalty is used. Conversely, if $\alpha = 0$ there is no individual feature penalty but only with group penalty. The first term $\sum_{j=1}^{q} ||\boldsymbol{\gamma}_j||_2$ gives sparsity for individual features, where $\boldsymbol{\gamma}_j = (\gamma_{j1}, ..., \gamma_{jK})$.

Following Huo and Tseng (2017), the second term is the overlapping group lasso penalty, which is defined as

$$\sum_{g=1}^{G} w_g \left\|\boldsymbol{m}_g \circ \boldsymbol{\gamma}\right\|_2$$

where G is the number of possible overlapping groups from prior biological knowledge. w_g is the group weight coefficient for group g, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q)^T$ is a vector of parameters where $\boldsymbol{\gamma}_j = (\gamma_{j1}, ..., \gamma_{jK})$, $\boldsymbol{m}_g = (\boldsymbol{m}_{g1}, \boldsymbol{m}_{g2}, ..., \boldsymbol{m}_{gJ})^T$ where $\boldsymbol{m}_{gj} = (m_{gj1}, ..., m_{gjK})$ is the design vector of gth group, and $\circ$ represents Hadamard product(i.e. element-wise product). Denote $\mathcal{J}_g$ the collection of features in group $g(1 \leq g \leq G)$, The design of $\boldsymbol{m}_{gj} = (m_{gj1}, ..., m_{gjK})$ can be specified as $m_{gj1} = ... = m_{gjK} = \{j \in \mathcal{J}_g\}/\sqrt{h(j)}$, where $h(j) = \sum_{g=1}^{G} I(j \in \mathcal{J}_g)$ denoting the frequency of feature j appearing in different groups.

Consider a toy example under the scenario of Figure 10(C). There are in total 8 genes $\{a, b, c, d, e, f, g, h\}$, two subgroups ($K = 2$) and two pathways $P_1 = \{a, b, c, d, e\}$ and $P_2 = \{c, d, e, f, g\}$. The the design vectors for the two pathways are $\boldsymbol{m}_1 = (\mathbf{1}, \mathbf{1}, \mathbf{0.5}, \mathbf{0.5}, \mathbf{0.5}, \mathbf{0}, \mathbf{0}, \mathbf{0})^T$ and $\boldsymbol{m}_2 = (\mathbf{0}, \mathbf{0}, \mathbf{0.5}, \mathbf{0.5}, \mathbf{0.5}, \mathbf{1}, \mathbf{1}, \mathbf{0})^T$, note that the elements are vectors of length K(i.e. $\mathbf{1} = (1, 1)$, $\mathbf{0.5} = (0.5, 0.5)$ and $\mathbf{0} = (0, 0)$).

To choose w_g, we defines the intrinsic feature set $\mathcal{I}$ (i.e. features that contribute to the underlying true subtyping). We choose the weight design $w_g = \sqrt{\sum_{j \in (J_g \cap \mathcal{I})} \sum_{k=1}^{K} 1/h_k(j)}$ following the unbiased feature selection property in Huo and Tseng (2017). Since the intrinsic feature set $\mathcal{I}$ is unknown in reality, in practice, the estimated intrinsic feature set $\hat{\mathcal{I}}$ is used from the set of features selected in the previous EM iteration. For the first EM iteration, $\hat{\mathcal{I}}$ is the set of features with nonzero $\boldsymbol{\gamma}$ initializations.

3.2.3 Numerical solution

A numerical method using EM algorithm is proposed for integrative ogClust parameter estimation with the proposed overlapping group lasso penalty. By introducing $z_{ik}, k = 1, \ldots, K$, as missing indicator variables, following the seminal idea in Dempster et al. (1977), the complete log likelihood function is the same as Equation (3), which can be written as

$$l_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ z_{ik} \log \pi_{ik} + z_{ik} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) \right\},$$

where $z_{ik} = 1$ if subject i belongs to subgroup k, and $z_{ik} = 0$ otherwise.

Since features are usually high dimensional, including features in $\mathcal{A}^c$ with non-predictive effect will introduce extra noise to the disease subtyping model and may produce irrelevant subtypes that are not necessarily related to the disease outcome of interest.

The penalized log-likelihood function is the same as Equation (4), which can be defined as

$$\tilde{l}_n^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ z_{ik} \log \pi_{ik} + z_{ik} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) \right\} - R(\boldsymbol{\gamma}),$$

where $R(\boldsymbol{\gamma})$ is defined in Equation (10). The irrelevant groups and features are removed by shrinking corresponding elements of $\boldsymbol{\gamma}_{[j]}$ to zero, thus a sub-model is automatically selected. This procedure performs group/feature selection and numerical estimation of parameters simultaneously.

Maximization of $\tilde{l}_n^c(\boldsymbol{\theta})$ can be achieved by sequentially and iteratively updating $\boldsymbol{\beta_0}$, $\boldsymbol{\beta}$, σ and $\boldsymbol{\gamma}$ in an EM-ADMM algorithm, which takes the following steps:

- The E step is the same as that in Section 2.2.2, which computes the conditional expectation of the function $\tilde{l}_n^c(\boldsymbol{\theta})$ with respect to z_{ik}, given the observed data y_i, $\boldsymbol{x}_i$ and the current parameter estimates $\boldsymbol{\theta}^{(m)}$,

$$Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log \pi_{ik} + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log f\left(y_i; \boldsymbol{x}_i, \beta_{0k}, \boldsymbol{\beta}, \sigma\right) - \lambda \sum_{j=1}^{q} R(\boldsymbol{\gamma}_j),$$

  where the posterior weights

$$w_{ik}^{(m)} = E\left(Z_{ik} | y_i, \boldsymbol{x}_i, \boldsymbol{\theta}^{(m)}\right) = \frac{\pi_{ik}^{(m)} f\left(y_i; \boldsymbol{x}_i, \beta_{0k}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\right)}{\sum_{l=1}^{K} \pi_{il}^{(m)} f\left(y_i; \boldsymbol{x}_i, \beta_{0l}^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\right)}. \tag{11}$$

- The M step on the $(m+1)$-th iteration maximizes the $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}\right)$ with respect to $\boldsymbol{\theta}$. By taking partial derivatives, it is easy to show that $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}$ and σ^2 are updated by the following updating equations which are the same as Equations (6) - (8):

$$\beta_{0k}^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(m)} \left(y_i - (\boldsymbol{\beta}^{(m)})^T \boldsymbol{x}_i\right)}{\sum_{i=1}^{n} w_{ik}^{(m)}}, \quad k = 1, \ldots, K, \tag{12}$$

$$\beta_{\ell}^{(m+1)} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} x_{i\ell} \left(y_i - \beta_{0k}^{(m+1)} - \sum_{h \neq \ell} \beta_h^{(m)} x_{ih}\right)}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} x_{i\ell}^2}, \quad \ell = 1, \ldots, p, \tag{13}$$

$$(\sigma^{(m+1)})^2 = \frac{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \left(y_i - \beta_{0k}^{(m+1)} - (\boldsymbol{\beta}^{(m+1)})^T \boldsymbol{x}_i\right)^2}{\sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)}}. \tag{14}$$

The updated estimates $\boldsymbol{\gamma}^{(m+1)}$ is obtained by minimizing the partial log likelihood function below:

$$\min \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)^2 \right\} + \lambda\alpha \sum_{j=1}^{q} \left\|\boldsymbol{\gamma}_j\right\|_2 + \lambda(1-\alpha) \sum_{g=1}^{G} w_g \left\|\boldsymbol{m_g} \circ \boldsymbol{\gamma}\right\|_2 \tag{15}$$

where the partial likelihood $\sum_{i=1}^{n} \sum_{k=1}^{K} \omega_{ik}^{(m)} \log \pi_{ik}$ is approximated by quadratic approximation $\sum_{k=1}^{K} \left\{ W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)^2 \right\}$, and $h_{ik} = \boldsymbol{g}_i^T \boldsymbol{\gamma}_k^{(m)} + \frac{w_{ik}^{(m)} - \pi_{ik}^{(m)}}{W_{ik}}$, $W_{ik} = \pi_{ik}^{(m)}(1 - \pi_{ik}^{(m)})$. To optimize Equation (15), we first transform the penalty $R(\boldsymbol{\gamma})$ such that the first and second terms are combined together, simplifying the penalty to overlapping group lasso, Then we perform ADMM algorithm to solve and estimate $\boldsymbol{\gamma}$. The detailed steps are discussed below:

1. We can rewrite the objective function as

$$\min \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \left\{ W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)^2 \right\} + \sum_{j=1}^{q} \left\|\lambda\alpha \boldsymbol{\phi_j} \circ \boldsymbol{\gamma}\right\|_2 + \sum_{g=1}^{G} \left\|\lambda(1-\alpha) w_g \boldsymbol{m_g} \circ \boldsymbol{\gamma}\right\|_2$$

, where $\boldsymbol{\phi}_j = (\boldsymbol{\phi}_{j1}, ..., \boldsymbol{\phi}_{jq})$, and $\boldsymbol{\phi}_{ji} = \{\phi_{ji1}, ..., \phi_{jiK}\}$. if $j = i$, $\boldsymbol{\phi}_{ji} = 1$, and if $j \neq i$, $\boldsymbol{\phi}_{ji} = 0$.

2. We can combine the q feature groups and G overlapping groups:

$$\min \tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{W_{ik}\left(h_{ik}-\boldsymbol{g}_i^T\boldsymbol{\gamma}_k\right)^2\right\}+\sum_{\mathrm{g}=1}^{q+G}\left\|\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}\right\|_2$$
$$\text{where } \beta_g=\begin{cases}\lambda\alpha\boldsymbol{\varphi}_j & \text{if } 1\le g\le q\\ \lambda(1-\alpha)w_g\boldsymbol{m}_g & \text{if } q+1\le g\le q+G\end{cases}$$

Therefore, optimizing the objective function is a convex problem with respect to $\boldsymbol{\gamma}$. We use ADMM algorithm in the next step to update $\boldsymbol{\gamma}$

3. We introduce an auxiliary variable $\boldsymbol{x}_g$ and write down the augmented Lagrange:

$$\min -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{W_{ik}\left(h_{ik}-\boldsymbol{g}_i^T\boldsymbol{\gamma}_k\right)^2\right\}+\sum_{\mathrm{g}=1}^{q+G}\left\|\boldsymbol{x}_g\right\|_2+\sum_{\mathrm{g}=1}^{q+G}\left\{\boldsymbol{s}_g^T\left(\boldsymbol{x}_g-\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}\right)+\frac{\rho}{2}\left\|\boldsymbol{x}_g-\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}\right\|_2^2\right\}$$

s.t. $\boldsymbol{x}_g=\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}$, and ρ is the augmented Lagrange parameter. This problem is equivalent to the original objective function, since any feasible terms added to the objective function is zero.

4. We define a new dual variable $\boldsymbol{u}_g=\frac{\boldsymbol{s}_g}{\rho}$. The scaled ADMM combines the linear and quadratic term in the augmented Lagrangian. Here the augmented Lagrange is minimized jointly with respect to two primal variables $\boldsymbol{x}_g,\boldsymbol{\gamma}$ and dual variable $\boldsymbol{u}_g$ in an alternating/sequential fashion. The updating equations can be written explicitly:

$$\begin{aligned}\boldsymbol{x}_g^+&=\underset{x_g}{\operatorname{argmin}}\left\|\boldsymbol{x}_g\right\|_2+\tfrac{\rho}{2}\left\|\boldsymbol{x}_g-\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}+\boldsymbol{u}_{gk}\right\|_2^2\\ \boldsymbol{\gamma}^+&=\underset{\gamma}{\operatorname{argmin}}\tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{W_{ik}\left(h_{ik}-\boldsymbol{g}_i^T\boldsymbol{\gamma}_k\right)^2\right\}+\sum_{\mathrm{g}=1}^{q+G}\tfrac{\rho}{2}\left\|\boldsymbol{x}_g^+-\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}+\boldsymbol{u}_g\right\|_2^2\\ \boldsymbol{u}_g^+&=\boldsymbol{u}_g+\boldsymbol{x}_g^+-\boldsymbol{\beta}_g\circ\boldsymbol{\gamma}^+\end{aligned}\tag{16}$$

The updating equation for $\boldsymbol{x}$ naturally decomposes across groups g and updates in parallel.

5. We can derive close form solution for $\boldsymbol{x}_g$ and $\boldsymbol{\gamma}$ (see Appendix B for detailed derivation) as below:

   a) The update for $\boldsymbol{x}_g$ is precisely a soft thresholding operation. $\boldsymbol{x}_g^+ = \left(1 - \frac{1}{\rho\|\boldsymbol{a}_g\|_2}\right)_+ \boldsymbol{a}_g$, where $\boldsymbol{a}_g = \boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g$.

   b) The updating equation for $\boldsymbol{\gamma}$ can decompose across k, which permits the parallel update of $\boldsymbol{\gamma}_k$:

$$\boldsymbol{\gamma}_k^+ = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left\{ W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k \right)^2 \right\} + \sum_{\mathrm{g}=1}^{q+G} \frac{\rho}{2} \left\| \boldsymbol{x}_{gk}^+ - \boldsymbol{\beta}_{gk} \circ \boldsymbol{\gamma} + \boldsymbol{u}_{gk} \right\|_2^2$$

   After some calculation, the explicit form for $\boldsymbol{\gamma}_k^+$ can be written as:

$$\boldsymbol{\gamma}_k^+ = (G^T D(W_k) G + \sum_{g=1}^{q+G} \rho B_{gk}^2)^{-1} (G^T D(W_k) h_k + \rho \sum_{g=1}^{q+G} D(\boldsymbol{\beta}_{gk})(x_{gk} + u_{gk}))$$

   where $D(\boldsymbol{x})$, $\boldsymbol{x} = \{x_1, ..., x_n\}$, is an $n \times n$ square matrix with diagonal elements equal $\boldsymbol{x}$ and off-diagonal are 0s.

The above algorithm is summarized into pseudo code as shown in Algorithm 2.

3.2.4 Stopping rules

Two stopping rules need to be set for the proposed optimization procedure. For the ADMM algorithm in the inner loop for estimating $\boldsymbol{\gamma}$, the l_2 norm of primal residual at the t^{th} iteration is calculated as $r^{(t)} = \sqrt{\sum_{g=1}^{q+G} \sum_{k=1}^{K} (\boldsymbol{x}_{gk}^{(t)} - \boldsymbol{\beta}_{gk}^{(t)} \circ \boldsymbol{\gamma}^{(t)})^2}$, and the l_2 norm of dual residual at the t^{th} iteration is calculated as $v^{(t)} = \sqrt{\sum_{g=1}^{q+G} \sum_{k=1}^{K} \boldsymbol{\beta}_{gk}^{(t)} \circ (\boldsymbol{\gamma}_k^{(t)} - \boldsymbol{\gamma}_k^{(t-1)})^2}$. We set the stopping rule for the ADMM algorithm to be $r < 10^{-7}$ and $s < 10^{-7}$ or the time of iteration is greater than 100. For the convergence of EM algorithm in the outer loop, the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}\}$ is updated iterative until $||\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}||_2 < 10^{-5}$ in the m^{th} iteration or the time of iterations m is greater than 200.

Algorithm 2 Pseudo code for integrative ogClust model estimation.

input: $\mathbb{Y}$, $\mathbb{X}$, $\mathbb{G}^{(1)}$,..., $\mathbb{G}^{(S)}$, $\mathbb{P}$ and K

Initialize $\boldsymbol{\theta}^{(0)}$ and set $m = 0$;

repeat

E-step: compute the posterior weights $w_{ik}^{(m)}$ by Equation (**??**);

M-step:

1. Update $\{\boldsymbol{\beta}_0^{(m)}, \boldsymbol{\beta}^{(m)}, \sigma^{(m)}\}$ to $\{\boldsymbol{\beta}_0^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \sigma^{(m+1)}\}$ by Equations (**??**)-(**??**);

2. Update $\boldsymbol{\gamma}^{(m)}$ to $\boldsymbol{\gamma}^{(m+1)}$ by ADMM algorithm:

Set t=0, $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}^{(m)}$;

repeat

Update $\boldsymbol{\gamma}^{(t)}$ to $\boldsymbol{\gamma}^{(t+1)}$ by Equation (16);

t=t+1;

until $r^{(t)} < 10^{-7} \& v^{(t)} < 10^{-7}$ or $t \geq 100$;

Set $\boldsymbol{\gamma}^{(m+1)} = \gamma^{(t)}$, $\boldsymbol{\theta}^{(m+1)} = \{\boldsymbol{\beta}_0^{(m+1)}, \boldsymbol{\beta}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}, \sigma^{(m+1)}\}$, $m = m + 1$;

until $||\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}|| < 10^{-5}$ or $m \geq 200$;

output: Parameter estimates $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(m)}$

3.2.5 Choice of tuning parameters

Choice of number of clusters K and penalty parameter λ. We use Bayesian information criterion (BIC) to determine the tuning parameter λ and the number of subgroups K in simulation. BIC is defined as $\ln(n)\mathrm{df}(\hat{\boldsymbol{\theta}}) - 2\ln(\mathrm{L}(\hat{\boldsymbol{\theta}}))$, where $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}\}$ and $\mathrm{df}(\hat{\boldsymbol{\theta}})$ is the number of non-zero estimated parameters. In the real application, because of potential data noises and violation of Gaussian assumption, BIC may fail to choose the correct K. To address this issue, we additionally plot the trend of root mean square error (RMSE) and R^2 as a function of K and identify the elbow point as the optimal number of clusters K as shown in Figure 11.

Choice of tuning parameter α. Parameter α balances individual feature penalty and group penalty. We set $\alpha = 0.5$ in this chapter and perform sensitivity analysis to examine the impact of α selection. The choice of α could be problem specific and up to user's preference, similar that described in Simon et al. (2013).

Choice of augmented Lagrangian parameter ρ. The augmented Lagrangian parameter ρ controls the convergence of ADMM algorithm. We follow the adaptive method proposed by He et al. (2000), and Wang and Liao (2001) to accelerates ADMM convergence: $\rho^{(t)} = 2\rho^{(t-1)}$ if $r^{(t-1)} > 10v^{(t-1)}$, $\rho^{(t)} = \rho^{(t-1)}/2$ if $v^{(t-1)} > 10r^{(t-1)}$, and $\rho^{(t)} = \rho^{(t-1)}$ otherwise.

3.3 Simulation

Simulation scheme

1. We simulate two omics datasets denoted by $s \in \{1, 2\}$, with 1000 features ($1 \leq q \leq 1000$) and 300 subjects ($1 \leq i \leq 300\}$) in each omics dataset. We assume the two omics datasets are paired, which means they are from the same 300 subjects. Let g_{qs} denotes the quantity levels of feature q in omics data s. In each omics dataset, simulate 30 feature modules ($1 \leq m \leq 30$) with 10 features in each module. So there are in total 2 omics datasets, 2000 features, 60 feature modules and 600 features within the modules.

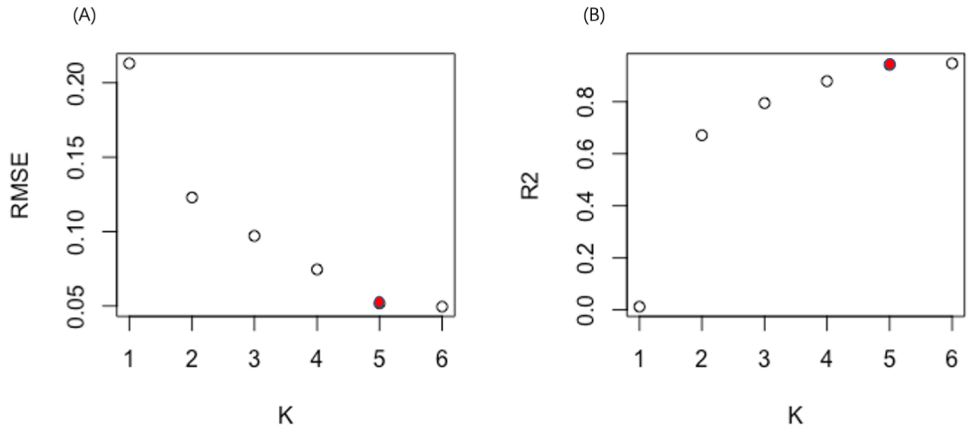


Figure 11: Plot of (A) RMSE and (B) R^2 against the number of clusters

2. The first 15 feature modules $(M_{s1}, ..., M_{s15})$ in each omics data define three outcome associated subgroups denoted by $k \in \{1, 2, 3\}$, where expression levels varies across groups. The remaining 15 feature modules $(M_{16s}, ..., M_{30s})$ define another three subgroups $k^{'} \in \{1, 2, 3\}$ independent of k and outcome $\boldsymbol{Y}$. The rest of the features are noise and their quantity levels are randomly drawn from standard normal distribution.
3. Simulate the expression levels of feature of the first 15 outcome associated feature modules$(M_{s1}, ..., M_{s15})$.
   - Simulate template expression $\mu_{skm} \sim N(0, 4)$ for omics s, subgroup k and module m, with constraint that $max_{k_1,k_2}|\mu_{sk_1m} - \mu_{sk_2m}| > 1$, where k_1 and k_2 denotes any two different subgroups.
   - Add biological variation to the template gene expression and simulate $X_{skmi} \sim N(\mu_{skm}, 1)$ for omics s, subgroup k, module m and subject i.
   - Generate covariance matrix Σ_{skm} from inverse Wishart distribution. First simulate $\Sigma^{'}_{skm} = W^{-1}(\phi, 100)$, where $\phi = 0.5I_{10\times10} + 0.5J_{10\times10}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all elements equal to 1. Then Σ_{skm} is calculated by normalizing $\Sigma^{'}_{mks}$ such that the diagonal elements are all 1s.
   - Simulate the level of features for omics s, subgroup k and module m by $(X_{skmi1}, ..., X_{skmi10}) \sim MVN(X_{skmi}, \Sigma_{skm})$.
4. Similarly simulate the expression levels for the remaining feature modules $M_{s16}, ..., M_{s30}$. We assume that the m^{th} modules in the first and second omics dataset are in the same group, i.e. $\{M_{1m}, M_{2m}\}$, where $m = \{1, 2, ..., 30\}$.
5. The features within modules have probability θ=0.3, 0.6, or 0.9 to be replaced by random noise simulated from standard normal distribution. We change the level of sparsity within groups by tuning the value of θ.
6. Generate latent subgroup index $\boldsymbol{Z}$, covariates $\boldsymbol{X}$ and outcome $\boldsymbol{Y}$.
   - Let $\boldsymbol{\gamma}_{sm} = (\boldsymbol{\gamma}_{s1m}, \boldsymbol{\gamma}_{s2m}, \boldsymbol{\gamma}_{s3m})^T$, where $\boldsymbol{\gamma}_{skm}$ denotes the effect of module m in omics dataset s on subgroup k in subtyping. For identifiability, we set $\boldsymbol{\gamma}_{s3m} = \boldsymbol{0}$. The active set for outcome-guided subtypes is restricted to the first 15 feature modules

(i.e., $M_{s1}, ..., M_{s15}$) in each omics data, in other words, $\boldsymbol{\gamma}_{s17} = \boldsymbol{\gamma}_{s17} =, ..., = \boldsymbol{\gamma}_{s30} = \mathbf{0}$. We set the level of $\boldsymbol{\gamma}_{skm}$ to be 0.5.

- Given $\boldsymbol{\mu}_{skm}$ and $\boldsymbol{\gamma}_{skm}$, we obtain $\pi_{ik} = \frac{\exp(\sum_{S=1}^{2}\sum_{m=1}^{15}\boldsymbol{\mu}_{skmi}^{T}\boldsymbol{\gamma}_{skm})}{\sum_{l=1}^{3}\sum_{S=1}^{2}\sum_{m=1}^{15}\exp(\boldsymbol{\mu}_{slmi}^{T}\boldsymbol{\gamma}_{slm})}$, where $k \in \{1, 2, 3\}$ and $i \in \{1, 2, ..., 300\}$. π_{ik} represent the probability of subject i belonging to the kth subgroup. Therefore, subgroup indicator Z_i is randomly drawn from a multinomial distribution with probability $\boldsymbol{p}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})$
- Independent covariates X_1 and X_2 are sampled from normal distributions $N(1, 1)$ and $N(2, 1)$ respectively. Recall that $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ is the set of regression coefficients of the two covariates and $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \beta_{03})^T$ represents the baseline mean of the three subgroups. We set $\boldsymbol{\beta} = (1, 1)^T$ and $\boldsymbol{\beta}_0 = (1, 4, 7)^T$.
- Given the latent subgroup index Z_i, the outcome for subject i can be simulated by $(Y_i|Z_i = k) = \beta_{0k} + \boldsymbol{X_i^T}\boldsymbol{\beta} + e_i$, where $e_i \sim N(0, \sigma^2)$ and we set $\sigma^2 = 1$.

We compared the performance of the following three methods (1)IS-Kmeans, integrative sparse K-means clustering proposed by Huo and Tseng (2017). (2) ogClust from Chapter 2 without incorporating group information from multi-omics or pathway information. (3Our proposed iogClust (integrative outcome-guided clustering) incorporating group information. We simulated 100 training dataset and 100 testing dataset. The three methods above are applied to the training datasets and then validated to testing datasets for evaluation. The performance of these three methods is evaluated in terms of subgrouping accuracy (adjusted Rand index (ARI)) in training and testing datasets), feature selection (the number of features selected, and the number of true positive features (TPs)), and outcome prediction (RMSE and R^2 in the testing datasets). The performance was compared under different inner group sparsity level (θ =0.3,0.6,0.9), which corresponds to approximately 30%, 60% and 90% of noise features within modules (groups).

The simulation results are summarized in Table 3. IS-Kmeans fails to select the correct outcome associated features/modules because it lacks outcome guidance, resulting in inferior performance in subgrouping and outcome prediction. The ogClust method without group information has much higher ARI and R^2 compared with IS-Kmeans due to outcome guidance. The iogClust method with overlapping group information further improves ogClust by selecting more true features, producing the highest clustering ARI, and achieving

the best outcome association in terms of the lowest RMSE and the highest R^2. By taking into account group information, iogClust method select more true positive features (TPs) than ogClust method, and much less false positive features than IS-Kmeans. Additionally, with outcome-guidance, iogClust encourage the selection of outcome associated overlapping groups, which results in improved feature selection, subgrouping and outcome prediction performance, especially when the inner group sparsity level θ is high.

Table 3: Performance comparison of integrative sparse K-means (ISKmeans), outcome-guided clustering with LASSO(ogClust), integrative outcome-guided clustering integrating groups information(iogClust) under different sparsity level θ.

Methods	ARI		Selected Genes		Outcome	
	Train	Test	Total	TPs	RMSE	R^2
$\theta = 0.3$						
ISKmeans	0.01	0.01	538.65	91.31	2.65	0.13
ogClust	0.94	0.86	17.71	12.84	0.93	0.89
iogClust	0.96	0.89	90.56	61.11	0.94	0.89
$\theta = 0.6$						
ISKmeans	0.00	0.00	541.34	54.61	2.69	0.11
ogClust	0.94	0.80	15.57	6.47	0.91	0.90
iogClust	0.96	0.83	70.32	27.10	0.90	0.90
$\theta = 0.9$						
ISKmeans	0.00	0.00	519.22	12.26	2.69	0.11
ogClust	0.95	0.72	9.84	0.88	0.94	0.89
igoClust	0.94	0.73	56.31	5.46	0.92	0.90

3.4 Real Data Application

3.4.1 Application to LGRC dataset

We applied our method to LGRC lung disease dataset with gene expression, miRNA expression and clinical information. The dataset contains 319 samples, covering 15,966 genes and 438 miRNAs. The dataset was preprocessed such that the lowly expressed and variable genes and miRNA were removed. We also removed the subjects with missing survival outcome, so there are 2000 highly variable genes, 246 miRNAs, and 234 subjects after preprocessing. The level of features were standardized to be mean 0 and standard deviation 1. We downloaded canonical pathway information from MSigdb (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#C2) as the prior group information to guide feature selection. There are 292, 186, 196, 1604 and 615 canonical pathways from BIOCARTA, KEGG, PID, REACTOME and WikiPathways respectively. The pathways with fewer than 5 genes and larger than 200 genes overlapping with selected 2000 genes were excluded, 551 pathways were left after filtering. Alternatively, we downloaded miRNA regulatory target genesets from MSigdb (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp#C3) as prior knowledge. A miRNA and its targeting genes were grouped together, we ended up with 107 such groups. We applied the proposed integrative ogClust method to LGRC dataset in the following three ways: (1) ogClust is the method proposed in Chapter 2. It deals with gene expression only without any prior knowledge. (2) iogClust (pathways) applied the integrative ogClust proposed in this chapter, it uses gene expression and considers canonical pathways as prior knowledge, . (3) iogClust (multi-omics) combines gene expression and miRNA expression, with miRNA and its regulating genes as prior group information.

As shown in Table 4, compared with the ogClust which selects 269 features, the iogClust applications select 434 and 471 features. For each method, we count the possible pairs of selected genes (e.g., C_2^{269} possible pairs of 269 selected genes by ogClust), and performed permutation to test whether gene pairs in the same pathway are more easily to be selected. For iogClust (multi-omics), we count the the miRNA and gene pairs that are in the same prior group, and similarly performed permutation to test whether the algorithm encourages

a miRNA and its targeting genes to be selected together. We also calculated the percentage of all pairs of selected features that are in the same group. Note that the percentage for iogClust (multi-omics) is 3.11% which is smaller than 3.45% for ogClust, however, the p value is much more significant. This is because pairs are counted differently in the two applications and the percentages have different null distributions. The selected features for iogClust (pathways) is significantly associated with the prior groups under permutation test with p value = 2.0×10^{-5} while for ogClust the test is not significant with p value = 0.68. This indicates that the prior groups are used to guide feature selection in iogClust. We also performed Kruskal-Wallis test to measure the association between outcome and subgroups, and Chi-square test to evaluate the association between subgroups and clinical diagnosis, all the three methods are significant with p values much smaller than 0.05. Additionally, RMSE and R^2 are calculate by 5 fold cross validation, iogClust (pathways) has slightly higher R^2 of 0.402 and lower RMSE of 0.165 among the three methods, while ogClust and iogClust (meta) are almost equally well with $R^2 = 0.166$ and 0.165 and RMSE=0.399 and 0.398 respectively. The expressions of selected features across subgroups for two iogClust applications are shown in Figure 12 (A), Figure 12 (B) shows that the distributions of the guiding outcome FEV1%prd are quite different across subgroups. We also compared the canonical pathway enrichment analysis results of applications (1)-(3). For a fair comparison, we tuned the parameter λ such that the number of selected features is around 200. Jitter plot of -log10(p) values of pathway enrichment analysis is shown in Figure 12 (C), ogClust has less significant pathways compared with iogClust applications if we set the p value cutoff to be 0.01, indicating that incorporating canonical pathways or multi-omics group information can improve the biological interpretation of selected features. The pathways and annotations with p value less than 0.01 in any of the three methods are listed in Table 5

3.5 Discussion

In this chapter, we propose an integrative outcome-guided clustering (integrative ogClust) framework for disease subtyping using multi-omics data or using prior biological

Table 4: Comparison of ogClust, iogClust (pathways), and iogClust (meta) when applied to the LGRC dataset. RMSE and R^2 measure outcome prediction performance. Kruskal-Wallis test measures outcome association with clusters. Chi-square exact test measures subgroup label association with clinical diagnosis. Permutation test measures the association between selected features and features in pathways

Methods	K	Number of features selected	Percentage (Permuation test)	Chisq test (diagnosis)	Kruskal-Wallis test	RMSE	R^2
ogClust	3	269	3.45%(0.68)	2.01×10^{-39}	2.10×10^{-32}	0.166	0.399
iogClust(pathways)	5	434	7.37%(2.0×10^{-5})	5.81×10^{-28}	2.83×10^{-61}	0.165	0.402
iogClust(meta)	5	471	3.11%(1.0×10^{-6})	1.11×10^{-31}	1.20×10^{-49}	0.165	0.398

knowledge, such as pathways. This is an extension of our previous work in Chapter 2, where only single omics data could be used. Integrative ogClust model also consists of disease subtyping model and outcome association model, which are linked through a latent cluster label Z. The disease subtyping model specifies an sparse overlapping group lasso penalty to incorporate multi-omics datasets as well as prior feature sets knowledge. The model is estimated by an EM-ADMM algorithm. From extensive simulations and a real data application on LGRC dataset, we demonstrate the ability of integrative ogClust for identifying outcome associated clusters (disease subtypes) that are otherwise easily masked by other outcome irrelevant feature structure. Interpretability of feature selection and reproducibility of subtyping can be improved by incorporating prior knowledge of features and combining multiple omics datasets.

In the current integrative ogClust model, univariate outcome variable Y is used to characterize the subtype Z. However, in many scenarios, it is desirable to use multiple variables to characterize the clinical outcome or characteristics of disease. Therefore, a multivariate outcome guided disease subgrouping model is appealing and could be a future direction.

Integrative ogClust parameter estimation procedure is implemented via an EM-ADMM

algorithm, providing a reasonable computing efficiency for high-dimensional data. In the lung disease example, the model fitting can be finished in 2.17 hours using 1 core for $n = 234$ patients, $q = 2000$ genes and $p = 2$ covariates and 551 pathways. To select tuning parameters K and λ, multiple runs are necessary. An R package will be freely available on https://github.com/liupeng2117/iogClust, along with all data and code to reproduce results in this chapter.

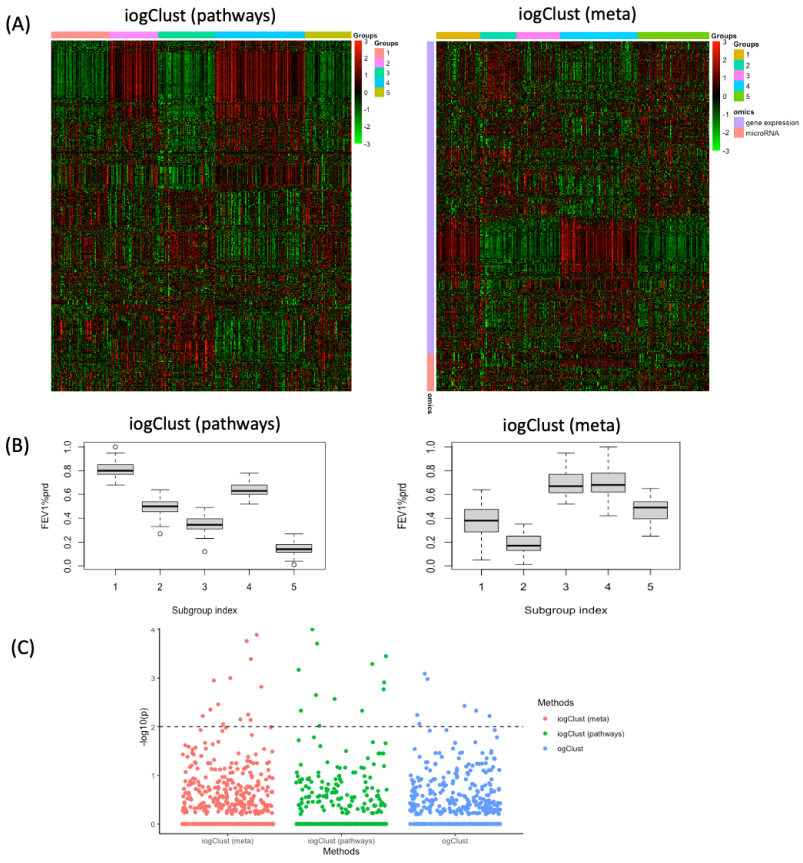


Figure 12: LGRC dataset application results. (A) Heatmaps of selected feature expression of iogClust (pathways) and iogClust (meta). (B) Boxplot of outcome Y for each subgroup in iogClust (pathways) and iogClust (meta). (C) Jitter plot of -log10(p) for canonical pathways enrichment analysis using IPA.)

Table 5: The top enriched canonical pathways of the three applications to LGRC dataset.

	Pathways	-log10(p)	Functional annotations	Method
1	Heparan Sulfate Biosynthesis	4.00	Metabolic Pathways	iogClust (pathways)
2	Tryptophan Degradation to 2-amino-3-carboxymuconate Semialdehyde	3.89	Amino Acids Degradation	iogClust (meta)
3	Axonal Guidance Signaling	3.76	Organismal Growth and Development	iogClust (meta)
4	Triacylglycerol Biosynthesis	3.71	Metabolic Pathways	iogClust (pathways)
5	Heparan Sulfate Biosynthesis (Late Stages)	3.45	Metabolic Pathways	iogClust (pathways)
6	Wnt/-catenin Signaling	3.39	cellular development	iogClust (meta)
7	Chondroitin Sulfate Biosynthesis	3.29	Metabolic Pathways	iogClust (pathways)
8	Dermatan Sulfate Biosynthesis	3.17	Organismal injury and abnormalities	iogClust (pathways)
9	Catecholamine Biosynthesis	3.09	Organismal injury and abnormalities	ogClust
10	NAD biosynthesis II (from tryptophan)	3.00	Metabolic Pathways	iogClust (meta)
11	Granulocyte Adhesion and Diapedesis	2.98	Cellular Immune Response	ogClust
12	Human Embryonic Stem Cell Pluripotency	2.95	Organismal Growth and Development	iogClust (meta)
13	Glutamate Receptor Signaling	2.91	Nervous System Signaling	iogClust (pathways)
14	Osteoarthritis Pathway	2.82	Organismal injury and abnormalities	iogClust (meta)
15	Chondroitin Sulfate Biosynthesis (Late Stages)	2.77	Metabolic Pathways	iogClust (pathways)
16	LPS/IL-1 Mediated Inhibition of RXR Function	2.65	Cellular Immune Response	iogClust (pathways)
17	GABA Receptor Signaling	2.57	Nervous System Signaling	iogClust (pathways)
18	Serine Biosynthesis	2.46	Metabolic Pathways	iogClust (meta)
19	Regulation of Cellular Mechanics by Calpain Protease	2.43	Organismal injury and abnormalities	ogClust
20	Agranulocyte Adhesion and Diapedesis	2.35	Cellular Immune Response	iogClust (meta)
21	Pathogenesis of Multiple Sclerosis	2.33	Inflammatory disease	ogClust
22	Glycine Cleavage Complex	2.33	Amino Acids Degradation	iogClust (pathways)
23	Tryptophan Degradation to 2-amino-3-carboxymuconate Semialdehyde	2.33	Amino Acids Degradation	iogClust (pathways)
24	Hepatic Fibrosis Signaling Pathway	2.25	chronic liver disease	iogClust (meta)
25	Osteoarthritis Pathway	2.24	Organismal injury and abnormalities	ogClust
26	Acute Phase Response Signaling	2.22	Cellular Immune Response	ogClust
27	Granulocyte Adhesion and Diapedesis	2.22	Cellular Immune Response	iogClust (meta)
28	Superpathway of Serine and Glycine Biosynthesis I	2.15	Metabolic Pathways	iogClust (meta)
29	Role of NANOG in Mammalian Embryonic Stem Cell Pluripotency	2.14	Organismal Growth and Development	iogClust (meta)
30	Agranulocyte Adhesion and Diapedesis	2.06	Cellular Immune Response	ogClust
31	Apelin Cardiac Fibroblast Signaling Pathway	2.05	Cardiovascular Signaling	iogClust (meta)
32	Dermatan Sulfate Biosynthesis (Late Stages)	2.02	Organismal injury and abnormalities	iogClust (pathways)

4.0 MethylSeqDesign: a Framework for Methyl-Seq Genome-Wide Power Calculation and Study Design Issues

4.1 Introduction

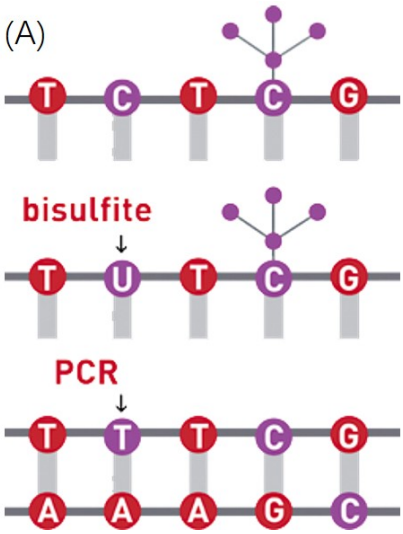
[1]DNA methylation is a chemical modification of DNA nucleotides when a methyl-group (CH_3) is attached at the 5^{th} position of cytosine (5mC). It is one of the best characterized and the most studied epigenetic markers, which has shown to control gene expression in both normal cell development and abnormal biological process such as cancer. Particularly in gene promoter regions, hyper-methylation is shown closely related to silencing gene expressions. In mammals, such as human, DNA methylation happens almost exclusively at cytosine site that follows with guanine known as CpG site. There are tens of thousands of regions with a high frequency of CpG sites in the whole genome that are classified as CpG islands, which typically exist at or near the transcription starting sites of genes. DNA methylation process has been found to link to many important biological processes, such as genomic imprinting, X-chromosome inactivation, repression of repetitive elements, aging and carcinogenesis (Li et al., 1993; Paulsen and Ferguson-Smith, 2001; Robertson, 2005). In cancer studies, aberrant DNA methylation changes are considered as one of the leading factors in developing tumors (Esteller, 2005; Baylin, 2005; Delpu et al., 2013; Licht, 2015).

Over the past couple decades, sodium bisulfite treatment has become widely used tool to study DNA methylation at the level of single nucleotide resolution. When DNA is treated with sodium bisulfite, the unmethylated cytosines are converted to uracil and amplified by polymerase chain reaction (PCR) as thymine while methylated cytosines remain protected from this conversion (see Figure 13(A)). The outcome of this treatment leads to identifying methylated and unmethylated Cytosines when the sequencing reads are mapped to reference genome using special mapping pipelines such as Bismark, which consider Thymine/Cytosine mismatch. Two major technologies have been developed to quantify the DNA methylation after bisulfite conversion. One is methylation microarray, which targets on pre-selected CpG

[1]This chapter has been published in Biostatistics (Liu et al., 2021).

sites in certain regions, mostly within CpG islands. The total number of targeted CpG sites is relatively small, for example, Illumina HumanMethylation27 and HumanMethylation450 Bead chips cover only about 27K and 480K CpG sites) compared to over 28 million CpG sites in human genome (Schumacher et al., 2006). Another recently developed technology is coupling bisulfite conversion with next generation sequencing to quantitatively query the methylation status across the whole genome. Whole genome bisulfite sequencing (WGBS) can provide accurate and quantitative estimates of the proportion of methylated cells in a population at each of the tens of millions of CpG sites across genome. However, accurate estimates of methylation level requires large number of reads to cover CpG sites of interest. Because of unevenly distributed CpG sites across the genome, a large proportion of sequencing reads do not contain any CpG sites, which results in high cost of WGBS. To overcome this disadvantage, reduced representation bisulfite sequencing (RRBS)(Meissner et al., 2005) was introduced to target on CpG rich regions, relying on restriction enzyme that can ensure the capture of at least one CpG site per sequencing read. Using RRBS, methylation levels of a portion of genome regions can be accurately obtained at much lower cost compared to WGBS. Here we use "Methyl-Seq" to refer to bisulfite sequencing technology including WGBS and RRBS. Due to the popularity of Methyl-Seq and the high sequencing cost with limited budget, sample size and power calculation methods become critical for design of such studies.

(A)



(B)

Single hypothesis testing	Methyl-Seq genome-wide screening
Type I error rate α	False discovery rate (FDR)
Power 1-β	Expected discovery rate (EDR)
Univariate effect size θ	Distribution of effect size of DML/DMRs

Figure 13: (A) An illustration of sodium bisulfite modification (B) Comparison of three elements between single hypothesis testing and Methyl-Seq genome-wide screening.

Traditional power calculation methods seek the statistical power ($1-\beta$; β here is type-II error) to detect the difference between groups by pre-specifying effect size (θ), type I error rate(α), and sample size(N). Alternatively, one can calculate the required sample size with pre-specified statistical power, θ and α. The effect size θ is the measure of group difference that is generally obtained from a pilot study or researchers' belief. Usually, the type I error rate is set to be 5% and the desired statistical power is 70-80% for a study design. This classical framework is based on performing a single hypothesis testing. For high-throughput genome-wide experimental data, however, many hypotheses are tested simultaneously to compare the methylation difference at the thousands of regions or millions of CpG sites. Therefore, genome-wide power calculation should be based on appropriately controlled type-I error rates. One widely used in Genomic study is false discovery rate (FDR; (Benjamini and Hochberg, 1995)) and in this chapter, we use FDR to control genome-wide type-I error rate. In addition, Gadbury et al. (2004b) introduced a useful concept, called expected discovery rate (EDR), to replace test power $1-\beta$ from single hypothesis to address genome-wide detection power. Since genome-wide screening considers the whole set of differentially methylated loci/regions (DML/DMRs), specifying a single effect size θ for power calculation

is no longer valid. Alternatively, the distribution of effect sizes of DML/DMRs is needed, which could be estimated from pilot data. Figure 13(B) shows changes of the essential elements in genome-wide screening, compared to traditional single hypothesis testing.

There are three unique characteristics inside Methyl-Seq data that should be considered in power and sample size calculation. First, it generates random binomial data for each sample at each CpG site. A model with discrete distributions is more suitable for Methyl-seq data and both sampling and biological variations should be considered. For this reason, the beta-binomial model (Dolzhenko and Smith, 2014; Feng et al., 2014; Park et al., 2014) has gained popularity over the binomial model. Second, for Methyl-Seq experiments, researchers have choices of different sequencing depth (R) for the design. In other words, one can choose to process one sample per lane, which results in roughly 250 million reads in Illumina HiSeq 2500 platform or three samples per lane each with 83 million reads for the same sequencing cost. Therefore the power calculation problem may need to consider both N and R. Finally, there are about 28 million CpG sites in human genome. Based on the current technology it is impossible to sequence most CpG sites with sufficient coverage even with ultra-deep sequencing depth. As a result, many CpG sites will have zero or almost zero reads in many subjects. We further discussed the coverage of single CpG sites and CpG region in Appendix C.1. Therefore, it is not realistic to study the differences of methylation levels at all CpG sites. To circumvent this difficulty, we restrict to methylation regions by aggregating methylation data across multiple CpG sites within a particular region, such as promoter regions, for power calculation.

Many power calculation methods have been developed for RNA-Seq data, such as RNASeqPower(Hart et al., 2013), Scotty(Busby et al., 2013), PROPER(Wu et al., 2015) and RNASeqDesign (Lin et al., 2019). For microarray methylation data, Tsai and Bell (2015) proposed method for study design and power calculation. To the best of our knowledge, for Methyl-seq data, no existing power calculation method has been developed so far in the literature. Here, we propose a statistical framework "MethylSeqDesign" for sample size and power calculation for studies with Methyl-seq data. The "MethylSeqDesign" R package is publicly available at https://github.com/liupeng2117/MethylSeqDesign.

The chapter is structured as follows. In Section 4.2, the statistical framework of MethylSe-

qDesign is proposed. In Section 4.3, we present comprehensive simulations and real data applications. Section 4.4 is real data application. Section 4.5 provides conclusion and discussion.

4.2 Model Specification

4.2.1 Notations and terminology

Consider D_0=$\{Y = (y_{gj})_{G\times(n_0+n_1)}, M = (m_{gj})_{G\times(n_0+n_1)}, X = (x_{jp})_{(n_0+n_1)\times P}\}$ ($1 \le g \le G$, $1 \le j \le n_0+n_1$) a pilot Methyl-Seq dataset, where y_{gj} and m_{gj} represent the methylated and total read counts for CpG region g of subject j respectively. Let X be a design matrix of dimension $(n_0+n_1)\times P$, which contains case/control group information and other continuous or discrete covariates. n_0 and n_1 are the number of controls and cases in the pilot data. Denote N_0 and N_1 the target number of controls and cases for power calculation. Let $R_j = \sum_{g=1}^{G} m_{gj}$ be the total number of reads observed in subject j (a.k.a. library size). We consider genome-wide power calculations under genome-wide type-I error control using FDR=E(number of claimed false positives/number of claimed positives). Following Gadbury et al. (2004b), we use expected discovery rate, EDR=E(number of claimed true positives/number of total true positives), as the genome-wide power. The methylation level is the proportion of methylated cells among all cells at a particular CpG site or region. Statistical power is impacted by both sample size and sequencing depth. Therefore, the statistical framework of MethylSeqDesign becomes to estimate the genome-wide power $\widehat{\text{EDR}}(N_0, N_1, R|D_0)$ based on the pilot data (D_0 with n_0 controls, n_1 cases and sequencing depth R_0) for designing a future experiment with N_0 controls, N_1 cases, sequencing depth R and under a prespecified FDR level (e.g. FDR=5%).

4.2.2 Three sequential steps for genome-wide Methyl-Seq power calculation

Park and Wu (2016) proposed the "DSS-general" method, a model-based method for detecting differentially methylated loci (DML) or regions (DMRs) based on beta-binomial

model with arcsine link function. The estimation procedure is based on generalized least square approach, which can significantly reduce the computation demands compared to other beta-binomial based methods (Dolzhenko and Smith, 2014; Feng et al., 2014). In addition, Park and Wu (2016) showed superior performance of their method in terms of DMR detection accuracy and type I error rate control. Therefore, in this chapter we will adopt Park and Wu (2016)'s approach for our power calculation tool.

Below we propose three sequential steps in MethylSeqDesign to estimate EDR. In step I, p-values and effect size distribution of all methylated regions from pilot data are obtained using DSS-general. In step II, a beta-uniform mixture (BUM) model is applied to characterize the genome-wide p-value distribution and to estimate the proportion of true DMRs. In step III, a parametric bootstrapping method based on DMR posterior probability is used to simulate and transform the genome-wide p-value distribution towards the targeted sample size and sequencing depth. The detailed description of our method is as follows.

Step I. Differential methylation analysis on pilot data. To account for both sampling and biological variation, denote by Y_{gj} the methylated read count for gene $g(g = 1, 2, ..., G)$ in sample j $(j = 1, 2, ..., (n_0 + n_1))$, let q_{gj} be the underlying methylation level for gene g and sample j, $Y_{gj} \sim \text{bin}(m_{gj}, q_{gj})$ and $q_{gj} \sim \text{beta}(\alpha_{gj}, \beta_{gj})$. Marginally, $Y_{gj} \sim \text{Beta-bin}\,(m_{gj}, \pi_{gj}, \phi_g)$, where π_{gj} and ϕ_g are the mean and dispersion parameter of beta distribution, such that $\pi_{gj} = E(q_{gj}) = \frac{\alpha_{gj}}{\alpha_{gj}+\beta_{gj}}$, $\phi_g = \frac{1}{\alpha_{gj}+\beta_{gj}+1}$ and we assume $\phi_{gj} = \phi_g$ for all $j = 1, 2, ..., (n_0 + n_1)$. Here we account for covariate effect as,

$$\arcsin\left(2\pi_{gj} - 1\right) = \boldsymbol{x}_j \boldsymbol{\beta}_g, \tag{17}$$

where $\boldsymbol{x}_j = (x_{j1}, x_{j2}, ..., x_{jp})$ is j^{th} subject's covariate, and $\boldsymbol{\beta}_g = (\beta_{g1}, \beta_{g2}, ..., \beta_{gp})^T$ is a vector of p covariate coefficients for g^{th} CpG region.

Denote $A_{gj} = \arcsin\left(2Y_{gj}/m_{gj} - 1\right)$. As shown in Park and Wu (2016), the expectation of A_{gj} can be approximated as $E\left(A_{gj}\right) \approx \arcsin\left[2E\left(Y_{gj}\right)/m_{gj} - 1\right] = \arcsin\left(2\pi_{gj} - 1\right) = \boldsymbol{x}_j \boldsymbol{\beta}_g$. Furthermore, the variance of A_{gj} can also be also approximated as $Var\left(A_{gj}\right) \approx \frac{1+(m_{gj}-1)\phi_g}{m_{gj}}$, which is approximately independent of the mean structure. Given dispersion parameter ϕ_g, the regression coefficients $\boldsymbol{\beta}_g$ can be estimated using generalized least square (GLS) method, i.e. $\hat{\boldsymbol{\beta}}_g = \left(X^T V_g^{-1} X\right)^{-1} X^T V_g^{-1} A$, where $V_g = diag\left(\frac{1+(m_{gj}-1)\phi_g}{m_{gj}}\right)$ is the covariance

matrix. The estimator of ϕ_g is given by $\hat{\phi}_g = \frac{(n_0+n_1)\left(\hat{\sigma}_g^2-1\right)}{\Sigma_j(m_{gj}-1)}$, where $\hat{\sigma}_g^2 = \frac{\Sigma_j m_{gj}(A_{gj}-\boldsymbol{x}_j\hat{\boldsymbol{\beta}}_g^{(0)})}{n_0+n_1-p}$, $\hat{\boldsymbol{\beta}}_g^{(0)}$ is the GLS estimator of β_g under $\hat{\phi}_g = 0$. The estimate of covariance structure is $\hat{V}_g = diag\left(\frac{1+(m_{gj}-1)\hat{\phi}_g}{m_{gj}}\right)$. Given $\hat{\phi}_g$ and $\hat{V}_g$, the estimator of variance of $\hat{\boldsymbol{\beta}}_g$ is $\hat{\Sigma}_g \equiv \hat{\text{var}}\left(\hat{\boldsymbol{\beta}}_g\right) = \left(X^T\hat{V}_g^{-1}X\right)^{-1}$.

Hypothesis testing for $H_0 : C^T\beta = 0$ vs. $H_A : C^T\beta_g \neq 0$ is based on Wald statistics

$$Z_g(\boldsymbol{C}) = \frac{\boldsymbol{C}^T\hat{\boldsymbol{\beta}}_g}{\sqrt{\boldsymbol{C}^T\hat{\Sigma}_g\boldsymbol{C}}},$$

where $\boldsymbol{C}$ can be any linear combination of the covariate effects. The statistic approximately follows a standard normal distribution under null hypothesis.

For simplicity, here we consider the study with two groups (case and control), i.e. p=2 with intercept and case/control effect. Let $n_0 = n_1 = n$ be the number of subjects in each group in pilot data, and $N_0 = N_1 = N$ be the target number of subjects in each group. Here we assume equal sample sizes for control and cases in pilot and target cohorts, while the method can be easily generalized for $n_0 \neq n_1$ and $N_0 \neq N_1$ later (see the leukemia application in Section 4.4.2). Then the model becomes $\arcsin(2\pi_{gj} - 1) = \beta_{0g} + x_{j2}\beta_{1g}$. In this case, $\boldsymbol{C} = (0, 1)$. Here the variance of $\hat{\beta}_{1g}$ is

$$\begin{aligned}
\widehat{Var}\left(\hat{\beta}_{1g}\right) &= \frac{\sum_{j=1}^{n_0+n_1}\frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}}}{\sum_{j_1=1}^{n_0}\frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}}\times\sum_{j_2=1}^{n_1}\frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}}} \\
&= \frac{\sum_{j=1}^{2n}\frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}}}{\sum_{j_1=1}^{n}\frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}}\times\sum_{j_2=1}^{n}\frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}}} \\
&= \frac{n\times\left(\frac{1}{n}\sum_{j_1=1}^{n}\frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}}+\frac{1}{n}\sum_{j_2=1}^{n}\frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}}\right)}{n^2\times\frac{1}{n}\sum_{j_1=1}^{n}\frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}}\times\frac{1}{n}\sum_{j_2=1}^{n}\frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}}} \\
&= \frac{1}{n}\frac{\bar{A}_g+\bar{B}_g}{\bar{A}_g\times\bar{B}_g} \\
&= \frac{1}{n}\Psi_g,
\end{aligned} \tag{18}$$

where $\bar{A}_g = \frac{1}{n}\sum_{j_1=1}^{n} \frac{m_{gj_1}}{1+(m_{gj_1}-1)\hat{\phi}}$ and $\bar{B}_g = \frac{1}{n}\sum_{j_2=1}^{n} \frac{m_{gj_2}}{1+(m_{gj_2}-1)\hat{\phi}}$. The Wald test statistic becomes

$$Z_g = \frac{\hat{\beta}_{1g}}{\sqrt{\hat{\text{var}}(\hat{\beta}_{1g})}} = \frac{\hat{\beta}_{1g}}{\sqrt{\frac{1}{n}\Psi_g}}. \tag{19}$$

where $\hat{\beta}_{1g}$ is the GLS estimator of β_{1g}.

Remark 1: A common over-dispersion parameter ($\hat{\phi}$) over all CpG regions is used which is the mean of all tag-wise dispersion parameters ($\hat{\phi}_g$) estimated from the procedure proposed by Park and Wu (2016). This is because when sample size is small, estimation of region-specific dispersion parameter is not precise, and region specific power calculation is very challenging.

Remark 2: One interesting finding is that the quantities $\bar{A}_g$ and $\bar{B}_g$ in Equation (18) are in the mean form that depends on coverage and dispersion parameter given a region g. When m_{gj} is small, it has negative correlation with Ψ_g, while $\bar{A}_g$ and $\bar{B}_g$ are roughly independent on m_{gj} when m_{gj} is large. If we assume the dispersion parameter stay constant, Ψ_g is mostly impacted by sequencing depth. In Appendix C.2.1, we further studied the property of the quantity Ψ_g by simulation.

Step II. Mixture model fitting for p-value distribution. A beta-uniform mixture (BUM) model (Allison et al., 2002) has been proposed to fit the p-value distribution. To be specific, we use a beta distribution $f_1(p|r,s)$ with shape parameter r and s ($0 < r < 1 \leq s$) for p-values of DMRs and a uniform distribution $f_0(p)$ for p-values of non-DMRs. The mixture density of overall p-value distribution is $f(p|r,s,\lambda) = \lambda f_0(p) + (1-\lambda) f_1(p|r,s)$, where λ is the proportion of non-DMRs. The constraints for r and s is used to have a proper shape for the p-value distribution of DMRs. A proper estimation of λ is essential in fitting a BUM model. We apply censored BUM (CBUM) proposed by Markitsis and Lai (2010) to reduce the impact of extremely small p-values, since our main purpose is to estimate the proportion of true DMRs for those with relatively larger p-values. The detailed comparisons of performance between BUM and CBUM methods are included in Appendix C.2.2. The shape parameters r and s can then be estimated using maximum likelihood approach using $\hat{\lambda}$ estimated from the CBUM method.

Step III. Parametric bootstrapping based on DMR posterior probability to estimate EDR.

Theoretically, the p-value distribution for non-DMRs with zero effect size follows a uniform distribution that does not change with the sample size. However, we expect that the p-values for those DMRs will be more significant as sample sizes increase. Equations (18) and (19) reveal a transformation of Z-statistics of DMRs from the pilot data with sample size n to the targeted sample size N. When the effect size $\hat{\beta}_{1g}$ and the common over-dispersion parameter $\hat{\phi}$ stay approximately unchanged, the Wald test statistics change by a factor of $\sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g^{'}}}$ (see Equation (18) and item (2) below; n is the pilot sample size, N is the targeted sample size, Ψ_g is the quantity under pilot sequencing depth R_0, and $\Psi_g^{'}$ is the quantity under the targeted sequencing depth R). Throughout this chapter, we assume sequencing depth of pilot data R_0 is deep enough and the targeted sequencing depth R does not exceed R_0 (i.e. $R \leq R_0$). Since Ψ_g is a function depending on pre-estimated dispersion parameter $\hat{\phi}_g$ and count data $\{m_{gi1}, m_{gj2}; 1 \leq g \leq G, 1 \leq j \leq (n_0 + n_1)\}$, it is readily calculated from pilot data. To estimate for $\Psi_g^{'}$, we can randomly subsample from pilot data to achieve sequencing depth R and derive $\Psi_g^{'}$ by definition based on subsampled counts $m_{gj}^{'}$ and $\hat{\phi}_g$. The influence of sequencing depth on Ψ_g is further discussed in Appendix C.2.1. We found that when the median coverage level is above 160, Ψ_g is roughly constant and the correction term for different sequencing depth is not necessary. Otherwise, the correction term $\sqrt{\frac{\Psi_g}{\Psi_g^{'}}}$ is needed.

Let I_g be the latent variable indicating region g a DMR (I_g=1) or non-DMR (I_g=0), and let p_g be the p-value of region g from the aforementioned Wald test in pilot data. The detailed parametric bootstrapping procedure is described as follows:

1. Calculate the posterior probability of the DMR indicator I_g with posterior probability

$$P(I_g = 1|\hat{\lambda}, \hat{r}, \hat{s}, p_g) = \frac{(1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})}{\hat{\lambda} + (1-\hat{\lambda})\hat{f}_1(p_g|\hat{r}, \hat{s})},$$

   where $\hat{\lambda}$, $\hat{r}$ and $\hat{s}$ are estimated in Step II. In the b-th parametric bootstrapping ($1 \leq b \leq B$), draw $I_g^{(b)}$ from $P(I_g|\hat{\lambda}, \hat{r}, \hat{s}, p_g)$ for $1 \leq g \leq G$.

2. Transform Z-statistics for DMRs using equation

$$Z_g^{(b)} = I_g^{(b)} \times Z_g \times \sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g^{'}}} + (1 - I_g^{(b)}) \times Z_g.$$

where we assume that the effect size $\hat{\beta}_{1g}$ and the common over-dispersion parameter $\hat{\phi}$ of a DMR in Equation (18) and (19) are roughly fixed. Therefore, when $I_g^{(b)} = 1$, region g is a DMR in the b-th parametric bootstrap and the Wald statistic is transformed to $Z_g \times \sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g'}}$. When $I_g^{(b)} = 0$, the Wald statistic remains unchanged.

3. Compute p-value based on the 2-sided test: $p_g^{(b)} = 2 \times (1 - \Phi(|Z_g^{(b)}|))$ for a DMR region ($I_g^{(b)} = 1$), where Φ is a cumulative density function of a standard normal distribution. When $I_g^{(b)} = 0$, $Z_g^{(b)} = Z_g$ and $p_g^{(b)} = p_g$ remain unchanged.
4. Control FDR at level α:
   a. In the b^{th} simulation, calculate $\text{FDR}^{(b)}(u) = \frac{\sum_{g=1}^{G}(1-I_g^{(b)})\cdot\chi_(p_g^{(b)}\leq u)}{\sum_{g=1}^{G}\chi_(p_g^{(b)}\leq u)}$ for a given p-value threshold u, where $\chi(\cdot)$ is an indicator function that takes value one when the statement is true and zero otherwise. Here, by definition, the denominator is the number of detected regions under p-value threshold u, and the numerator is the number of non-DMRs among those detected regions.
   b. Let $u^{(b)} = \underset{u}{\text{argmax}}(\text{FDR}^{(b)}(u)) \leq \alpha)$, where $u^{(b)}$ is the p-value threshold to keep FDR at α level for the b^{th} simulation.
5. Obtain the estimated EDR for the b^{th} simulation with $\widehat{\text{EDR}}^{(b)} = \frac{\sum_{g=1}^{G} I_g^{(b)}\cdot\chi(p_g^{(b)}<u^{(b)})}{\sum_{g=1}^{G} I_g^{(b)}}$. Here, by definition, the denominator is the number of total DMRs and the numerator is the number of detected true DMRs.
6. Repeat step (1) to (5) for B times and the robust estimator of EDR from the B simulations is $\widehat{\text{EDR}}(N|D_0) = \underset{b}{median}(\widehat{\text{EDR}}^{(b)})$. The first and third quantile of bootstrapped EDRs can also be derived to account for the variability of EDR estimation.

4.3 Simulation

4.3.1 Simulation scheme

We simulated data based on parameters estimated directly from the mouse pregnancy dataset (see details in Section 4.4.1 (Katz et al., 2015)). We empirically drew the total number of reads and baseline methylation level of control group from the data. Effect size

(methylation level difference between two groups) was either fixed or randomly generated from $U(0.1, 0.2)$. In total, 10,000 regions were simulated, and we assigned 10% regions as DMRs. The common dispersion parameter was set to $\phi = 0.048$, which was the mean dispersion parameters estimated from the data.

The steps to simulate pilot data with sample size n and sequencing depth R_0, and targeted data with sample size N and sequencing depth R are shown below.

1. Draw total read m_{gj} for region g and sample j randomly from the mouse pregnancy data, and baseline methylation level q_g for each CpG region from the empirical distribution estimated from the mouse pregnancy data.
2. To simulate pilot data with sequencing depth R_0, we directly use the total reads drawn from step 1. When simulating the targeted data with sequencing depth $R \neq R_0$, we downsample the matrix of total reads based on the ratio of $\frac{R}{R_0}$.
3. DM index: Generate random number I_g from $U(0, 1)$ for each region. If $I_g \leq 0.1$ the g-th region is DMR and I_g=1. Otherwise, it is non-DMR and I_g=0. This generates roughly 10% DMRs.
4. Effect size $\triangle$: Draw effect size from $U(0.1, 0.2)$ for each DMR. The effect size for non-DMRs is set to 0.
5. Generate the number of methylated reads: If the g-th region is non-DMR, then $y_{gj} \sim$ Beta-bin(m_{gj}, q_g, ϕ). If the g-th region is DMR, then in control group $y_{gj} \sim$ Beta-bin(m_{gj}, q_g, ϕ) while in case group, the methylated counts $y_{gj} \sim$ Beta-bin(m_{gj}, q'_g, ϕ), where $q'_g = q_g + \triangle_g$ if $q_g \leq 0.5$ and $q'_g = q_g - \triangle_g$ otherwise.
6. Follow above steps to simulate pilot data and the targeted data.

4.3.2 Performance comparison with other hypothesis testing methods

We compared the statistical power of our proposed test statistic (Equation (19)) with other three methods: Beta value$(2Y_{gj}/m_{gj})$ with t-test, M value$(\text{logit}\,(2Y_{gj}/m_{gj}))$ with t-test, and A value$(A_{gj} = \arcsin(2Y_{gj}/m_{gj} - 1))$ with t-test.

To compare the performance, we conducted the analysis by stratifying the baseline methylation proportion in control group into three categories: low $(0 < q_g < 0.2)$, medium

$(0.2 < q_g < 0.8)$, and high $(0.8 < q_g < 1)$. In each baseline group, we simulated 20 times independent analysis, in which pilot data had 10 subjects in each group (i.e., $n_0 = n_1 = 10$), and 10,000 regions (10% are DMRs). As shown in Figure 14, we compared the power based on how many true DMRs could be declared among different numbers of top declared DMRs. As a result, the result clearly shows better performance of using our arcsin transformation and Wald statistics compared to other approaches. Furthermore, we observe that the power of each method is stronger in either low or high baseline group and relatively weaker in medium baseline group, which is reasonable because the effect size is at methylation level scale and for binomial distribution, the same difference is easier to detect when methylation is close to boundary 0 or 1. Overall, the results justifies the need of using arcsine transformation and Wald test stististics for our power calculation framework.

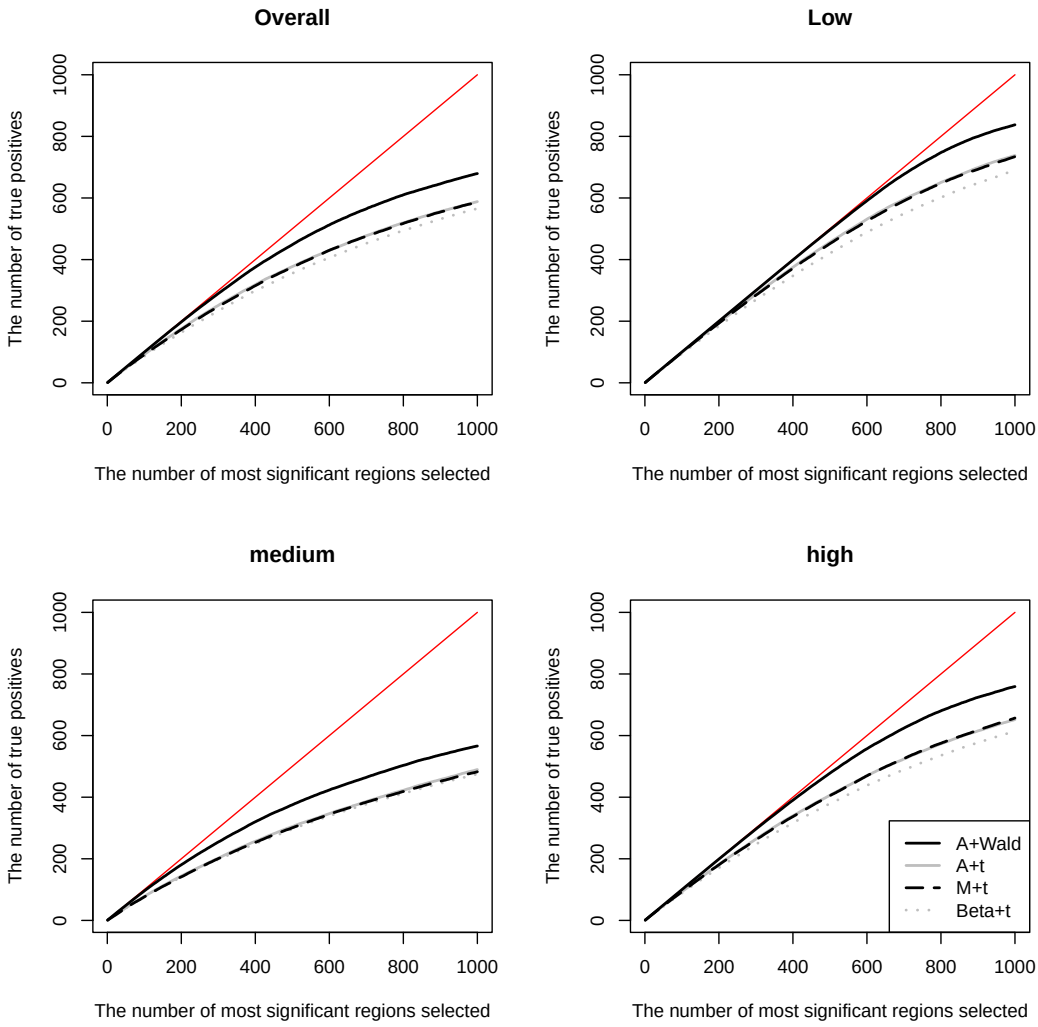


Figure 14: Comparison of hypothesis testing performance of different testing methods stratified by different baseline methylation level (low, medium and high). Different line types represent different methods as shown in the legend (Beta values with t-tests in dotted lines, M values with t-tests in dashed lines, arcsine transformed Z statistics with t-tests in grey lines, and arcsine transformed Z statistics with Wald tests in solid black). X-axis is the number of top declared DMRs and Y-axis is the number of true DMRs among selected. Over all conditions, the Wald test with arcsine transformed Z statistics performs the best.

4.3.3 Performance evaluation

We simulated B=10 pilot datasets ($b = 1, 2, ...B$) with pilot sample size n_0= 2, 4, 6, 8, 9 and 10 when R_0 are 5 million reads. For each pilot dataset with (n_0, R_0), the projected power for targeted sample size N_i=2, 6, 10, 15, 25, 50 ($i = 1, 2, ..., 6$) and $R_j = 0.25$, 0.5, 1, 2, 3, 4 million reads (j =1,2,...,6) from a power calculation method is denoted by $\widehat{\text{EDR}}(N_i, R_j; n_0, R_0)$. Since the underlying truth is known, the true EDR for each (N_i, R_j) can be estimated as $\widehat{EDR}(N_i, R_j) = \frac{\sum_{b=1}^{B} \widehat{EDR}^{(b)}(N_i, R_j)}{B}$ where $\widehat{EDR}^{(b)}(N_i, R_j)$ is the actual EDR in the b-th simulation when sample size N_i and R_j are simulated. We propose the following benchmarks based on root mean squared error (RMSE) to evaluate performance of different power calculation methods:

1. Consider two-dimensional power calculation from (n_0, R_0) to (N_i, R_j) ($i = 1, 2, ..., 6$ and $j = 1, 2, ..., 6$). The RMSE of estimated EDR from power calculation is

$$\text{RMSE} = \sqrt{\frac{\sum_{b=1}^{B}\sum_{i=1}^{6}\sum_{j=1}^{6}\left[\widehat{EDR}^{(b)}(N_i, R_j; n_0, R_0) - \widehat{EDR}(N_i, R_j)\right]^2}{B \times 6 \times 6}}$$

We first performed a stratified analysis based on different level of effect size, as we already know it will impact the EDR. $\triangle$ was set as 0.1, 0.14, and 0.18. In each setting, we generated the same number of regions to compare the performance (Figure 15 for $\triangle$=0.14, Figure C.2.3 and C.2.4 for $\triangle$=0.1 and 0.18). Table 6 shows the RMSEs and computing time of Figure 15, C.2.3 and C.2.4. As shown in Figure 15, similarly in Figure C.2.3 and C.2.4, the estimated true EDR increases as the sequencing depth increases, however, the gain of EDR decreases as the sequencing depth increases. When the ratio of targeted sequencing depth to the pilot sequencing depth > 0.4 (i.e. prop > 0.4), increasing sequencing depth has almost no effect on the true EDR. The observed trend of true EDR is consistent with the trend of predicted EDR as described in Appendix C.2.1. This consistency indicates that our method can estimate EDR prediction well when sequencing depth changes. We also observed that the predicted EDR curves from MethylSeqDesign are close to the true EDR curves, and the performance improves as the sample size of pilot data (n_0) increased. The

result of Table 6 shows affordable computing time (2-5 minutes using a regular laptop) for one run under this simulation setting. Secondly, to mimic real situation, we generated $\triangle$ from $U(0.1, 0.2)$ and compared the predicted versus true curves as shown in Figure C.2.5. We observed results similar to that in the case of fixed $\triangle$.

Table 6: Performance evaluation in simulation study stratified by different effect sizes. Performance evaluation based on RMSE of $\hat{\text{EDR}}(D; D_0)$ in simulation analysis. Results based on different pilot sample size (n_0 = 2, 4, 6, 8, 9, and 10) are shown in different rows. In the first three columns, stratified analysis is performed as $\triangle$ =0.1, 0.14, and 0.18. In the last column, "Overall" refers to generating $\triangle$ from $U(0.1, 0.2)$.

	RMSE (computing time in seconds)			
Pilot n_0	$\triangle = 0.1$	$\triangle = 0.14$	$\triangle = 0.18$	Overall
2	0.27(332)	0.12(176)	0.04(171)	0.10(198)
4	0.16(190)	0.04(175)	0.04(172)	0.03(179)
6	0.08(150)	0.02(164)	0.02(144)	0.01(148)
8	0.05(153)	0.02(149)	0.01(154)	0.02(144)
9	0.03(153)	0.02(151)	0.02(158)	0.01(150)
10	0.02(116)	0.01(121)	0.03(118)	0.01(120)

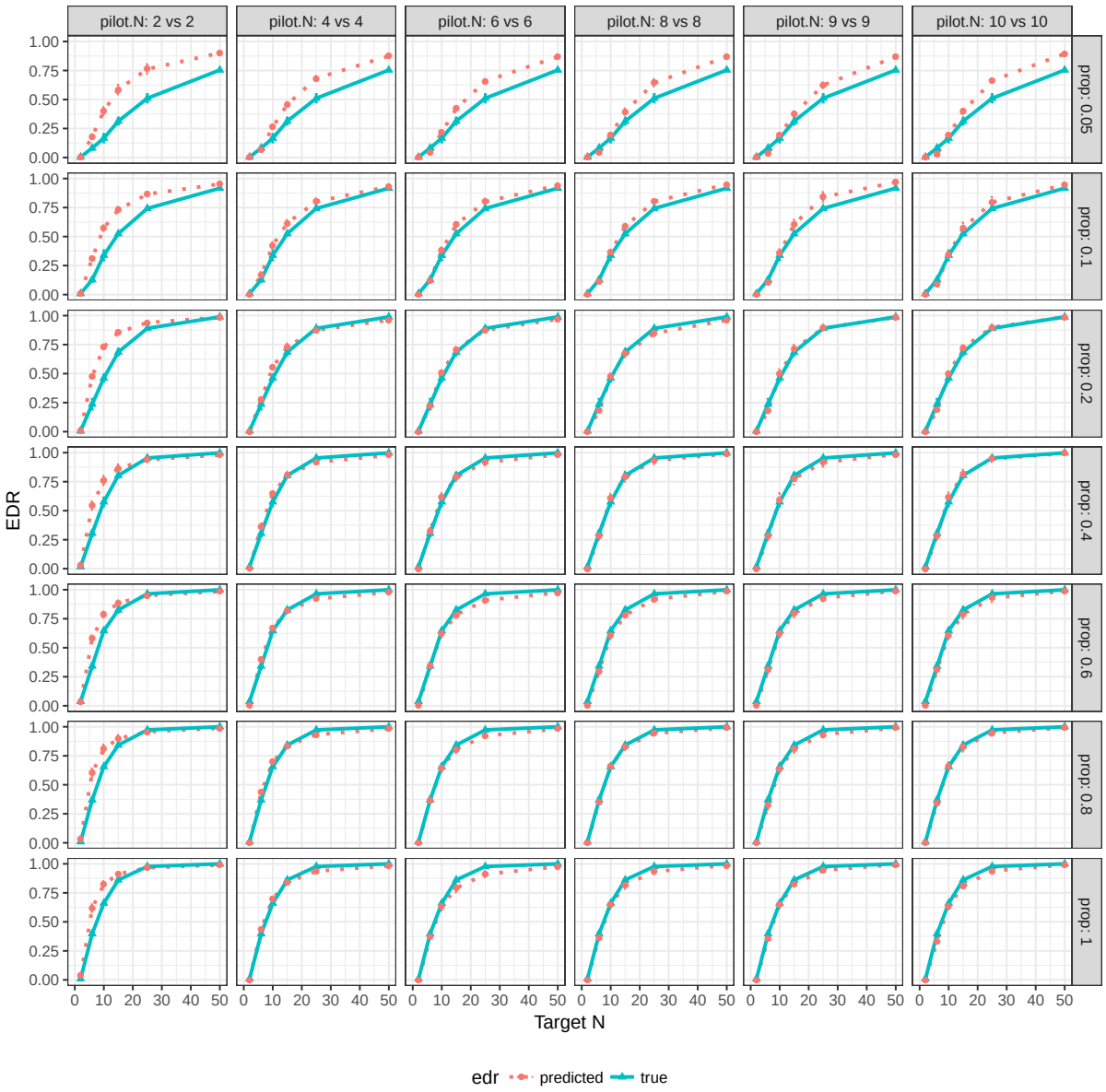


Figure 15: EDR prediction from MethylSeqDesign compared with true EDR under different pilot data sample sizes and sequencing depth. Effect size $\triangle$ is fixed at 0.14. The pilot data sample size per group varied from 2 to 10, and $R_0 = 5M$ is fixed. The predicted EDRs by the pilot data are shown by the dotted red curves. The targeted data sample size per group varied from 2 to 50, and the ratio of targed sample sequencing depth to that of pilot is $prop = \frac{R_j}{R_0}$, which varied from 0.05 to 1. The estimated true EDRs by targeted data are in blue solid curves.

4.3.4 Cost and benefit analysis and study design

In this subsection, we illustrated two scenarios where our method can guide Methyl-Seq study design. In scenario 1, a desired level of EDR was given, and we would like to find the optimal combination of N and R that achieves the desired EDR with the minimal budget. Whereas in scenario 2, a fixed budget limit was given, and we would like to find the optimal combination of N and R which spends within the budget and maximizes the EDR. We obtained the sequencing cost information from Sequencing and Microarray Facility core at MD Anderson for this example. Price per lane (P) is \$1500 dollars when the total number of reads per lane (D) is set at 250M with alignment rate (A) of 50%. Library preparation cost per sample (X) including bisulfite conversion treatment is \$300. The total cost can be written as:

$$\begin{aligned} \text{Total cost} &= \frac{R \cdot 2 \cdot N}{D \cdot A} \cdot P + X \cdot 2 \cdot N \\ &= \frac{R \cdot 2 \cdot N}{250 \cdot 0.5} \cdot 1500 + 300 \cdot 2 \cdot N \end{aligned} \tag{20}$$

Based on Equation (20), we can calculate the cost for any combination of N and R. Given this extra information, the optimal combinations of N and R in both scenarios can be derived. In scenario 1, the optimal N and R combination corresponds to the design with the lowest cost among those with the desired level of EDR; in scenario 2, the optimal design is the one with the highest EDR with costs at most the given budget.

We simulated pilot data ($n_0 = 4$ and $R_0 = 1/4$ lane) based on the settings described in Section 4.3.1. The targeted sample size $N = 4$ to 50 by a gap of 2, and targeted sequencing depth $R = 1/10, 1/8, 1/6, 1/4$ of one lane. The EDR for each N and R combination is estimated from pilot data and the corresponding cost for each design is calculated. . The optimal design can be identified according to different constraints: (1) in scenario 1, we want to achieve at least 80% EDR; (2) a limited budget \$20,000 dollars is given in scenario 2. For any given design (i.e. combination of N and R), if there is no other design achieving higher EDR with lower cost than the current one, it is called an admissible design (colored in black), otherwise it is an inadmissible design (colored in grey). Figure 16 shows the resulting N-R and cost-EDR corresponding plots. The optimal design is highlighted and circled in red. In

Scenario 1, the optimal design to achieve at least 80% EDR is to perform N=12 and R=1/8 lane and the design will cost $11,700 to achieve EDR=0.807 (see the left two plots in Figure 16). In Scenario 2, with maximal budget of $20,000, the optimal design is N=20 and R=1/10 lane, which will cost $18,000 and achieve EDR=0.922 (the two plots on the right of Figure 16).

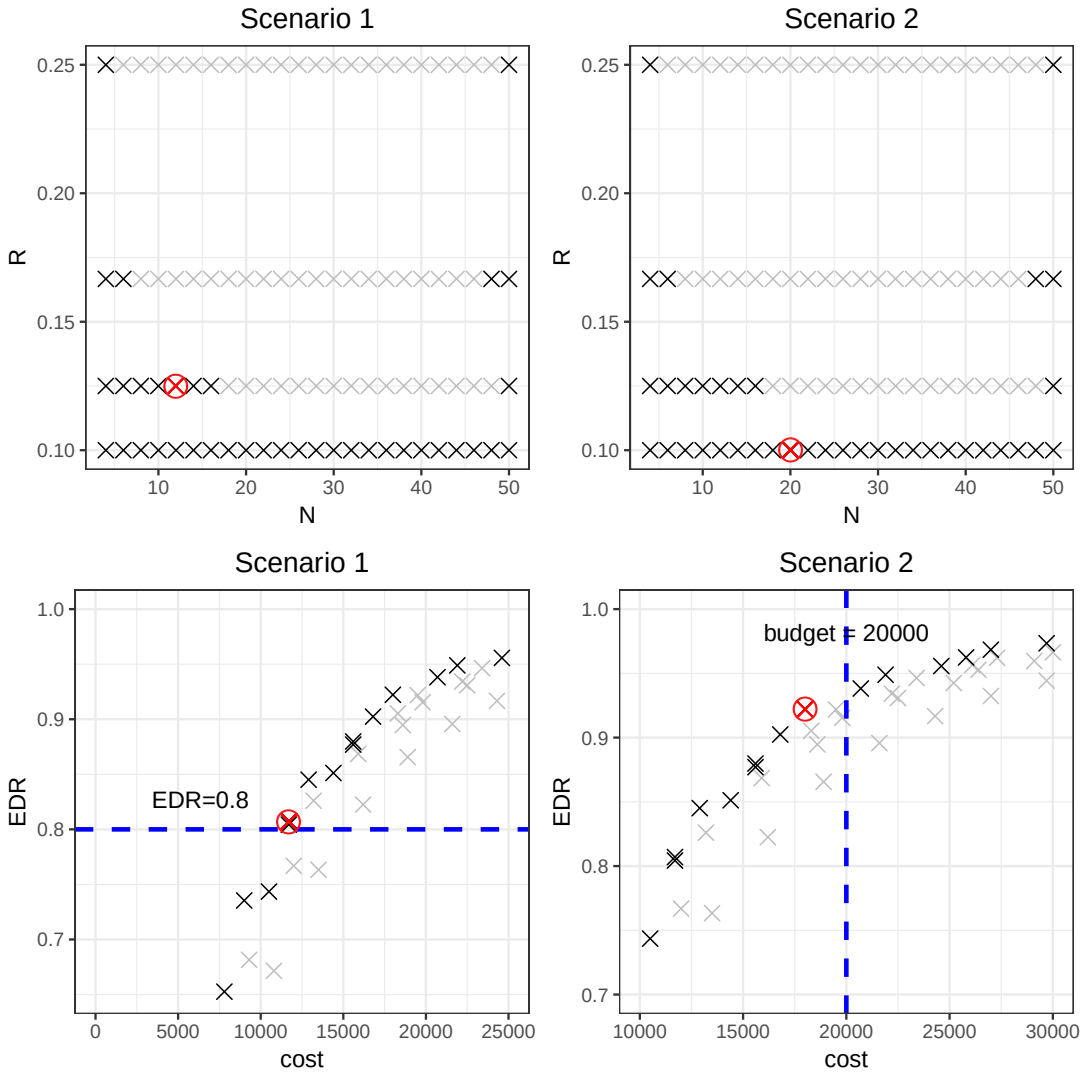


Figure 16: Illustration of study design optimization in two scenarios. The first row plots all N and R combinations. The optimal N and R combination is highlighted and circled in red, the admissible combinations are in black and inadmissible ones are in grey. The second row plots the corresponding Budget and EDR relation for N and R combinations. The blue dashed line marked the targeted EDR in scenario 1 and budget limit in scenario 2.

4.4 Real Data Application

4.4.1 Breast cancer mouse data

In this subsection, we demonstrated the performance of MethylSeqDesign using the Katz's data (Katz et al., 2015), which was used to investigate the protective risk effect of pregnancy toward breast cancer in a mouse model. The DNA methylation data were from the mammary gland tissue. The sample library was prepared using Agilent SureSelectXT Mouse Methyl-Seq Kit. The kit design covered 109 Mb of Ensemble regulatory features (promoters, promoter flanking regions, enhancers, etc.), CpG islands, known tissue-specific DMR, and open regulatory elements. The dataset was generated with mm9 mouse reference genome (Kent et al., 2002). Aligned reads outside of the targeted regions (provided by Agilent SureSelectXT Mouse Kit) were removed. Data preprocessing was performed by R package "MethyKit". We only used samples from batch one, which harvested from mammary gland tissure immediately after involutions including 5 parous and 5 non-parous mice.

A total of $G = 297,773$ methylation regions of interest (ROI) are pre-defined from the Agilent SureSelectXT kit. We randomly subsampled 2 vs. 2, 3 vs. 3 and 4 vs. 4 samples from the full data as the pilot data, and used MethylSeqDesign to calculate the predicted EDR. We repeated this procedure for 10 times. The predicted EDR from the subsampled data was compared with the reference EDR calculated using the full data. As shown in Figure 17, although the underlying true EDR is unknown, as the sample size of subsampled data increases, the predicted EDR from subsampled data converges to the predicted EDR from the full 5 vs 5 samples.

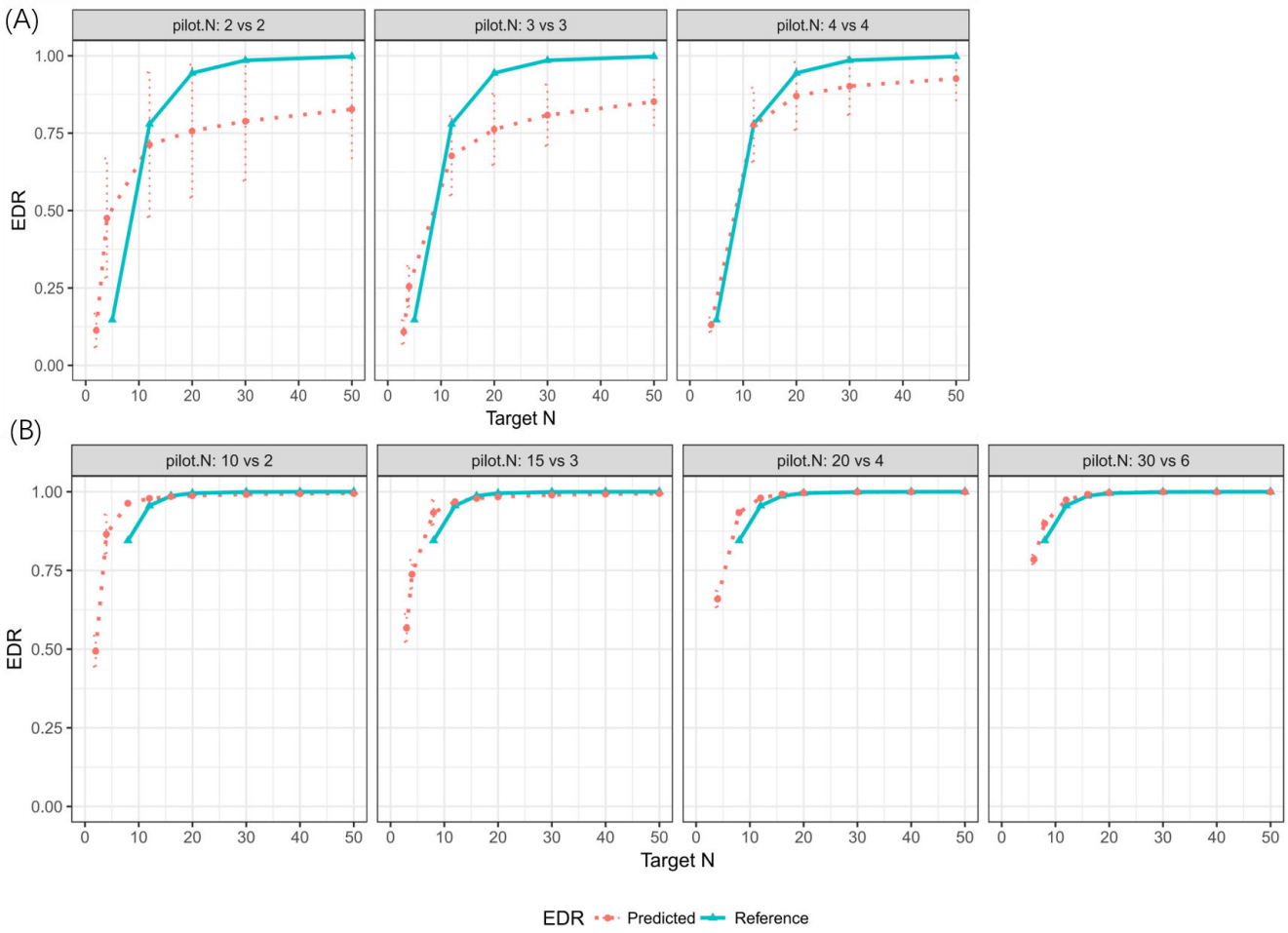


Figure 17: (A) Real data application using mouse pregnancy dataset. (B) Real data application using CLL dataset. The mean and 95% CI of the predicted EDR from subsampled data is shown in red, and the blue curve is the reference EDR from the full data. As sample size of subsampled data increases, the predicted EDR becomes closer to the reference.

4.4.2 Chronic lymphocytic leukemia data

Kushwaha et al. (2016) studied hypomethylated and hypermethylated regions and how the methylation changes affect gene expression in the oncogenesis of chronic lymphocytic leukemia(CLL) (GEO accession number GSE66167). RRBS was performed for a genome-

wide DNA methylation analysis in 43 tumors and 8 controls. We implemented MethylSeqDesign to the dataset using targeted regions defined as 250bp tiling windows with at least 10 read counts.

Similar to the previous example, the true underlying EDR is unknown in real data. We instead showed the performance of our method by comparing the predicted EDR from smaller sample size to full sample size. Since the sample size in control and tumor groups were unbalanced (number of tumor samples is roughly 5 times more than that of controls), we kept this ratio and randomly subsampled $(n_0, n_1) = (2,10)$, (4,20) and (6,30) from full dataset to treat as pilot data and repeated independent subsampling for 10 times for each (n_0, n_1) pair (see formulation for unbalanced design in Appendix C.3).

For full data$(n_0, n_1) = (8, 43)$, we also derived predicted EDR and treated it as a reference to compare with predicted EDR from smaller pilot data (shown in Figure 17). Although no underlying truth was available for this application, predicted EDR from our method gave reasonably accurate results, where increased sample size in pilot data generated less variation in predicted EDR curves and converged to the result from large pilot data. In this example, power calculation using $(n_0, n_1) = (3, 15)$ pilot data is roughly sufficient. The required larger sample size is reasonable due to unbalanced design and small sample size in n_0.

4.5 Discussion and Conclusion

NGS-based bisulfite sequencing is an increasingly important high-throughput technology to measure genome-wide methylation patterns. An important goal of Methyl-Seq is to detect differentially methylated regions (DMRs), such as promoter regions and transcription biding sites. During the study design stage, it is essential to accurately estimate study power based on appropriate method, particularly when pilot data exist. Given thousands of targeted methylation regions are considered simultaneously to detect differential methylation, it is imperative that the power calculation method is able to appropriately control genome-wide type I error rate, to evaluate genome-wide statistical power and to account for varying DMR effect sizes. To our knowledge, there is no existing method for this purpose. In this chapter,

we proposed a MethylSeqDesign statistical framework to accommodate all three elements mentioned above with FDR, EDR and estimating effect sizes from pilot data. This method uses a beta-binomial model to account for variations in the Methyl-Seq count data that is due to sampling variations and biological heterogeneities between subjects. We use FDR to control genome-wide type I error rate, and EDR as the genome-wide power. In addition, the use of Wald test statistic enables the transformation of statistics from pilot data to targeted sample size and sequencing depth, which allows two-way power calculation and saves computing time. Our method utilizes the pilot data to estimate the genome-wide distribution of methylation level difference between two groups (effect size) and the proportion of true DMRs, which can be efficiently avoid arbitrary guesses by researchers. Finally, with the specified cost function, we demonstrated how our method guides the selection of proper study designs in two scenarios.

The MethylSeqDesign framework needs a pilot dataset as input. It is crucial that the pilot data is technically similar to the targeted data as possible. If no pilot dataset is available in the local lab, existing datasets on the public domain with similar biological and technical setting (e.g. similar tissue, disease and sequencing protocols) are the appropriate alternative. In general, a pilot data with larger sample size would yield a superior estimate of EDR. Although pilot sample size required for accurate power calculation depends on biological and experimental variability in each project, from our experience, $n_0 = n_1 = 5$ is usually sufficient for accurate power calculation. Our second real example shows that unbalanced n_0 and n_1 design requires larger pilot sample size.

In this chapter, we restricted the power calculation framework to pre-defined targeted regions since only small proportion of CpG sites (5-20%) is available for differential methylation analysis. When the sequence depth is sufficiently deep, the effect on identifying DMRs is minimal, so does on power calculation. However, sequencing depth can still play important roles when sequencing depth is not deep enough. In this chapter, our method allows a two-dimensional power calculation by considering both sample size and sequencing depth.

In summary, we proposed a MethylSeqDesign framework to deal with the study design and power calculation issues for epigenetic studies of DNA methylation where the regions of interest are prespecified. As technology advances and sequencing cost decreases, the emer-

gence of more large-scale Methyl-Seq studies will lead to increased demand for Methyl-Seq study design and power calculation. An R package "MethylSeqDesign" is publicly available at https://github.com/liupeng2117/MethylSeqDesign.git and all code and data used in this chapter are available at https://github.com/liupeng2117/MethylSeqDesign_data_code.git.

5.0 Discussion and Future Direction

Chapter 2 and 3 proposed the ogClust framework for disease subtyping for high-dimensional omics data with outcome guidance. A robust estimation procedure is proposed to guard against possible violations of assumption. This model has been modified to use an accelerated failure time model to use time-to-event outcome. In the second part, we leverage prior biological information and multi-omics data by integrative ogClust method to achieve a more accurate, robust, and interpretable disease subtyping.

The methods identify outcome-associated clusters and provide convenient tools for capturing biologically meaningful disease subtypes. In addition, the proposed methods are generative, which means they could be applied to future patients to predict their disease subtypes. Unlike hard (deterministic) assignment in hierarchical clustering or K means, the prediction is a soft assignment with classification probability, reflecting each patient's confidence of subtyping prediction.

Chapter 4 proposed a MethylSeqDesign framework that utilizes pilot data for power calculation and experimental design for Methyl-Seq experiments. Firstly, this method models discrete data distributions properly, considering both sampling and biological variation. Secondly, it considers false discovery rate to control genome-wide type-I error and expected discovery rate to be the genome-wide power. Thirdly, by incorporating pilot data, our method can transform the test statistics from pilot data to statistics in targeted experiments, avoiding time-consuming simulations for power calculation. Lastly, the power calculation is two dimensional, which accounts for both sample size and sequencing depth. Our method can provide the optimal combination of sample size and sequencing depth to achieve the maximum power given a budget limit and pre-specified unit sequencing cost.

Currently, both ogClust and iogClust methods incorporate a single outcome. We will extend towards multiple outcomes or even multi-types of outcomes (e.g., continuous, survival, and categorical) to guide disease subtyping. Additionally, it is desirable to tune the level of outcome guidance in subgrouping. A joint modeling framework of outcome Y and features G given a common latent subgroup structure Z could be a future direction. Finally, we are

interested in extending our current framework using a sparse Gaussian graphical model to target subtype-specific network dynamics.

Appendix A for Chapter 2

A.1 Visual illustration of γ and δ

As shown in Figure A.1.1 (A), when γ increases from 1 to 3, the clusters become more tight in the space of subgroup assignment probabilities π. This is caused by the increased level of seperation in omics space and $\pi_{ik}|\boldsymbol{\gamma} = \frac{\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_k)}{\sum_{l=1}^{K}\exp(\boldsymbol{g}_i^T\boldsymbol{\gamma}_l)}$, where $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_k, 1 \leq k \leq K\}$. In Figure A.1.1 (B), as δ increases, the seperation of clusters in outcome Y becomes clearer. This is because δ represents the difference of intercept β_{0k} between adjacent clusters, increasing δ will increase the difference of mean outcome between different subgroups.

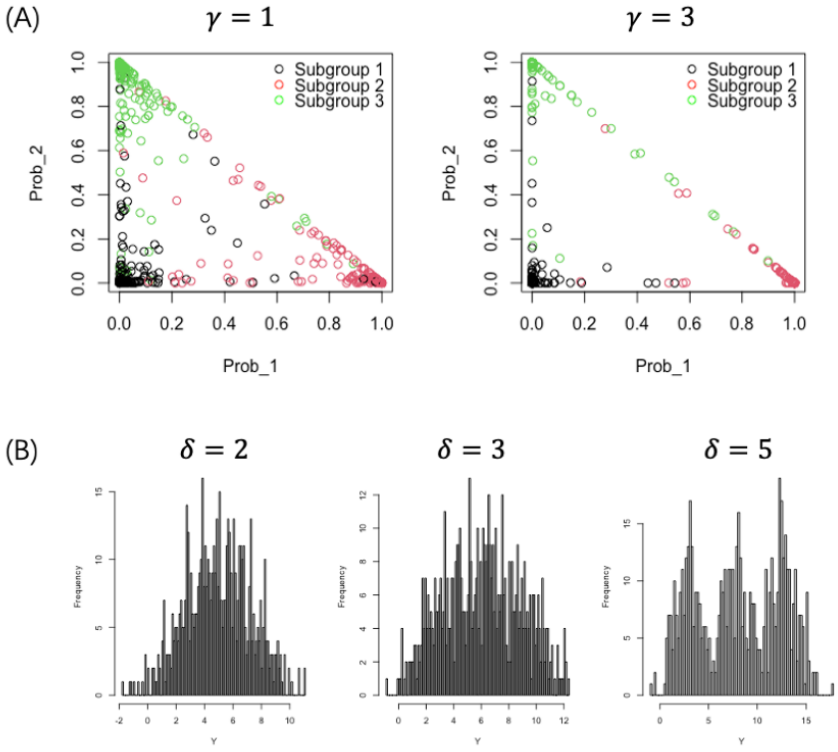


Figure A.1.1: (A) The subgroup assignment probabilities π_1 and π_2 under $\gamma = 1$ or $\gamma = 3$. (B) The distribution of simulated outcome Y when δ=2, 3, or 5

Table A.1.1: Comparison of sparse k-means (SKM), penalized model based clustering (PMBC), supervised clustering (SC) and outcome-guided clustering (ogClust) under four simulation model settings with 600 observations and 2 baseline covariates, 1000 genes and $G_{j|j\in\mathcal{A}_2} \sim N(3,1)$ in 100 repetitions.

Methods	Estimated K			ARI	Selected Genes		Outcome	
	2	3	>3		FPs	FNs	RMSE	R^2
Model 1: $\gamma = 1; \delta = 2$								
SKM	37	63	0	0.00	286.2	9.9	1.94	0.24
PMBC	0	97	3	0.00	78.7	13.4	1.93	0.24
SC	100	0	0	0.35	45.3	4.8	1.59	0.46
ogClust	41	59	0	0.45	5.8	3.0	1.55	0.51
Model 2: $\gamma = 1; \delta = 3$								
SKM	37	63	0	0.00	286.2	9.9	2.68	0.14
PMBC	0	95	5	0.00	85.3	13.4	2.68	0.13
SC	100	0	0	0.36	28.5	5	2.13	0.45
ogClust	2	98	0	0.86	14.6	0	1.90	0.55
Model 3: $\gamma = 1; \delta = 5$								
SKM	37	63	0	0.00	286.2	9.9	4.25	0.05
PMBC	0	98	2	0.00	92.0	13.2	4.27	0.04
SC	100	0	0	0.36	32.4	4.9	3.25	0.42
ogClust	1	99	0	0.91	14.4	0	2.72	0.61
Model 4: $\gamma = 3; \delta = 3$								
SKM	38	62	0	0.00	406.1	8.5	2.67	0.14
PMBC	0	100	0	0.00	82.0	13.8	2.67	0.13
SC	100	0	0	0.41	17.5	5	2.20	0.41
ogClust	2	98	0	0.88	12.1	0	1.74	0.63

Table A.1.2: Comparison of sparse k-means (SKM), penalized model based clustering (PMBC), supervised clustering (SC) and outcome-guided clustering (ogClust) under four simulation model settings with 600 observations and 2 baseline covariates, 1000 genes and $G_{j|j\in\mathcal{A}_2} \sim N(0.5, 1)$ in 100 repetitions.

Methods	Estimated K			ARI	Selected Genes		Outcome	
	2	3	> 3		FPs	FNs	RMSE	R^2
Model 1: $\gamma = 1; \delta = 2$								
SKM	100	0	0	0.05	794.0	1.6	1.94	0.24
PMBC	73	1	26	0.33	182.7	4.4	1.93	0.24
SC	100	0	0	0.35	47.9	4.8	1.59	0.48
ogClust	66	31	3	0.45	14.0	3.3	1.55	0.51
Model 2: $\gamma = 1; \delta = 3$								
SKM	100	0	0	0.05	794.0	1.6	2.66	0.13
PMBC	74	0	26	0.30	172.0	5.9	2.68	0.13
SC	100	0	0	0.36	51.0	4.8	2.09	0.47
ogClust	2	97	1	0.86	21.3	0.1	1.90	0.56
Model 3: $\gamma = 1; \delta = 5$								
SKM	100	0	0	0.05	794.0	1.6	4.22	0.05
PMBC	69	1	30	0.28	171.8	6.5	4.24	0.04
SC	100	0	0	0.36	47.5	4.8	3.21	0.46
ogClust	0	100	0	0.91	5.1	0.1	2.70	0.61
Model 4: $\gamma = 3; \delta = 3$								
SKM	100	0	0	0.06	769.0	2.0	2.62	0.15
PMBC	77	1	22	0.33	187.8	6.4	2.64	0.15
SC	100	0	0	0.41	17.3	5.0	2.02	0.51
ogClust	0	100	0	0.88	2.5	0.1	1.75	0.63

Appendix B for Chapter 3

B.1 Detailed Derivations for ADMM Updating Equations

1. Derivation for $\boldsymbol{x}_g^+$

$$\boldsymbol{x}_g^+ = \underset{x_g}{\operatorname{argmin}} \|\boldsymbol{x}_g\|_2 + \frac{\rho}{2} \left\|\boldsymbol{x}_g - \boldsymbol{\beta}_g \circ \boldsymbol{\gamma} + \boldsymbol{u}_{gk}\right\|_2^2$$

   a. When $\|\boldsymbol{x}_g\| = 0$, $\frac{\partial \|\boldsymbol{x}_g\|_2}{\partial \boldsymbol{x}_g}$ is the set of sub gradients which satisfy $\{\boldsymbol{z} : \|\boldsymbol{z}\|_2 \leq 1\}$. Therefore,

$$\frac{\partial \|\boldsymbol{x}_g\|_2}{\partial \boldsymbol{x}_g} + \rho\{\boldsymbol{u}_g - \boldsymbol{\beta}_g \circ \boldsymbol{\gamma}\} = 0$$

   .

$$\rho\{\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\} = \frac{\partial \|\boldsymbol{x}_g\|_2}{\partial \boldsymbol{x}_g}$$

   We take the ℓ_2 norm for both side of the equation

$$\rho\|\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\| \leq 1$$

   b. When $\|\boldsymbol{x}_g\| \neq 0$, take derivative with respect to $\boldsymbol{x}_g$ and set it to zero, we have

$$\frac{\boldsymbol{x}_g}{\|\boldsymbol{x}_g\|} + \rho \boldsymbol{x}_g = \rho(\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g) \tag{21}$$

   We take ℓ_2 norm on both side

$$1 + \rho\|\boldsymbol{x}_g\| = \rho\|\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\|$$

   where $\rho\|\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\| > 1$ as $\rho > 0$. Then, we have

$$\|\boldsymbol{x}_g\| = \|\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\| - \frac{1}{\rho}$$

   Plug in (21), this gives the solution for $\boldsymbol{x}_g^+$

$$\boldsymbol{x}_g^+ = (1 - \frac{1}{\rho\|\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g\|}(\boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g))$$

c. Integrating the results from (a) and (b), we can get the solution for $\boldsymbol{x}_g^+$:

$$\boldsymbol{x}_g^+ = \left(1 - \frac{1}{\rho \left\|\boldsymbol{a}_g\right\|_2}\right)_+ \boldsymbol{a}_g, \text{ where } \boldsymbol{a}_g = \boldsymbol{\beta}_g \circ \boldsymbol{\gamma} - \boldsymbol{u}_g$$

2. Derivation for $\boldsymbol{r}_k^+$

$$\boldsymbol{\gamma}_k^+ = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left\{ W_{ik} \left(h_{ik} - \boldsymbol{g}_i^T \boldsymbol{\gamma}_k\right)^2 \right\} + \sum_{\mathrm{g}=1}^{q+G} \frac{\rho}{2} \left\| \boldsymbol{x}_{gk}^+ - \boldsymbol{\beta}_{gk} \circ \boldsymbol{\gamma} + \boldsymbol{u}_{gk} \right\|_2^2$$

The objective function can be reformatted as

$$\boldsymbol{\gamma}_k^+ = \underset{\gamma}{\operatorname{argmin}} \frac{1}{2} \| D(\boldsymbol{W}_k)^{\frac{1}{2}} \left(\boldsymbol{h}_k - G\boldsymbol{\gamma}_k\right) \|_2^2 + \sum_{\mathrm{g}=1}^{q+G} \frac{\rho}{2} \left\| \boldsymbol{x}_{gk}^+ - D(\boldsymbol{\beta}_{gk})\boldsymbol{\gamma} + \boldsymbol{u}_{gk} \right\|_2^2$$

, where $D(\boldsymbol{x})$, $\boldsymbol{x} = \{x_1, ..., x_n\}$, is a $n \times n$ square matrix with diagonal elements equal $\boldsymbol{x}$ and off-diagonal elements are 0s.

Take derivative with respect to $\boldsymbol{\gamma}_k$ and set it to 0, we have

$$G^T D(\boldsymbol{W}_k)(G\boldsymbol{\gamma}_k - \boldsymbol{h}_k) + \sum_{g=1}^{q+G} \rho D(\boldsymbol{\beta}_{gk})^2 \boldsymbol{\gamma}_k - \sum_{g=1}^{q+G} D(\boldsymbol{\beta}_{gk})(\boldsymbol{x}_{gk} + \boldsymbol{u}_{gk}) = 0$$

Therefore,

$$\boldsymbol{\gamma}_k^+ = (G^T D(\boldsymbol{W}_k) G + \sum_{g=1}^{q+G} \rho D(\boldsymbol{\beta}_{gk})^2)^{-1} (G^T D(\boldsymbol{W}_k) \boldsymbol{h}_k + \rho \sum_{g=1}^{q+G} D(\boldsymbol{\beta}_{gk})(\boldsymbol{x}_{gk}^+ + \boldsymbol{u}_{gk}))$$

Appendix C for Chapter 4

C.1 Coverage of Methyl-Seq Data

In this section, we showed the coverage of three types of Methyl-Seq data at CpG site level. The three Methyl-Seq types are WGBS, RRBS and Agilent SureSelect. We selected an example dataset for each type to illustrate (Smallwood et al., 2014; Bouschet et al., 2016; Katz et al., 2015). The coverage for these types of Methyl-Seq data is summarized in Table C.1.1, Table C.1.2 and Table C.1.3 respectively. As shown in Table C.1.1- C.1.3, the mapping rate of WGBS is around $30 \sim 50\%$, which is relatively lower compared with that of RRBS (>70%) and Agilent SureSelect ($60 \sim 70$ %). For RRBS and Agilent SureSelect we used the combined CpG sites of all samples as the total CpG set. We observed that only less than 20% (<1% in WGBS, <10% in RRBS and <20% in Agilent SureSelect) of the CpGs have coverage greater than 10, while 10 is an extremely low coverage for a good estimate of genome-wide power. However, when CpG sites are combined into regions, more than 60% regions of interest (ROI) have coverage greater than 10 in Agilent SureSelect dataset. The coverage of regions increased to a level where power calculation is feasible.

Table C.1.1: Summary of coverage at single CpG site level for WGBS data. Five samples are listed in the table below. Bulk MII is bulk WGBS data with ultra deep sequencing depth, the other four samples are single-cell WGBS data. The columns from left to righ are: sample ID, the total number of reads, % of reads mapped, total number of CpG site, % of CpG sites with coverage >0, >5 and >10, the number of lanes, and the cost.

Sample ID	# Reads	% Mapped	total # CpGs	% >0	% >5	% >10	# Lanes	Cost($)
MII #2	27,712,173	35.1	21,342,779	26.1	0.9	0.1	0.1	450
MII #5	13,370,171	40.1	21,342,779	23.5	0.3	0.02	0.05	375
MII #2 deep	54,185,479	32.1	21,342,779	35.8	3.6	0.7	0.2	600
MII #5 deep	49,015,151	36.2	21,342,779	45.1	4.6	0.8	0.2	600
Bulk MII	874,735,536	51.6	21,342,779	59.5	5.5	0.6	3.5	5550

Table C.1.2: Summary of coverage at single CpG site level for RRBS data. The columns from left to righ are: sample ID, the total number of reads, % of reads mapped, total number of CpG site, % of CpG sites with coverage >0, >5 and >10, the number of lanes, and the cost.

Sample ID	# Reads	% Mapped	total # CpGs	% >0	% >5	% >10	# Lanes	Cost($)
ESCs d0 e14 rep1	38,572,054	73.8	24,539,081	14	9	9	0.16	540
ESCs d0 e14 rep2	58,797,704	73.0	24,539,081	15	10	9	0.24	660
ESCs d0 BJ1 rep1	49,808,291	71.6	24,539,081	18	11	9	0.20	600
ESCs d0 BJ1 rep2	42,114,981	70.8	24,539,081	16	10	9	0.17	555
ESCs d0 JB6 rep1	66,369,673	71.3	24,539,081	16	11	9	0.27	705
ESCs d0 JB6 rep2	60,846,336	70.4	24,539,081	15	10	9	0.24	660

Table C.1.3: Summary of coverage for Agilent SureSelect data. The columns from left to righ are: sample ID, the total number of reads, % of reads mapped, total number of CpG site, % of CpG sites with coverage >0, >5 and >10, the number of lanes, the cost, % of regions with coverage >10, and mean coverage of regions.

Sample ID	# Reads	% Mapped	Total # CpGs	% >0	% >5	% >10	# Lanes	Cost($)	% ROI* >10	Mean ROI cov
P2	61,583,442	62.6	11,250,286	39.8	21.6	17.1	0.25	675	63.8	407.6
P3	33,822,020	63.1	11,250,286	31.7	17.8	13.0	0.14	510	60.7	239.7
P5	69,291,406	66.8	11,250,286	25.6	21.2	17.4	0.28	720	62.0	494.8
P12	146,380,998	57.5	11,250,286	25.7	20.9	19.3	0.58	1170	60.6	755.3
P13	58,523,029	50.0	11,250,286	40.5	18.6	14.3	0.24	660	62.7	283.6
V1	80,415,885	66.5	11,250,286	47.4	24.4	19.8	0.32	780	65.5	561.4
V2	48,455,148	61.3	11,250,286	38.2	20.1	15.7	0.19	585	62.8	321.7
V3	54,430,968	62.8	11,250,286	45.1	21.8	17.8	0.22	630	63.5	402.1
V4	56,599,418	65.9	11,250,286	23.6	19.1	15.6	0.23	645	60.7	396.0
V9	38,017,592	62.9	11,250,286	34.3	16.1	11.7	0.15	525	61.2	237.0

*Region of interest (ROI)

C.2 Simulations

C.2.1 The value of Ψ_g as sequencing depth changes

Ψ_g denotes the value of $\frac{\bar{A}_g+\bar{B}_g}{\bar{A}_g\times\bar{B}_g}$ in Equation (18) of main manuscript. In step III of our method, the test statistic is transformed by a factor of ratio $\sqrt{\frac{N}{n}} \times \sqrt{\frac{\Psi_g}{\Psi_g'}}$, where $\sqrt{\frac{\Psi_g}{\Psi_g'}}$ is mainly influenced by sequencing depth. Thus the trend of the quantity Ψ_g as the change of sequencing depth is of our interest, it also helps us understand the trend of EDR prediction. We simulated methylation count data with sample size $n_0 = n_1$= 3, 6, 9, 12 for one targeted region to examine this assumption. The simulation was as follows: (1) We empirically drew the coverage data for sample s (i.e. m_s,$1 \leq s \leq n_0+n_1$) from the mouse pregnancy data and calculated the total reads $R = \Sigma_s m_s$. The median of m_s is about 330. (2) By downsampling, the total reads changed from R_0 to R_j ($j = 1, 2, ..., 10$), where R_j/R_0 = 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1. In other words, mean coverage $\bar{m} \approx$ 15, 30, 60, 120, 180, 240, 330 respectively. (3) Ψ_g and $\Psi_g{}'$ were calculated given coverage level m_s, $m_s^{'}$ respectively, and given the dispersion ϕ=0.048. We repeated steps (1) to (3) for 100 times and calculated the median and interquantile range (IQR) of Ψ_g under each (n_0, n_1).

As shown in Figure C.2.1, as sequencing depth increases, the value of Ψ_g decreases, however, when $R_j/R_0 > 0.5$ (i.e. coverage is greater than about 160) sequencing depth has little influence on Ψ_g, thus it has little influence on the prediction of EDR. This shows that when mean coverage level of pilot data is above $\sim$ 160, downsampling can account for all the necessary searching space for sequencing depth.

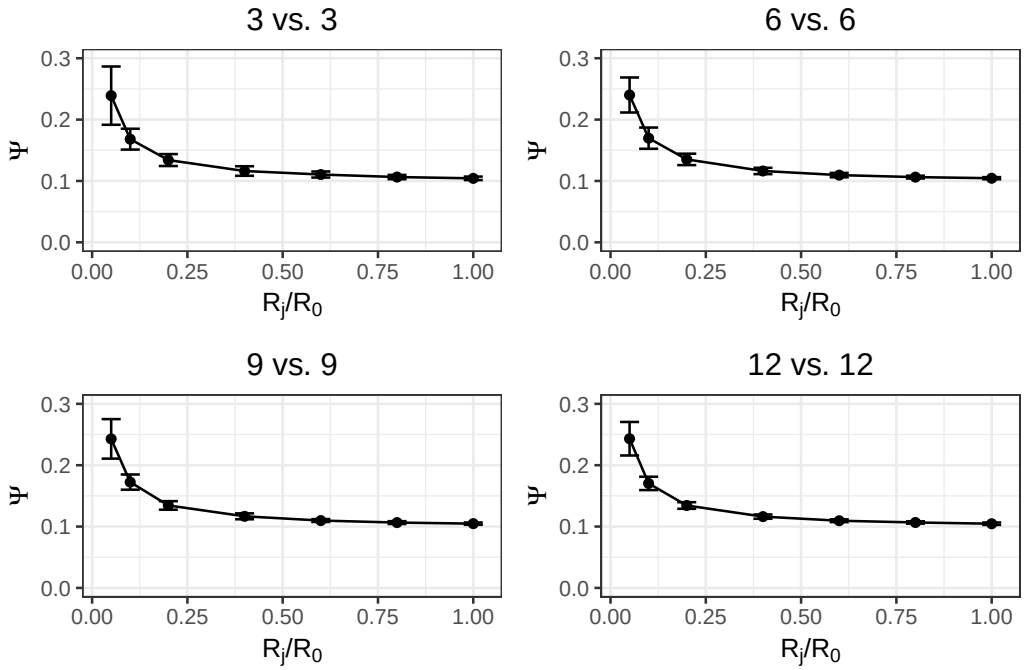


Figure C.2.1: The mean and 95% CI of the quantity $\Psi_g = \frac{\bar{A}_g + \bar{B}_g}{\bar{A}_g \times \bar{B}_g}$ as sample size and sequencing depth changes.

C.2.2 Estimation of λ

We evaluated the performance of BUM and CBUM methods in estimation of λ under various power of pilot data. The BUM method fits a Beta-uniform mixture model and estimates λ by optimizing log likelihood. This method is implemented by the optim function of stats package in R. The CBUM method proposed by Markitsis and Lai (2010) fits a censored beta-uniform mixture model to reduce the impact of extremely small p-values. This method is implemented by the CBUM function of pi0 package in R.

The power of pilot data is varied by sample size n_0 and effect size $\triangle$. n_0 varies from 2 to 10 and $\triangle$ is either fixed ($\triangle = 0.1, 0.14, 0.18$) or follows $U(0.1, 0.2)$. The pilot data was simulated following the simulation scheme in Section 4.3.1 of main manuscript. True value of λ is 0.9. The result is shown in Figure C.2.2. Overall, as the power of pilot data increase, the performance of both methods becomes better, however, the estimates of CBUM are close to 0.9 and more accurate than that of BUM.

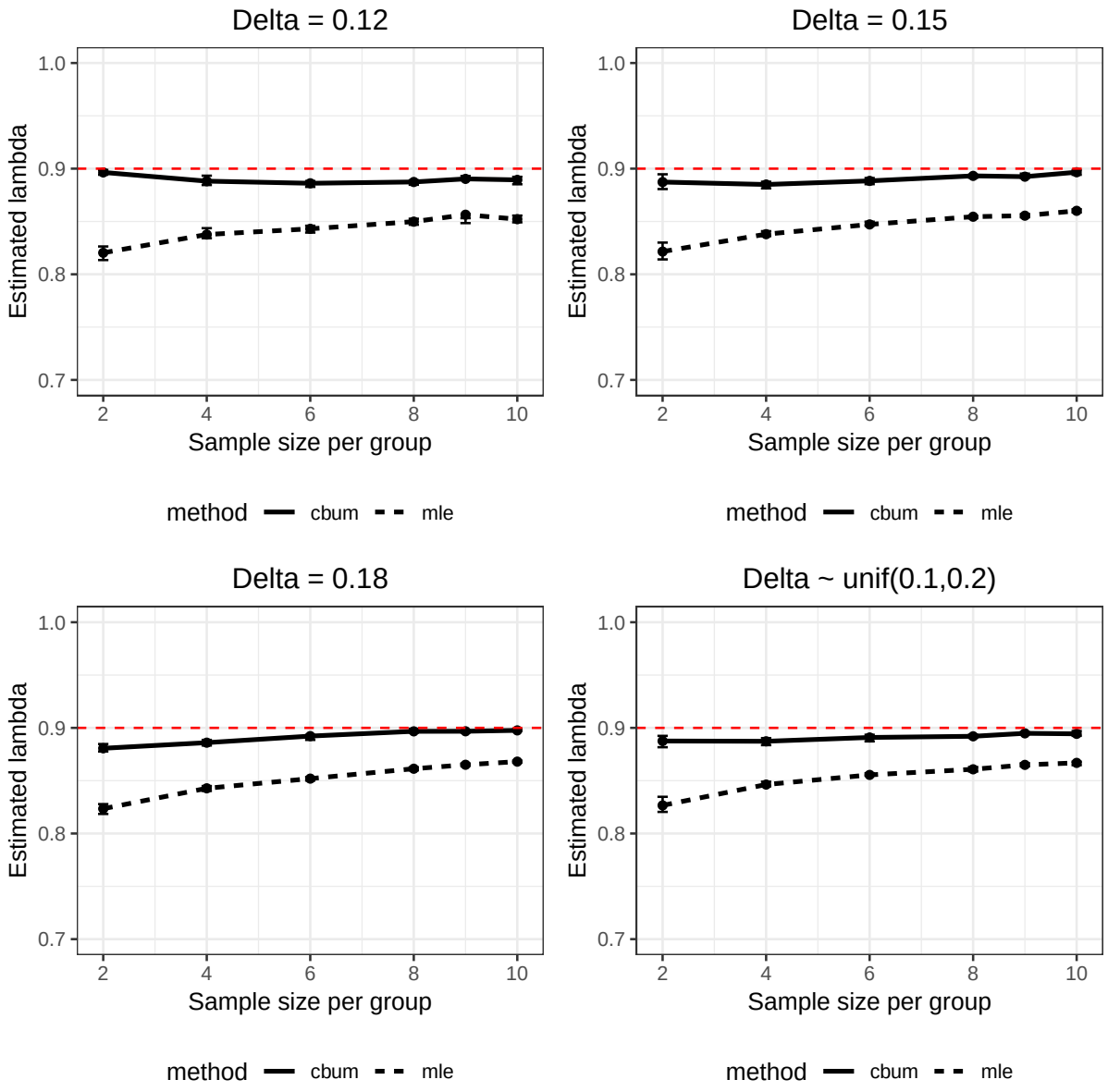


Figure C.2.2: The λ estimates of BUM (solid line) and CBUM (dashed line) under various power level of pilot data. The pilot data sample size varied from 2 to 10, and the effect size delta is either fixed (delta=0.1, 0.14, 0.18) or randomly drawn from $U(0.1, 0.2)$.

C.2.3 Performance evaluation

We stratified the performance evaluation based on different level of effect size $\triangle$. We first simulated B=10 pilot datasets ($b = 1, 2, ...B$) with pilot sample size n_0= 2, 4, 6, 8, 9

and 10 when R_0 are 2 million reads. Then for each pilot dataset with (n_0, R_0), the projected power $\widehat{\text{EDR}}(N_i, R_j; n_0, R_0)$ for target sample size $N_i = 5, 10, 15, 20, 30, 50$ $(i = 1, 2, ..., 6)$ and $R_j = 0.1, 0.2, 0.4, 0.8, 1.2, 1.6$ million reads(j =1,2,...,6) is estimated using our method. At the same time, the true EDR for each $(N_i,\ R_j)$ can be estimated as $\widehat{EDR}\,(N_i, R_j) = \frac{\sum_{b=1}^{B} \widehat{EDR}^{(b)}(N_i, R_j)}{B}$ where $\widehat{EDR}^{(b)}\ (N_i, R_j)$ is the actual EDR in the b-th simulation when sample size N_i and R_j are simulated. We compared the estimated EDR using our method and the estimated true EDR under different effect size $\triangle$. Results are shown in Figure C.2.3, C.2.4 and C.2.5 (Figure C.2.3 for $\triangle = 0.1$, Figure C.2.4 for $\triangle = 0.18$ and Figure C.2.5 for $\triangle$ drawn from $U(0.1, 0.2)$).

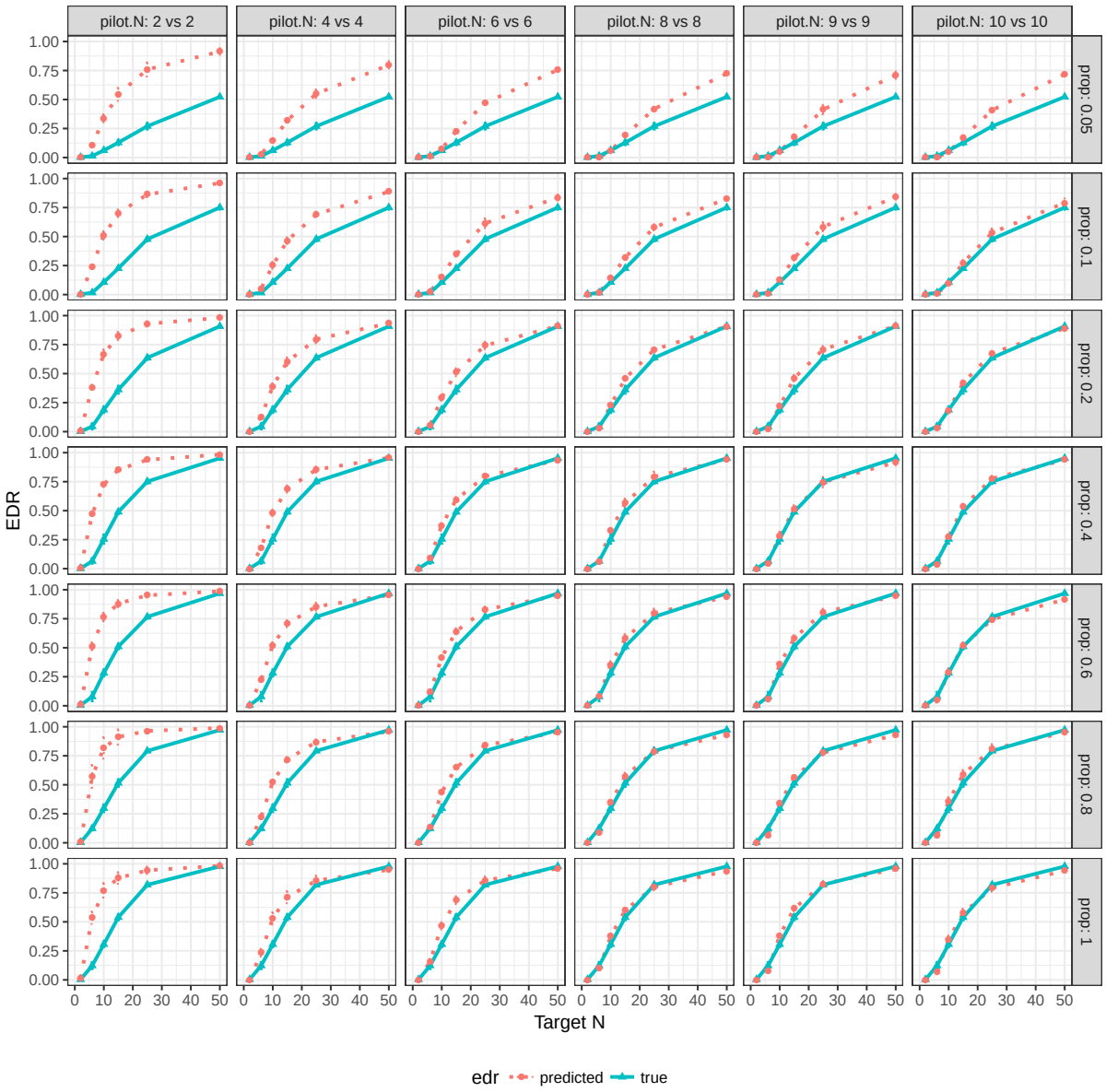


Figure C.2.3: The EDR prediction from MethylSeqDesign compared to the true EDR. Effect size $\triangle = 0.1$. The pilot data sample size per group varied from 2 to 10, and $R_0 = 5M$ is fixed. The predicted EDRs by the pilot data are shown by the dotted red curves. The targeted data sample size per group varied from 2 to 50, and the ratio of targed sample sequencing depth to that of pilot is $prop = \frac{R_j}{R_0}$, which varied from 0.05 to 1. The estimated true EDRs by targeted data are in blue solid curves.

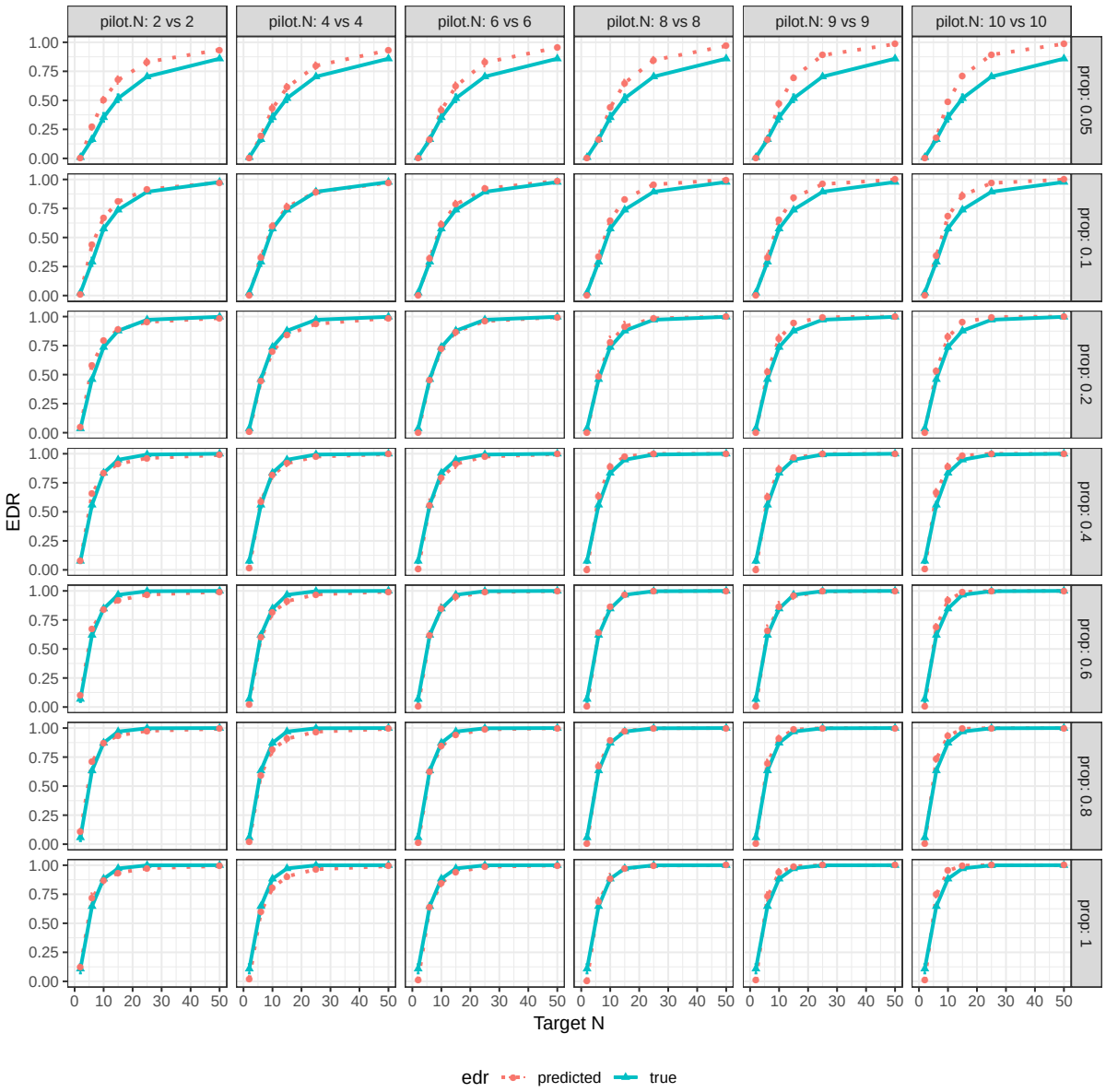


Figure C.2.4: The EDR prediction from MethylSeqDesign compared to the true EDR. Effect size $\triangle = 0.18$. The pilot data sample size per group varied from 2 to 10, and $R_0 = 5M$ is fixed. The predicted EDRs by the pilot data are shown by the dotted red curves. The targeted data sample size per group varied from 2 to 50, and the ratio of targed sample sequencing depth to that of pilot is $prop = \frac{R_j}{R_0}$, which varied from 0.05 to 1. The estimated true EDRs by targeted data are in blue solid curves.

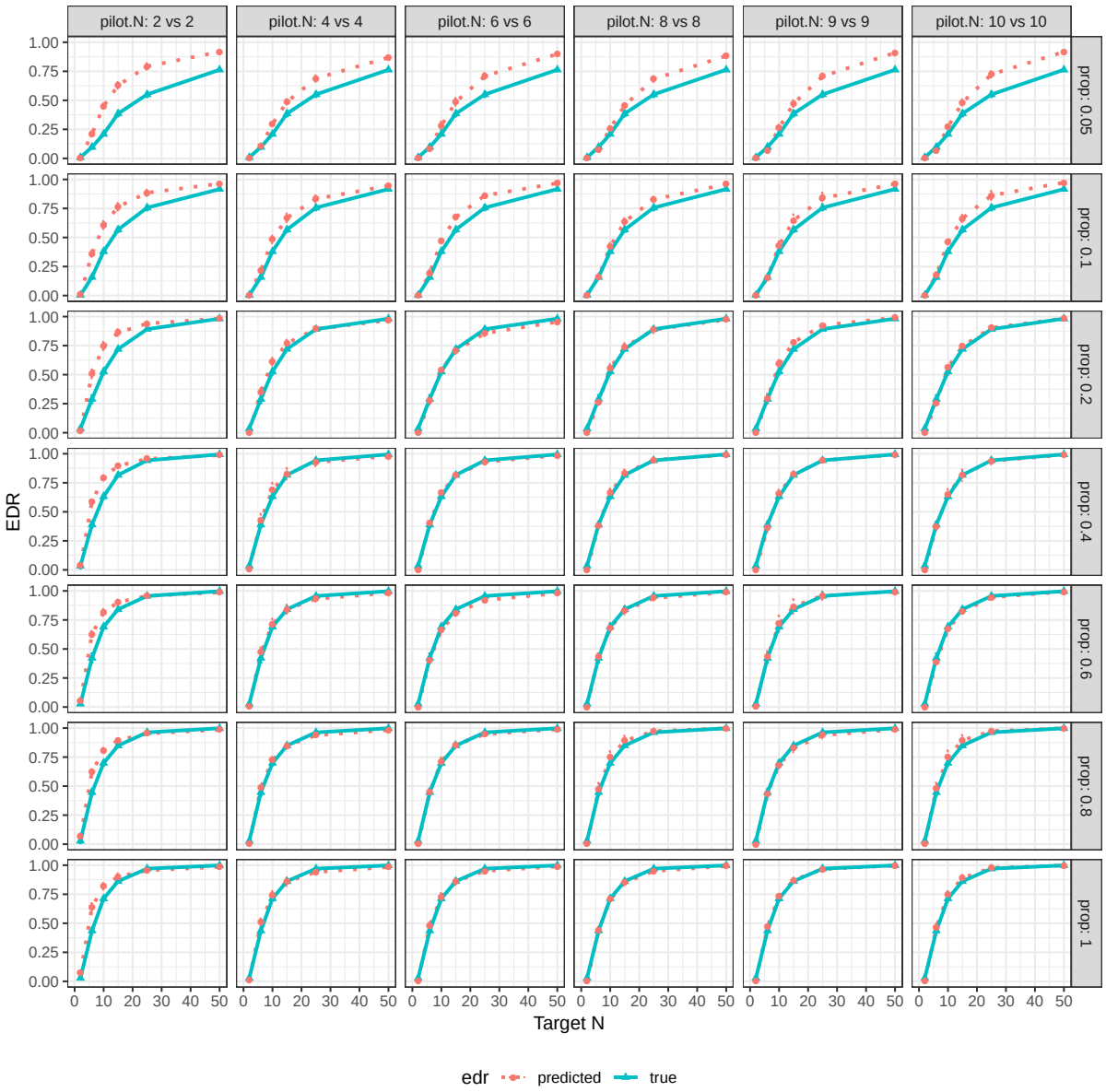


Figure C.2.5: The EDR prediction from MethylSeqDesign compared to the true EDR. Effect size $\triangle$ were drawn from $U(0.1, 0.2)$. The pilot data sample size per group varied from 2 to 10, and $R_0 = 5M$ is fixed. The predicted EDRs by the pilot data are shown by the dotted red curves. The targeted data sample size per group varied from 2 to 50, and the ratio of targed sample sequencing depth to that of pilot is $prop = \frac{R_j}{R_0}$, which varied from 0.05 to 1. The estimated true EDRs by targeted data are in blue solid curves.

C.3 Unbalanced Design

When there are two groups and $n_0 \neq n_1$ and $N_0 \neq N_1$, and $n_0 \neq n_1$. The model can be written as $\arcsin(2\pi_{gj} - 1) = \beta_{0g} + x_{j2}\beta_{1g}$ and the variance of $\hat{\beta}_{1g}$ can be written as

$$\begin{aligned}
\widehat{\text{Var}}\left(\hat{\beta}_{1g}\right) &= \frac{\sum_{j=1}^{n_0+n_1} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}}{\sum_{j_1=1}^{n_0} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g} \times \sum_{j_2=1}^{n_1} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}} \\
&= \frac{(n_0+n_1) \times \frac{1}{n_0+n_1} \sum_{j=1}^{n_0+n_1} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}}{n_0 \times n_1 \times \frac{1}{n_0} \sum_{j_1=1}^{n_0} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g} \times \frac{1}{n_1} \sum_{j_2=1}^{n_1} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}} \\
&= \frac{n_0+n_1}{n_0 \times n_1} \frac{\bar{X}}{\bar{X}_1 \times \bar{X}_2},
\end{aligned} \tag{22}$$

where $\bar{X} = \frac{n_0+n_1}{n_0 \times n_1} \sum_{j=1}^{n_0+n_1} \frac{m_{gj}}{1+(m_{gj}-1)\hat{\phi}_g}$, $\bar{X}_1 = \frac{1}{n_0} \sum_{j_1=1}^{n_0} \frac{m_{gj_1}}{1+\left(m_{gj_1}-1\right)\hat{\phi}_g}$ and $\bar{X}_2 = \frac{1}{n_1} \sum_{j_2=1}^{n_1} \frac{m_{gj_2}}{1+\left(m_{gj_2}-1\right)\hat{\phi}_g}$. When Sample size change from (n_0, n_1) to (N_0, N_1), the Wald test statistic Z_g change by a factor of $\sqrt{\frac{N_0 N_1}{n_0 n_1} \times \frac{n_0+n_1}{N_0+N_1}}$

Bibliography

Ahmad, A. and Fröhlich, H. (2017). Towards clinically more relevant dissection of patient heterogeneity via survival-based bayesian clustering. *Bioinformatics*, 33(22):3558–3566.

Allison, Gadbury, Heo, Fernandez, Lee, Prolla, Weindruch, DB, A., GL, G., M, H., JR, F., C-K, L., TA, P., and R., W. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39:1–20.

Bair, E. (2013). Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5):349–361.

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Bair, E. and Tibshirani, R. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2(4):e108.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

Baylin, S. B. (2005). Dna methylation and gene silencing in cancer. *Nature Reviews. Clinical Oncology*, 2(S1):S4.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bouschet, T., Dubois, E., Reynès, C., Kota, S. K., Rialle, S., Maupetit-Méhouas, S., Pezet, M., Le Digarcher, A., Nidelet, S., Demolombe, V., et al. (2016). In vitro corticogenesis from embryonic stem cells recapitulates the in vivo epigenetic control of imprinted gene expression. *Cerebral Cortex*, 27(3):2418–2433.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78.

Bowen, E. F. W., Burgess, J. L., Granger, R., Kleinman, J. E., and Rhodes, C. H. (2019). Correction: DLPFC transcriptome defines two molecular subtypes of schizophrenia. *Translational Psychiatry*, 9(1).

Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21(1):33–37.

Burstein, M. D., Tsimelzon, A., Poage, G. M., Covington, K. R., Contreras, A., Fuqua, S. A., Savage, M. I., Osborne, C. K., Hilsenbeck, S. G., Chang, J. C., et al. (2015).

Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. *Clinical Cancer Research*, 21(7):1688–1698.

Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R., and Marth, G. T. (2013). Scotty: A web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, 29(5):656–657.

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282.

Chen, X., Li, J., Gray, W. H., Lehmann, B. D., Bauer, J. A., Shyr, Y., and Pietenpol, J. A. (2012). TNBCtype: A subtyping tool for triple-negative breast cancer. *Cancer Informatics*, 11:CIN.S9983.

Chi, Y., Li, Y., Zhang, H., and Liang, Y. (2019). Median-truncated gradient descent: A robust and scalable nonconvex approach for signal estimation. In *Compressed Sensing and Its Applications*, pages 237–261. Springer.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the american statistical association*, 83(401):173–178.

Dean, N. and Raftery, A. E. (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62(1):11.

Delpu, Y., Cordelier, P., Cho, W. C., and Torrisani, J. (2013). DNA methylation and cancer diagnosis.

Demkow, U. and Wolańczyk, T. (2017). Genetic tests in major psychiatric disorders—integrating molecular medicine with clinical psychiatry—why is it so difficult? *Translational Psychiatry*, 7(6):e1151–e1151.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Desantis, S. M., Andrés Houseman, E., Coull, B. A., Nutt, C. L., and Betensky, R. A. (2012). Supervised bayesian latent class models for high-dimensional data. *Statistics in Medicine*, 31(13):1342–1360.

DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., and Betensky, R. A. (2007). A penalized latent class model for ordinal data. *Biostatistics*, 9(2):249–262.

DeSantis, S. M., Houseman, E. A., Coull, B. A., Stemmer-Rachamimov, A., and Betensky, R. A. (2008). A penalized latent class model for ordinal data. *Biostatistics*, 9(2):249–262.

Dobbin, K. and Simon, R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, 6(1):27–38.

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC bioinformatics*, 15(1):215.

Esteller, M. (2005). Aberrant DNA methylation as a cancer-inducing mechanism. *Annual review of pharmacology and toxicology*, 45:629–56.

Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69.

Ferreira, J. A. and Zwinderman, A. (2006). Approximate Sample Size Calculations with Microarray Data: An Illustration. *Statistical Applications in Genetics and Molecular Biology*, 5(1).

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1.

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., Zack, D. J., Arakawa, S., Cipriani, V., Ripke, S., et al. (2013). Seven new loci associated with age-related macular degeneration. *Nature genetics*, 45(4):433–439.

Furgal, A. K., Sen, A., and Taylor, J. M. (2019). Review and comparison of computational approaches for joint longitudinal and time-to-event models. *International Statistical Review*, 87(2):393–418.

Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J. D., and Allison, D. B. (2004a). Power and sample size estimation in high dimensional biology. *Statistical methods in medical Research*, 13(4):325–338.

Gadbury, G. L., Page, G. P., Edwards, J., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J. D., and Allison, D. B. (2004b). Power and sample size estimation in high dimensional biology. *Statistical Methods in Medical Research*, 13(4):325–338.

Gaynor, S. and Bair, E. (2013). Identification of relevant subtypes via preweighted sparse clustering. *arXiv preprint arXiv:1304.3760*.

Gaynor, S. and Bair, E. (2017). Identification of relevant subtypes via preweighted sparse clustering. *Computational statistics & data analysis*, 116:139–154.

Guinney, J., Dienstmann, R., Wang, X., De Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 21(11):1350–1356.

Guo, J., Wall, M., and Amemiya, Y. (2006). Latent class regression on latent factors. *Biostatistics*, 7(1):145–163.

Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A., and Kocher, J.-P. (2013). Calculating sample size estimates for RNA sequencing data. *J Comput Biol*, 20(12):970–8.

He, B., Yang, H., and Wang, S. (2000). Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications*, 106(2):337–356.

Houseman, E. A., Coull, B. A., and Betensky, R. A. (2006). Feature-specific penalized latent class analysis for genomic data. *Biometrics*, 62(4):1062–1070.

Huber, P. J. (2004). *Robust statistics*, volume 523. John Wiley & Sons.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *Journal of the American Statistical Association*, 111(513):27–42.

Huo, Z. and Tseng, G. (2017). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The annals of applied statistics*, 11(2):1011.

Hurd, P. J. and Nelson, C. J. (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics and Proteomics*, 8(3):174–183.

Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., and Thun, M. J. (2009a). Cancer statistics, 2009. *CA: A Cancer Journal for Clinicians*, 59(4):225–249.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., and Thun, M. J. (2009b). Cancer statistics, 2009. *CA: A Cancer Journal for Clinicians*, 59(4):225–249.

Jiang, Y., Shi, X., Zhao, Q., Krauthammer, M., Rothberg, B. E. G., and Ma, S. (2016). Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. *Genomics*, 107(6):223–230.

Jung, S.-H. (2005). Sample size for FDR-control in microarray data analysis. *Bioinformatics*, 21(14):3097–3104.

Jung, S.-H. and Young, S. S. (2012). Power and sample size calculation for microarray studies. *Journal of biopharmaceutical statistics*, 22(1):30–42.

Katz, T. A., Liao, S. G., Palmieri, V. J., Dearth, R. K., Pathiraja, T. N., Huo, Z., Shaw, P., Small, S., Davidson, N. E., Peters, D. G., Tseng, G. C., Oesterreich, S., and Lee, A. V. (2015). Targeted DNA methylation screen in the mouse mammary genome reveals a parity-induced hypermethylation of igf1r that persists long after parturition. *Cancer Prevention Research*, 8(10):1000–1009.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006.

Kim, S., Herazo-Maya, J. D., Kang, D. D., Juan-Guardela, B. M., Tedrow, J., Martinez, F. J., Sciurba, F. C., Tseng, G. C., and Kaminski, N. (2015). Integrative phenotyping framework (ipf): integrative clustering of multiple omics data identifies novel lung disease subphenotypes. *BMC Genomics*, 16(1):924.

Kim, S., Oesterreich, S., Kim, S., Park, Y., and Tseng, G. C. (2017). Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. *Biostatistics*, 18(1):165–179.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38.

Koestler, D. C., Marsit, C. J., Christensen, B. C., Karagas, M. R., Bueno, R., Sugarbaker, D. J., Kelsey, K. T., and Houseman, E. A. (2010). Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, 26(20):2578–2585.

Ku, C. S., Naidoo, N., Wu, M., and Soong, R. (2011). Studying the epigenome using next generation sequencing. *Journal of medical genetics*, 48(11):721–730.

Kushwaha, G., Dozmorov, M., Wren, J. D., Qiu, J., Shi, H., and Xu, D. (2016). Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia. *Human genomics*, 10(2):18.

Lanza, S. T. and Rhoades, B. L. (2013). Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*, 14(2):157–168.

Larsen, K. (2004). Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics*, 60(1):85–92.

Lee, M.-L. T. and Whitmore, G. A. (2002). Power and sample size for DNA microarray studies. *Statistics in Medicine*, 21(23):3543–3570.

Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., and Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*, 121(7):2750–2767.

Li, C.-I., Samuels, D. C., Zhao, Y.-Y., Shyr, Y., and Guo, Y. (2018). Power and sample size calculations for high-throughput sequencing-based experiments. *Briefings in bioinformatics*, 19(6):1247–1255.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–5.

Li, W., Dai, C., Zhou, X. J., Tseng, G., Ghosh, D., and Zhou, X. J. (2015). Integrative analysis of many biological networks to study gene regulation. *Integrating Omics Data*, 68.

Licht, J. D. (2015). DNA Methylation Inhibitors in Cancer Therapy: The Immunity Dimension. *Cell*, 162(5):938–939.

Lin, C.-W., Liao, S. G., Liu, P., Lee, M.-L. T., Park, Y. S., and Tseng, G. C. (2019). Rnaseqdesign: a framework for ribonucleic acid sequencing genomewide power calculation and study design issues. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):683–704.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97(457):53–65.

Linnekamp, J. F., Wang, X., Medema, J. P., and Vermeulen, L. (2015). Colorectal cancer heterogeneity and targeted therapy: A case for molecular disease subtypes. *Cancer Research*, 75(2):245–249.

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nature genetics*, 46(2):200.

Liu, P. and Hwang, J. G. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, 23(6):739–746.

Liu, P., Lin, C.-W., Park, Y., and Tseng, G. (2021). Methylseqdesign: a framework for methyl-seq genome-wide power calculation and study design issues. *Biostatistics*, 22(1):35–50.

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616.

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics*, 7(1):523.

Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., Lin, C.-W., Liu, S., Wang, L., Liu, P., et al. (2019). Metaomics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics*, 35(9):1597–1599.

Mankoo, P. K., Shen, R., Schultz, N., Levine, D. A., and Sander, C. (2011). Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PloS one*, 6(11):e24709.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141.

Markitsis, A. and Lai, Y. (2010). A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes. *Bioinformatics*, 26(5):640–646.

Masuda, H., Baggerly, K. A., Wang, Y., Zhang, Y., Gonzalez-Angulo, A. M., Meric-Bernstam, F., Valero, V., Lehmann, B. D., Pietenpol, J. A., Hortobagyi, G. N., et al. (2013). Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clinical cancer research*, 19(19):5533–5540.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164.

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). MethylSig: A whole genome DNA methylation analysis pipeline. *Bioinformatics*, 30(17):2414–2422.

Park, Y. and Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453.

Paulsen, M. and Ferguson-Smith, A. C. (2001). DNA methylation in genomic imprinting, development, and disease. *The Journal of pathology*, 195(1):97–110.

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., et al. (2000). Molecular portraits of human breast tumours. *nature*, 406(6797):747–752.

Planey, C. R. and Gevaert, O. (2016). Coincide: A framework for discovery of patient subtypes across multiple datasets. *Genome medicine*, 8(1):27.

Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical methods in medical research*, 23(1):74–90.

Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. *Biostatistics*, 10(3):535–549.

Raghavachari, N. and Garcia-Reyero, N. (2018). Overview of gene expression analysis: transcriptomics. In *Gene Expression Analysis*, pages 1–6. Springer.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med*, 5(9):e184.

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics.

Robertson, K. D. (2005). Dna methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610.

Rojas, V., Hirshfield, K., Ganesan, S., and Rodriguez-Rodriguez, L. (2016). Molecular characterization of epithelial ovarian cancer: Implications for diagnosis and treatment. *International Journal of Molecular Sciences*, 17(12):2113.

Schumacher, A., Kapranov, P., Kaminsky, Z., Flanagan, J., Assadzadeh, A., Yau, P., Virtanen, C., Winegarden, N., Cheng, J., Gingeras, T., and Petronis, A. (2006). Microarray-based DNA methylation profiling: Technology and applications. *Nucleic Acids Research*, 34(2):528–542.

Shen, K. and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics*, 26(10):1316–1323.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912.

Shi, C., Fan, A., Song, R., and Lu, W. (2018). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817.

Stessman, H. A. F., Turner, T. N., and Eichler, E. E. (2016). Molecular subtyping and improved treatment of neurodevelopmental disease. *Genome Medicine*, 8(1).

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *The international journal of biostatistics*, 4(1).

Sugasawa, S. and Noma, H. (2019). Estimating individual treatment effects by gradient boosting trees. *Statistics in medicine*, 38(26):5146–5159.

Sun, J., Herazo-Maya, J. D., Molyneaux, P. L., Maher, T. M., Kaminski, N., and Zhao, H. (2019a). Regularized latent class model for joint analysis of high-dimensional longitudinal biomarkers and a time-to-event outcome. *Biometrics*, 75(1):69–77.

Sun, Q., Zhou, W.-X., and Fan, J. (2019b). Adaptive huber regression. *Journal of the American Statistical Association*, pages 1–24.

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., Diana, E., Lehmann, G., Toren, D., Wang, J., et al. (2018). Human ageing genomic resources: new and updated databases. *Nucleic Acids Research*, 46(D1):D1083–D1090.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266.

Tsai, P.-C. and Bell, J. T. (2015). Power and sample size estimation for epigenome-wide association scans to detect differential dna methylation. *International Journal of Epidemiology*, 44(4):1429–1441.

Tseng, G., Ghosh, D., and Zhou, X. J. (2015). *Integrating omics data*. Cambridge University Press.

van Iterson, M., Hoen, P. '., Pedotti, P., Hooiveld, G., den Dunnen, J., van Ommen, G., Boer, J., and Menezes, R. (2009). Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*, 10(1):439.

Vermunt, J. K. and Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, 41(3-4):531–537.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, volume 1, pages 577–584.

Wang, L., Zheng, C., Zhou, W., and Zhou, W.-X. (2020a). A new principle for tuning-free huber regression. *to appear, Statistic Sinica*.

Wang, M., Yao, T., and Allen, G. I. (2020b). Supervised convex clustering. *arXiv preprint arXiv:2005.12198*.

Wang, S. and Liao, L. (2001). Decomposition method with a variable parameter for a class of monotone variational inequality problems. *Journal of optimization theory and applications*, 109(2):415–429.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2013). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

Wang, Z., Gerstein, M., and Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Wu, H., Wang, C., and Wu, Z. (2015). PROPER: Comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics*, 31(2):233–241.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., and Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, 71(3):645–653.

Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L.-J. (2013). Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association*, 108(502):527–539.

Zhao, Q., Shi, X., Xie, Y., Huang, J., Shia, B., and Ma, S. (2015). Combining multidimensional genomic measurements for predicting cancer prognosis: observations from tcga. *Briefings in bioinformatics*, 16(2):291–303.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhu, R., Zhao, Y.-Q., Chen, G., Ma, S., and Zhao, H. (2017). Greedy outcome weighted tree learning of optimal personalized treatment rules. *Biometrics*, 73(2):391–400.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.