THE UNIVERSITY OF CHICAGO

STATISTICAL METHODS FOR GENOMICS DATA WITH CLUSTERED STRUCTURES AND MISSING VALUES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
JIEBIAO WANG

CHICAGO, ILLINOIS
AUGUST 2017

To My Wife, Xu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I especially thank my advisors, Dr. Lin Chen and Dr. Robert Gibbons, for their selfless mentorship during my study at the University of Chicago. Dr. Chen led me to the colorful world of genetics, taught me how to become a researcher, and provided essential and invaluable guidance that made this dissertation possible. Dr. Gibbons taught me numerous applications and the philosophy of mixed-effects models, which is one of the main themes of these proposed methods in this dissertation. I thank Dr. Donald Hedeker for spending time and energy to answer many questions from me. It was fun to sit in his classroom. I appreciate the help and contribution from Dr. Dan Nicolae and Dr. Pei Wang by serving on my committee. I also appreciate the help and suggestions from Dr. Ronald Thisted. I thank Kevin Gleason for helping proofread this dissertation. In addition, I would like to thank all members of the Department of Public Health Sciences for their sincere help.

Finally, I would like to thank my wife Xu Qin for her consistent support and companionship throughout my study in statistics/biostatistics. I am grateful for the unconditional love from my family and friends.

ABSTRACT

In the era of "big data," with the rapidly evolving high-throughput technology in genomics, massive amounts of genomic data have been generated to measure a variety of genomic features. Each of these data types has unique characteristics while sharing some commonalities. Some arising issues in these data types pose new challenges to traditional statistical methods in data analyses.

In this dissertation, I developed tailored statistical methods and extended these methods to more general frameworks for related statistical problems. The methods developed in this work was motivated by three related but distinct areas of genomics research: multi-tissue gene expression and expression-quantitative-trait-loci (eQTL) studies, quantitative proteomics studies, and gene-environment interaction in genetic association studies. The proposed statistical methods accounted for the unique data structures in each data type. In particular, some samples or measurements were naturally or experimentally clustered and may have ignorable or non-ignorable missing values. It has been shown that failing to account for these data characteristics may result in unfaithful conclusions or biased/inefficient estimation. Furthermore, with the goal of detecting individually weak but collectively strong effects of interest, I proposed multivariate analysis methods that jointly analyze multiple functionally related genomic features, as complementary approaches to standard univariate analyses.

All of the methods developed here are computationally efficient and can be scaled up for high-dimensional data analysis. I developed R packages for each method. I conducted extensive simulation studies to examine the performance of the proposed methods and compared each with existing relevant approaches. I also applied the proposed statistical methods to each of the motivating data sets and obtained new biological results. In the end of each of the three methods chapters, I discussed general applicability of these methods and potential future directions.

CHAPTER 1

INTRODUCTION

1.1 Overview

Rapidly evolving high-throughput technology has generated massive amounts of omics data, measuring a variety of genomic features in the genome. These big omics data allow researchers to examine biological information at the cellular level, and properly analyzing these data will ultimately improve our understanding of biology and further develop precision medicine (Collins and Varmus, 2015).

With comprehensive measurements of genomic features, it is natural for certain features/samples to be correlated and for certain data to have specific structures. In other cases, when samples are processed in batches in the data generating process, measurements can also be experimentally clustered. These complex and clustered data structure created statistical challenges for traditional methods, in particular, when analyzing high-dimensional "big data." On the other hand, properly handling these data characteristics and clustering structures could also help to borrow strength across functionally- and biologically-related measures, and enhance the efficiency of data analyses. A major focus of this dissertation is to develop tailored statistical methods and computational tools for three related but distinct genomic research areas, accounting for each of the unique data characteristics and feature structures.

It is often the case that not all of the data are collectible or are collected but fail quality control, resulting in large amounts of missing data. Ignoring these missing data may result in loss of efficiency, or may lead to biased estimation and unfaithful inference if the missing-data mechanism is non-ignorable (Little and Rubin, 2002). Two chapters in this dissertation develop methods to handle these unique missing data problems.

Another major theme in this dissertation is to develop multivariate statistical methods.

In standard analysis of high-throughput genomic data, the common practice is to examine each genomic feature one at a time, using univariate analysis approaches. However, it is well known that due to the high-dimensionality of genomic data, those univariate approaches can only detect the effects of interest when the individual effects survive stringent multiple testing adjustments given the current sample size. One may use multivariate analysis methods as a complementary approach, to examine those individually small to moderate but aggregately strong effects. Moreover, when multiple genomic features are grouped into functional sets as analysis units, the multiple testing problem is alleviated, the results are directly interpretable based on the functionality of the sets, and one may also borrow strength across multiple related features to improve the power. In this dissertation, I propose methods to jointly analyze multiple phosphopeptides from a phosphoprotein, multiple phosphoproteins from a biological pathway or protein set in proteomics studies, and gene-based gene-environment interaction (GxE) tests across multiple single nucleotide polymorphisms (SNPs) from a gene or SNP set in a meta-analysis of genetic association studies.

When jointly analyzing multiple outcomes and when the dimensionality of outcome or predictor is high, I propose new statistical methods and models to regularize the likelihood-based approaches and circumvent estimation issues. Computation efficiency is emphasized throughout since the development of powerful and efficient computational tools will propagate their adoption and use.

The overarching goal of this dissertation is to develop statistical methods and computational tools for analyzing big omics data with complex data structures and missing values. The proposed methods are motivated by three related but distinctive genomic research fields, each with a set of unique yet generalizable statistical challenges. I develop tailored statistical methods for each motivating research field. In this chapter, I describe the background, challenges, and existing literature of these three fields, as well as some of the unique motivating data sets.

1.2 Background and motivations of the research questions

1.2.1 A multi-tissue gene expression and eQTL study – the GTEx project

The expression of genes in a particular cell determines what the cell can do in a biological system. It has been shown that many gene expression levels are regulated by genetic factors. Unlike genotypes, the expression of messenger RNA (mRNA) levels of genes can vary substantially across different cell types (i.e., in different tissue types). Many of the existing gene expression and eQTL studies are based on whole blood samples collected on cohorts of interests. It is known that whole blood samples are heterogeneous and contain a mixture of different cell types. The gene expression measured from whole blood samples can be poorly correlated with the gene expression levels measured from the target tissue types. As such, eQTL studies conducted on whole blood samples may not inform the eQTLs in the other tissue types. It is important to understand how genotype information regulates the transcriptome in various human tissue types.

The GTEx project

In order to synthesize new knowledge about gene expression across human tissues, the GTEx (Genotype-Tissue Expression) program (Lonsdale et al., 2013) has generated rich transcriptome data in a wide variety of human tissue types as well as whole genome sequencing data from a large number of donors. In 2015, GTEx released pilot data including transcriptome measurements in 44 human tissue types and sequencing data on 175 donors (The GTEx Consortium, 2015). To date, the recent release of GTEx data is version 6p in 53 tissues types from 544 donors. Among them, 7,051 tissues from 449 donors have both measured genotype and multi-tissue expression.

One major challenge in the GTEx tissue sample collection, which is also a challenge in conducting similar types of studies in other cohorts, is tissue accessibility. Certain tissue

types are inaccessible and are non-regenerative or difficult to obtain. It would be valuable if the expression data in the uncollected tissues can be accurately imputed by harnessing observed data and the rich resource in GTEx. With multi-tissue imputation, one may reanalyze and recapitalize on existing single-tissue expression data, or design future multi-tissue expression studies with limited resources.

In this dissertation, I used data from the GTEx pilot study. There are nine tissue types prioritized for RNA sequencing (The GTEx Consortium, 2015). The nine tissue types are whole blood, muscle (skeletal), lung, tibial artery, thyroid, skin (Sun exposed lower leg), adipose (subcutaneous), tibial nerve, and heart (left ventricle). The nine tissue types in the GTEx pilot data have been collected on more than 80 donors, with the remainder yielding tissue-specific sample sizes less than 40. The average uncollected tissue rate is 27%.

Literature review

In order to impute gene expression levels in the genome, there were dozens of existing methods (Liew et al., 2011). However, most of them were developed to impute missing gene expression levels based on observed gene expression levels in a single tissue type. They may not be directly applicable to the GTEx multi-tissue imputation problem to impute the entire transcriptome of an uncollected tissue. Here I discuss a few widely used imputation methods.

The k-Nearest Neighbors (k-NN) method (Troyanskaya et al., 2001) is a commonly used method in gene expression imputation. It finds the k nearest neighbors with distance defined by a Euclidean metric for each gene with missing values, and imputes the missing values by averaging the values of the observed neighbors. It is fast and simple, but it tends to impute missing values towards the mean and is less flexible in incorporating covariates. It is not suitable for multi-tissue imputation because the intention of multi-tissue imputation is to consider tissue-specific eQTL effects (as covariates) and characterize tissue-specific expression variations, while with k-NN the missing expression levels in the uncollected tissues for a gene

in an individual tend to be imputed to the same mean.

The missForest (Stekhoven and Bühlmann, 2012) algorithm is a non-parametric imputation method based on random forests. A random forest constructs a multitude of regression trees, each of which uses a subset of samples and a subset of predictors. A random forest then imputes the missing values by averaging over many trees. The missForest algorithm treats each gene with missing values as the response and the (observed and imputed) values of all the other genes as predictors. It imputes each gene in the genome one by one and repeats the procedure until the imputed values converge. When imputing the entire genome with thousands of genes, the computation is intensive. Despite its robustness, missForest also does not account for covariates and tissue-specific expression variations.

The multivariate imputation by chained equations (MICE; Van Buuren and Groothuis-Oudshoorn, 2011) algorithm utilizes a Gibbs sampler to successively draw samples from the distribution of one variable with missing values conditional on all the other variables. It is widely used for general imputation. The missing values are imputed with their posterior expectation. When applied to multi-tissue imputation, one would pool all tissue types and use observed expression levels to impute the missing expression levels.

In summary, most existing gene expression imputation methods impute the unobserved expression levels by correlated observed gene expression levels within a single tissue (Liew et al., 2011). These single-tissue imputation methods largely rely on gene-gene correlations, which can be unstable and vary by phenotypic conditions. Existing single-tissue imputation approaches do not fully utilize the available information in the GTEx data. With multi-tissue expression data, as characterized by the GTEx data, there are tissue-tissue correlations, major variations by sample/tissue characteristics, and robust eQTLs that can be used in the imputation.

The proposed method

In Chapter 2, I propose a method to impute each gene's expression levels in the uncollected or inaccessible tissues by harnessing the rich information in GTEx, including the clustered expression levels of the genes of interest in other observed tissue types (tissue-tissue correlation), the measured and unmeasured sample/tissue characteristics, and the eQTL information (Brem et al., 2002). Inspired by a mixed-effects regress tree (Sela and Simonoff, 2012), we propose a mixed-effects random forest to handle a large number of predictors, model non-linear effects of the selected predictors and interactions among predictors, and account for the clustered data structure (the expression levels of the same gene from multiple tissue types are naturally clustered and are correlated). Note that related methods have recently been proposed by Hajjem et al. (2014) and Stephan et al. (2015), but our method is more flexible and tailored for multi-tissue gene expression imputation.

With simulation studies and real data analyses, I illustrate that one may use the imputed GTEx data to enhance the power of different types of subsequent analyses. The imputed data can serve as secondary and supplemental data to the primary and observed data, to strengthen the evidence of biological results and conclusions.

1.2.2 Labeling-based quantitative proteomics studies

Description of labeling-based MS experiments and data characteristics

To date, MS (mass spectrometry)- based platforms still serve as the workhorses in quantitative proteomics research. In the traditional shotgun MS experiments, samples were processed one at a time. Processing one sample involves extensive fractionation and weeks of MS time. The time and cost required for such experiments greatly limit the scale of most quantitative proteomics studies.

To improve the efficiency of MS-based protein quantification, the iTRAQ (isobaric Tag

for Relative and Absolute Quantitation) and TMT (Tandem Mass Tag) technique were introduced about a decade ago (Ross et al., 2004). This labeling based technique enables researchers to compare up to 4 or 8 (for iTRAQ) and 11 (for TMT) different samples in one MS-based experiment with multiple channels. As an example, in an iTRAQ 4-plex MS based experiment, samples are first grouped into batches of 4 samples, and each batch is processed by one iTRAQ experiment. In each experiment, intact proteins are first enzymatically digested into smaller segments of amino acid sequences, i.e., peptides. Peptides from different samples in one batch are then labeled with different isotope-coded covalent tags before they are mixed and introduced into the MS instruments. These peptides from all samples in the same batch will be quantified together and then will be mapped back to proteins to yield protein/peptide quantification. In this way, multiple samples can be processed together, greatly reducing overall quantification time and cost. The enhancement of throughput with iTRAQ has greatly advanced proteomics research. However, since samples are processed in batches by different MS experiments and there is large experimental variation among different experiments, data generated by iTRAQ experiments suffer from severe batch-effects.

Another unique challenge in the labeling-based proteomics data analysis is the substantial amount of batch-level non-ignorable missing data. It is well known that in the general MS experiments, the lower the abundance of a given peptide, the more likely the peptide is missing in the output data (Chen et al., 2014; Wang et al., 2006). With iTRAQ-MS experiments, since all samples in a batch are processed together, a given peptide is either detected and quantified or missing simultaneously in all samples from the same batch. The missing-data probability of the peptide largely depends on the combined abundances of the peptide from all of the batch samples in the experiment (the batch-level abundance). We term the missing-data mechanism as batch-level or cluster-level non-ignorable missingness (Chen et al., 2014, 2017b). In order to properly analyze iTRAQ-based proteomics data, it is highly desirable to develop methods accounting for severe batch effects and batch-level non-

ignorable missingness. Table 1.1 shows an illustration of the iTRAQ data for one peptide and its missing-data pattern. The samples in the same 4-plex experiment tend to be missing or observed together, and the missingness of an experiment relates to the peptide abundance level.

Table 1.1: **An illustration of iTRAQ data for the k-th peptide.** Let $\mathbf{y}_{N\times 4,k}$ represent the abundance data for the k-th peptide based on 4-plex iTRAQ experiments. A total of $3N$ tumor samples are processed. The samples in the same 4-plex experiment tend to be missing or observed together, and the missingness relates to the peptide abundance level.

Batch	Ref sample	Sample 1	Sample 2	Sample 3
1	y_{1Rk}	y_{11k}	y_{12k}	y_{13k}
2	?	?	?	?
$\vdots$		$\cdots$		
i	y_{iRk}	y_{i1k}	y_{i2k}	y_{i3k}
$\vdots$		$\cdots$		
N	y_{NRk}	y_{N1k}	y_{N2k}	y_{N3k}

In order to improve the ability to diagnose, treat and prevent cancer, the National Cancer Institute launched the Clinical Proteomic Tumor Analysis Consortium (CPTAC, http://proteomics.cancer.gov) to systematically identify proteins that are derived from alterations in cancer genomes (Paulovich et al., 2010; Ellis et al., 2013b). The CPTAC program has recently conducted global proteome and phosphoproteome profiling of a subset of breast, colon and ovarian cancer samples that have been extensively characterized in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) (The Cancer Genome Atlas Network, 2012). This is the first attempt to characterize protein activities in cancer samples using sophisticated proteomics experiments on a large scale. In 2014, CPTAC released the largest-ever breast cancer proteome data set. A total of 108 breast cancer tumor samples were analyzed with 36 iTRAQ 4-plex experiments, each processing three tumor samples and one common reference sample, with the goal of identifying proteins related to breast cancer clinical variables and outcomes.

Literature review

The standard practice to analyze labeling-based proteomics data

As discussed earlier, although the iTRAQ-based batch processing greatly reduces the cost and improves the efficiency of data generation, the consequent batch-effects are substantial due to the dynamic nature of MS instrument. To alleviate this problem, a standard practice is to include a common reference sample in each batch. For example, in the 4-plex iTRAQ experiments of the CPTAC breast cancer study, each batch consisted of three breast tumor samples and a common reference sample created by combining 40 tumor samples. The corresponding data analysis is usually performed based on the relative abundances of proteins/peptides in the target samples relative to the common reference sample in the same batch (Luo et al., 2009; Mertins et al., 2012; Karp et al., 2010). This strategy helps to account for the variation across different iTRAQ multiplex experiments to a certain extent. However, due to the complicated process of protein/peptide identification and quantification in the MS instruments, there is a large variation among the measurements of the common reference sample across different experiments/batches, and the target samples and the reference sample could be subjected to different variances (Karp et al., 2010). The relative abundance measures cannot fully capture these data features.

Using mixed-effects models to account for the batch-processed data with non-ignorable missingness

The batch-processed data created clustered data structure. Mixed-effects models are natural ways to account for the correlations among multiple samples from the same batch. There are several existing mixed-effects models for analyzing clustered data with non-ignorable missing responses (Little, 2008; Ibrahim and Molenberghs, 2009).

Based on the factorization of the joint distribution of the outcome ($\mathbf{y}_i$), random effects ($\mathbf{b}_i$) and missing-data indicator ($\mathbf{r}_i$), there are only three sensible classes of models:

selection models (Diggle and Kenward, 1994),

$$f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i) = f(\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i) f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{b}_i),$$

pattern-mixture models (Little, 1995; Hedeker and Gibbons, 1997),

$$f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i) = f(\mathbf{r}_i) f(\mathbf{b}_i|\mathbf{r}_i) f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{r}_i),$$

and mixed-effects hybrid models (Yuan and Little, 2009),

$$f(\mathbf{y}_i, \mathbf{b}_i, \mathbf{r}_i) = f(\mathbf{b}_i) f(\mathbf{r}_i|\mathbf{b}_i) f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{r}_i).$$

The three general classes of models are not directly applicable to data with cluster-level missingness because completely missing clusters would be simply discarded. If the completely missing clusters are discarded, it is essentially applying mixed-effects models by treating the missingness as ignorable. The consequence is that only responses with higher abundance values are used in the modeling, and the resulting estimation and inference would be biased (Saha and Jones, 2005).

The proposed methods and data analysis

Motivated by the cluster-level non-ignorable missing data in the CPTAC breast cancer phosphoproteomic data, in our previous work (Chen et al., 2017b), we proposed to incorporate the cluster-level non-ignorable missing-data mechanism into the mixed-effects models. We developed a univariate mixed-effects model for cluster-level non-ignorable missing data in analyzing one phosphopeptide at a time. Instead of ignoring the missing batches of samples, we made use of missing batches to account for the non-ignorable missing-data mechanism.

In Chapter 3 of this dissertation, motivated by the phosphoproteomics data from the

CPTAC breast cancer study, I develop methods to analyze data with clustered data structure and non-ignorable missingness. They can simultaneously analyze multiple phosphopeptides in a phosphoprotein or multiple phosphoproteins in a functional pathway. The application of the proposed methods to the CPTAC breast cancer data identifies phosphoproteins and pathways associated with breast cancer and will ultimately provide important insights into breast cancer etiology and help to identify protein biomarkers.

1.2.3 Gene-based gene-environment interaction tests in meta-analyses

In genetic association studies, there has been a tremendous success in identification of genetic variants associated with risks of complex diseases (Welter et al., 2014). However, a large proportion of the estimated heritability for most complex diseases remains unexplained (Manolio et al., 2009). Gene-environment interactions (GxEs) were believed to explain part of the missing heritability (Thomas, 2010).

It is cost-efficient to reanalyze data from existing genome-wide association or next generation sequencing studies for the purpose of detecting GxE effects. However, it has been shown that the detection of an interaction effect requires a sample size at least four times larger than that required for the detection of a main effect of comparable effect sizes (Smith and Day, 1984). In other words, reanalyzing existing association data for detecting GxE would be underpowered. It is highly desirable to develop new and more powerful methods for GxE analysis.

In fact, power or sample size is a major challenge not only for GxE analyses but also for detecting genetic effects with moderate to small effect sizes or effects of rare variants (McCarthy et al., 2008). In order to overcome this problem in individual studies, large consortia are formed, and meta-analysis methods and approaches are widely used for combining results of different cohorts with the same complex diseases (Hunter et al., 1982). These analyses are often based on summary statistics from individual studies. By borrowing information across

studies, one may increase statistical strength and precision, and improve statistical power.

In addition to meta-analyses combining samples from different studies based on different cohorts, another way to borrow strength is to aggregate the effects of multiple genetic variants in the same gene or genomic region — set-based analyses (Wu et al., 2011; Neale et al., 2011; Chen et al., 2012). These multivariate analyses approaches serve as complementary methods to genome-wide scans of individual genetic or GxE effects, aiming to detect those individually weak but aggregately strong effects of interests.

Motivated by these questions and challenges in reanalyzing genetic association data for GxE effects, I propose meta-analyses approaches for detecting gene-based GxE effects. This work extends an idea that was previously proposed (Liu et al., 2016) for detecting GxE effects in one individual study. In GxE analyses with case-control studies, genetic and environmental (G-E) factors are often assumed to be independent in the general population (i.e., the controls), but when GxE is present, G-E would appear to be correlated. G-E associations are often used to filter out the genetic variants that do not show promise with GxE effects (Murcray et al., 2009; Dai et al., 2012).

In Chapter 4, I propose meta-analyses filtering methods, and incorporate meta-filtering into the meta-analysis approach to analyzing each SNP set. I discuss the models to aggregate GxE effects from individual variants to the set in meta-analyses, and propose a new GxE meta-analysis test that is robust to study heterogeneity. I apply this meta-analysis approach to two existing breast cancer GWAS data sets (Ahsan et al., 2014; Huo et al., 2016), both of which collected samples from multiple sites/subgroups and have many early-onset breast cancer (EOBC) cases, to identify the genes showing potential gene by age interactions.

1.3 Summary

In this dissertation, motivated by the state-of-the-art technology and research questions in big omics data analysis, I propose tailored statistical methods accounting for the clustered

data structure, ignorable and non-ignorable missing values, and correlations among multiple related genomic features in three related but distinct genomic research areas. The questions of interest and the models and methods developed also share many similarities. For example, when analyzing high-dimensional omics data, penalized likelihood-based approaches are often proposed and used. Our models and methods are always developed with the goal of potentially scaling up to analyses with even higher dimensionality or much larger sample sizes. Chapters 3 and 4 propose multivariate analysis methods, aiming to detect individually weak but collectively strong effects of interests. Chapters 2 and 3 handle ignorable and non-ignorable missing data problems with clustered data structures. All the three main chapters deal with clustered data sets. The models and methods proposed here are motivated by real data and challenging questions, which can be generalized to other similar problems with similar complex data structures.

CHAPTER 2

A MIXED-EFFECTS RANDOM FOREST FOR IMPUTING GENE EXPRESSION IN UNCOLLECTED TISSUES WITHIN AND BEYOND GTEX

2.1 Introduction

Studies of gene expression in peripheral whole blood, skin, liver, and other tissues have revealed that gene expression and its regulation depend on cell context (Schadt et al., 2012). The expression of a given gene can vary substantially across tissue types and the genetic variants that regulate gene expression — eQTLs (Brem et al., 2002; Rockman and Kruglyak, 2006) — can have effects on transcription that also vary across tissue types (Flutre et al., 2013; Torres et al., 2014; Li et al., 2013; Raj et al., 2014). A careful examination of gene expression across human tissues and within target tissues would not only help to answer a wide range of scientific questions related to transcriptional variation, but also inform other fundamental aspects of biology, and prioritize therapeutic gene targets in the development of precision medicine (Collins and Varmus, 2015). The challenge is that many tissues are not regenerative and are difficult to collect (hereinafter referred to as "inaccessible" tissues). To date, most large-scale gene expression studies have been conducted using RNA extracted from peripheral blood cells, or their derivatives such as lymphoblastoid cell lines. The blood samples are generally heterogeneous and contain a mixture of different cell types. The expression in the blood cells may not directly inform the expression and its regulatory mechanisms in other target cell types from other tissues.

The National Institutes of Health Common Fund's GTEx program has generated rich transcriptome data in a wide variety of human tissue types as well as genome sequencing data from a large number of post-mortem donors, thus allowing researchers to generate knowledge about gene expression across human tissues and also characterize the regulatory role of

genetic variation from both cross-tissue and tissue-specific perspectives (Lonsdale et al., 2013; Keen and Moore, 2015; The GTEx Consortium, 2015). In May 2015, GTEx released pilot data, including transcriptome measurements in 44 reference human tissue types and sequencing data on 175 donors (The GTEx Consortium, 2015). The GTEx project provides a unique opportunity to systematically evaluate the relationships among transcriptomes of different tissues and inform the design of future studies of multi-tissue gene expression.

One major challenge of conducting similar types of analyses in studies beyond GTEx is tissue accessibility. Despite the importance of obtaining specific target tissues from additional cohorts of interest, multi-tissue expression data might be difficult to collect in many studies. For example, the collection of inaccessible tissues is neither possible nor ethical from living study participants; certain samples in some existing expression studies may not be available for additional data collection; or certain samples may have only limited tissue biopsies available, etc. In these cases, it would be desirable if the expression data in the uncollected or inaccessible tissues could be accurately imputed by harnessing available information on the target samples and the rich resource in GTEx. With multi-tissue imputation, we may reanalyze and leverage existing single-tissue expression data, or design future multi-tissue expression studies with limited resources. In comparison with single-tissue expression data, multi-tissue expression data provide a more comprehensive and systematic view of the underlying biological mechanisms. Moreover, the expression levels of a gene in functionally related tissues often show coordinated expression patterns, reflecting shared developmental and genetic factors. By jointly analyzing expression data from multiple tissue types, one may enhance the power to identify biomarkers for complex diseases/traits and facilitate the development of precision medicine.

In this chapter, we propose to impute the expression data in uncollected or inaccessible tissues by harnessing eQTLs and tissue-tissue expression level correlations. We propose a multi-tissue imputation algorithm and its extension based on a mixed-effects model (Laird

and Ware, 1982) that treats the expression measures from multiple tissues as the outcome, and considers as predictors the eQTL genotypes, known covariates, and the estimated tissue-specific top principal components (PCs) of expression data. By borrowing information across genes and across related tissues, the proposed method not only captures the genetic factors influencing gene expression in tissues but also major developmental and environmental factors. We conduct simulation studies to show the superior imputation performance of the proposed methods over existing imputation approaches (Celton et al., 2010; Liew et al., 2011; Donner et al., 2012; Stekhoven and Bühlmann, 2012; Troyanskaya et al., 2001; Brock et al., 2008; Liao et al., 2014) in multi-tissue expression imputation, as well as the utility of the imputed multi-tissue expression data. Moreover, based on cross-validation (CV) analyses of GTEx pilot data, we demonstrate the feasibility of imputing expression in uncollected GTEx tissues and using GTEx data as a reference to impute expression in inaccessible tissues from samples beyond GTEx.

2.2 Methods

2.2.1 A mixed-effects model for multi-tissue imputation

To impute the expression levels of a gene in uncollected or inaccessible tissues, structured information is uniquely available in the GTEx data, including the expression levels of the gene of interest in observed tissues, the *cis*- (local) and *trans*- (distal) eQTLs, and sample characteristics (gender, age, etc.) shared across genes. In GTEx data, we measure the expression levels in multiple tissues from each individual, and the multi-tissue expression measures are naturally clustered within individuals.

A natural model to account for these features is a mixed-effects model with expression levels from multiple tissues of a gene as the response, eQTLs and other cross-tissue or tissue-specific covariates as predictors, and random effects (here a random intercept) for each

individual:

$$y_{it} = \mu_t + \boldsymbol{\beta}_t^T \mathbf{x}_i + \boldsymbol{\alpha}_t^T \mathbf{c}_i + \gamma_i + \epsilon_{it}; \tag{2.1}$$

where y_{it} is the expression level of a gene in tissue type t ($t = 1, \ldots, T$) of individual i ($i = 1, \ldots, N$); μ_t is the tissue-specific mean expression; $\mathbf{x}_i$ is the genotype vector of length K in individual i for K selected eQTLs ($\mathbf{x}_i$ is the same across tissues); $\boldsymbol{\beta}_t$ is a vector of length K representing the tissue-specific eQTL effects in tissue type t; γ_i is the random intercept for individual i with $\gamma_i \sim N(0, D)$; $\mathbf{c}_i$ is the vector of covariates for individual i with $\boldsymbol{\alpha}_t$ as the corresponding coefficients in tissue type t; and ϵ_{it} is the error term.

In model (2.1), the effect of each eQTL may vary across tissues. Some eQTLs consistently regulate the expression of a gene across multiple tissues and are considered cross-tissue eQTLs, while others only show eQTL effects in certain tissue types and are considered tissue-specific (Flutre et al., 2013; Torres et al., 2014; Li et al., 2013; Gamazon et al., 2015). Even for cross-tissue eQTLs, the effect sizes $\boldsymbol{\beta}_t$ can vary by tissue type (similar to an interaction effect of eQTL and tissue type).

To estimate the tissue-specific eQTL effects, we need to estimate a total of $T \times K$ parameters in model (2.1). To reduce the number of parameters, we further employ an adaptive weighting scheme (Tukey, 1949; Chatterjee et al., 2006): we regress the gene expression in tissue type t on the k-th eQTL and let the marginal eQTL effect be the adaptive weight, w_{kt}. This strategy implicitly assumes that the tissue-specific eQTL effects in different tissues in model (2.1) are proportional to the marginal tissue-specific eQTL effects. The pre-specified adaptive weights in the following model allow us to account for tissue-specific eQTL effects with only one parameter θ_k for the k-th eQTL, thereby reducing the total number of parameters for eQTL effects from $T \times K$ to K:

$$y_{it} = \mu_t + \sum_k \theta_k \cdot (w_{kt} x_{ki}) + \boldsymbol{\alpha}_t^T \mathbf{c}_i + \gamma_i + \epsilon_{it}. \tag{2.2}$$

2.2.2 A mixed-effects random forest

To obtain the predicted values of y_{it} with weighted genotypes and other covariates as predictors, we propose a Mixed-model-based Random Forest approach (MixRF). Random forest is an ensemble learning method that operates by constructing a multitude of regression trees (Breiman et al., 1984), each of which considers a subset of model predictors and a subset of samples. To learn a regression tree for a continuous outcome based on some predictors, one can employ a recursive binary partitioning algorithm (Friedman, 1977). At each partitioning, the algorithm splits the response variable based on a binary (or dichotomized) predictor in the current node such that the reduction in the sum of squares for values in the node is maximized. The split continues until the tree is too complex or the number of observations in the current node is too small. A regression tree is a non-linear model that predicts the value of a target variable. Predictions based on a single regression tree can be unstable. By aggregating over many regression trees, a random forest approach intrinsically constitutes a multiple imputation scheme (Stekhoven and Bühlmann, 2012) and provides a more robust prediction that minimizes the overall CV prediction (i.e., imputation) errors (Breiman et al., 1984; Friedman, 1977; Sela and Simonoff, 2012).

Most existing random forest approaches (Breiman, 2001; Liaw and Wiener, 2002) ignore the clustered data structure. With the proposed MixRF algorithm, we obtain the predictive values using the following steps: for each gene, we obtain the externally defined eQTLs or select the eQTLs based on the current data, and assign the adaptive weight to each eQTL genotype in each tissue type. We set the initial values of $\gamma_i^{(0)} = 0$. Given the estimated random effects at the j-th iteration, we build a random forest with $u_{it}^{(j)} = y_{it} - \hat{\gamma}_i^{(j)}$ as the response and with weighted genotypes in each tissue type and other covariates as predictors, $u_{it}^{(j)} = f\left(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i\right) + \delta_{it}$ where δ_{it} is the error term. We obtain the predicted value $\hat{u}_{it}^{(j)}$. In re-estimating the random effects, we let $\omega_{it}^{(j)} = y_{it} - \hat{u}_{it}^{(j)}$ and fit a linear random-effect model with $\omega_{it}^{(j)} = \gamma_i^{(j)} + \epsilon_{it}$ to obtain the estimated random effect $\hat{\gamma}_i^{(j)}$. The

proposed MixRF algorithm iterates through estimating the random effect γ_i in the linear mixed-effects model (Laird and Ware, 1982) and constructing a random forest (Breiman, 2001) for the new response variable u_{it} until the change in the likelihood at successive iterations is small (< 0.001). The proposed MixRF often converges quickly in a few iterations and the prediction is not sensitive to the specified initial values. We summarize the MixRF algorithm in Algorithm 1.

Algorithm 1 MixRF: A Mixed-effects Random Forest for multi-tissue expression imputation

1. For each gene, use externally defined eQTLs or select eQTLs based on the currently **observed** data. Obtain the adaptive weights (w_{kt}) for each eQTL in each tissue type.
2. Initialize the random effects estimate in model (2.2), $\hat{\gamma}_i^{(0)} = 0$.
3. At the j-th iteration, let $u_{it}^{(j)} = y_{it} - \hat{\gamma}_i^{(j-1)}$. Build a random forest with $u_{it}^{(j)}$as the response and weighted genotypes in each tissue and other covariates ($\mathbf{c}_i$) as predictors,

$$u_{it}^{(j)} = f\left(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i\right) + \delta_{it}.$$

   Obtain the estimated/predicted value $\hat{u}_{it}^{(j)}$.
4. Let $\omega_{it}^{(j)} = y_{it} - \hat{u}_{it}^{(j)}$. Fit a linear random-effects-only model with $\omega_{it}^{(j)}$ as the response,

$$\omega_{it}^{(j)} = \gamma_i^{(j)} + \epsilon_{it}.$$

   Obtain the estimated random effect $\hat{\gamma}_i^{(j)}$.
5. Iterate through Steps 3-4 until the change in the likelihood is small.

Our random-forest-based prediction model is a non-linear function of the predictors in model (2.2): $\hat{y}_{it} = \hat{f}\left(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i\right) + \hat{\gamma}_i$. It can automatically capture the potential non-linear effects of the predictors and the interaction effects among the predictors on the outcome. In the multi-tissue expression GTEx data, we observed that the eQTL effects on gene expression levels could be additive, dominant, or recessive (with 58%, 38% or 4% of the eQTL-expression pairs better fitting an additive, a dominant or a recessive eQTL model, respectively).

In addition, we also observed eQTL-eQTL interaction effects and gender-specific eQTLs

(gender-eQTL interactions) (Dimas et al., 2012) on many genes. The proposed random-forest-based prediction model would be helpful in capturing these effects and would improve the imputation performance. Moreover, since the random-forest-based prediction model allows higher-order interactions among the predictors, it is more flexible than lasso-type penalized regression-based predictions and would not induce biased prediction (Tibshirani, 1994).

2.2.3 An extension to capture the effects of major developmental and environmental factors in the imputation

We further propose an extension — MixRF+iPC. Specifically, we propose to: (1) impute selected gene expression levels with multiple eQTLs ($\sim$1000 genes with $\geq$ 3 eQTLs) using MixRF with adaptively weighted genotypes and other known covariates as predictors; (2) construct tissue-specific PCs by performing singular value decomposition (SVD) on the combined observed and imputed expression data on the selected genes within each tissue type, keeping the top 5 PCs for each tissue type; and (3) incorporate the tissue-specific PCs with adaptively weighted genotypes and other known covariates as predictors in MixRF+iPC for imputing or re-imputing gene expression levels in the genome.

Most of the differences in gene expression among tissues and much of the correlations in gene expression across tissues are driven by the sets of genes that are not expressed in many of the same tissues but are expressed in other tissues. Their expression levels are so correlated across tissues not because of shared genetic architecture, but because they are completely and invariantly not expressed in so many of the same tissues. Human developmental profiles are invariantly shared within our species and major developmental information is important information that augments the genetic information. By borrowing information across genes, the top PCs within each tissue type partially capture major developmental factors, as well as the tissue-specific effects of major environmental factors. By incorporating the top PCs

from each tissue type as predictors, the extension MixRF+iPC improves the multi-tissue imputation for genes with no eQTLs or low heritability. We summarize the MixRF+iPC algorithm in Algorithm 2.

Algorithm 2 MixRF+iPC: An extension of MixRF incorporating PCs of expression data

0. Select eQTLs.
1. For each tissue type, construct the top PCs of combined observed and imputed expression data on selected genes to capture unknown sample characteristics and tissue-specific major developmental patterns and environmental effects.

    i. Impute the selected gene expression levels (here we impute the expression levels of ∼1,000 genes with at least 3 eQTLs) using MixRF with adaptively weighted eQTL genotypes and other known covariates as predictors.

    ii. For each tissue type, perform SVD on the combined observed and imputed data on the selected genes and keep the top 5 PCs. (Note that the results based on the top 10 PCs are similar.)

2. Apply MixRF to each gene, with gene expression as the response, and adaptively weighted eQTL genotypes, other known covariates and the constructed tissue-specific PCs from Step 1 as predictors.

In addition to the predicted values of multi-tissue expression levels, MixRF and MixRF+iPC provide an imputation quality measure — the estimated imputation correlation ($\hat{r}_{imp}$). It is estimated based on a 10-fold CV of the currently observed data. One splits the data into 10 folds, each time using 9 folds as training data and the rest of as testing data. One then applies MixRF to the training data to impute the testing data and repeats until all the data have been imputed once. In the end, one calculates the correlation of the observed expression levels and the imputed expression levels for each gene.

Based on simulation studies, we suggest excluding the imputed expression levels for genes with estimated imputation correlations less than 0.3 in the subsequent analyses, although there is no universal cutoff value for post-imputation exclusion/filtering. The appropriate threshold for a specific analysis may differ. With parallel computing, imputing 10,000 genes in nine tissue types from about 150 individuals and obtaining the 10-fold CV-based imputation quality measures could be completed within 30 hours using a 40-node cluster (3.0 GHz

Intel Xeon E7 processor) with a 16.5 GB memory usage.

The overall computation time of MixRF and MixRF+iPC increases linearly with the number of genes, and the number of eQTLs/covariates. The computation complexity of random-forest-based approaches is also dependent on the total number of observed tissues, N_T, a summation of observed tissues for all individuals. The runtime of MixRF and MixRF+iPC scales with a complexity of $\mathcal{O}(N_T \log N_T)$ in the total number of tissues (Stephan et al., 2015). The computation is highly parallelizable.

2.2.4 Selection of eQTLs

To obtain the eQTLs for each gene, one may use the reported eQTL lists from other independent data. However, most of the published eQTLs are mapped in whole blood or lymphoblastoid cell lines and may not show eQTL effects in other tissues. In our CV analyses, we did not use the eQTLs reported in the GTEx project (Lonsdale et al., 2013; The GTEx Consortium, 2015). Those eQTLs were calculated based on all of the GTEx tissues, while in each round of our CV analyses, we treated a certain proportion of GTEx tissues as "uncollected," imputed the expression in those tissues, and evaluated the imputation performance. Using eQTLs that were calculated based on all tissues to impute the expression in the "uncollected" tissues would have overestimated the imputation performance.

We propose to select eQTLs for each gene based on the observed data (for example, the training data in the cross-validation analysis). The selection of eQTLs may affect the predictors used in the imputation and therefore the imputation performance. Nevertheless, the selection can be viewed as a pre-screening of predictors before imputation, and this step will not lead to biased imputation assessment yet will greatly reduce the computational burden. When using GTEx data as a reference to impute expression in the uncollected tissues from other studies, one may combine the GTEx data with data from non-GTEx samples to obtain the eQTLs used in the imputation.

In each round of CV of our data analyses, we calculated and selected eQTLs based on only the "observed" (i.e., training) data. Given the limited sample size in the GTEx pilot project, we selected only the cross-tissue *cis*- and *trans*-eQTLs, and ignored the tissue-specific ones due to low power to detect the latter.

Most of the *cis*-eQTLs are cross-tissue (Flutre et al., 2013; Torres et al., 2014; Li et al., 2013) and can potentially be replicated in different cell contexts or even across ethnicities (Pierce et al., 2014; Stranger et al., 2012). To obtain the cross-tissue *cis*-eQTLs, we calculated the tissue-specific *cis*-eQTL effects using MatrixEQTL (Shabalin, 2012), combined the Z-statistics from the nine tissue types using Stouffer's method (Stouffer et al., 1949), and selected the *cis*-eQTLs with Stouffer's p-values being less than 10^{-6}. For *trans*-eQTLs, we selected the *trans*-eQTLs with tissue-specific p-value ≤ 0.05 in at least 8 out of 9 tissues. These selected cross-tissue *trans*-eQTLs have Stouffer's p-values of less than 10^{-8}. The omission of tissue-specific *trans*-eQTLs in our analysis may hurt the imputation performance, but this can be improved with the later phase of GTEx data, in which the project will scale up donor collection to 900 and all the 44 tissue types will have reasonably large sample sizes.

2.2.5 GTEx data processing

Our GTEx data analyses focused on the expression data from 9 tissue types that have ≥ 80 collected samples. We restricted the analyses to the 150 samples with at least 4 observed tissues, such that in each fold of the CV data each individual had at least two observed tissues.

We applied standard data preprocessing and quality control procedures to both DNA and RNA-Seq data. We considered only the 10,919 genes that were expressed in all the 9 tissues, with tissue-specific log2 (mean expression level) significantly greater (based on one-sided t-test) than log2 (5 read counts). We normalized each gene expression in each tissue and removed the batch effects. For genotype data, we excluded the single nucleotide variants

(SNVs) with minor allele frequencies of less than 5% or with p-values of Hardy-Weinberg equilibrium test ≤ 0.001, and pruned the SNVs with a linkage disequilibrium (LD) threshold of 0.5 using PLINK (Purcell et al., 2007). After filtering and pruning, there were 282,295 variants being considered as potential eQTLs in the imputation analyses.

2.3 Simulations

2.3.1 Methods comparison on imputation performance

In order to evaluate the imputation performance of our proposed methods and other competing imputation methods, we simulated gene expression data for 150 individuals and 9 tissue types based on equation (2.2). We simulated 1,000 gene expression levels each with 0, 1, 2, 5, and 10 eQTLs. We examined the imputation performance of competing methods when the "heritability" (the percentage of expression variation explained by the eQTLs) ranged from 0 to 80%, a wide range as commonly observed in eQTL studies (Gamazon et al., 2015). We simulated the random intercept $\gamma_i \sim N\left(0, 1.26^2\right)$ and the error term $\epsilon_{it} \sim N\left(0, 3^2\right)$. Given the standard deviations of γ_i and ϵ_{it}, the intra-class correlation was 0.15. Additionally, we simulated two cross-tissue covariates $\mathbf{c}_i$s with various effects on the simulated gene expression levels. The input parameters for the simulations, including eQTL count, eQTL effect sizes, tissue-tissue expression level correlations, and covariate effect sizes, reflect what we have observed in the real data from the GTEx pilot project.

We randomly treated 30% of all tissues as "uncollected" and set their gene expression data as "missing." We applied eight imputation methods to the simulated data set to impute the missing gene expression data. These eight competing methods were k-NN (Troyanskaya et al., 2001), missForest (Stekhoven and Bühlmann, 2012), MICE (Van Buuren and Groothuis-Oudshoorn, 2011), linear regression (lm), linear mixed-effects model (lmer) (Laird and Ware, 1982), REEMtree (Sela and Simonoff, 2012), MixRF, and MixRF+iPC. The true eQTLs were

used as predictors in the five regression-based methods, lm, lmer, REEMtree, MixRF, and MixRF+iPC. The imputation performance was evaluated by the median of gene-level true imputation correlations of the 1,000 genes. Note that here the gene-level true imputation correlation was calculated as the Spearman's correlation of the true versus the imputed values of a given gene in a specific tissue type.

As shown in Figure 2.1A, our proposed methods MixRF and MixRF+iPC outperformed other imputation methods and MixRF+iPC showed an advantage over MixRF for imputing gene expression with zero eQTLs. The five regression-based methods incorporated eQTL effects and performed better than other methods. The imputation methods, k-NN, missForest, and MICE, were designed for single-tissue imputation — using selected gene expression levels to impute the rest of expression levels from the same tissues, and performed less competitively in the multi-tissue imputation.

In Figure 2.1B, we simulated another setting, in which each expression level was affected by two eQTLs and an interaction effect between them (a gene-gene interaction effect). In this setting, we simulated 1,000 gene expression levels each for 5 varying "heritability" levels from 15% to 87%. Our proposed methods MixRF and MixRF+iPC showed more obvious advantages over other competing methods when the heritability was low. The likely reasons for the observed advantages were because that our methods were based on random forest and thus capable of capturing the non-linear effects of predictors and their interactions with minor extra computation burdens.

2.3.2 Incorporating imputed data to improve the power for detecting phenotype-expression correlations

When directly collecting certain tissues in a specific cohort is challenging and when resources are available, one may impute expression data on inaccessible tissues using available information and potentially GTEx as a reference. We argue that the imputed data can be treated

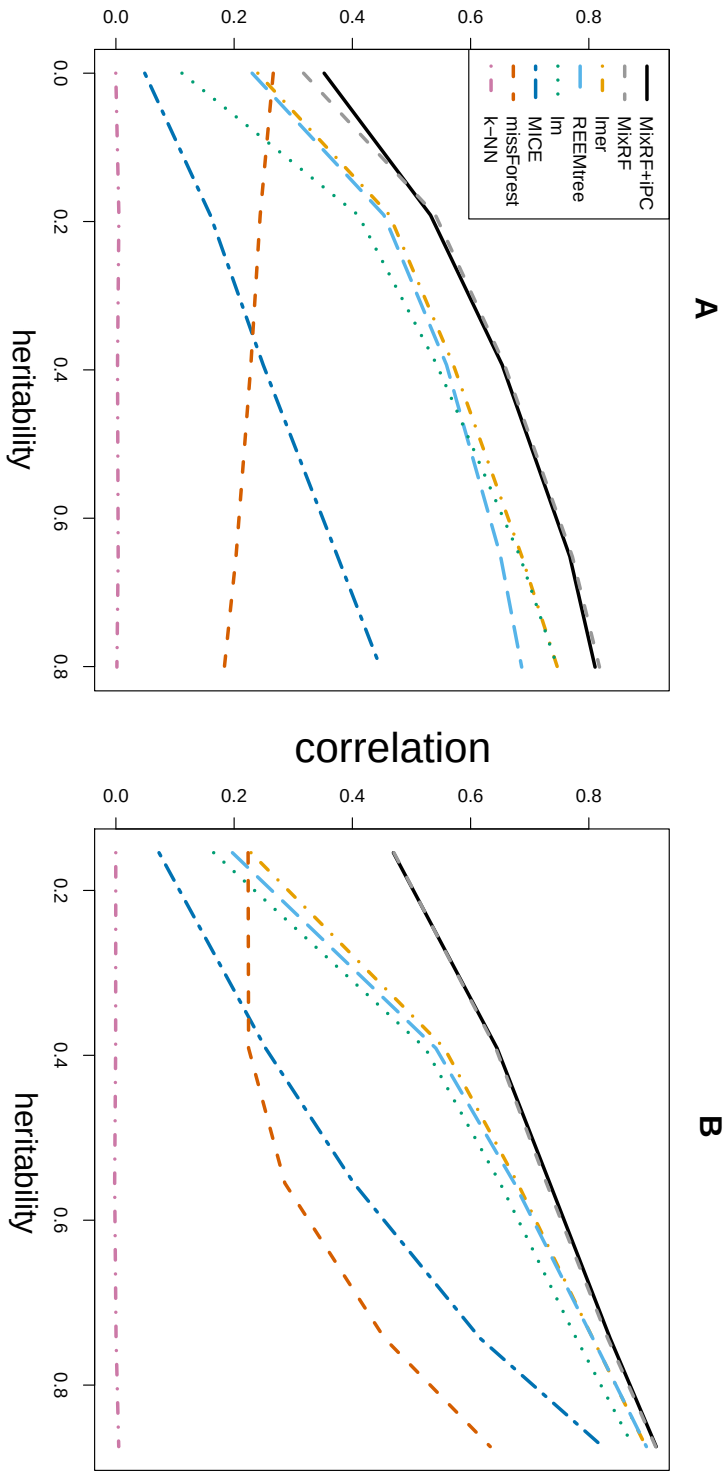

Figure 2.1: **Methods comparison on imputation performance based on simulations.** (A) We simulated 1,000 gene expression each with 0, 1, 2, 5, and 10 eQTLs. (B) Each expression level was simulated to be affected by two eQTLs and their interaction. We simulated 1,000 gene expression each for 5 varying heritability levels.

as supplementary data or supporting data to enhance the primary analysis based on the observed expression data. To support this claim, we took the expression data on the whole blood as well as adipose and nerve tissues and the genotype data in the GTEx pilot project, and then simulated phenotypes that were correlated with gene expression levels in the nerve tissues with correlations of 0.25 and 0.3. We treated 50% of the nerve tissues as "uncollected" and set the expression levels in those tissues as missing.

By applying the proposed MixRF+iPC method to the 10,919 genes in the observed data (with blood + adipose + 50% nerve tissues) and estimating the imputation correlation for each gene, we obtained 1,537, 762, and 324 genes with estimated imputation correlations ($\hat{r}_{imp}$) greater than 0.3, 0.4, and 0.5, respectively. At the significance thresholds of 5 and 10% false discovery rates (FDRs), we compared the power to detect the phenotypes associated with the nerve expression levels based on (1) only the observed nerve expression data (50% of the complete nerve data), (2) the combined observed and imputed nerve expression data with varying imputation quality ($\hat{r}_{imp} \geq$ 0.3, 0.4, and 0.5), and (3) the complete GTEx nerve expression data with 95 samples.

The results are presented in Table 2.1. Incorporating reasonably imputed data helped to improve the power to detect phenotype-expression correlations even when the phenotype-expression correlations were not strong and/or when the imputed data quality was not superb. As the imputation quality improved, the power improvement became more substantial. Analyses based on poorly imputed genes may not help or may hurt the power of the analyses. Although after excluding the poorly imputed expression levels, there may be only a small proportion of imputed gene expression levels retained in the subsequent analyses, those genes are often affected by multiple eQTLs and/or related to other factors in functional pathways, and as such, often of biological interest.

Table 2.1: **Power comparison for detecting phenotype-expression correlations based on the observed, the observed+imputed, and the complete expression data**. More specifically, the three expression data sources are: only the observed nerve expression data (with 50% of GTEx nerve tissue missing), the observed+imputed data with varying imputation quality, and the complete GTEx nerve expression data. The significance thresholds are 5% and 10% FDR. We assessed the power comparison when the phenotype-expression correlations are 0.25 and 0.3 for three groups of genes with estimated imputation correlations of at least 0.3, 0.4, and 0.5 (representing fair, moderate and good imputation quality).

# Genes passing estimated imputation correlation ($\hat{r}_{imp}$) thresholds	phenotype-expression correlations	FDR	Observed data only	Observed + Imputed data	Complete data
1537 genes ($\hat{r}_{imp} \geq 0.3$)	0.25	0.05	0.269	0.377	0.953
		0.1	0.455	0.586	1
	0.3	0.05	0.548	0.738	1
		0.1	0.729	0.898	1
762 genes ($\hat{r}_{imp} \geq 0.4$)	0.25	0.05	0.230	0.652	0.933
		0.1	0.391	0.839	1
	0.3	0.05	0.613	0.734	1
		0.1	0.778	0.887	1
324 genes ($\hat{r}_{imp} \geq 0.5$)	0.25	0.05	0.454	0.534	1
		0.1	0.688	0.744	1
	0.3	0.05	0.676	0.784	1
		0.1	0.883	0.926	1

2.4 GTEx data analyses

2.4.1 Imputing uncollected GTEx tissues

The GTEx project is collecting 44 human tissue types and most of them are difficult to access. In the pilot data, only 9 tissue types were collected in more than 80 out of 175 donors with the remainder yielding tissue-specific sample sizes of less than 40. We sought to impute the uncollected GTEx tissues using MixRF+iPC (Figure 2.2A). We conducted a 10-fold CV analysis focusing on the 9 tissue types to evaluate the imputation performance within GTEx. Specifically, we randomly split the GTEx transcriptome data on the 9 tissue types into 10 folds with each fold containing data on 1/10 of the collected tissues from each tissue type. In each round of CV, we treated one fold of the transcriptome data as unobserved/uncollected and the other 9 folds as observed/collected. For each gene, we imputed the unobserved expression levels in uncollected GTEx tissues using the expression levels in the collected tissues, and illustrated the imputation scheme in Figure 2.2A. We repeated the exercise for each fold of data, combined the imputed data, and evaluated the true tissue-specific imputation correlations.

The imputation performance of MixRF was generally comparable with its extension MixRF+iPC, though the latter performed better in imputing gene expression with no eQTLs or in the blood tissue (Table 2.2). We also compared the true imputation correlations by MixRF+iPC in different tissue types with the standard practice of using blood expression as a surrogate for target tissue expression (hereafter referred to as "blood surrogate") (Figure 2.3). The imputation performance of MixRF+iPC largely relied on the heritability and tissue-tissue expression level correlations for each gene, both of which were directly related to the number of cross-tissue eQTLs. For genes with 5 or more combined *cis*- and *trans*- eQTLs, the median true imputation correlation was 0.48. For genes with 10+ and 30+ eQTLs, the median true imputation correlation increased to 0.55 and 0.63, respectively. Note that al-

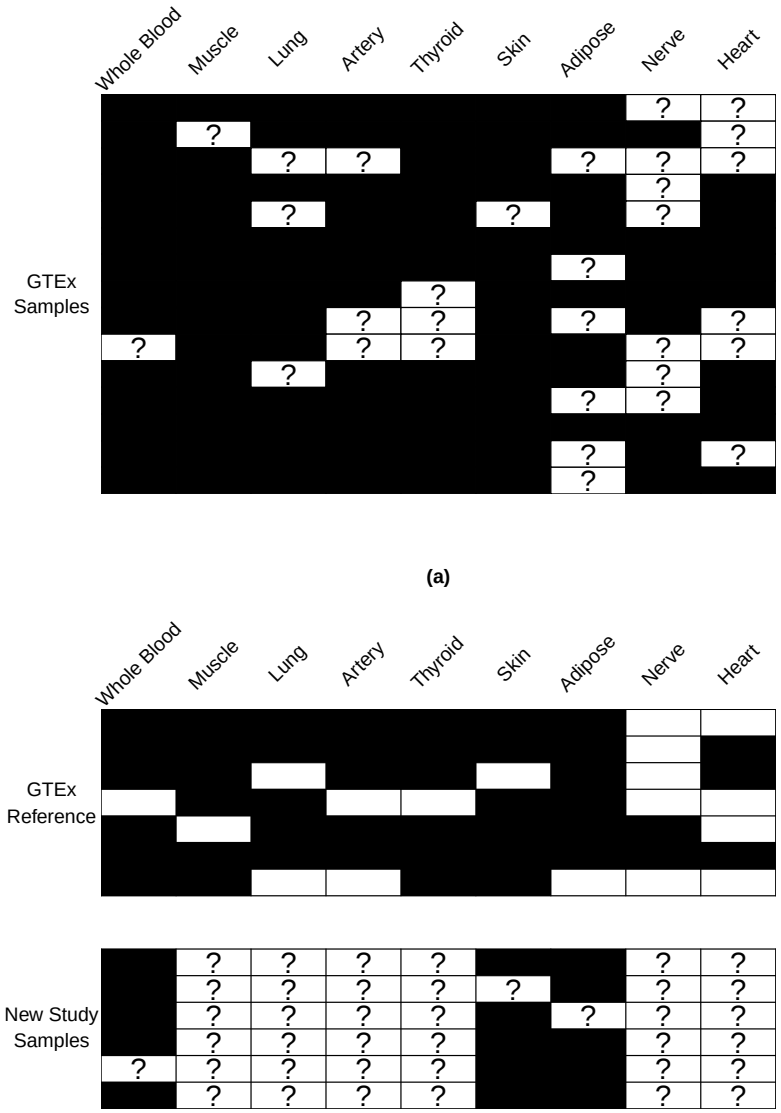

Figure 2.2: **Illustrations of two imputation scenarios.** In both scenarios, one can apply the proposed methods to impute the expression in the uncollected or inaccessible tissues of interest. Each row is one individual and each column is one tissue type. The collected and measured tissues are shown in black and the uncollected or inaccessible ones are in white. The tissues with question marks are the ones of interest. (A) Imputing expression in the uncollected GTEx tissues (with question marks) based on expression in the collected/measured GTEx tissues. (B) Using GTEx as a reference, imputing expression in the uncollected tissues (with question marks) including inaccessible tissue types based on collected tissues in a new expression study.

though we performed LD pruning (with an LD threshold of 0.5 and a window size of 50 base pairs) on the SNVs by PLINK (Purcell et al., 2007), there could still be moderate LD remaining among the eQTLs. Among the 10,919 expressed genes that we considered, the genes with 5+, 10+ and 30+ eQTLs in at least one fold of data numbered 1,065 (9.8%), 465 (4.3%) and 170 (1.6%), respectively.

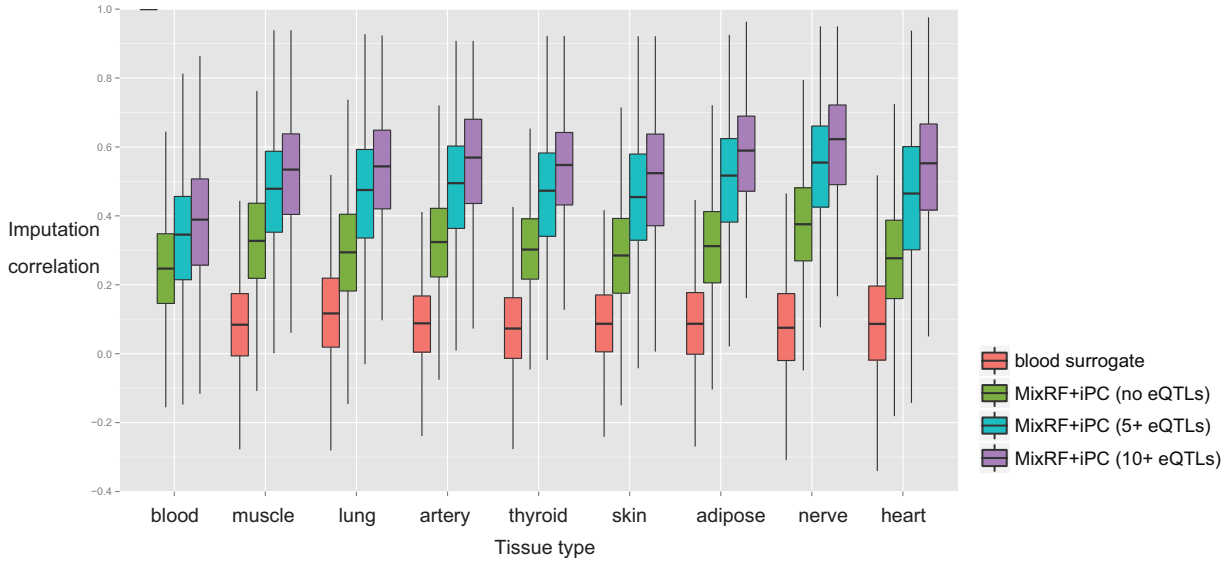

Figure 2.3: **Boxplots of gene-level true imputation correlation by tissue type.** The results are based on a 10-fold CV analysis within GTEx tissues. Specifically, we randomly split the GTEx transcriptome data into 10 folds, with each fold of data containing 1/10 of the collected tissues from each tissue type. In each round of CV, we treated one fold of the transcriptome data as unobserved and the other 9 folds as observed. We then imputed the unobserved data. Figure 2.2A illustrates the imputation scheme in one round of the CV analysis. We repeated the analysis for each fold of data. Each correlation was calculated as the correlation of the observed versus the combined imputed values of a given gene in the current tissue. We compared the true imputation correlations based on MixRF+iPC for genes with 0, 5+, and 10+ eQTLs with the correlations based on blood surrogate. Note that the eQTLs for each gene can be in moderate LD.

Generally, the expression in whole blood is weakly correlated (with a median correlation of 0.1) with the expression in other tissues and is a poor surrogate for the latter. We compared the imputation performance of the proposed methods with additional competing imputation methods for genes with different numbers of eQTLs (Table 2.2).

Table 2.2: **The comparison of the median gene-level true imputation correlations by different methods.** We compared the median gene-level true imputation correlations for 8 competing methods including the proposed MixRF and MixRF+iPC for three groups of genes – those with 0, 5+, and 10+ eQTLs in any fold of data. The numbers of genes for the three groups are 9,240, 1,065, and 465 respectively. We did not show the imputation results by linear regression (lm) for genes with no eQTLs due to a lack of predictor.

	blood			muscle			lung		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	-	-	-	0.08	0.21	0.27	0.10	0.25	0.32
k-NN	0.02	0.05	0.06	0.14	0.19	0.18	0.10	0.14	0.16
missForest	0.25	0.27	0.26	0.33	0.36	0.33	0.27	0.32	0.31
MICE	0.02	0.09	0.12	0.08	0.22	0.28	0.07	0.22	0.29
lm	-	0.18	0.27	-	0.30	0.42	-	0.32	0.43
lmer	0.09	0.28	0.33	0.26	0.47	0.53	0.27	0.49	0.56
REEMtree	0.10	0.28	0.33	0.26	0.46	0.53	0.27	0.48	0.55
MixRF	0.09	0.27	0.33	0.26	0.45	0.50	0.27	0.47	0.54
MixRF+iPC	0.25	0.35	0.39	0.33	0.48	0.53	0.29	0.48	0.54
	artery			thyroid			skin		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	0.07	0.21	0.25	0.06	0.22	0.29	0.07	0.22	0.28
k-NN	0.14	0.16	0.15	0.24	0.21	0.18	0.11	0.14	0.14
missForest	0.27	0.32	0.30	0.30	0.35	0.32	0.25	0.27	0.24
MICE	0.10	0.26	0.33	0.07	0.22	0.30	0.06	0.22	0.28
lm	-	0.34	0.47	-	0.30	0.42	-	0.32	0.42
lmer	0.34	0.52	0.59	0.27	0.49	0.55	0.25	0.47	0.54
REEMtree	0.35	0.51	0.58	0.27	0.48	0.55	0.26	0.47	0.53
MixRF	0.34	0.51	0.57	0.27	0.48	0.54	0.25	0.45	0.51
MixRF+iPC	0.32	0.49	0.57	0.30	0.47	0.55	0.28	0.45	0.52
	adipose			nerve			heart		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	0.08	0.24	0.30	0.05	0.22	0.30	0.07	0.23	0.30
k-NN	0.13	0.15	0.16	0.20	0.20	0.19	0.15	0.15	0.16
missForest	0.28	0.33	0.30	0.34	0.38	0.34	0.30	0.33	0.32
MICE	0.10	0.27	0.35	0.12	0.29	0.37	0.07	0.23	0.30
lm	-	0.36	0.48	-	0.35	0.49	-	0.31	0.43
lmer	0.33	0.55	0.60	0.39	0.58	0.63	0.28	0.49	0.56
REEMtree	0.33	0.54	0.59	0.40	0.58	0.63	0.28	0.49	0.56
MixRF	0.33	0.54	0.58	0.40	0.57	0.62	0.28	0.49	0.55
MixRF+iPC	0.31	0.52	0.59	0.38	0.55	0.62	0.28	0.46	0.55

To evaluate the impact of sample size on imputation, we performed a 3-fold CV analysis within GTEx tissues and compared the results with those obtained from the 10-fold CV analysis (Table 2.3). In each round of the 3-fold and the 10-fold CV analyses, 2/3 and 9/10 of the data were treated as "observed," yielding average tissue-specific sample sizes of 73 and 99, respectively. We found that sample size substantially affected imputation performance largely because sample size substantially affected the power to detect cross-tissue eQTLs. With a 36% sample size increase in the 10-fold CV analysis and at the same significance criteria, we detected 65% more cross-tissue *cis*-eQTLs (5,332 versus 8,792) and 225% more cross-tissue *trans*-eQTLs (3,976 versus 12,884). As a result, the median true imputation correlation across the genome improved from 0.305 to 0.349. Additional simulations are presented in Table 2.4 to further demonstrate the impact of sample size on imputation. When more GTEx samples become available, we expect further improvement in imputation performance.

Table 2.3: **The comparison of a 10-fold versus a 3-fold CV analysis within GTEx tissues shows the impact of sample size.** The increase in sample size affects the power to detect cross-tissue eQTLs and subsequently imputation results. The number of genes with x eQTLs is counted as the number of genes with x eQTLs in at least one fold of data. For example, genes may have no eQTLs in one or several folds of data and have 1 or 2 eQTLs in other folds of data. We calculated the true imputation correlation for genes with no eQTLs by only considering the folds of data in which the gene has no eQTLs. As such, there is overlapping among genes with 0, 1 or 2 eQTLs, etc.

	10-fold CV			3-fold CV		
Average sample size across tissues	98.9			73.2		
Average # *cis*-eQTLs	8,792			5,332		
Average # *trans*-eQTLs	12,884			3,976		
Average median true imputation correlation	0.349			0.305		
	# genes	median true imputation correlation	# genes with true imputation cor. ≥ 0.5 (%)	# genes	median true imputation correlation	# genes with true imputation cor. ≥ 0.5 (%)
Genes with no eQTLs	9,240	0.307	207 (2.2)	10,063	0.271	88 (0.9)
Genes with 1 eQTLs	5,062	0.338	250 (4.9)	2,521	0.317	62 (2.5)
Genes with 2 eQTLs	2,486	0.368	228 (9.2)	767	0.355	63 (8.2)
Genes with 3 eQTLs	1,394	0.386	191 (13.7)	371	0.390	44 (11.9)
Genes with 4 eQTLs	883	0.412	169 (19.1)	225	0.401	48 (21.3)
Genes with 5-9 eQTLs	839	0.430	191 (22.8)	275	0.454	90 (32.7)
Genes with $\geq$10 eQTLs	465	0.521	270 (58.1)	185	0.578	128(69.2)

Table 2.4: **The impact of sample size on imputation performance.** We simulated the gene expression levels for 1000 genes in 9 tissue types. Each gene expression level is affected by two cross-tissue eQTLs and their interactions, with average heritability of 0.2, 0.3, and 0.5. As sample size (# individual) increases from 150, 300, 500, to 1000, we observed improved power to detect true eQTLs and improved imputation performance.

True heritability	Sample size (#individual)	Median est. heritability	Median imp. cor. (r_{imp})	% genes with $r_{imp} \geq 0.3$	% genes with $r_{imp} \geq 0.5$	% genes with $r_{imp} \geq 0.7$
0.20	150	0.001	0.294	48.6	10.7	0.5
	300	0.042	0.335	59.0	13.6	0.6
	500	0.077	0.359	67.8	15.9	0.8
	1000	0.129	0.383	75.8	18.6	0.9
0.30	150	0.014	0.347	59.4	22.0	2.9
	300	0.148	0.403	72.0	29.0	4.0
	500	0.204	0.434	79.6	34.5	5.1
	1000	0.271	0.451	85.7	37.8	6.0
0.50	150	0.216	0.463	73.4	44.3	14.7
	300	0.413	0.529	86.4	55.4	18.9
	500	0.478	0.566	91.9	63.4	22.6
	1000	0.562	0.581	94.6	67.4	24.5

Overall, both eQTL and tissue-tissue expression level correlation play a major role in multi-tissue imputation. The average estimated heritability for expressed genes was reported as $0.14 \sim 0.26$ for different tissue types in other studies (Grundberg et al., 2012; Wright et al., 2014), which roughly corresponds to an imputation correlation of $0.37 \sim 0.51$ if the appropriate SNVs were selected in the imputation. According to our results, the median true imputation correlation based on linear regressions that use only eQTLs as predictors (see "lm" in Table 2.2) was much lower (~ 0.2), indicating that the current imputation results could be improved if sample size were to increase and more eQTLs were detected and used in the imputation. Additional comparison of linear regression versus mixed-effects models (Table 2.2) showed that information on tissue-tissue expression level correlation helped improve the absolute median imputation correlation by $0.1 \sim 0.3$ for genes with 5+ or 10+ eQTLs. For genes with no eQTLs, the median imputation correlation with mixed-effects models was nearly 0.3 and was improved compared to using blood expression as a surrogate.

2.4.2 Using GTEx as a reference to impute other studies

Tissue accessibility often limits the feasibility and scale of multi-tissue expression studies in specific cohorts. multi-tissue expression imputation would be helpful, when direct measurements in specific tissues are limited or not available and when expression data on related tissues are existing or accessible. Incorporating expression in the secondary and related tissue types with the primary data may enhance the power to detect differentially expressed genes under different phenotypic conditions and provide insights into disease etiology from a multi-tissue perspective. Multi-tissue imputation could impute expression in the uncollected tissues, which could be used in the secondary data analysis as supplementary data to be combined with the primary observed data. In those imputation scenarios, one may use GTEx as a reference and impute gene expression in the uncollected tissues or tissue types in non-GTEx samples. Figure 2.2B shows an example of such imputation scenarios.

We conducted another 10-fold CV analysis to evaluate the feasibility of such imputation. In contrast to the 10-fold CV analysis conducted in the previous section, here we split the GTEx individuals into 10 folds. In each round of the current CV, we treated 9 folds of the samples as the "GTEx reference," and the other fold as testing samples from a new study. In the new study samples, we only observed the transcriptome data in the 3 accessible tissues, and used the data on the 3 tissues with GTEx as a reference to impute the expression in the uncollected tissues in the new samples (Figure 2.2B).

We evaluated the tissue-specific gene-level true imputation correlations in the six inaccessible tissue types by MixRF+iPC (Figure 2.4). Blood surrogate could achieve a median correlation of only ~ 0.1. In contrast, even for genes with no eQTLs, MixRF+iPC could achieve a median true imputation correlation of 0.17 to 0.27 in different tissue types. For genes with 10+ eQTLs, the median true imputation correlation increased to ~ 0.4 across tissue types. The imputation performance was better in the nerve than other tissues, with a median correlation of 0.37, 0.42, and 0.54 for genes with 5+, 10+, and 30+ eQTL, respec-

tively. This may be attributable to the relatedness of the nerve to adipose and skin, or to its reaction to stimuli.

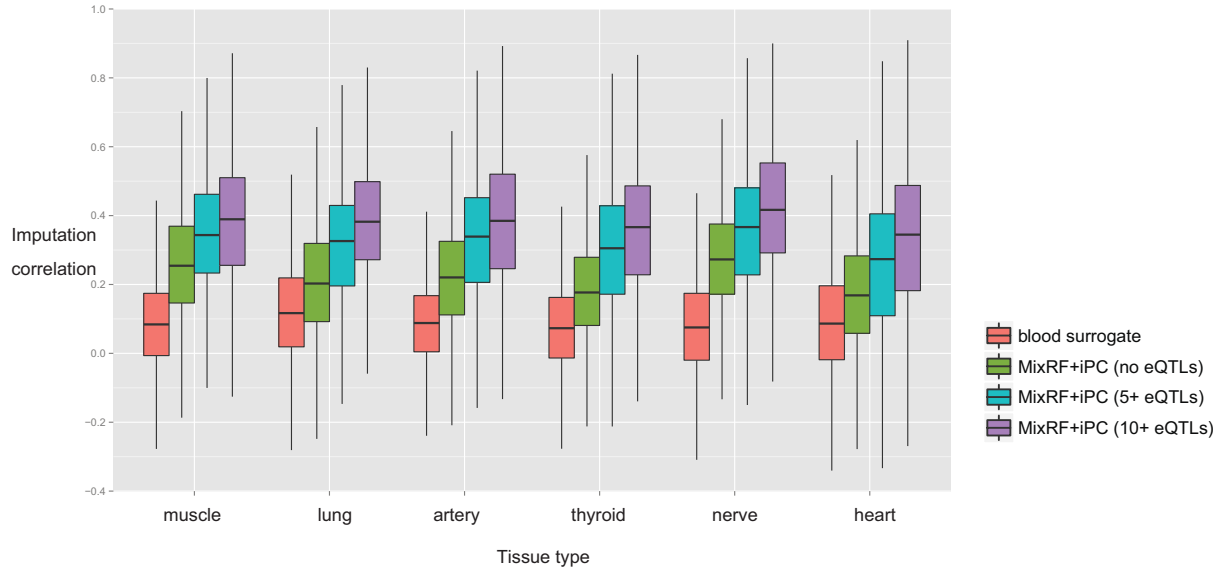

Figure 2.4: **Boxplots of gene-level true imputation correlation in inaccessible tissues in a new study.** The results are based on a 10-fold CV analysis of imputing uncollected tissues in the new samples while using GTEx as a reference. Specifically, we split the GTEx individuals into 10 folds. In each round of CV, we used 9 folds of the samples as reference samples and treated the other fold of the sample as new study samples. With GTEx data as a reference, we imputed the transcriptome data in the inaccessible tissues based on the accessible ones (blood, skin, and adipose) in the new samples. Figure 2.2B illustrates the imputation scheme in one round of CV. We repeated the analysis for each fold of data. We compared the true imputation correlations based on MixRF+iPC for genes with 0, 5+, and 10+ eQTLs with the correlations based on blood surrogate.

Additionally, one may also use multi-tissue imputation to build on existing single-tissue expression and eQTL data. One may collect the tissues of interest in a small set of new samples in the specific cohorts as the learning tissues and then use those tissues together with the GTEx reference to impute the samples with expression data only on a single tissue and not available for additional data collection.

The multi-tissue imputation strategy can also be used in designing future multi-tissue expression studies in certain populations/ethnicities, or with specific phenotypes. One may utilize the GTEx resource and conduct CV analyses on the GTEx tissues, leveraging tissue

availability and predictability to select the tissue types that are most relevant and predictive for the target tissue types.

2.4.3 Using GTEx as a reference in the presence of potential study heterogeneity and a validation analysis

The performance of the proposed multi-tissue imputation methods primarily depends on the predictive ability of eQTLs and the tissue-tissue expression level correlations. When using GTEx as a reference to impute other non-GTEx samples with potential study heterogeneity, we suggest including a reference sample indicator variable in the MixRF as a covariate. When the eQTL effects or effects of other covariates are sufficiently different among the GTEx reference and the non-GTEx samples, the interaction terms of the reference indicator and the eQTLs/covariates will be selected in building the random forest. As such, in the presence of study heterogeneity, the estimation of eQTL effects in the non-GTEx samples will be based primarily on the non-GTEx samples only.

Recent studies have shown that the predictive ability of eQTLs can be replicated across GTEx and other studies (Gamazon et al., 2015), and the expression patterns of many pharmacogenes investigated by the Pharmacogenomics Research Network project can also be validated in the GTEx samples (Chhibber et al., 2016).

To further validate the utility of the proposed methods and of GTEx data as a reference in multi-tissue imputation for non-GTEx samples, we applied MixRF to a study of insulin sensitivity – the IS-MA (Insulin-Sensitivity Muscle-Adipose) study (Elbein et al., 2012). Fifty-nine samples at the tails of the distribution of insulin sensitivity were selected in the study. The expression levels in adipose and muscle tissues and genotype data are available on those 59 samples. We considered 229 genes with preserved Ensemble IDs in both GTEx and the IS-MA study. Such genes are likely to have completely preserved gene structure across the two data sets. We normalized the gene expression levels of each gene within each

study. We focused on imputing the gene expression levels of those 229 genes in the muscle tissues from the IS-MA samples.

We compared the performance of the following analyses to impute the muscle tissue expression levels in the IS-MA samples: A) Imputing with GTEx reference + adipose expression levels and eQTLs from the IS-MA study, B) Imputing with GTEx reference + adipose expression levels from the IS-MA study, no eQTLs, and C) Impute based only on the eQTLs from the IS-MA study, no GTEx reference. We calculated the imputation correlations of measured muscle tissue expression levels and the imputed values based on the three sets of analyses. Figure 2.5 shows the Quantile-Quantile plot of the three sets of imputation correlations versus the null correlations. Including GTEx as a reference greatly improves the imputation performance, and the mean imputation correlation of those 229 genes based on Analysis A is 0.313. When imputing based only on eQTL genotypes or imputing based only on tissue-tissue expression level correlations, the imputation correlations substantially deviate from the null correlations. That implies that both eQTL genotypes and tissue-tissue expression level correlations help in the multi-tissue imputation. MixRF with GTEx as a reference combines the two sources of information and improves the overall imputation.

2.5 Discussion

The joint analysis of transcriptome data from multiple tissues would enhance the power of expression data analysis and ultimately improve our understanding of biological mechanisms from a systems perspective. The bottleneck that limits the feasibility and scale of multi-tissue expression studies is tissue accessibility. When a tissue is not accessible in an individual, the gene expression levels in that tissue are not available and are considered "missing". We propose a multi-tissue imputation algorithm and an extension for imputing multi-tissue expression data. The proposed approaches can be used in imputing expression on uncollected tissues in the GTEx project to facilitate downstream analyses, and moreover for imputing

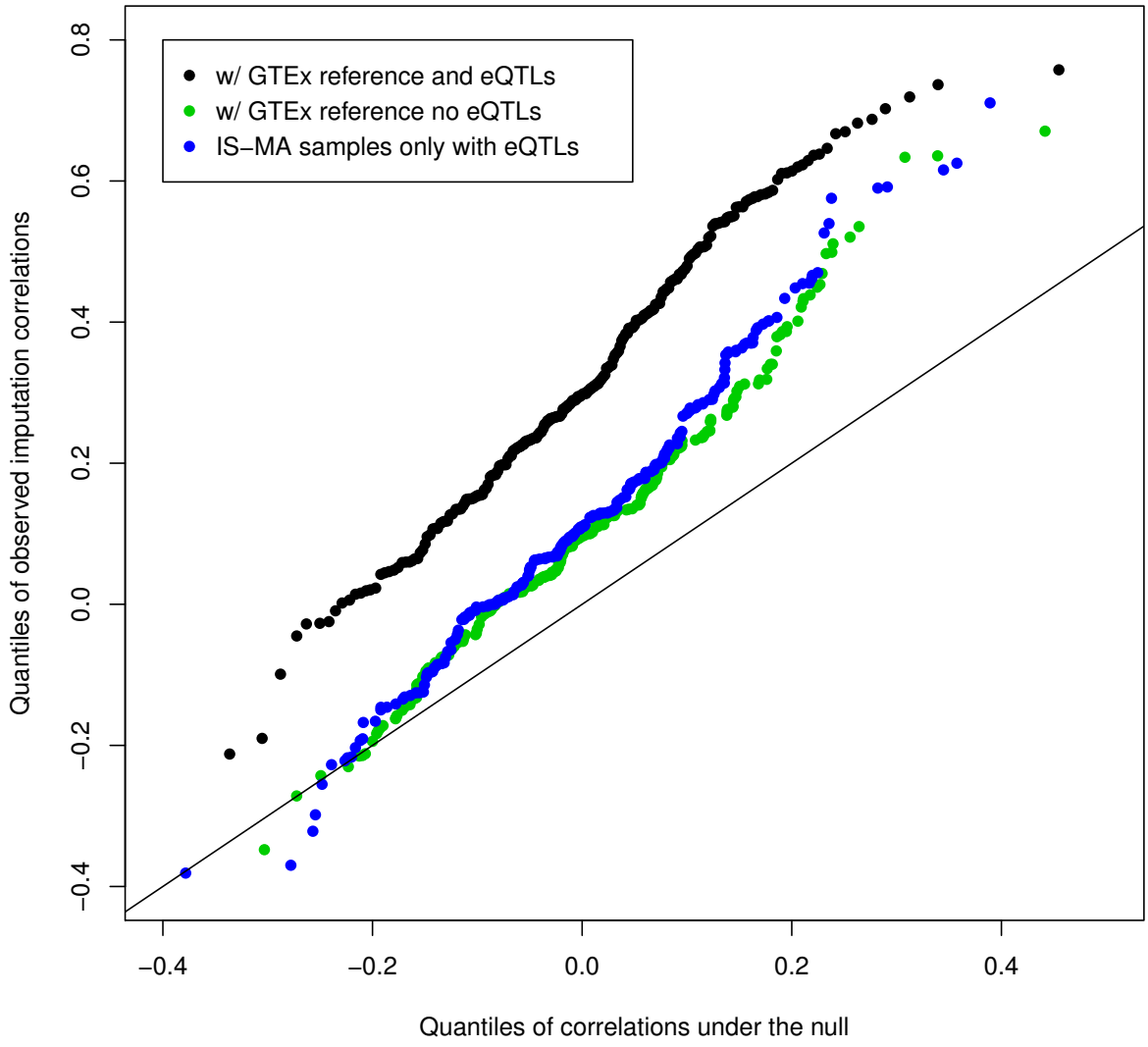

Figure 2.5: **Quantile-quantile plots of the observed imputation correlations versus the null correlations.** The three sets of observed imputation correlations are based on the analysis with GTEx reference and eQTLs (black), the analysis with GTEx reference and no eQTLs (green), and the analysis only with eQTLs without GTEx reference (blue).

inaccessible tissues in other expression studies using GTEx as a reference. Different than methods that predict expression levels based on eQTL information (Gamazon et al., 2015), our proposed methods impute multi-tissue expression levels based on eQTLs, tissue-tissue expression level correlations, and tissue-specific PCs of expression data, and harness genetic factors, major developmental biological factors, and environmental factors. Additionally, our mixed model based random-forest approach captures the dominant and recessive eQTL effects, as well as the interactions among eQTLs, tissue types, and other factors. Most existing single-tissue imputation methods rely on gene-gene correlations, which can be unstable. Our methods outperform existing imputation methods in multi-tissue imputation.

Multi-tissue imputation can be helpful when direct measurements in the desired tissues are uncollected or difficult to collect, and one may use the imputed data as supplement data to support scientific findings from observed data. Within the GTEx project, we can impute the expression in the uncollected tissues and use imputed expression data to enhance the detection of protein-QTLs or facilitate the construction of integrative genomics networks. More importantly, by using GTEx as a reference, we can potentially impute inaccessible tissues in other expression studies, impute and recapitalize on existing data, design effective multi-tissue expression studies in other populations or ethnicities, and further inform disease-related tissues. As a multi-tissue imputation method, we anticipate that our work will initiate research on methods development and enable the discovery of scientific findings using multi-tissue expression data within and beyond the GTEx project.

One caveat of the current analyses is that we used only cross-tissue eQTLs in the imputation. The sample size in the GTEx pilot data limits the power to detect tissue-specific eQTLs. We believe that a larger sample size in the later phase of GTEx data will bring increased power to detect both cross-tissue and tissue-specific eQTLs, thereby substantially improving imputation performance. An alternative strategy to select eQTLs is to combine the eQTLs reported in other studies, which ideally involve multiple tissue types.

We anticipate that the later phase of GTEx data will bring additional challenges to methods development, e.g., the scalability of the approaches, and the selection of the accessible tissues for maximizing imputation accuracy. In addition to multi-tissue imputation, it is desirable to develop methods to account for observed and imputed expression values in the subsequent disease/trait-related analyses and to enable multi-tissue network and integrative analyses.

CHAPTER 3

A MULTIVARIATE MIXED-EFFECTS SELECTION MODEL FRAMEWORK FOR LABELING-BASED PROTEOMICS DATA WITH NON-IGNORABLE MISSINGNESS

3.1 Introduction

In quantitative proteomics research, mass spectrometry (MS) experiments are widely used to quantify the abundances of proteins and peptides. In these experiments, proteins are digested into smaller sequences of amino acid, i.e., peptides. After the abundances of peptides are derived from MS experiments, one may analyze each peptide as the analysis unit or summarize multiple peptides in a protein and analyze each protein as the analysis unit. Traditional shotgun MS experiments process each sample one by one, and thus greatly limit the scale of quantitative proteomics research. More recently, mass tag labeling techniques, such as isobaric tags for relative and absolute quantitation (iTRAQ) and tandem mass tags (TMT), have been widely adopted in MS experiments (Wiese et al., 2007). These techniques allow proteins/peptides from multiple samples of a batch being quantified in a single MS experiment, and thus greatly enhance the efficiency of data generation. However, the batch processing of samples results in severe batch effects in the output data, and the missing data also occur at the batch level, i.e., a peptide abundance is either all observed or all missing in the samples processed by the same experiment.

To correct for batch effects, a shared reference sample is often included in all batches. Conventional data analyses are performed based on the observed relative abundances of peptides in the target samples to the common reference sample. However, it is known that the experimental variation is large across different MS experiments. Measurements for the same peptide in the common reference sample may vary across experiments. As such, analyses of relative abundances of peptides/proteins cannot fully correct for batch effects.

Furthermore, most existing analyses are based on only the observed abundances and ignore the missing data. Here the missing data are non-ignorable (Little and Rubin, 2002) because the missing probability of a peptide/protein in a batch of samples largely depends on its average abundance levels of all samples in the batch. Ignoring the missing data may lead to biased estimation and inference.

To analyze labeled (batch-processed) proteomics data, Chen et al. (2017b) recently proposed a univariate mixed-effects model accounting for the batch-processing design and the batch-level non-ignorable missing-data mechanism. This univariate model can be used to directly analyze each individual peptide. When using this model to analyze each protein, one may take the average abundance level of peptides from the same protein as the protein abundance level. For unlabeled proteomics experiments (in which samples are processed one by one), Clough et al. (2009) showed that by jointly modeling multiple peptides of the same protein, one can gain improved precision and power than analyses based on averaging peptide abundances. For labeled proteomics experiments, to our knowledge, there is no such comparison, and this is part of the goal of our work.

In this work, we consider the problem of jointly analyzing multiple correlated features / outcomes in labeled proteomics data. Note that here "outcome" refers to response variable, not clinical outcome. Feature and outcome may be used interchangeably throughout the work. A desirable model for the data type should consider the batch design, the non-ignorable missing-data mechanism, and the correlations among multiple features. Existing methods have been proposed for multivariate analysis within the mixed-effects model framework. Shah et al. (1997) proposed a bivariate random-effects model for data with possible ignorable missing data. Roy and Lin (2002) proposed a latent variable selection model for analyzing multivariate longitudinal data with missing values caused by dropouts. More recently, Liu et al. (2015) developed a selection model for bivariate outcomes with missing values occurring sporadically. However, none of those models can be directly applied to labeled proteomics

data with the unique clustered outcome-dependent missingness. Additionally, these existing methods can only analyze multivariate outcomes in low dimensionality, whereas in proteomics data, the number of peptides in a protein and the number of proteins in a pathway can be up to hundreds and may exceed the sample size.

These challenges motivated us to develop a multivariate model incorporating the batch design and the batch-level (or cluster-level) missingness. We termed this model framework as the multivariate MIxed-effects SElection (mvMISE) model. In addition to accounting for the batch design, our method employed tailored modeling for different correlation structures of different types of features and can achieve efficient estimation of parameters for high-dimensional features. Specifically, we developed two multivariate models: Model I – mvMISE$_b$ – for jointly analyzing multiple peptides from each protein, and Model II – mvMISE$_e$ – for jointly analyzing multiple proteins from *a priori* defined pathways. In both models, we considered the estimation feasibility and computational efficiency when the number of features is very large. Via simulation studies, we demonstrated the advantages of the proposed mvMISE methods. We applied the proposed methods to an iTRAQ-based breast cancer proteomic data set from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Mertins et al., 2016). The proposed framework can also be extended to analyze general clustered data with ignorable and non-ignorable missing data.

3.2 Methods

3.2.1 A multivariate mixed-effects model for clustered data

Let $\mathbf{Y}_i$ be an $n_i \times K$ matrix, denoting the abundance levels of K peptides in a protein (or K proteins in a pathway) from the i-th experiment (i.e., the i-th batch of samples), $i = 1, 2, \ldots, N$. Let n_i be the number of samples in the i-th batch including the reference sample. In the CPTAC breast cancer proteomics project, every three randomly selected

target tumor samples and one common reference sample were grouped into a batch and were analyzed by a four-plex (i.e. four-channel) iTRAQ experiment. A total of 108 tumor samples were analyzed in 36 iTRAQ experiments. The samples in the same batch were processed together, and the quantitations of abundance levels from these samples were related. Let the vector $\mathbf{y}_i = \text{vec}\,(\mathbf{Y}_i)$ denote all the abundance levels of K-variate features in the i-th batch. A natural model for the batch-processed data is a mixed-effects model with multivariate abundance levels as the outcome variables:

$$\mathbf{y}_i = \mathbf{X}_i\,\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{3.1}$$

where $\mathbf{X}_i$ is an $n_iK \times p$ design matrix with fixed effects $\boldsymbol{\alpha}$, and $\mathbf{Z}_i$ is an $n_iK \times q$ design matrix with random effects $\mathbf{b}_i$. We assume that the random effects $\mathbf{b}_i$ follow a normal distribution $\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D})$, where the matrix $\mathbf{D}$ captures the covariances of random effects. Let $\mathbf{e}_i = \text{vec}\,(\mathbf{E}_i)$, where $\mathbf{E}_i$ is the $n_i \times K$ error matrix. We assume that $\mathbf{E}_i$ follows a matrix normal distribution $\mathbf{E}_i \sim MN_{n_i,K}(\mathbf{0}, \mathbf{S}_i, \boldsymbol{\Sigma})$, and thus $\mathbf{e}_i$ follows a multivariate normal distribution $\mathbf{e}_i \sim N_{n_iK}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S}_i)$, where $\otimes$ denotes the Kronecker product. The matrix $\mathbf{S}_i = \text{diag}\left(\sigma_0^2, \sigma^2, \ldots, \sigma^2\right)$ is an $n_i \times n_i$ diagonal matrix with the first diagonal element (σ_0^2) corresponding to the variance of the common reference sample, and the rest of the diagonal elements (σ^2) being the variance of the target tumor samples. The variance for the reference is different and generally smaller than that of other tumor samples because the reference sample is created by combining multiple tumor samples. The matrix $\boldsymbol{\Sigma}$ captures the error (or unexplained) covariances among K features. The model can be used to estimate the common fixed effects averaged across K features and can also be used to estimate feature-specific effects. To estimate the common fixed effects, one may let $\mathbf{X}_i = \mathbf{1}_K \otimes \mathbf{X}_i^*$, where $\mathbf{1}_K$ is a vector of 1s and $\mathbf{X}_i^*$ is a $n_i \times p$ covariates matrix shared among K features (for example, the tumor subtype indicators). To estimate feature-specific effects, one may include feature-specific covariates or the interactions of the predictors of interest and the indicators

of features in the design matrix, for example, let $\mathbf{X}_i = \mathbf{I}_K \otimes \mathbf{X}_i^*$ with $\mathbf{I}_K$ being an identity matrix.

3.2.2 The non-ignorable batch-level missing-data mechanism

A major challenge in analyzing proteomic data is the substantial amount of non-ignorable missing data. It is known that the probability of abundance levels of a peptide/protein being missing largely depends on the values themselves, and thus the missing data are non-ignorable. For batch-processed proteomics data, another complication is that the abundance levels of a peptide are generally either all missing or all observed in the samples from the same batch (i.e., the same experiment). The missing probability of a peptide in a batch largely depends on its average abundance levels in the batch.

Chen et al. (2017b) termed this type of missing data as the "batch-level abundance-dependent" missing-data mechanism and proposed to model it with an exponential function of the abundance values. The exponential function can be naturally integrated with the normal density function, facilitating the estimation and computation of the employed Expectation-Maximization (EM) algorithm. Following the idea, we modeled the missing probability of the abundance levels for each of the k-th features ($k = 1, 2, \ldots, K$) in the i-th batch as a function of the average abundance levels of feature k in batch i and other covariates

$$\Pr\left(r_{ik} = 1 | \mathbf{y}_{ik}\right) = g^{-1}\left(\phi_0 + \phi_1/n_i \cdot \mathbf{1}_{n_i}^{'}\mathbf{y}_{ik} + \boldsymbol{\phi}_2^{'}\mathbf{c}_{ik}\right), \tag{3.2}$$

where $\mathbf{y}_{ik}$ is a vector of abundance levels of the k-th feature for the samples from the i-th batch, $g(\cdot)$ is the link function, r_{ik} is a missing indicator with $r_{ik} = 1$ if all of the n_i values in $\mathbf{y}_{ik}$ are missing, and $\mathbf{c}_{ik}$ denotes the average of covariates for the k-th outcome of the i-th batch, if there is any. The parameters ϕ_0, ϕ_1 and $\boldsymbol{\phi}_2$ control the missing-data mechanism. When $\phi_1 = 0$, the missing probability does not depend on the abundance values and the missing data is missing at random (Little and Rubin, 2002). The multivariate mixed-effects

model (3.1) and the missing-data model (3.2) compose the proposed multivariate Mixed-Effects Selection (mvMISE) models.

For the link function in model (3.2), one may use logit or probit links to model the binary missing indicator. However, parameter estimation with such likelihood functions can be computationally prohibitive when analyzing high-dimensional data. Following Chen et al. (2017b), we used a log link function

$$\Pr\left(r_{ik}=1|\mathbf{y}_{ik}\right)=\exp\left(\phi_0+\phi_1/n_i\cdot\mathbf{1}_{n_i}^{'}\mathbf{y}_{ik}\right). \tag{3.3}$$

Further estimation details are discussed in Section 3.3.1.

3.2.3 The mvMISE models with different correlation structures tailored for different high-dimensional features

In order to model the correlations among multivariate or even high-dimensional features, it is important to understand the nature of correlation structures. In this section, we considered two distinct applications in proteomics data analyses and developed tailored models for different correlation structures among features in different applications. These models consider the scalability and computation feasibility for jointly analyzing high-dimensional features.

The mvMISE model with correlated random effects (mvMISE$_b$) for analyzing multiple peptides from a protein

In the MS proteomics experiments, large proteins are digested into smaller peptides, and peptides are measured by the instruments. Multiple peptides from the same protein have very similar abundance levels and are highly correlated. To model this type of correlations, we allow peptide-specific random effects to be correlated using a non-diagonal covariance

matrix $\mathbf{D}$. Additionally, we assume $\mathbf{\Sigma} = \mathbf{I}_K$. That is, the error terms of multiple peptides from the same protein are uncorrelated after accounting for the correlations of random effects and the covariates, i.e., there are no unexplained correlations among peptides in a protein.

When the number of peptides (K) in a protein is large, it becomes computationally prohibitive to estimate the above models with an unstructured $\mathbf{D}$ matrix. Due to the highly correlated nature of those peptides from the same protein, we propose to employ a factor-analytic random-effects structure for the correlated random effects (Liu and Hedeker, 2006). We termed this model as the mvMISE model with correlated random effects, mvMISE$_b$, and wrote it as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{Z}_i \boldsymbol{\tau} b_i + \mathbf{e}_i, \tag{3.4}$$

where $\mathbf{Z}_i = \mathbf{I}_K \otimes \mathbf{1}_{n_i}$, $\boldsymbol{\tau}$ is a $K \times 1$ vector for the peptide-specific variance components corresponding to the random effect b_i, and b_i is a standard normal random variable. In model (3.4), only K rather than $K(K+1)/2$ parameters need to be estimated for the covariance matrix of random effects, which substantially speeds up the computation. The model assumes a latent variable underlying the peptides in the same proteins and implies correlations equal one among random effects. Since the peptides from a protein are segments of the same molecule, the correlations are not only high but also similar in magnitude. This simplifying assumption is reasonable.

The mvMISE model with correlated error terms (mvMISE$_e$) for pathway analysis

In protein pathway analysis, one jointly analyzes multiple proteins in *a priori* defined functional pathways or protein-sets. One may first obtain the protein-abundance levels by averaging the peptide abundances mapped to each protein, and then treat multiple protein abundances in a pathway as the multivariate outcome measures. A major challenge in this

analysis is to model the biological correlations among multiple proteins. And those correlations are often unstructured. Here we propose to model the unstructured biological correlations among proteins via the error covariance matrix $\mathbf{\Sigma}$ while assuming the random effects are independent among features. We term this model as the mvMISE model with correlated error terms, mvMISE$_e$.

Protein abundances in a functional pathway can be generally correlated, while their interdependence (i.e., network or conditional correlation) structures are often sparse (Baladandayuthapani et al., 2014). Therefore, we introduced a graphical lasso penalty on the error precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ (Danaher et al., 2014). We proposed to obtain the estimates by

$$\text{maximize}_{\mathbf{\Theta}} \quad 2l\left(\mathbf{\Theta}\right) - \lambda N \left|\mathbf{\Theta}\right|_1, \tag{3.5}$$

where λ is a tuning parameter and $\left|\mathbf{\Theta}\right|_1 = \sum_{i \neq j} \left|\theta_{ij}\right|$. The graphical lasso penalty term ($\lambda \sum_{i \neq j} \left|\theta_{ij}\right|$) imposes regularization on the off-diagonal elements of the error precision matrix and as such ensures a sparse dependence structure. This will greatly facilitate the estimation of covariance and precision matrices for high-dimensional features.

3.3 Model Estimation

3.3.1 An EM algorithm for the mvMISE$_b$ model

In this section, we derived an EM algorithm for the proposed mvMISE$_b$ model with correlated outcome-specific random effects and an exponential missing-data mechanism function. The maximum likelihood estimation consists of iterating between two steps, the E and the M steps. The conditional expectations of the sufficient statistics in the E-step were derived and presented in the next subsection. In short, the non-ignorable missing-data mechanism function in (3.3) was integrated with the multivariate normal distribution function and imposed a bias-correction for the estimated conditional expectation of the missing outcome

values.

The E-step of the EM algorithm

Let $\boldsymbol{\gamma}$ be the set of all parameters. The expected complete-data log-likelihood for the i-th experiment ($l_i(\boldsymbol{\gamma})$) conditional on the observed data and current parameter estimates ($\boldsymbol{\gamma}^{(t)}$) is given by

$$\begin{aligned}
& Q_i\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right) = \mathrm{E}\left(l_i(\boldsymbol{\gamma})|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \\
= & \int\int \log f(\mathbf{y}_i|b_i) f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
& + \int\int \log f(b_i) f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
& + \int\int \log f(\mathbf{r}_i|\mathbf{y}_i) f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
= & const - \frac{1}{2}K\left(\log\sigma_0^2 + (n_i-1)\log\sigma^2\right) - \frac{1}{2}\int\int \mathbf{e}_i'\mathbf{R}_i^{-1}\mathbf{e}_i f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
& - \frac{1}{2}\int\int b_i' b_i f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
& + \int\left(\sum_{k=1}^{K} \log f(r_{ik}|\mathbf{y}_{ik})\right) f\left(\mathbf{y}_i^m|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}\mathbf{y}_i^m \\
= & const - \frac{1}{2}K\left(\log\sigma_0^2 + (n_i-1)\log\sigma^2\right) - I_1 - I_2 + I_3,
\end{aligned}$$

where $\mathbf{r}_i$ is the missing-data indicator and $\mathbf{R}_i = \boldsymbol{\Sigma} \otimes \mathbf{S}_i$. We let $\mathbf{y}_i$ denote the complete data $\mathbf{y}_i = \left(\mathbf{y}_i^{o\prime}, \mathbf{y}_i^{m\prime}\right)'$, where $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$ are the observed and missing values in the outcomes. The matrices $\mathbf{X}_i^o$ and $\mathbf{X}_i^m$ are the design matrices corresponding to $\mathbf{y}_i^o$ and $\mathbf{y}_i^m$, and both matrices are fully observed.

Specifically, to calculate

$$\begin{aligned}
I_1 &= \int\int \mathbf{e}_i'\mathbf{R}_i^{-1}\mathbf{e}_i f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathrm{d}b_i \mathrm{d}\mathbf{y}_i^m \\
&= \mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) \mathbf{R}_i^{-1} \mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{tr}\left(\mathbf{R}_i^{-1}\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\right),
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) - \mathbf{X}_i\boldsymbol{\alpha} - \mathbf{Z}_i\boldsymbol{\tau}\mathrm{E}\left(b_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right), \\
\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathbf{R}_i^{(t)} - \mathbf{R}_i^{(t)}\boldsymbol{\Sigma}_i^{-(t)}\mathbf{R}_i^{(t)} + \mathbf{R}_i^{(t)}\boldsymbol{\Sigma}_i^{-(t)}\mathrm{var}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\boldsymbol{\Sigma}_i^{-(t)}\mathbf{R}_i^{(t)},
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \begin{pmatrix} \mathbf{y}_i^o \\ \mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) \end{pmatrix}, \\
\mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathbf{X}_i^m\boldsymbol{\alpha}^{(t)} \\
&\quad +\boldsymbol{\Sigma}_{i,mo}^{(t)}\boldsymbol{\Sigma}_{i,oo}^{-(t)}\left(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\alpha}^{(t)}\right) + \phi_1^{(t)}\mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{1}/n_i, \\
\mathrm{var}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \begin{pmatrix} 0 & 0 \\ 0 & \mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) \end{pmatrix}, \\
\mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \boldsymbol{\Sigma}_{i,mm}^{(t)} - \boldsymbol{\Sigma}_{i,mo}^{(t)}\boldsymbol{\Sigma}_{i,oo}^{-(t)}\boldsymbol{\Sigma}_{i,om}^{(t)},
\end{aligned}$$

and $\mathrm{E}\left(b_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$ is defined below. For the multivariate mixed-effects model with factor-analytic random effects, we let $\boldsymbol{\Sigma}_i = \mathrm{var}\left(\mathbf{y}_i\right) = \mathbf{Z}_i\boldsymbol{\tau}\boldsymbol{\tau}'\mathbf{Z}_i' + \mathbf{R}_i$ and $\boldsymbol{\Sigma}_i^{(t)} = \mathbf{Z}_i\boldsymbol{\tau}^{(t)}\boldsymbol{\tau}'^{(t)}\mathbf{Z}_i' + \mathbf{R}_i^{(t)} = \begin{pmatrix} \boldsymbol{\Sigma}_{i,oo}^{(t)} & \boldsymbol{\Sigma}_{i,om}^{(t)} \\ \boldsymbol{\Sigma}_{i,mo}^{(t)} & \boldsymbol{\Sigma}_{i,mm}^{(t)} \end{pmatrix}$.

Similarly, to calculate

$$\begin{aligned}
I_2 &= \int\int b_i' b_i f\left(\mathbf{y}_i^m, b_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathrm{d}b_i\mathrm{d}\mathbf{y}_i^m \\
&= \mathrm{E}\left(b_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathrm{E}\left(b_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) + \mathrm{var}\left(b_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right),
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{E}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) &= \mathbf{B}_i\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) - \mathbf{X}_i\boldsymbol{\alpha}^{(t)}\right\}, \\
\mathrm{var}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) &= \mathrm{var}\left(b_i|\mathbf{y}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathbf{B}_i\mathrm{var}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\mathbf{B}_i',
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{B}_i &= \mathrm{var}\left(b_i|\mathbf{y}_i, \boldsymbol{\gamma}^{(t)}\right)\boldsymbol{\tau}'^{(t)}\mathbf{Z}_i'\mathbf{R}_i^{-(t)}, \\
\mathrm{var}\left(b_i|\mathbf{y}_i, \boldsymbol{\gamma}^{(t)}\right) &= \left(1 + \boldsymbol{\tau}'^{(t)}\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\mathbf{Z}_i\boldsymbol{\tau}^{(t)}\right)^{-1}.
\end{aligned}$$

For the missing-data mechanism, we have

$$\begin{aligned}
I_3 &= \int\left(\sum_{k=1}^{K}\log f\left(r_{ik}|\mathbf{y}_{ik}\right)\right) f\left(\mathbf{y}_i^m|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\mathrm{d}\mathbf{y}_i^m \\
&= K_i^m\phi_0 + \phi_1/n_i\cdot\mathbf{1}'\mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) + \sum_{k\in\mathbf{O}_i}\log\left\{1 - \exp\left(\phi_0 + \phi_1/n_i\cdot\mathbf{1}'\mathbf{y}_{ik}\right)\right\},
\end{aligned}$$

where K_i^m is the number of peptides missing in experiment i, and $\mathbf{O}_i$ denotes the index set of the observed columns (peptides) of $\mathbf{Y}_i$.

The expected complete-data log-likelihood for all experiments is thus given by

$$Q\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right) = \sum_{i=1}^{N} Q_i\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right).$$

The M-step of the EM algorithm

In the M-step, we maximized the expected complete-data log-likelihood $Q\left(\boldsymbol{\gamma}\,|\,\boldsymbol{\gamma}^{(t)}\right)$ derived in the E-step with respect to each parameter ($\boldsymbol{\gamma} = \{\boldsymbol{\alpha}, \boldsymbol{\tau}, \sigma_0^2, \sigma^2, \phi_0, \phi_1\}$) to find solutions of the maximum likelihood estimates (MLEs).

We obtained the estimates of the variance components for the random effects as

$$\begin{aligned}\boldsymbol{\tau}^{(t+1)} &= \left\{\sum_{i=1}^{N} \mathrm{E}^2\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\mathbf{Z}_i\right\}^{-1} \\ &\quad \sum_{i=1}^{N}\left[\mathrm{E}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) - \mathbf{X}_i\boldsymbol{\alpha}^{(t)}\right\}\right].\end{aligned}$$

The estimates of the fixed effects were given by

$$\begin{aligned}\boldsymbol{\alpha}^{(t+1)} &= \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{R}_i^{-(t)}\mathbf{X}_i\right)^{-1} \\ &\quad \sum_{i=1}^{N}\left[\mathbf{X}_i'\mathbf{R}_i^{-(t)}\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) - \mathbf{Z}_i\boldsymbol{\tau}^{(t)}\mathrm{E}\left(b_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\right\}\right].\end{aligned}$$

The variance estimates of the error term were

$$\begin{aligned}\sigma_0^{2(t+1)} &= \frac{1}{NK}\sum_{i=1}^{N}\mathbf{V}_{i,11}^{(t)}, \\ \sigma^{2(t+1)} &= \frac{1}{K\left(\sum_{i=1}^{N} n_i - N\right)}\sum_{i=1}^{N}\mathrm{tr}\left(\mathbf{V}_{i,-1,-1}^{(t)}\right),\end{aligned}$$

where $\mathbf{V}_i^{(t)} = \mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)\mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right) + \mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)$. Here $\mathbf{V}_{i,11}^{(t)}$ was the element among the n_i diagonal elements of $\mathbf{V}_i^{(t)}$ corresponding to the reference sample. The rest of the elements $\mathbf{V}_{i,-1,-1}^{(t)}$ were corresponding to the other tumor samples.

We can estimate the parameters of the missing-data mechanism, ϕ_0 and ϕ_1, by directly maximizing the expected log-likelihood function based on the complete data through an optimization algorithm. However, convergence issues may arise when the estimated probability in the exponential model (3.3) equals or exceeds one. To avoid this issue, following Lumley et al. (2006), we used a Poisson working model specifically for the purpose of estimating

missing-data mechanism parameters and obtained consistent point estimates.

To monitor the convergence of the proposed EM algorithm, we derived the observed-data log-likelihood function. We set convergence criteria as the relative change in the log-likelihood being smaller than 0.00001. The observed-data log-likelihood for our model is given by

$$\begin{aligned} l^o(\boldsymbol{\gamma}) &= const + \sum_{i=1}^{N} \sum_{k \in \mathbf{O}_i} \log \left\{ 1 - \exp\left(\phi_0 + \phi_1 / \mathrm{n_i} \cdot \mathbf{1}^{'} \mathbf{y}_{\mathrm{ik}} \right) \right\} \\ & - \frac{1}{2} \sum_{i=1}^{N} \log \left| \boldsymbol{\Sigma}_{i,oo} \right| - \frac{1}{2} \sum_{i=1}^{N} \left(\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\alpha} \right)^{'} \boldsymbol{\Sigma}_{i,oo}^{-1} \left(\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\alpha} \right) \\ & + \sum_{i=1}^{N} \left[K_i^m \phi_0 + \phi_1 / n_i \cdot \mathbf{1}^{'} \left\{ \mathrm{E}\left(\mathbf{y}_i^m | \mathbf{y}_i^o \right) + \phi_1 / n_i \cdot \mathrm{var}\left(\mathbf{y}_i^m | \mathbf{y}_i^o \right) \mathbf{1} / 2 \right\} \right], \end{aligned}$$

where $\mathbf{O}_i$ denoted the set of indices for the observed features/outcomes (peptides) of $\mathbf{Y}_i$, and K_i^m is the number of missing peptides in the experiment i.

At the convergence of the EM algorithm, we derived the variance for the estimated fixed effects from the estimated information matrix of the observed-data log-likelihood function,

$$\widehat{\mathrm{var}}\left(\hat{\boldsymbol{\alpha}} \right) = \left(\sum_{i=1}^{N} \mathbf{X}_i^{o'} \boldsymbol{\Sigma}_{i,oo}^{-1} \mathbf{X}_i^o \right)^{-1}.$$

One may construct the Wald statistics for testing the fixed effects.

3.3.2 A penalized EM-ADMM algorithm for the mvMISE$_e$ model

In this section, we derived an EM algorithm for the proposed mvMISE$_e$ model with correlated error terms. The penalized estimation of the precision matrix only affected the M-step of the penalized EM algorithm. We employed an alternating direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) for estimating the precision matrix and provided the

estimation details within the EM algorithm framework.

The E-step of the EM algorithm

Conditional on the observed data and current parameter estimates $\boldsymbol{\gamma}^{(t)}$, the expected complete-data log-likelihood for the i-th batch is

$$\begin{aligned} Q_i\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{E}\left(\ell_i\left(\boldsymbol{\gamma};\mathbf{y}_i,\mathbf{b}_i,\mathbf{r}_i\right)|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) \\ &= \int\int \log\left[f\left(\mathbf{y}_i|\mathbf{b}_i\right)\right] f\left(\mathbf{y}_i^m,\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) d\mathbf{b}_i d\mathbf{y}_i^m \\ &\quad + \int\int \log\left[f\left(\mathbf{b}_i\right)\right] f\left(\mathbf{y}_i^m,\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) d\mathbf{b}_i d\mathbf{y}_i^m \\ &\quad + \int\int \log\left[f\left(\mathbf{r}_i|\mathbf{y}_i\right)\right] f\left(\mathbf{y}_i^m,\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) d\mathbf{b}_i d\mathbf{y}_i^m \\ &= I_1 + I_2 + I_3. \end{aligned}$$

To calculate I_1, I_2, I_3, we have

$$\begin{aligned} \mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}^{(t)}\right) \\ \mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right) &= \left(\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\mathbf{Z}_i + \mathbf{D}^{-(t)}\right)^{-1}, \end{aligned}$$

where $\mathbf{R}_i = \boldsymbol{\Sigma}\otimes\mathbf{S}_i$. Let $\mathbf{y}_i = \left(\mathbf{y}_i^o,\mathbf{y}_i^m\right), \mathbf{Z}_i\mathbf{D}^{(t)}\mathbf{Z}_i' + \mathbf{R}_i^{(t)} = \begin{pmatrix} \mathbf{V}_{i,oo}^{(t)} & \mathbf{V}_{i,om}^{(t)} \\ \mathbf{V}_{i,mo}^{(t)} & \mathbf{V}_{i,mm}^{(t)} \end{pmatrix}$.

It follows that

$$\begin{aligned} \mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\boldsymbol{\gamma}^{(t)}\right) + \phi_1^{(t)}\mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\boldsymbol{\gamma}^{(t)}\right)\mathbf{1}/n_i \\ \mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\boldsymbol{\gamma}^{(t)}\right) \\ \mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\boldsymbol{\gamma}^{(t)}\right) &= \mathbf{X}_i^m\boldsymbol{\beta}^{(t)} + \mathbf{V}_{i,mo}^{(t)}\mathbf{V}_{i,oo}^{-(t)}\left(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta}^{(t)}\right) \\ \mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\boldsymbol{\gamma}^{(t)}\right) &= \mathbf{V}_{i,mm}^{(t)} - \mathbf{V}_{i,mo}^{(t)}\mathbf{V}_{i,oo}^{-(t)}\mathbf{V}_{i,om}^{(t)}, \end{aligned}$$

thus

$$\begin{aligned}
\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)-\mathbf{X}_i\boldsymbol{\beta}^{(t)}\right\}\\
\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right) &= \mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{Z}_i'\mathbf{R}_i^{-(t)}\mathrm{var}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\\
&\quad \mathbf{R}_i^{-(t)}\mathbf{Z}_i\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right),
\end{aligned}$$

where $\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)=\begin{pmatrix}\mathbf{y}_i^o\\ \mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\end{pmatrix}$, and

$\mathrm{var}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)=\begin{pmatrix}0 & 0\\ 0 & \mathrm{var}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\end{pmatrix}$.

Then we can obtain the following components of the log-likelihood function

$$\begin{aligned}
I_1 &= const-\frac{1}{2}\log|\mathbf{R}_i|-\frac{1}{2}\mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{R}_i^{-1}\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\\
&\quad -\frac{1}{2}\mathrm{tr}\left\{\mathbf{R}_i^{-1}\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\right\},\\
I_2 &= const-\frac{1}{2}\log|\mathbf{D}|-\frac{1}{2}\mathrm{tr}\left\{\mathbf{D}^{-1}\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\right\}\\
&\quad -\frac{1}{2}\mathrm{E}\left\{\mathrm{E}\left(\mathbf{b}_i'|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{D}^{-1}\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right\},\\
I_3 &= K_i^m\phi_0+\phi_1/n_i\cdot\mathbf{1}'\mathrm{E}\left(\mathbf{y}_i^m|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)+\sum_{k\in\mathbf{O}_i}\log\left\{1-\exp\left(\phi_0+\phi_1/\mathrm{n_i}\cdot\mathbf{1}'\mathbf{y}_{\mathrm{ik}}\right)\right\},
\end{aligned}$$

where $\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)=\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)$, $\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$ is defined in Section 3.3.1, and $\mathbf{O}_i$ denotes the set of indices of the observed columns of $\mathbf{Y}_i$.

The expected complete-data log-likelihood for all batches is given by

$$Q\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right)=\sum_{i=1}^{N}Q_i\left(\boldsymbol{\gamma}|\boldsymbol{\gamma}^{(t)}\right).$$

The M-step of the EM algorithm

In the M-step, we first obtained the MLEs for the missing-data mechanism parameters by directly maximizing the expected complete-data log-likelihood through a Poisson working model (Lumley et al., 2006), as discussed in Section 3.3.1. We then obtained the estimates of the covariance matrix for the random effects

$$\mathbf{D}^{(t+1)} = \frac{1}{N}\sum_{i}\left\{\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathrm{E}\left(\mathbf{b}_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)+\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)+\mathrm{var}\left(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\gamma}^{(t)}\right)\right\},$$

and the estimates of the fixed effects

$$\boldsymbol{\alpha}^{(t+1)} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{R}_i^{-(t)}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{N}\left[\mathbf{X}_i'\mathbf{R}_i^{-(t)}\left\{\mathrm{E}\left(\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)-\mathbf{Z}_i\mathrm{E}\left(\mathbf{b}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\right\}\right],$$

where $\boldsymbol{\gamma}=\{\boldsymbol{\alpha},\mathbf{D},\sigma_0^2,\sigma^2,\boldsymbol{\Sigma},\phi_0,\phi_1\}$ here.

The estimates of the error variances were obtained from the derivative of the log-likelihood function associated with $\mathbf{R}_i$ with respect to σ_0^2,σ^2 and $\boldsymbol{\Sigma}$,

$$\begin{aligned} l\left(\sigma_0^2,\sigma^2,\boldsymbol{\Sigma}\right) &= const-\frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{R}_i|-\frac{1}{2}\sum_{i=1}^{N}\mathrm{E}\left(\mathbf{e}_i'|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\mathbf{R}_i^{-1}\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\\ &\quad -\frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left\{\mathbf{R}_i^{-1}\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)\right\}. \end{aligned}$$

The derivative of the term associated with the quadratic form of $\mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$ can be calculated by following Glanz and Carvalho (2013). To obtain closed-form solutions for variance parameters, with Kronecker products ($\mathbf{R}_i=\boldsymbol{\Sigma}\otimes\mathbf{S}_i$) involved in I_1, we rewrote $\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)$ in the form of the Kronecker product singular value decomposition (Van Loan, 2000). Specifically, $\mathrm{var}\left(\mathbf{e}_i|\mathbf{y}_i^o,\mathbf{r}_i,\boldsymbol{\gamma}^{(t)}\right)=\sum_k\delta_{ik}\mathbf{G}_{ik}\otimes\mathbf{H}_{ik}$. In this way, based on the properties of the Kronecker product, we derived the closed-form estimated variances

in the error term as

$$\begin{aligned}\sigma_0^{2(t+1)} &= \frac{1}{NK}\sum_{i=1}^{N}\tilde{\mathbf{V}}_{i,11}^{(t)},\\ \sigma^{2(t+1)} &= \frac{1}{K\left(\sum_{i=1}^{N} n_i - N\right)}\sum_{i=1}^{N}\mathrm{tr}\left(\tilde{\mathbf{V}}_{i,-1,-1}^{(t)}\right),\end{aligned}$$

where $\tilde{\mathbf{V}}_i^{(t)} = \tilde{\mathbf{E}}_i\boldsymbol{\Sigma}^{-(t)}\tilde{\mathbf{E}}_i' + \sum_k \delta_{ik}\mathrm{tr}\left(\mathbf{G}_{ik}\mathbf{S}_i^{-(t)}\right)\mathbf{H}_{ik}'$, and $\tilde{\mathbf{E}}_i$ is an $n_i \times K$ matrix with $\mathrm{vec}(\tilde{\mathbf{E}}_i) = \mathrm{E}\left(\mathbf{e}_i|\mathbf{y}_i^o, \mathbf{r}_i, \boldsymbol{\gamma}^{(t)}\right)$. As noted in Glanz and Carvalho (2013), there is a non-identifiability issue in the error covariance matrix $\boldsymbol{\Sigma}\otimes\mathbf{S}_i$ introduced by the Kronecker product. To resolve this issue, we scaled $\sigma^{2(t+1)}$ with $\sigma_0^{2(t+1)}$ and set $\sigma_0^{2(t+1)} = 1$.

The major challenge in the mvMISE$_e$ model is the error precision matrix, $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$. It has a dimensionality equaling to the number of outcome variables (K) analyzed simultaneously in the multivariate model, and thus needs special treatments when K is large. To estimate $\boldsymbol{\Theta}$, we first rewrote the log-likelihood function as a function of the error precision matrix:

$$\begin{aligned}l(\boldsymbol{\Theta}) &= const + \frac{1}{2}\sum_{i=1}^{N} n_i \log|\boldsymbol{\Theta}| - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left(\tilde{\mathbf{E}}_i'\mathbf{S}_i^{-1}\tilde{\mathbf{E}}_i\boldsymbol{\Theta}\right)\\ &\quad -\frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left\{\boldsymbol{\Theta}\otimes\mathbf{S}_i^{-1}\left(\sum_k \delta_{ik}\mathbf{G}_{ik}\otimes\mathbf{H}_{ik}\right)\right\}. \qquad (3.6)\end{aligned}$$

We then added a graphical lasso penalty (Friedman et al., 2008) on $\boldsymbol{\Theta}$ to the log-likelihood function and maximized the penalized log-likelihood in equation (3.5).

Following Boyd et al. (2011) and Danaher et al. (2014), we used an ADMM algorithm to solve this problem. The problem in (3.5) is equivalent to

$$\mathrm{minimize}_{\boldsymbol{\Theta}} \quad -2l(\boldsymbol{\Theta}) + \lambda N|\mathbf{T}|_1$$

subject to $\mathbf{T} = \mathbf{\Theta}$. Different from the method of Lagrange multipliers, with an additional penalty term (the augmentation) and a scaled Lagrange multiplier matrix $\mathbf{U}$ by a penalty parameter ρ, we wrote the scaled augmented Lagrangian (Boyd et al., 2011) as

$$L_\rho(\mathbf{\Theta}, \mathbf{T}, \mathbf{U}) = -2l(\mathbf{\Theta}) + \lambda N |\mathbf{T}|_1 + \frac{\rho N}{2} \|\mathbf{\Theta} - \mathbf{T} + \mathbf{U}\|_F^2 - \frac{\rho N}{2} \|\mathbf{U}\|_F^2,$$

where we used $\rho = 1$. The idea is to minimize $L_\rho(\mathbf{\Theta}, \mathbf{T}, \mathbf{U})$ with respect to $\mathbf{\Theta}$ and $\mathbf{T}$ respectively and then update $\mathbf{U}$ at each iteration. The detailed ADMM algorithm is described in Algorithm 3. We used it to re-estimate the regularized $\mathbf{\Theta}$ within each iteration of the M-step. The tuning parameter (λ) can be selected via the Akaike information criterion (Danaher et al., 2014).

3.4 Simulations

3.4.1 Comparing the biases and mean squared errors of fixed effect estimates

In this section, we conducted simulations to assess the biases and mean squared errors (MSEs) of the proposed models versus competing methods. We compared the biases and MSEs of the fixed effect estimates based on: 1) the standard practice in the current proteomics literature using linear regression (`lm`) on the relative abundance measures of observed abundances relative to the reference sample abundances; 2) the univariate mixed-effects model (`mixEMM`) proposed by (Chen et al., 2017b); and 3) the proposed multivariate selection models (`mvMISE`).

Algorithm 3 The ADMM algorithm for re-estimating regularized $\mathbf{\Theta}$ within the M-step

1. Initialize with $\mathbf{\Theta} = \mathbf{I}, \mathbf{U} = \mathbf{T} = \mathbf{0}$.
2. Minimize
$$-2l\left(\mathbf{\Theta}\right) + \frac{\rho N}{2}\left\|\mathbf{\Theta} - \mathbf{T} + \mathbf{U}\right\|_F^2$$
with respect to $\mathbf{\Theta}$. Let $\mathbf{\Lambda\Omega\Lambda}^{'}$ be the eigendecomposition of $\frac{1}{\sum_{i=1}^N n_i}\sum_{i=1}^N \mathbf{W}_i + \frac{\rho N\left(\mathbf{U}^{(t)} - \mathbf{T}^{(t)}\right)}{\sum_{i=1}^N n_i}$, where $\mathbf{W}_i = \tilde{\mathbf{E}}_i^{'}\mathbf{S}_i^{-(t)}\tilde{\mathbf{E}}_i + \sum_k \delta_{ik}\mathrm{tr}\left(\mathbf{H}_{ik}\mathbf{S}_i^{-(t)}\right)\mathbf{G}_{ik}^{'}$. We have the estimate $\mathbf{\Theta}^{(t+1)}$ as $\mathbf{\Lambda}\tilde{\mathbf{\Omega}}\mathbf{\Lambda}^{'}$, where $\tilde{\mathbf{\Omega}}$ is a diagonal matrix with the i-th diagonal element as
$$\frac{\sum_{i=1}^N n_i}{2\rho N}\left(-\omega_{ii} + \sqrt{\omega_{ii}^2 + \frac{4\rho N}{\sum_{i=1}^N n_i}}\right),$$
where ω_{ii} is the i-th diagonal element of $\mathbf{\Omega}$.
3. Minimize
$$\lambda N\left|\mathbf{T}\right|_1 + \frac{\rho N}{2}\left\|\mathbf{T} - \mathbf{\Theta} - \mathbf{U}\right\|_F^2$$
with respect to $\mathbf{T}$, where $|\mathbf{T}|_1 = \sum_{i\neq j}\left|\mathbf{T}_{ij}\right|$. Let $\mathbf{A} = \mathbf{\Theta} + \mathbf{U}$. We have $Z_{ii}^{(t+1)} = A_{ii}^{(t)}, i = 1, \ldots, K$, for diagonal elements, and for $i \neq j$
$$Z_{ij}^{(t+1)} = \mathrm{sgn}\left(A_{ij}^{(t)}\right)\left(\left|A_{ij}^{(t)}\right| - \lambda/\rho\right)_+.$$
4. Update
$$\mathbf{U}^{(t+1)} = \mathbf{U}^{(t)} + \mathbf{\Theta}^{(t+1)} - \mathbf{T}^{(t+1)}.$$
5. Iterate Step 2-4 until convergence.

The biases and MSEs for the mvMISE$_b$ model

To assess the performance of competing methods in analyzing multiple highly correlated peptide abundances from the same protein, we simulated multivariate incomplete data based on models (3.3) and (3.4), with a fixed intercept (α_0) and a fixed effect of interest (α_1). Here we assume different peptides in the same protein share a common α_1. We simulated a total of 108 samples from 36 clusters/batches, where each batch consists of 4 samples including one common reference sample. This is the same sample size as in the motivating CPTAC data. We set the parameters as $\alpha_0 = 10$, $\phi_0 = 0$ and $\phi_1 = -0.025$ to mimic the data structure that has been observed in the CPTAC data. The average rate of missing values is 37.5%. We simulated a binary predictor of interest. We specified the variance components of factor-analytic random effect $\boldsymbol{\tau}$ to have elements $\tau = \sqrt{7U/3}$, where $U \sim \text{Unif}\left[\sigma_0^2 - 1, \sigma_0^2 + 1\right]$.

To assess the biases and MSEs of α_1 estimates in different settings, we vary the number of outcomes (K), the variance components (σ_0^2, σ^2), and the true value of α_1. In each setting, we simulate 1000 datasets for replication. Note that we kept the variance of the reference sample (σ_0^2) to be half of the variance of the tumor samples (σ^2). This is because the reference sample is usually the same biological sample across different batches and there should be no biological variation, other than experimental perturbation across batches. The ratio of the two variances is similar to what we have observed in the CPTAC data. As both variance components increase, the total variation of the data becomes larger, and the estimation biases are larger. Also note that when using the `lm` and `mixEMM` methods to analyze the data, since both approaches are univariate analysis methods, we took the average of the abundance levels of multiple peptides from each protein in each sample and analyzed its association with the predictor using `lm` and `mixEMM`. Table 3.1 shows that the proposed mvMISE$_b$ method has consistently smaller biases and MSEs in different settings as compared to the standard practice based on relative abundances (`lm`) and the `mixEMM` method. The methods with the smallest biases or MSEs are shown in bold in the table.

Table 3.1: The biases and MSEs of the fixed effect estimates for α_1 based on data that were generated with correlated random effects. We varied the number of outcomes (K, which here refers to the number of peptides in a protein), the variance components and total variation (σ_0^2, σ^2), and the effect size of α_1. We compared the proposed mvMISE$_b$ with `lm` (linear regression based on relative abundance) and `mixEMM` (Chen et al., 2017b). We set $\alpha_0 = 10$, $\phi_0 = 0$, $\phi_1 = -0.025$ and the average missing rate is 37.5%. The results were based on 1000 replications. The smallest biases or MSEs in different settings are shown in boldface.

				bias			MSE		
K	σ_0^2	σ^2	α_1	lm	mixEMM	mvMISE$_b$	lm	mixEMM	mvMISE$_b$
5	1	2	0.7	0.014	0.022	**0.011**	0.046	0.036	**0.031**
5	2	4	0.7	0.003	0.022	**0.002**	0.097	0.076	**0.064**
5	3	6	0.7	-0.004	0.026	**0.000**	0.159	0.127	**0.104**
5	4	8	0.7	0.004	0.032	**-0.003**	0.228	0.176	**0.142**
5	1	2	0	0.016	0.022	**0.012**	0.046	0.036	**0.031**
5	2	4	0	**0.016**	0.030	0.017	0.100	0.081	**0.066**
5	3	6	0	**0.003**	0.029	0.011	0.155	0.132	**0.107**
5	4	8	0	**0.008**	0.051	0.012	0.191	0.159	**0.124**
5	1	2	0	0.016	0.022	**0.012**	0.046	0.036	**0.031**
10	1	2	0	0.002	0.008	**0.000**	0.023	0.017	**0.016**
20	1	2	0	0.002	0.009	**0.001**	0.011	**0.008**	**0.008**
50	1	2	0	**0.001**	0.009	**0.001**	0.005	0.004	**0.003**
5	1	2	0.2	0.015	0.023	**0.011**	0.046	0.036	**0.031**
10	1	2	0.2	0.002	0.008	**0.000**	0.023	0.017	**0.016**
20	1	2	0.2	0.002	0.009	**0.000**	0.011	**0.008**	**0.008**
50	1	2	0.2	**0.001**	0.009	**0.001**	0.005	0.004	**0.003**

The biases and MSEs for the mvMISE$_e$ model

Next, we simulated multivariate correlated outcomes with a sparse inverse covariance matrix to compare the performance of the proposed mvMISE$_e$ model with competing methods in analyzing multiple proteins from a *a priori* defined functional pathway. Each simulation below was repeated 1000 times. In each of the simulated data set, we simulated K proteins based on the model (3.1) with a random intercept and error term $\mathbf{e}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{S}_i)$, where $\boldsymbol{\Sigma}$ captures the biological correlation among proteins. Following Peng et al. (2009) and Danaher et al. (2014), we simulated a sparse biological correlation structure from a power-law degree distribution. We first generated a network with a degree sequence simulated from a power-law distribution $f(d) = d^{-2.3}$, where d is the node degree. The network was then converted to an adjacency matrix $\tilde{\boldsymbol{\Delta}}$, with diagonal elements and elements corresponding to edges denoted as 1, otherwise denoted as 0. We then multiplied the nonzero off-diagonal elements with a random number simulated from Unif $([-1, -0.5] \cup [0.5, 1])$. To ensure positive definiteness, we rescaled $\tilde{\boldsymbol{\Delta}}$ by dividing each off-diagonal element over 1.5 fold of the sum of the absolute value of off-diagonal elements in that row. We then let $\boldsymbol{\Delta} = \left(\tilde{\boldsymbol{\Delta}} + \tilde{\boldsymbol{\Delta}}'\right)/2$ to assure symmetry. The ij-th element of the error covariance matrix $\boldsymbol{\Sigma}$ is determined as $\boldsymbol{\Delta}^{-1}_{ij} / \sqrt{\boldsymbol{\Delta}^{-1}_{ii} \boldsymbol{\Delta}^{-1}_{jj}}$.

We simulated a fixed intercept (α_0) and a fixed effect of interest (α_1) that is common to all proteins within a pathway. Note that the fixed effects for different proteins in a pathway need not be the same. This specification is only for the ease of comparing biases and MSEs of different methods in the current setting. We set the parameters in the models (3.1) and (3.3) to be $\alpha_0 = 10$, $\phi_0 = 0$, and $\phi_1 = -0.025$. The predictor of interest is simulated the same way as in the previous section.

To assess the biases and MSEs of different estimates for α_1, we varied the number of outcomes/proteins (K) in each pathway and the true value of α_1. Table 3.2 shows that the proposed mvMISE$_e$ model has the smallest MSEs as compared to competing methods in all

of the settings. Its biases are also small in most of the settings.

3.4.2 Comparing the type I error rates and power in testing for fixed effects

In this section, we compared the type I error rates and power in testing for fixed effects based on the proposed mvMISE methods versus the `lm` and the `mixEMM` methods.

The type I error rates and power for the mvMISE$_b$ model

To assess the type I error rates, we simulated the data similarly as in Section 3.4.1. We set $\phi_0 = 0$, $\phi_1 = -0.025$, and $\boldsymbol{\alpha} = (10, 0)^{'}$, and varied σ_0^2, σ^2 and K in different settings. In each simulation setting, we repeated the simulation 5,000 times to ensure accuracy.

We simulated two settings: (a) when $K = 5$, we increased the total variation; and (b) with fixed variance components, we increased the number of peptides in each protein (K) from 5 to 50. When applying the proposed mvMISE$_b$ model to the abundance levels of K simulated peptides, we first obtained the fixed effects estimates and then calculated the standard errors using the observed information matrix from the EM algorithm, and obtained the p-values based on the Wald Z-test (see Section 3.3.1 for more details). When using `lm` and `mixEMM` to analyze multiple peptides in each protein, again we first took the average of multiple peptide abundances in each sample, and analyzed the average abundances using the `lm` and the `mixEMM` methods. Both univariate methods ignored the correlations among multiple peptides.

Table 3.3 shows that the type I error rates for our mvMISE$_b$ model are well controlled at the nominal level (0.05). However, the linear regression analysis based on relative abundances has a deflated type I error rate in most cases, resulting in a conservative test. The univariate mixEMM model ignores the correlations among multiple peptides and has inflated type I error rates in some settings. As such, we suggest calculating permutation-based p-values for the two univariate analysis approaches, and we did so to compare the power of the methods.

With the permutation-based p-value calculation, the type I error rates of both univariate methods were controlled at the nominal level.

To compare the power, (a) when $K = 5$ with varying variation (σ_0^2, σ^2), we set $\alpha_1 = 0.7$; (b) with increasing K, we set $\alpha_1 = 0.2$. Table 3.4 shows that the proposed mvMISE$_b$ substantially improved the power for testing the fixed effect α_1 as compared to the univariate methods (including the permutation-based methods) in all settings.

As a summary, in comparison with univariate methods based on the average of multiple correlated outcomes, the proposed mvMISE$_b$ model accounts for the outcome variability and the structured correlations among the outcomes in jointly analyzing multiple peptides in each protein, and as such enjoys smaller biases/MSEs, well-controlled type I error rates and improved power in testing for fixed effects.

The type I error rates and power for the mvMISE$_e$ model

To assess the type I error rates and power of the mvMISE$_e$ model in pathway analyses of multiple proteins, we simulated the data similarly as in the Section 3.4.1. Different than before, here we simulated each protein in a pathway to have a protein-specific effect. All of the results were based on 5,000 replications.

In the analyses of the univariate methods, `lm` and `mixEMM`, we first took the average peptide abundances as the protein abundances, estimated the protein-specific effects using the corresponding methods, and calculated the protein-level p-values of the Wald statistics. We then combined those protein-level p-values for proteins in a pathway using Fisher's method. The p-values for each pathway are based on 5,000 permutations.

In the mvMISE$_e$ analysis, we also took the average peptide abundances as the protein abundances while jointly considered multiple proteins in the same pathway. We simultaneously estimated the Wald statistic for each protein in the pathway in an mvMISE$_e$ model and calculated the p-values. We then combined those protein-level p-values for proteins in

Table 3.2: The biases and MSEs of the fixed effect estimates for α_1 based on data that were generated with correlated error terms (in protein pathway analysis). We varied the pathway sizes (K) and the true values of α_1. We compared the proposed mvMISE$_e$ model with two univariate analysis methods, `lm` based on relative protein abundances and `mixEMM` (Chen et al., 2017b). We set $\alpha_0 = 10, \phi_0 = 0, \alpha_1 = -0.025$. The tuning parameter λ was fixed at 0.1. The results were based on 1000 replications. The smallest biases or MSEs in different settings are shown in boldface.

		bias			MSE		
K	α_1	lm	mixEMM	mvMISE$_e$	lm	mixEMM	mvMISE$_e$
5	0	**0.001**	0.010	**0.001**	0.038	0.028	**0.020**
10	0	0.007	0.013	**0.000**	0.026	0.021	**0.016**
20	0	**0.001**	0.008	-0.002	0.010	**0.008**	**0.008**
50	0	**0.001**	0.007	-0.002	0.004	**0.003**	**0.003**
5	0.25	**0.000**	0.010	0.001	0.038	0.027	**0.020**
10	0.25	0.006	0.012	**-0.001**	0.026	0.021	**0.017**
20	0.25	**0.001**	0.008	**-0.001**	0.010	0.008	**0.007**
50	0.25	**0.001**	0.007	-0.002	0.004	**0.003**	**0.003**

Table 3.3: The type I error rates of testing for fixed effect α_1 based on data that were generated with correlated random effects. We compared `lm`, `mixEMM`, their permutation-based tests (`lm_perm` and `mixEMM_perm`), and the proposed mvMISE$_b$ method in testing for fixed effects (α_1). We varied K and total variation (σ_0^2, σ^2). We set $\phi_0 = 0$, $\phi_1 = -0.025$, and $\boldsymbol{\alpha} = (10, 0)'$. The results were based on 5,000 replications in each setting. The significance level was 0.05.

K	σ_0^2	σ^2	lm	mixEMM	lm_perm	mixEMM_perm	mvMISE$_b$
5	1	2	0.036	0.058	0.045	0.046	0.056
5	2	4	0.036	0.058	0.049	0.052	0.051
5	3	6	0.041	0.060	0.053	0.055	0.053
5	4	8	0.037	0.055	0.052	0.041	0.049
10	1	2	0.037	0.052	0.049	0.049	0.046
20	1	2	0.036	0.057	0.050	0.057	0.050
50	1	2	0.040	0.063	0.056	0.057	0.053

the same pathway using Fisher's method. The p-values for each pathway are based on 5,000 permutations.

Note that here we used Fisher's method as the pathway analysis approach to aggregate protein-level p-values to pathways. One may use a different gene-set analysis approach, but it does not affect the conclusions of our comparisons.

To assess the type I error rates, we set all protein-specific effects to be zero. We calculated the permutation-based p-values for mvMISE$_e$ with a fixed tuning parameter of $\lambda = 0.05$ for all permutations. Table 3.5 shows that with permutations the type I error rates for `lm`, `mixEMM` and the proposed `mvMISE`$_e$ models were all well-controlled at the nominal level (0.05).

Table 3.5 also shows the power comparison. When comparing power, we simulated two settings: in the first setting, half of the proteins in each pathway were associated with the predictor of interest with effects $\alpha_k = 0.6$, and the rest of the proteins were not associated; and in the second setting, all of the proteins were associated with the predictor of interest with effect sizes sampled from a uniform distribution, $\alpha_k \sim \text{Unif}\,[-0.6, 0.6]$. The highest power in each setting is shown in boldface in Table 3.5. The proposed mvMISE$_e$ model is most powerful in all of the simulated settings.

3.5 Applications to the CPTAC breast cancer data

The CPTAC (http://proteomics.cancer.gov) is a comprehensive and coordinated effort launched by the National Cancer Institute (Paulovich et al., 2010; Ellis et al., 2013a; Mertins et al., 2016). The overall goal of CPTAC is to improve our ability to diagnose, treat and prevent cancer through the application of robust, quantitative, proteomic technologies and workflow. The consortium has recently conducted global proteome and phosphoproteome profiling of a subset of breast, colon and ovarian cancer samples that have been extensively characterized in The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov) (The Cancer Genome Atlas Network, 2012).

Table 3.4: The power of testing for fixed effect α_1 based on data that were generated with correlated random effects. We compared `lm`, `mixEMM`, their permutation-based tests (`lm_perm` and `mixEMM_perm`), and the proposed mvMISE$_b$ method in testing for fixed effects (α_1). Permutations are used to control the type I error rates for lm and mixEMM. Note that our mvMISE$_b$ method calculated the parametric p-values without using permutations. The results were based on 5,000 replications in each setting. The significance level was 0.05. The highest power in each setting is shown in boldface. Note that the `mixEMM` method without using permutation has inflated type I error rates, despite the slightly higher power in settings with larger K's.

K	α_1	σ_0^2	σ^2	lm	mixEMM	lm_perm	mixEMM_perm	mvMISE$_b$
5	0.7	1	2	0.855	0.954	0.884	0.915	**0.973**
5	0.7	2	4	0.553	0.750	0.595	0.692	**0.800**
5	0.7	3	6	0.374	0.575	0.426	0.502	**0.623**
5	0.7	4	8	0.297	0.465	0.344	0.392	**0.491**
5	0.2	1	2	0.121	0.205	0.141	0.179	**0.215**
10	0.2	1	2	0.224	**0.358**	0.265	0.321	**0.358**
20	0.2	1	2	0.418	**0.639**	0.453	0.570	0.621
50	0.2	1	2	0.837	**0.953**	0.865	0.905	0.950

Table 3.5: The type I error rates and power of Fisher's method-based pathway analyses. Data were simulated to have correlated error terms among multivariate outcomes. We set $\alpha_0 = 10, \phi_0 = 0, \phi_1 = -0.025, \sigma_0^2 = 1, \sigma^2 = 2$. We let $\alpha_k = 0$ when assessing the type I error rates. When comparing power, we simulated two settings: in the first setting, half of the proteins in each pathway were associated with the predictor of interest with effects $\alpha_k = 0.6$, and the rest of the proteins were not associated; and in the second setting, all of the proteins were associated with the predictor of interest with effect sizes sampled from a uniform distribution, $\alpha_k \sim \text{Unif}\,[-0.6, 0.6]$. All of the results were based on 5,000 replications. In analyzing the simulated data, we first obtained the permutation-based protein-level p-values for each method, combined those p-values for proteins in a pathway using Fisher's method, and then calculated the p-values for each pathway with 5,000 permutations. The highest power in each setting is shown in boldface. The proposed mvMISE$_e$ model is most powerful in all of the simulated settings.

	type I error rate			power: 50% signal			power: $\alpha_k \sim \text{Unif}\,[-0.6, 0.6]$		
K	lm	mixEMM	mvMISE$_e$	lm	mixEMM	mvMISE$_e$	lm	mixEMM	mvMISE$_e$
5	0.052	0.051	0.047	0.288	0.370	**0.440**	0.180	0.216	**0.283**
10	0.045	0.047	0.053	0.407	0.515	**0.602**	0.269	0.334	**0.422**
20	0.056	0.061	0.052	0.638	0.764	**0.841**	0.419	0.530	**0.638**
50	0.047	0.045	0.050	0.921	0.974	**0.987**	0.721	0.847	**0.915**

In this work, we focused on analyzing the phosphoproteomics data from the CPTAC breast cancer study (Mertins et al., 2016). Phosphorylation is a key post-translational modification and plays major roles in many biological processes. On the one hand, different phosphorylation sites (phosphopeptides) of one protein could induce different biological activities. On the other hand, the phosphopeptides from the same phosphoprotein could be highly correlated in terms of their abundances, because all of them are fragments of the same phosphoprotein. Most existing analyses were done by averaging the abundances of phosphopeptides mapped to the same phosphoprotein as the protein abundance level and analyzing the protein abundance level using univariate approaches. In our Analysis I, we treated the abundances of multiple phosphopeptides from a phosphoprotein as multivariate features and analyzed them using the mvMISE$_b$ method, and identified the phosphoproteins associated with TNBC subtype. In Analysis II, we jointly analyzed multiple phosphoproteins from each KEGG pathway using the mvMISE$_e$ method to identify the pathways enriched with phosphoproteins associated with TNBC subtype.

In the motivating CPTAC breast cancer data set (Mertins et al., 2016), a total of 108 tumor samples from 104 women with breast cancer and aged 26-90 were randomly assigned to 36 batches and were processed by 36 four-plex (i.e., four-channel) iTRAQ experiments. In each batch of samples, there were 3 tumor samples and one common reference sample created by combining 40 tumor samples. After quality control, in the following analyses, we focused on 77 tumor samples with superior data quality (Mertins et al., 2016). After standard data preprocessing (including log transformation, global normalization, etc.), we analyzed the abundance data for 25,961 phosphopeptides of high quality (Chen et al., 2017b), which correspond to 6,078 phosphoproteins. Figure 3.1(a) shows the histogram of the number of phosphopeptides for each phosphoprotein. There were 2,261 phosphoproteins with only one phosphopeptide. Most phosphoproteins had no more than five phosphopeptides, but the number of phosphopeptides can typically reach 25. There were two extremely large

phosphoproteins with 164 and 195 phosphopeptides. Figure 3.1(b) shows the average missing rate for each phosphoprotein. The missing rate typically ranged from 0 to 50%, with an average of 20%.

3.5.1 Analysis I: using mvMISE$_b$ to jointly analyze multiple phosphopeptides from one phosphoprotein

In this analysis, for each phosphoprotein, we treated the abundance levels of multiple phosphopeptides of this phosphoprotein as the multivariate feature measure. We used an indicator for TNBC tumor sample as the predictor of interest and also included an indicator for the reference sample. We applied the proposed mvMISE$_b$ method to each of the 6,078 phosphoproteins. We obtained p-values based on the Wald tests for non-zero effects of the TNBC indicator.

Figure 3.2 shows the Manhattan plot of the mvMISE$_b$-based p-values for the 6,078 phosphoproteins. At the Bonferroni-adjusted p-value threshold of 8.23×10^{-6}, there were 119 phosphoproteins detected to associate with TNBC subtype. The top ten significant protein names were also shown in Figure 3.2. Among the 119 significant proteins, 77 were uniquely identified by our method and were not identified by the standard practice of linear regression based on relative abundance protein measures using 10,000 permutations. The top protein is AHNAK, with a p-value of 1.93×10^{-53}. The gene *AHNAK* negatively regulates cell growth and acts as a tumor suppressor of the TGF-β signaling pathway. More recently, Chen et al. (2017a) also suggested that AHNAK suppresses tumor proliferation and invasion by targeting multiple pathways, including the MAPK signaling pathway, in triple-negative breast cancer. In our data, the protein AHNAK contains 195 phosphopeptides, and about half of them are individually significant at the p-value threshold of 0.05 if analyzed by `mixEMM`. This suggested that, when analyzing large proteins with many phosphopeptides, our approach showed an advantage over simply averaging all of the phosphopeptide abundances which

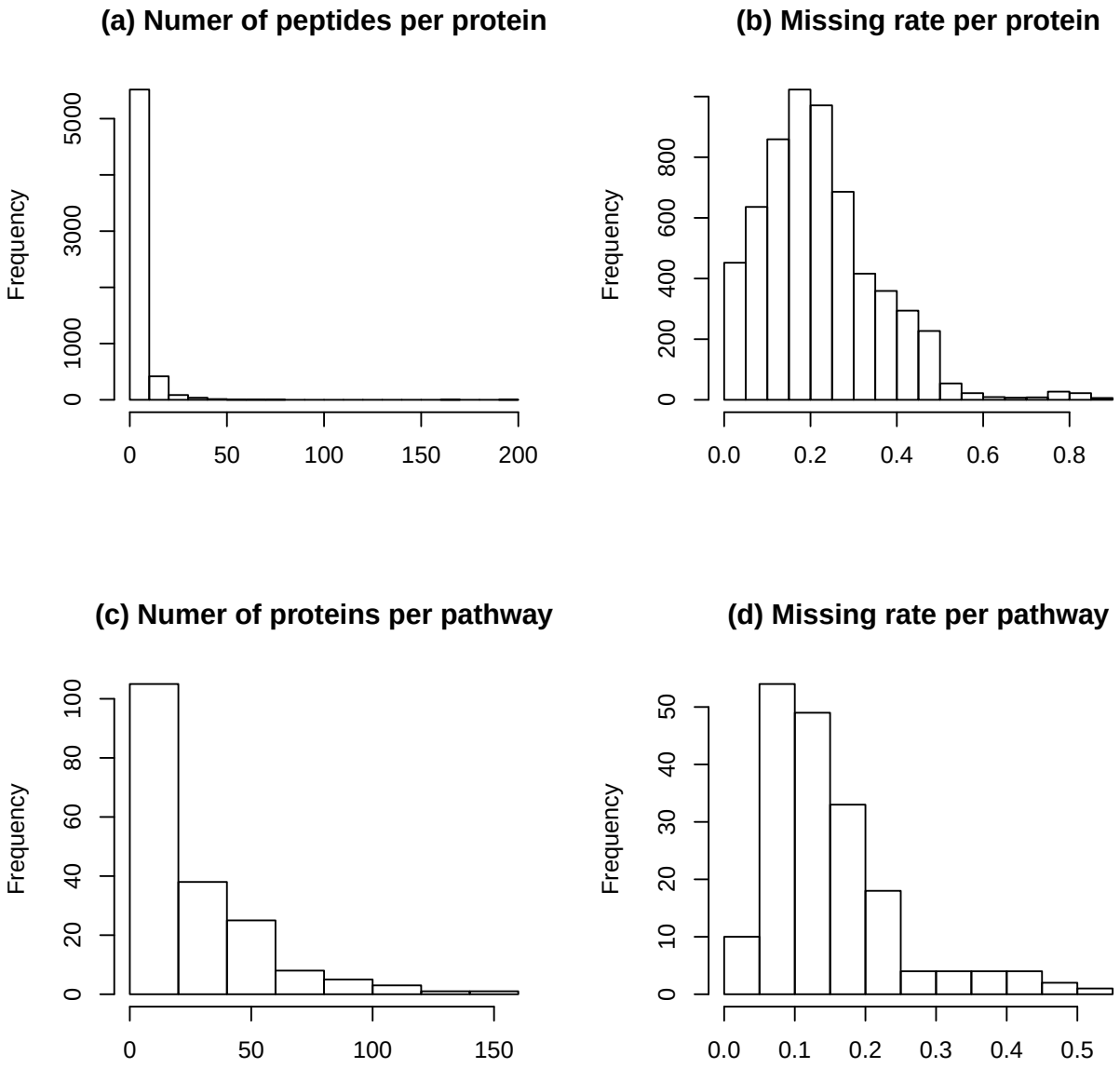

Figure 3.1: A summary of the CPTAC breast cancer data. (a) The histogram of the number of phosphopeptides in each phosphoprotein. There were two very large phosphoproteins with 164 and 195 phosphopeptides. (b) The histogram of the average missing rate for each phosphoprotein. The missing rate typically ranged from 0 to 50%. (c) The histogram of the number of proteins in each KEGG pathway. (d) The histogram of the average missing rate in each KEGG pathway.

may include many noisy abundances. Note that in this analysis, using the `mixEMM` method while controlling the type I error rate would require a large number of permutations, and each takes substantial computation time. As such, `mixEMM` is not pursued here. In contrast, our mvMISE$_b$ based on the parametric p-value calculation can well control the type I error rate and is computationally efficient.

3.5.2 Analysis II: using mvMISE$_e$ for protein pathway analyses

We applied the mvMISE$_e$ method to the KEGG human disease pathways (Kanehisa et al., 2016). Among the 186 KEGG pathways, there were 183 pathways that have at least one phosphoprotein being mapped in the CPTAC breast cancer data. Figure 3.1(c) shows the histogram of the number of phosphoproteins in each mapped KEGG pathway. Figure 3.1(d) shows the average missing rate for phosphoproteins in each pathway. The average missing rate ranged from almost 0 to over 50%. In this analysis, we focused on the 150 pathways that have 5 or more mapped phosphoproteins.

We first derived protein-level abundances as described earlier, and then apply the proposed mvMISE$_e$ method to jointly analyzes multiple phosphoprotein abundances. After standardizing protein abundances and treating them as multivariate features, we estimated the protein-specific effects on TNBC subtype by introducing the interactions between the protein indicators and the TNBC indicator and obtaining the protein-specific p-values based on Wald tests. For a pathway having K phosphoproteins, we estimated an intercept, the effect of the reference sample, and K protein-specific TNBC effects. Here we used Fisher's method as the pathway analysis method to aggregate the protein-level p-values.

Table 3.6 lists the significant KEGG pathways detected by the mvMISE$_e$ method at the Bonferroni-adjusted p-value cutoff of $0.1/150=6.67\times10^{-4}$. Among those pathways, the MAPK signaling pathway is known to be related to breast cancer risk and in particular TNBC. Giltnane and Balko (2014) reviewed the evidence supporting clinical trials of targeted

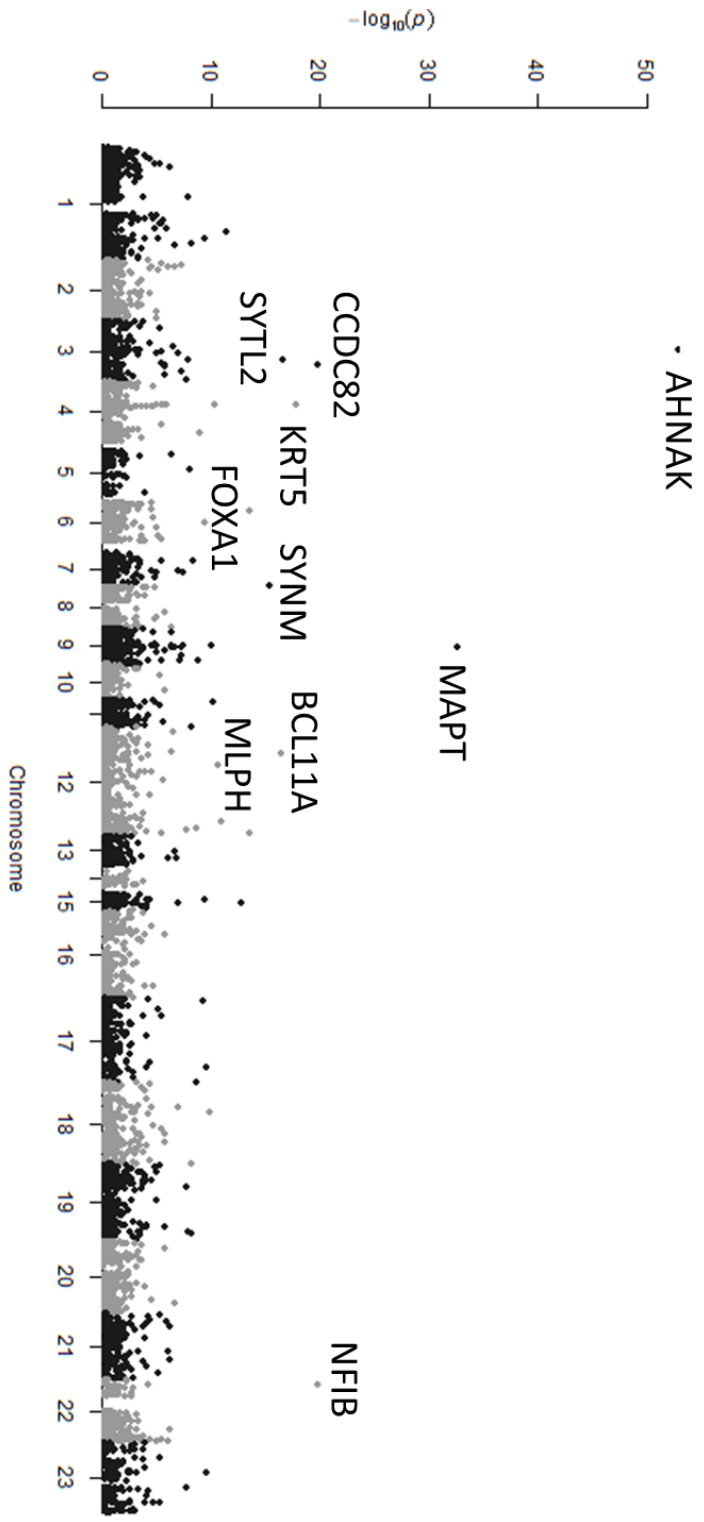

Figure 3.2: The Manhattan plot of the phosphoproteins identified by the proposed mvMISE$_b$ method. The top ten most significant protein names were listed.

inhibitors of the MAPK pathway in TNBC. Eighteen out of 130 proteins in this pathway showed marginally significant evidence of associations with TNBC.

3.6 Discussion

In this work, we proposed multivariate mixed-effects selection models for simultaneously analyzing multiple outcomes with clustered data and non-ignorable missing data. The proposed framework is flexible and generalizable, and we discussed two specific models tailored for two different types of multivariate analyses for labeling-based proteomics data. In the analyses of multiple phosphopeptides from a phosphoprotein, we introduced an $\texttt{mvMISE}_b$ model that allowed the random effects to be correlated to account for the correlations among multiple phosphopeptides. We adopted a factor-analytic outcome-specific random effects structure, which assumes identity correlations among random effects and greatly reduces the number of parameters to be estimated. This model is reasonable for this analysis because different phosphopeptides are different segments digested from the same protein, and are highly correlated in abundances. In the protein pathway analysis, the abundances among multiple proteins in a pathway were treated as multivariate outcomes, and the correlation structures among different proteins are often unstructured. We proposed an $\texttt{mvMISE}_e$ model, with a graphical lasso penalty on the error precision matrix to regularize the estimation of the precision matrix of multiple proteins from a function pathway. We derived an EM algorithm and a penalized EM-ADMM algorithm, respectively, for the estimations of the two models. Through simulations, we demonstrated the advantages of the proposed methods in reducing estimation biases, controlling type I error rates, and improving power in testing for fixed effects, in comparison with competing methods.

We applied the methods to the CPTAC breast cancer proteomic data set. We identified significant phosphoproteins that are related to TNBC risk and would otherwise not be detected by competing methods. We also identified pathways enriched with phosphoproteins

with differential abundances in TNBC tumors versus other tumors. The proposed multivariate methods can serve as complementary methods to the univariate analyses, and the analyses may provide additional insights into breast cancer and subtype etiology such as identifying new protein biomarkers.

In this work, we explicitly modeled the non-ignorable batch-level missing-data mechanism using an exponential function and incorporated the function in the likelihood. The modeling of this missing-data mechanism well characterized the missing-data mechanisms observed in labeling-based proteomic studies. For other types of missing-data mechanisms, one may modify the missing-data model and adjust the estimation in the EM estimation.

One caveat of the $\texttt{mvMISE}_e$-based pathway analysis is that we ignored the high correlations among phosphopeptides in each protein and took the average phosphopeptide abundances as the protein-level abundance measures; we then applied the $\texttt{mvMISE}_e$ model to multiple protein abundances in a pathway. We used permutations to calculate p-values in pathway analyses, so this is less of an issue for significance control. An alternative approach is to directly model the hierarchical structures of multiple phosphopeptides nested in each protein and multiple proteins nested in a pathway. The development of this hierarchical model requires the joint modeling of correlated random effects among multiple phosphopeptides from each protein and correlated error terms among multiple proteins in a pathway. It is challenging to estimate the large number of parameters in such a model. Some flexible Bayesian models may serve the purpose and those will be explored in future work.

For the CPTAC proteomic data analyses, we have a good understanding of the missing-data mechanism and as such we fit a selection model with a feature-dependent missing-data function. In other studies beyond labeling-based proteomic studies, there may be little information regarding the missing-data mechanism. One may also develop multivariate mixed-effects models that deal with clustered data with other types of non-ignorable missingness, such as shared-parameter models (Albert and Follmann, 2008), pattern-mixture models (Lit-

tle, 1994; Hedeker and Gibbons, 1997) and mixed-effects hybrid models (Little, 2008; Yuan and Little, 2009), for sensitivity analysis and model selection.

The cluster-level non-ignorable missing data may occur beyond quantitative proteomic studies. It may also happen in other areas with outcome vector dependent sampling (Schildcrout et al., 2013). The current work may also be modified and applied to general multivariate analyses based on clustered data with clustered outcome-dependent missingness.

Table 3.6: The significant KEGG pathways detected by the proposed mvMISE$_e$ method after Bonferroni-correction (p-value $\leq 0.1/150 = 6.67\times10^{-4}$). The results were based on 10,000 permutations.

KEGG human disease pathway name	No. of proteins	No. (%) proteins significant	mvMISE$_e$ p-values	The marginally significant proteins in the pathway
MAPK signaling pathway	130	18 (13.85)	0.0006	DUSP9, EGFR, FGFR3 FLNB, HSPA8, MAP2K4 MAP2K5, MAP3K1, MAP3K11 MAP3K4, MAPK11, MAPK8IP3 MAPT, MYC, NLK PDGFRB, PRKCA, TP53
non-small cell lung cancer	34	9 (26.47)	0.0005	AKT2, EGFR, ERBB2 FOXO3, PIK3R1, PIK3R5 PLCG2, PRKCA, RXRA
inositol phosphate metabolism	31	8 (25.81)	0.0004	IMPA1, INPP4A, INPP5E INPP5J, PIK3CG, PIP4K2B PLCG2, SYNJ1
thyroid cancer	17	3 (17.65)	0.0006	CTNNB1, MYC, RXRA

CHAPTER 4

A META-ANALYSIS APPROACH WITH FILTERING FOR IDENTIFYING GENE-LEVEL GENE-ENVIRONMENT INTERACTIONS WITH GENETIC ASSOCIATION DATA

4.1 Introduction

Complex traits or diseases arise as the consequence of many factors, including genetic factors, environmental exposures, and the interplay among them. The identification of gene-environment interaction (GxE) effects can help to further detect genetic risk factors contributing to the "missing heritability" as well as elucidate trait and disease aetiology (Thomas, 2010). Moreover, the identification of individuals or cohorts susceptible to specific environmental hazards is essential to the development of precision medicine (Collins and Varmus, 2015).

Despite enthusiasm for detecting GxE effects via genome-wide scans of existing genome-wide association (GWA) or sequencing data, it is shown that such endeavors have been underpowered (Smith and Day, 1984; Murcray et al., 2009). Only a limited number of GxE effects were detected with data from individual association studies at the stringent genome-wide significance thresholds (Thomas, 2010; Rothman et al., 2010). There is also growing recognition that much larger sample sizes are required in order to identify genetic variants with moderate-to-small effect sizes, are rare in the population, and/or have effects modified by the environment (Smith and Day, 1984; McCarthy et al., 2008). Large consortia on many complex traits and diseases have been formed to share data and resources, with the goal of further uncovering genetic risk factors with the combined large samples (Hunter et al., 1982; Hu et al., 2013; Ahsan et al., 2014; Huo et al., 2016). Novel and powerful meta-analysis approaches are needed to address the new challenges arising from the analysis of consortium data.

With data from a single GWA study, a common strategy to improve power in detecting GxE effects is to first filter out the majority of unpromising genetic variants and only test GxE on the remaining ones. The filtering test and the GxE test are required to be independent under the null to preserve the overall genome-wide error rates (Dai et al., 2012). Such two-stage or multi-stage analysis strategies relax the stringent genome-wide significance thresholds and improve the power for detecting variants with GxE effects (Kooperberg and LeBlanc, 2008; Murcray et al., 2009). In addition to variant-level analysis, gene-based or set-based approaches are proposed to detect those individually-weak-but-collectively-strong GxE effects within a gene or a set. Many gene-based association tests (Wu et al., 2011; Chen et al., 2012) can be generalized to test for gene-level GxE effects. Chatterjee et al. (2006) proposed the Tukey's 1-degree-of-freedom test, which assumes that GxE effects are proportional to the main effects of variants in a gene. Jiao et al. (2013) proposed a gene-based GxE test in which the gene-environment (G-E) correlations are used to filter variants showing no promise of GxE effects. More recently, Liu et al. (2016) developed a unified set-based GxE testing framework that simultaneously considers filtering on individual variants and testing on the GxE effects from a set of variants (e.g., the variants in a gene). They calculated the estimated optimal filtering threshold for each set of variants (e.g., each gene) and showed that the unified test with adaptive filtering threshold generally improves the power for gene-based GxE analysis.

These aforementioned methods were proposed for GxE analyses in a single association study, and there is only limited discussion in the literature on methods for meta-analysis of GxE with consortium data. Lee et al. (2013) proposed gene-based association tests for meta-analyses with rare variants, and the tests can also be applied to meta-analysis of GxE. In their work, they considered both fixed-effects and random-effects meta-analysis tests to aggregate all the genetic variants in a gene.

In this work, we proposed to introduce filtering in the gene-based GxE tests for meta-

analysis. We proposed to first perform variant-level "meta-filtering" tests that combine the filtering statistics across multiple studies for each individual variant, and then used meta-analysis to test for gene-based GxE effects on only the retained variants. We studied both the fixed-effects and random-effects meta-analysis approaches, and proposed to combine the strengths of both to a unified test, the omnibus filtering-based GxE meta-analysis (ofGEM). We compared the proposed ofGEM approach with alternative approaches including the gene-based meta-analysis test with no variant-level filtering. As an extension, we also suggested a gene-based GxE test for studies with more than one ethnic group. We applied the proposed approach to two breast cancer GWA data sets (Ahsan et al., 2014; Huo et al., 2016), with samples of European and African ancestries respectively, to identify the genes harboring variants interacting with age. Both studies included several sub cohorts or sub-populations, and as such a meta-analysis rather than a pooled analysis would better account for the sample heterogeneity for both.

4.2 Methods

In this work, we consider meta-analysis methods for jointly testing GxE effects of k variants in a gene or a set. Let θ_j be the interaction effect of the j-th genetic variant and the environmental exposure. We are interested in testing the null hypothesis of no interaction effect versus the alternative hypothesis of at least one variant in the gene having a non-zero interaction effect.

$$H_0 : \theta_j = 0 \text{ for all } j \ (j = 1, \ldots, k). \text{ vs. } H_1 : \text{at least one } \theta_j \neq 0.$$

4.2.1 Testing gene-based GxE effects with data from a single study

We first consider the gene-based GxE effects with variant-level filtering and test statistics from a single GWA study. A unique characteristic of GxE effects in association studies is

that those effects are very sparse in the genome. By filtering out the variants that show no promise of GxE effects in a gene, one may increase the ratio of variants with GxE effects versus null variants and improve the power to detect genes with GxE effects (Jiao et al., 2013). Genes with different sizes may have different optimal filtering thresholds. Liu et al. (2016) further proposed a general framework for gene-based GxE test with adaptive filtering based on the following statistic:

$$T = \sum_{j=1}^{k} w_j Z_j^2 \cdot \mathbb{1}\{|X_j| \geq z_{\eta/2}\},$$

where w_j is the weight for variant j, X_j is the filtering statistic, Z_j is the statistic testing for GxE effects, $z_{\eta/2}$ is the $(1-\eta/2) \times 100$-th quantile of $N(0,1)$, and η is the adaptive filtering threshold for the gene being tested. The optimal filtering threshold primarily depends on the number of variants with GxE effects (k_1) and the total number of variants in a gene, i.e., the gene size (k).

In case-control studies, a commonly-used variant-level filtering test for GxE is to test for gene-environment correlations in the combined case-control samples (Murcray et al., 2009). For quantitative traits, one may test for equal variances across genotype groups to filter out the less-promising variants (Paré et al., 2010).

4.2.2 Meta-analysis approaches for gene-based tests

To combine data from multiple studies, Lee et al. (2013) proposed and discussed a general framework for meta-analysis of gene-based tests in association studies. The proposed framework can also be applied to gene-based GxE tests. Assuming homogeneous GxE effects across multiple studies, they proposed a fixed-effects approach that first combines the score-statistics for each variant across different studies and then aggregates the squared collapsed

score statistics of all variants in a gene:

$$Q_{hom-meta-SKAT} = \sum_{j=1}^{k} \left(\sum_{s=1}^{S} w_{js} Z_j^{(s)} \right)^2, \tag{4.1}$$

where w_{js} is the weight and $Z_j^{(s)}$ is the GxE test statistic for variant j in study s.

With heterogeneous GxE effects, they proposed a random-effects approach that combines the squared score statistics of all variants in different studies.

$$Q_{het-meta-SKAT} = \sum_{j=1}^{k} \left(\sum_{s=1}^{S} w_{ks}^2 Z_j^{(s)2} \right). \tag{4.2}$$

4.2.3 Meta-analysis approaches for gene-based GxE tests with filtering

The meta-analysis tests proposed by Lee et al. (2013) aggregated the effects of all variants in a gene across studies, and did not fully utilize the information that may help to filter out unpromising variants in GxE tests. In this section, we proposed a meta-analysis approach with variant-level filtering and gene-level GxE testing based on a total of S studies. Our method also considered different levels of heterogeneity of GxE effects.

A meta-filtering strategy

When the sample size of each individual study is moderate to small, the power of filtering test can be low and many variants with potential GxE effects may be filtered out. We proposed a meta-filtering strategy that combines samples across studies. Let $\{X_j^{(s)} : j = 1, \ldots, k; s = 1, \ldots, S\}$ be the filtering statistic for variant j in the s-th study.

For the fixed-effects model (4.1) assuming homogeneous GxE effects, we proposed a

meta-filtering statistic for variant j across multiple studies:

$$X_j^{\text{MF-fixed}} = \sum_{s=1}^{S} w_{js} X_j^{(s)}, \tag{4.3}$$

where MF stands for meta-filtering, and $w_{.s}$ is the weight for each study. One may use the weight, $w_{.s} = \sqrt{n_s/N}$, and n_s is the sample size for the s-th study, and $N = n_1 + \ldots + n_S$. Here we used equal weights for variants. One may also impose the variant weights being proportional to the minor allele frequencies (MAFs) and those options have been extensively discussed elsewhere (Lee et al., 2013).

For the random-effects model (4.2) assuming heterogeneous GxE effects, we proposed a random-effects meta-filtering statistic as

$$X_j^{\text{MF-random}} = \sum_{s=1}^{S} w_{js}^2 X_j^{(s)2}. \tag{4.4}$$

An alternative analysis strategy is to conduct filtering within each individual study, obtain the gene-level GxE statistics/p-values, and then combine the gene-level significances across studies. Given the relatively small sample size in each individual study and potential study heterogeneity, the alternative strategy may have lower filtering power and miss out the variants with potential GxE effects but failing to pass filtering in the single study analyses.

The meta-analysis tests with meta-filtering

For meta-analysis assuming homogeneous GxE effects, we proposed a gene-level fixed-effects statistic with the meta-filtering in model (4.3) as

$$T_{\text{hom-MF-fixed}} = \sum_{j=1}^{k} \left(\sum_{s=1}^{S} w_s Z_j^{(s)} \right)^2 \cdot \mathbb{1}\{|X_j^{\text{MF-fixed}}| \geq z_{\eta/2}\}, \tag{4.5}$$

where $z_{\eta/2}$ is the $(1-\eta/2)\times 100$-th quantile of $N(0,1)$ and η is the filtering threshold of p-values.

For meta-analysis assuming heterogeneous GxE effects across studies, we proposed the gene-level random-effects statistic with the meta-filtering in model (4.4) as

$$T_{\text{het-MF-random}} = \sum_{j=1}^{k}\left(\sum_{s=1}^{S} w_s^2 Z_j^{(s)2}\right)\cdot \mathbb{1}\{X_j^{\text{MF-random}} \geq q_\eta\}, \tag{4.6}$$

where q_η is the $(1-\eta)\times 100$-th quantile of a weighted χ^2 distribution under the null (Davies, 1980).

In comparison with the statistics in (4.1) and (4.2), the tests in (4.5) and (4.6) incorporated variant-level meta-filtering in the gene-based meta-analysis tests.

The omnibus filtering-based GxE meta-analysis (ofGEM)

The fixed- or random-effects tests have been shown to enjoy better power when the effects of interest are homogeneous or heterogeneous across studies (Lee et al., 2013). In practice, the study heterogeneity is often unknown and may vary for different genes in the genome.

Considering that the homogeneous fixed-effects test and the heterogeneous random-effects test are two complementary tests, one may combine the p-values of the two tests and obtain a new test that is powerful regardless of whether the true effects of interest are homogeneous or heterogeneous across studies. One may take the minimum of the two p-values and adjust the significance by Bonferroni correction. More recently, Soave et al. (2015) proposed to use Fisher's method to combine two relatively independent and complementary tests. Here we adopted Fisher's method to combine the p-values based on tests (4.5) and (4.6), and we termed this approach as the omnibus filtering-based GxE meta-analysis (ofGEM).

An extension: the meta-analysis GxE test for data from two or more ethnic groups

Some very large consortia may have samples and studies from different ethnic groups. In analyzing those data, it is expected that certain studies and sub-cohorts within the same ethnic group are more homogeneous and studies across different ethnic groups are more heterogeneous.

Here we extended the previous tests and proposed a grouped fixed-effects GxE test for data from B ethnic groups. The grouped fixed-effects test first calculated the fixed-effects statistic with meta-filtering on studies within each ethnic group and then aggregated the statistics from multiple ethnic groups. The test statistic is given by

$$T_{\text{grp-MF-fixed}} = \sum_{b=1}^{B} T^{b}_{\text{hom-MF-fixed}} = \sum_{b=1}^{B}\sum_{j=1}^{k}\left(\sum_{s\in\Omega_b} w_s Z_j^{(s)}\right)^2 \cdot \mathbb{1}\{|X_{jb}^{\text{MF-fixed}}| \geq z_{\eta/2}\}, \quad (4.7)$$

where Ω_b contains the study indices for studies from the b-th group, and $X_{jb}^{\text{MF-fixed}}$ is the meta-filtering statistic for variant j based on the studies from the b-th group. This grouped test applied the fixed-effects test in (4.5) to each ethnic group and then combined the group statistics as the overall statistic for a gene being tested.

Similarly, we proposed the grouped random-effects tests, $T_{\text{grp-MF-random}}$. This test can be formulated by applying the random-effects tests in (4.6) to studies from each ethnic group and combining the group statistics for the gene of interest. Additionally, one may further obtain the grouped ofGEM test, $T_{\text{grp-MF-ofGEM}}$, by combining the grouped fixed and random-effects p-values using Fisher's method.

4.2.4 The filtering thresholds

One major innovation of our proposed tests is to incorporate filtering into the meta-analysis of gene-level GxE tests. It has been shown that in the analyses of individual studies, filtering led to an increase in the proportion of non-null variants in a gene and as such improved the power to detect gene-level GxE effects (Jiao et al., 2013; Liu et al., 2016). The filtering thresholds may also affect the power. When the filtering threshold is too stringent, many of the variants with potential GxE effects may be filtered and power could be hurt. When the filtering threshold is too liberal, the power may not improve much compared to tests without filtering.

For gene-level GxE tests in a single association study, Jiao et al. (2013) employed a fixed filtering threshold of 0.1 for most of the genes in the genome. Liu et al. (2016) proposed to adaptively calculate the optimal filtering threshold for each gene in the genome. The calculated optimal filtering thresholds largely depend on the gene size.

For the fixed-effects model with meta-filtering in (4.5), the optimal filtering threshold can be calculated by directly using the formula for a single study proposed in Liu et al. (2016), by specifying the assumed homogeneous effect sizes for all studies. This is due to the finding that the GxE effects are homogeneous across studies, the power of the fixed-effects meta-analysis test with no filtering is almost identical to the power of the joint analysis by pooling all samples together (Lee et al., 2013). The calculation of the optimal filtering threshold proposed in Liu et al. (2016) is based on the pooled analysis and can be used to approximate the optimal threshold for meta-analysis with homogeneous GxE effects.

For the random-effects model with meta-filtering in (4.6), the calculation of the optimal filtering threshold would require the specification of the heterogeneous effect sizes for different studies. The level of heterogeneity may heavily influence the choice of the calculated filtering threshold. In reality, the heterogeneity across variants and studies is often hard to specify, and the misspecification of those parameters could lead to a sub-optimal filtering

threshold, which may not improve the power to detect GxE in comparison with a fixed filtering threshold. Therefore in the current work, for all the meta-analysis regardless of fixed- or random-effects, we followed Jiao et al. (2013) and adopted a fixed and liberal filtering threshold of 0.1.

4.2.5 Significance evaluation

The presence of linkage disequilibrium (LD) in real data makes it challenging to evaluate the significance of each gene using the proposed tests. Moreover, most consortia would only share summary statistics of individual studies, and the raw genotype data are often not available. Permutation-based p-value calculation is not only computational prohibitive but also impractical.

Here we adopted the sequential sampling procedure proposed in Liu et al. (2016). Specifically, with centered genotype and environmental data, if one uses the Wald statistics (or other t- or normal statistics) based on generalized linear models as the filtering and test statistics, we can approximately sample the null variant-level statistics of each gene from S studies $\{(X_1^{(s)}, ..., X_k^{(s)}), s = 1, \ldots, S\}$ and $\{(Z_1^{(s)}, ..., Z_k^{(s)})\}$ from a multivariate normal distribution $N(0, \mathbf{R})$, where $\mathbf{R} = (R_{ij})_{k\times k}$ and $R_{ij} = \text{cor}(G_i, G_j)$ for any i, j. The $\mathbf{R}$ matrix is the LD matrix of genotype correlations and can be estimated from the HapMap data or based on subsets of the samples in the current consortia (Yang et al., 2012; Hu et al., 2013).

To further improve computational efficiency, for a given gene, if no variant passes the meta-filtering, the p-value is set as 1 and no sampling is needed. For genes with at least one SNP passing the meta-filtering, we first draw $L = 100$ sets of null filtering and test statistics from $N(0, \mathbf{R})$. If the p-value calculated based on these 100 sets of null statistics is less than 0.1, then we draw 10 times the number of sets of null statistics (i.e. $L = 1000$) to obtain a p-value with higher precision. We repeat this procedure until the p-value is greater than $10/L$ or the total number of sampling exceeds 10^6 – a precision needed for the genome-wide

gene-level significance of 10^{-5}.

4.3 Simulations: power comparison

In this section, we compared the power of 1) our proposed method ofGEM, 2) the fixed- and random-effects methods with no filtering, 3) the fixed- and random-effects methods with filtering by individual studies, and 4) the fixed- and random-effects methods with meta-filtering. In particular, we examined the power comparison when there is no to mild study heterogeneity and when there is moderate to strong study heterogeneity.

In each simulated data set, we simulated multiple case-control studies each with 1000 cases and 1000 controls from a population with case prevalence of 0.05. We simulated a binary environmental factor with 50% probability of being 1. We simulated each genetic variant as an independent binomial variable with an MAF of 0.2. For each simulated data set below, we simulated 1000 genes with different number of variants.

In the following 576 sets of simulation data, we varied the key parameters that may affect the power of the meta-analyses. Specifically, we have: the GxE effect size θ being log(1.2) or log(1.3); the number of studies (S) varying from 3 to 10; the number of variants in the gene (k) being 30 or 50; the number of variants with GxE effects (k_1) being 3 or 5; and the level of study homogeneity (ϕ) being 20% to 100%. Here the level of study homogeneity refers to the percentage of total studies that have the non-null GxE effects for a non-null GxE variant in a gene. As an example, when $\theta = \log(1.3)$, $S = 8$, $k = 30$, $k_1 = 5$ and $\phi = 50\%$, we simulated 1000 genes in 8 studies, each gene having 30 variants, 5 out of 30 variants having GxE effects being log(1.3) in 4 out of 8 studies (i.e., 50%), and the rest of the variants having no GxE effects. The 192 simulated data sets with ($\phi \leq 40\%$) were considered as data sets with moderate to strong heterogeneity.

In Figure 4.1, we compared the power of the proposed methods and competing methods at the significance levels of 0.05. We also repeated the comparison at the significance level

of 0.001, and the conclusions were unchanged (and not shown here). Each dot in the figure represents the power of two comparing methods at the significance level based on one simulated data set. The data sets with $\phi \leq 40\%$, i.e., the heterogeneous data, were plotted in red and the rest relatively more homogeneous data were plotted in black. We first compared the power of fixed- and random-effects models with that of meta-filtering versus other filtering options, no filtering or filtering by individual study. For both homogeneous and heterogeneous data sets, the power of fixed-effects model with meta-filtering was much higher than that of the fixed-effects models with no filtering or filtering with individual study, as shown in Figure 4.1A and Figure 4.1B. For random-effects models, the power of meta-filtering was also substantially improved over that of no filtering (Figure 4.1C). When comparing the random-effects model with meta-filtering versus filtering by individual study in Figure 4.1D, we observed some but less power improvement. The power improvement came from aggregating concerted GxE effects across studies in the filtering test. The less power improvement was because the random-effects model accounts for some degrees of study heterogeneity. For the extremely heterogeneous studies, the meta-filtering approach may have reduced filtering power than filtering by individual studies and in those cases, meta-analyses would not be more powerful than analyses of individual studies.

Regardless of using fixed- or random-effects models and the level of study heterogeneity, filtering out the unpromising variants improves the proportion of variants with GxE effects (k_1/k) and the power to detect gene-level GxE effects, as shown in Figure 4.1A and Figure 4.1C. Since individual studies had limited sample sizes, by combining all studies to conduct filtering, we may still improve the filtering power relative to filtering by individual studies and as such may improve the power of gene-level GxE testing. The power improvement largely depended on the study heterogeneity and the model choice.

In Figure 4.2, we compared the power of the proposed ofGEM method versus the fixed- and random-effects models with meta-filtering for homogeneous and heterogeneous data sets.

In both settings, the power of ofGEM was comparable or even higher than the more powerful one of the fixed- and random-effects models. It has been shown in Soave et al. (2015) that when applying Fisher's method to combine two complementary tests, if only one of the tests has power, then the combined test may have comparable but slightly reduced power; and if both tests show some power, then the combined test can be much more powerful than either individually. This is also what we have observed in our simulation studies.

In reality, the level of study heterogeneity is often unknown, and may also vary by different genes. It is desirable to have a test that is powerful regardless of homogeneous or heterogeneous effect sizes.

4.4 Applications: identifying gene-by-age interactions in breast cancer women

The risk of breast cancer is relatively low in young women aged 50 or younger. But younger women with early onset breast cancer (EOBC) are more likely to have more advanced stage cancers at diagnosis in comparison with older women with breast cancer. According to the American Cancer Society, the absolute risks of developing breast cancer in the next 10 years for women aged 20, 30, and 40 are 0.05%, 0.4%, and 1.5%, respectively, and up to age 85, the chance of developing breast cancer over a lifetime is up to 12 percent (American Cancer Society, 2013). The risk of breast cancer increases with age. Women with certain known genetic risk factors, for example, mutations in *BRCA1* and *BRCA2*, are at an increased risk of developing breast cancer at early ages. For those reasons, it is desirable to detect the genetic risk factors that interact with age and affect the risk of breast cancer. In this section, we analyzed existing GWA data from two studies by Ahsan et al. (2014) and Huo et al. (2016) with European and African ancestries, respectively, to detect the genes potentially interacting with age and affecting breast cancer risk.

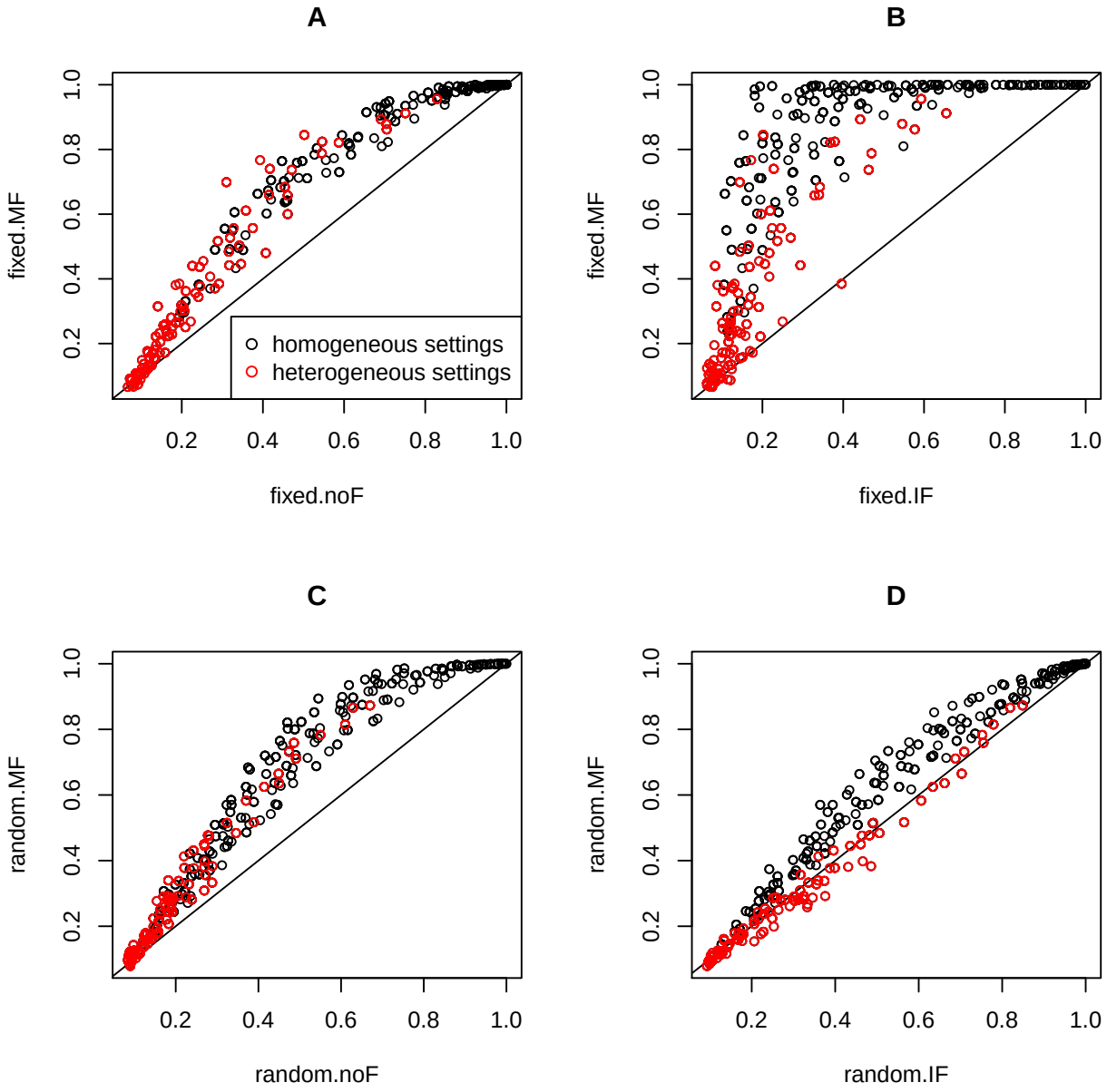

Figure 4.1: Power comparison of different filtering methods for fixed- and random-effects models. We simulated 576 data sets, each with 1000 genes. We varied the effect sizes, the gene sizes, the numbers of studies, and the level of study heterogeneity. The red dots denoted the more heterogeneous data sets and the black ones were the more homogeneous data sets. We compared the power of A) the fixed-effects model with meta-filtering (fixed.MF) versus no filtering (fixed.noF); B) the fixed-effects model with meta-filtering (fixed.MF) versus filtering by individual study (fixed.IF); C) the random-effects model with meta-filtering (random.MF) versus no filtering (random.noF); and D) the random-effects model with meta-filtering (random.MF) versus filtering by individual study (random.IF).

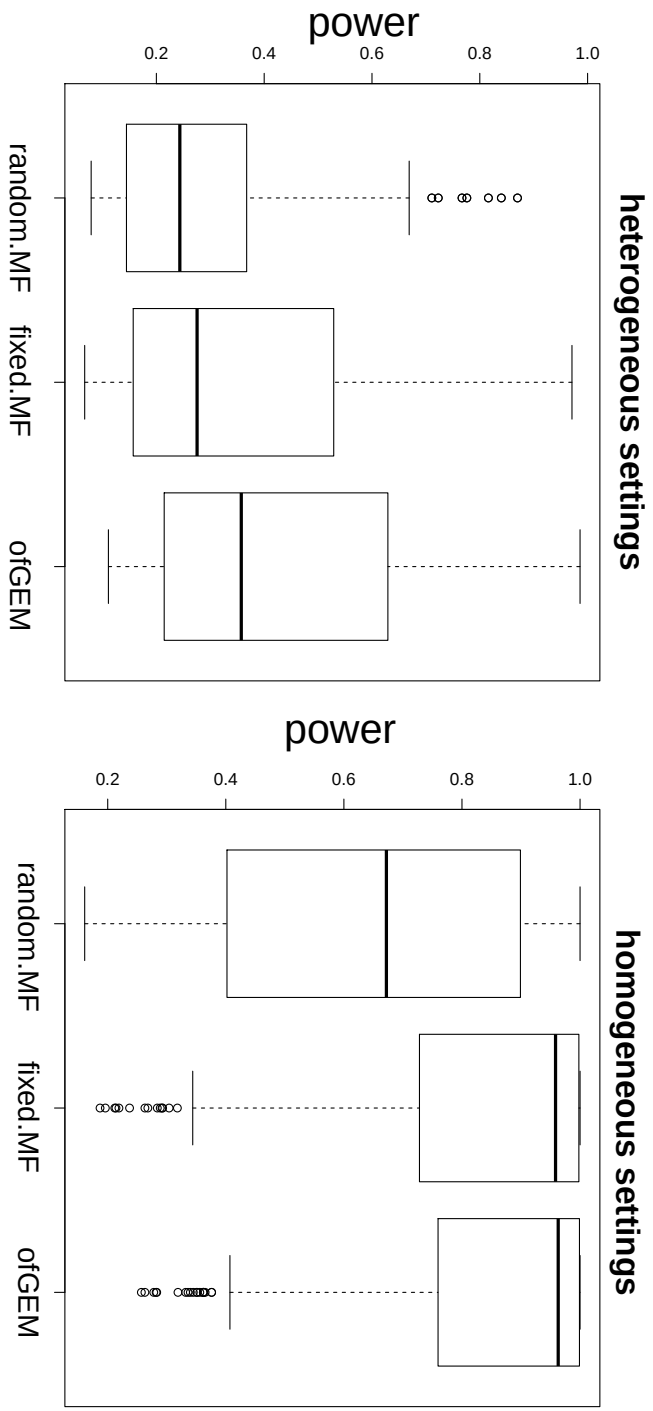

Figure 4.2: Power comparison of fixed- and random-effects models with meta-filtering versus the ofGEM method. Based on the 576 simulated data sets described in Section 4.3, we compared the power of fixed- and random-effects models versus the proposed ofGEM methods for the heterogeneous (left) and the homogeneous (right) settings. The ofGEM method is more powerful than either method in most simulated data sets.

4.4.1 Detecting gene-by-age interaction effects in an EOBC study

In the study by Ahsan et al. (2014), a total of 3,523 cases and 2,702 controls of non-Hispanic White (NHW) women were recruited. The cases were NHW women diagnosed with invasive breast cancer aged 20 to 50 and not known to carry pathogenic mutations in *BRCA1* or *BRCA2*. The controls were NHW women age 20 to 50 years without a history of breast cancer. Among the eight sites in the study, we focused on the six sites with both controls and cases. Those women were recruited from the Breast Cancer Family Registry (BCFR) in Australia, Canada, and the United States; the Genetic Epidemiologic Study of Breast Cancer (GECBC) in Germany; the Surveillance, Epidemiology, and End Results (SEER) Program; and the Long Island Breast Cancer Study Project. After excluding the subjects with missing age, we restricted our analysis to 2,540 cases and 2,429 controls. Table 4.1 shows the number of cases and controls and the study sites of this EOBC study.

Table 4.1: The data summary for the EOBC and ROOT consortia.

Consortium	Study	Location	Cases	Controls	Total
EOBC	BCFR (AUS)	Australia	593	250	843
	BCFR (NCA)	Northern California, USA	204	156	360
	BCFR (Ontario)	Ontario, Canada	668	395	1,063
	GESBC	Germany	553	1,071	1,624
	LI	Long Island, New York, USA	225	229	454
	Seattle	Seattle, Washington, USA	297	328	625
ROOT	NBCS	Ibadan, Nigeria	711	623	1,334
	BNCS	Barbados	92	229	321
	RVGBC	Philadelphia & Detroit, USA	145	257	402
	BBCS	Baltimore, Maryland, USA	95	102	197
	CCPS	Chicago, Illinois, USA	394	387	781
	SCCS	Southern United States	220	430	650

After genotype imputation and quality control, we obtained a total of 1,060,790 genetic variants on Chromosomes 1 to 22. For the gene-based test, we mapped the variants to genes based on the definition of the Single Nucleotide Polymorphism database (dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/), and focused on genes with at least five variants. The dbSNP assigned an SNP to a gene within 2,000 base pairs upstream and 500

base pairs downstream. For each gene, we then performed LD pruning at the threshold of $r^2 > 0.9$ to remove the variants in high LD. This resulted in a total of 14,635 genes for the analysis.

We applied the fixed- and random-effects models with meta-filtering and the ofGEM methods to this data set. We calculated the Wald statistics with age being the response variable and the genotypes as the predictors based on the combined control and case data as the filtering statistics. We calculated the Wald statistics for testing gene-by-age interaction effects based on logistic regression models as the test statistics. In calculating the filtering and test statistics, we adjusted for the top ten principal components calculated from the genotype matrix. The summary statistics from each study were then used in the meta-analyses. All the p-values were calculated based on the sequential sampling procedure described in Section 4.2.5.

Table 4.2 shows the genes with suggestive evidence of gene-age interactions reached the gene-level genome-wide significance threshold of 10^{-5} by at least one of the three methods. One can also see that if the p-value was small for only one of the fixed- or random-effects models, the ofGEM method would also report a small p-value, and if the p-value was small for both the fixed- and random-effects models, the ofGEM method would report a more significant p-value than either, similar to what we have observed in the simulation studies.

Table 4.2: The p-values of significant genes in the meta-analysis of gene-age interactions for the EOBC data, the ROOT data, and the combined data analysis, with a gene-level genome-wide significance level of 10^{-5}.

Consortium	Sample size (Case/Control)	Gene	Chr	Gene size	p_random.MF	p_fixed.MF	p_ofGEM
EOBC	4,969 (2,540/2,429)	*HLA-DMB*	6	9	8.40×10^{-5}	9.39×10^{-4}	1.37×10^{-6}
ROOT	3,685 (1,657/2,028)	*ATP6V1D*	14	5	1.00×10^{-5}	1.00×10^{-6}	2.63×10^{-10}
		EIF2S1	14	5	1	3.00×10^{-6}	4.12×10^{-5}
		FAM71D	14	14	1	4.00×10^{-6}	5.37×10^{-5}
		MPP5	14	10	1	2.00×10^{-6}	2.82×10^{-5}
Combined	8,654 (4,197/4,457)	*TMEM206*	1	6	1.71×10^{-2}	2.10×10^{-5}	5.69×10^{-6}

4.4.2 Detecting gene-by-age interaction effects in the ROOT data

The ROOT consortium consists of 3,686 women of African ancestry from six studies: the Nigerian Breast Cancer Study (NBCS), Barbados National Cancer Study (BNCS), Racial Variability in Genotypic Determinants of Breast Cancer Risk Study (RVGBC), Baltimore Breast Cancer Study (BBCS), Chicago Cancer Prone Study (CCPS), and Southern Community Cohort (SCCS). Table 4.1 shows the sample size within each study. There was one control from the NBCS study with age missing and was removed from our analyses. Among those 3,685 women, 1,657 were cases and 2,028 were controls. Those women aged from 18 to 92, and 2,114 (54.0%) were 50 years old or younger. This study also has a large proportion of women with breast cancer occurred at younger ages.

After quality control and filtering out the variants with MAFs of less than 5%, we focused on 1,394,070 variants from Chromosome 1 to 22. Similar to the analysis of the EOBC data, we restricted the analysis to the 17,524 genes with at least five variants and then performed an LD pruning at $r^2 > 0.9$ for each gene.

We applied the fixed- and random-effects models with meta-filtering and the ofGEM method to the ROOT data. Four genes showed some suggestive evidence of age interaction at the p-value threshold of 10^{-5} by at least one of the three methods.

Additionally, we jointly analyzed the EOBC and the ROOT data using the grouped tests for multiple ethnic groups proposed in Section (4.2.3). We identified one gene with potential gene-age interaction effects. However, those grouped effects were mostly driven by data from one of the two ethnic groups but not both. This result echoed the well-known disparity of breast cancer etiology among women with European and African ancestries and further suggested the age interaction effects on breast cancer are quite different in those two groups.

4.5 Discussion

In this work, we proposed a gene-based meta-analysis test with filtering to detect gene-environment interactions with association data. We first proposed to conduct filtering test in a meta-analysis of GxE by combining all samples in the consortia data. The proposed meta-filtering showed improved power relative to no filtering or filtering by individual study. We then proposed an ofGEM test that combines the strengths of the fixed- and random-effects models with meta-filtering. With simulation studies, we showed that the proposed methods and tests are more powerful than competing methods in a wide range of settings and regardless of the level of study heterogeneity. We further extended the proposed approaches to analyze data from multiple ethnic groups.

We applied the proposed methods to GWA data from two breast cancer consortia of European and African ancestries, to identify genes potentially interacting with age on breast cancer risk. we identified several genes showed suggestive evidence of gene-age interactions in the EOBC, the ROOT data, and in both studies across different ancestries.

The proposed methods and tests can be applied to other existing GWA or sequencing data from large consortia of complex diseases to detect interaction effects. Those analyses are based on summary statistics and can be a cost-efficient way to recapitalize on existing data.

CHAPTER 5

SUMMARY AND FUTURE PLANS

5.1 Summary of the work

In this dissertation, I developed tailored statistical methods for 1) imputing missing or unmeasured gene expression levels in multi-tissue expression and eQTL studies, 2) testing for multivariate outcomes and their associations with clinical variables with data from labeling-based quantitative proteomics studies, and 3) testing gene-based GxE in meta-analyses of genetic association studies.

The methods developments were motivated by different areas in genomic research, though they share some common challenges. One challenge is the handling of correlations among samples. In the GTEx multi-tissue imputation project, tissue samples from the same donor are naturally clustered. In the proteomics research project, samples processed by the same iTRAQ or TMT experiment are experimentally correlated. To model those correlations among samples, mixed-effects models were extensively used and studied in my work. I developed univariate and multivariate mixed-effects models and integrated them with the modeling of other data characteristics in parameter estimation, inference, and model prediction.

Another challenge is the handling of various types of missing values. In the GTEx project, we developed multi-tissue imputation methods to impute the missing values. We assumed the missing-data mechanism is ignorable and proposed to impute the missing gene expression values from a rich collection of genomic information including tissue-tissue expression correlations, developmental factors, and eQTL effects. We also showed that the imputed data with good quality can be used as supporting or secondary data, to enhance the power of the primary data analyses based on the observed data. They can facilitate subsequent analyses that require a complete data set. In the quantitative proteomics project, we en-

countered a unique cluster-level non-ignorable missing-data mechanism. We proposed to model the clustered-level missing-data mechanism using an exponential function that links the probability of missing a cluster with the values of the cluster. We incorporated this exponential function into the multivariate normal density function and derived the solutions of the EM algorithm to correct for the biases with the observed data. We showed that the modeling of clustered missing-data mechanism largely improved the accuracy and efficiency of the statistical methods. We applied the proposed methods to the CPTAC breast cancer data and identified meaningful biological results.

One area of focus for this dissertation is the development of multivariate methods. In both Chapters 3 and 4, for analyses of quantitative proteomics data and meta-analysis methods for GxE, we proposed to jointly analyze functional sets of genomic features. Those multivariate analysis methods are not replacements of univariate analysis methods. Instead, they serve as informative complementary approaches. It is known that for complex hypotheses, there is no uniformly most powerful test. The power of the analyses depends on the type of data and the strength/structure of the effects of interest. The use of multivariate analysis approaches aims to detect individually weak but collectively strong sets of effects. Furthermore, the functional sets themselves are often biologically more interpretable; the number of tests performed is reduced, and thus alleviates the multiple testing burden and improves the power. In this dissertation, I analyzed multiple phosphopeptides from the same phosphoproteins, and multiple phosphoproteins from KEGG pathways in quantitative proteomics; and I analyzed multiple genetic variants from the same gene or gene regions in GxE meta-analyses with genetic association data.

When handling high-dimensional predictors or high-dimensional outcomes, I employed machine learning based methods and penalized likelihood approaches to regularize the problem and improve the efficiency of the analyses. I proposed a random-forest-based imputation model considering the mixed-effects in Chapter 2 to impute expression values based on po-

tentially high-dimensional predictors. In Chapter 3, I used a factor-analytic random-effects structure to model highly correlated experimental correlations when jointly analyzing high-dimensional phosphopeptides from the same phosphoprotein. Also, I developed a penalized likelihood approach to regularize the elements in the high-dimensional precision matrix of multivariate outcomes and used an ADMM algorithm to estimate the sparse precision matrix. In Chapter 4, I introduced meta-filtering in gene-based tests to filter out the less promising features in a set, improve the signal-to-noise ratio in a set, and improve the power of the analysis.

Computational feasibility and efficiency are always within the scope of my work. I developed multiple R software packages: MixRF (Wang et al., 2016; Wang and Chen, 2016), mixEMM (Chen et al., 2017b,c), mvMISE (https://github.com/randel/mvMISE), and ofGEM (https://github.com/randel/ofGEM), for the methods development presented in this dissertation work. Those R packages are currently all publicly available.

5.2 Future directions

In the future, I will continue to expand my research interests in developing statistical methods and computational tools for analyzing genomics data with clustered structures and missing values. All of the new statistical methods and computational tools will be optimized for large-scale data. Here I discuss some possible research directions and projects, including an imputation method for multi-tissue gene expression data incorporating pedigree information and a generalized machine learning method to binary clustered outcomes. Additionally, the multivariate mixed-effects models can also be applied to other high-dimensional data settings.

5.2.1 Pedigree-based imputation of gene expression in uncollected tissues

In Chapter 2, motivated by the GTEx project, we proposed MixRF (Wang et al., 2016) to impute the gene expression levels in uncollected tissues from unrelated post-mortem donors in the GTEx project. When the study samples consist of related individuals from a few extended families with known pedigrees, incorporating the pedigree information may further improve the imputation accuracy.

In another project, a motivating data set is collected from 786 members of 26 pedigrees from Costa Rica and Colombia (Fears et al., 2014). All individuals have blood expression data, while about 400 additionally have fibroblast expression data. I propose a pedigree-based mixed-effects random forest to impute fibroblast expression with blood expression and eQTLs (Peterson et al., 2016). With the imputed expression data, we can test the association between the fibroblast expression and hundreds of neurocognitive and activity phenotypes (Fears et al., 2014; Pagani et al., 2016) or conduct other downstream analyses.

More specifically, a mixed-effects random forest can be written in matrix form,

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the expression level of a gene in tissues of individuals, $\boldsymbol{X}$ is the genotype and covariates, $f(\cdot)$ is the function form of a random forest model, $\boldsymbol{Z}$ is the random effect covariate matrix with $\boldsymbol{Z} = \boldsymbol{1}$ for a random-intercept-only model, $\boldsymbol{u}$ is the random intercept for each individual, and $\boldsymbol{\epsilon}$ is the error term. The random intercept $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)'$ is usually assumed to be independent across individuals. For pedigree data, we could allow $\boldsymbol{u}$ to be correlated (Vazquez et al., 2010) with a variance-covariance matrix

$$\text{var}(\boldsymbol{u}) = \boldsymbol{A}\sigma_u^2,$$

where $\boldsymbol{A}$ is a known additive relationship (kinship) matrix. With the Cholesky decomposition

$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$ and $\boldsymbol{Z}^* = \boldsymbol{Z}\boldsymbol{L}, \boldsymbol{u}^* = \boldsymbol{L}^{-1}\boldsymbol{u}$, we can fit the pedigree-based model as a standard mixed-effects model

$$\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{Z}^*\boldsymbol{u}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{u}^* \sim N(0, \boldsymbol{I}\sigma_u^2)$. The flexible function $f(\cdot)$ can also be estimated using machine learning methods other than random forest.

In summary, we can change the design matrix of random effects in MixRF to incorporate the known pedigree information. We expect that this may increase the imputation accuracy and thus improve the power in the downstream testing.

5.2.2 Mixed-effects random forest for binary clustered data

To further enhance the data-driven decision making for clustered/multilevel/longitudinal data, I plan to further develop machine learning methods for clustered data, in addition to the proposed mixed-effects random forest in Chapter 2. The general EM-type framework can incorporate methods such as neural networks and gradient boosting. Furthermore, I will generalize the proposed methods to work for discrete outcomes, such as count data and binary data. Here I propose to use the penalized quasi-likelihood (PQL) to linearize nonlinear models.

I applied the proposed generalized mixed-effects random forest to an ecological momentary assessment (EMA) dataset of adolescent smokers smoking, to identify factors related to their reported adolescent smoking events. The motivating EMA data have 11,855 observations from 247 subjects, each with 48 repeated measures on average (Colvin and Mermelstein, 2010). The goal of this analysis is to predict the smoking event with all the 44 covariates.

For binary response y_{ij}, we assume $y_{ij} \sim \text{Bernoulli}\left(p_{ij}\right)$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. We have $E\left(y_{ij}\right) = p_{ij}, \text{var}\left(y_{ij}\right) = \sigma^2 p_{ij}\left(1 - p_{ij}\right)$, where σ^2 is the dispersion parameter. We

propose a mixed-effects random forest with a logit link function,

$$\text{logit}\left(p_{ij}\right) = f\left(\mathbf{x}_{ij}\right) + \mathbf{z}_{ij}^{'}\mathbf{b}_i,$$

where $f\left(\cdot\right)$ is a random forest, and the random effect has $\text{var}\left(\mathbf{b}_i\right) = \mathbf{D}$. The nonlinear link function makes estimation approaches not readily available.

Inspired by the generalized mixed-effects regression trees (Hajjem et al., 2017), we use an EM-like algorithm with PQL for the parameter estimation. The idea of PQL is to linearize the nonlinear model with a first-order Taylor series expansion, and then repeatedly fit a weighted linear model until convergence. With the linearized model, we can estimate random forest and variance components of random effects within an additive model.

A linearized pseudo-model can be written as

$$\begin{aligned} y_{ij}^{l} &= f\left(\mathbf{x}_{ij}\right) + \mathbf{z}_{ij}^{'}\mathbf{b}_i + \left(y_{ij} - p_{ij}\right) / \left[p_{ij}\left(1 - p_{ij}\right)\right], \\ &= \text{logit}\left(p_{ij}\right) + \left(y_{ij} - p_{ij}\right) / \left[p_{ij}\left(1 - p_{ij}\right)\right]. \end{aligned}$$

Let $e_{ij} = \left(y_{ij} - p_{ij}\right) / \left[p_{ij}\left(1 - p_{ij}\right)\right]$. We can fit a weighted mixed-effects random forest with weight $w_{ij} = p_{ij}\left(1 - p_{ij}\right)$,

$$\begin{aligned} y_{ij}^{l} &= f\left(\mathbf{x}_{ij}\right) + \mathbf{z}_{ij}^{'}\mathbf{b}_i + e_{ij}, \\ \sqrt{w_{ij}}y_{ij}^{l} &= \sqrt{w_{ij}}f\left(\mathbf{x}_{ij}\right) + \sqrt{w_{ij}}\mathbf{z}_{ij}^{'}\mathbf{b}_i + \sqrt{w_{ij}}e_{ij}. \end{aligned}$$

Now the weighted error term $\sqrt{w_{ij}}e_{ij}$ has a constant variance σ^2 and there is no heterogeneity. It can be another extension of the linear mixed-effects random forest that I proposed in Chapter 2 and similar estimation approaches may be used.

5.3 Ending remarks

Overall, my research interests are to develop statistical methods and computational tools for analyzing big genomics data with clustered structures and complex missing-data mechanisms. Those methods developments are always motivated by challenges in real data problems. My dissertation work consists of methods/software development and data analyses in big "omics" data from multi-tissue gene expression and eQTL studies, quantitative proteomics studies, and genetic association and gene-environment interaction studies. Those methods and tools are tailored for specific challenges arising in each motivating data set, yet are generalizable to a broader area of statistical analyses of similar genomics data and data with similar structures.

REFERENCES

H. Ahsan, J. Halpern, M. G. Kibriya, B. L. Pierce, L. Tong, E. Gamazon, V. McGuire, A. Felberg, J. Shi, F. Jasmine, S. Roy, R. Brutus, M. Argos, S. Melkonian, J. Chang-Claude, I. Andrulis, J. L. Hopper, E. M. John, K. Malone, G. Ursin, M. D. Gammon, D. C. Thomas, D. Seminara, G. Casey, J. A. Knight, M. C. Southey, G. G. Giles, R. M. Santella, E. Lee, D. Conti, D. Duggan, S. Gallinger, R. Haile, M. Jenkins, N. M. Lindor, P. Newcomb, K. Michailidou, C. Apicella, D. J. Park, J. Peto, O. Fletcher, I. D. S. Silva, M. Lathrop, D. J. Hunter, S. J. Chanock, A. Meindl, R. K. Schmutzler, B. Müller-Myhsok, M. Lochmann, L. Beckmann, R. Hein, E. Makalic, D. F. Schmidt, Q. M. Bui, J. Stone, D. Flesch-Janys, N. Dahmen, H. Nevanlinna, K. Aittomäki, C. Blomqvist, P. Hall, K. Czene, A. Irwanto, J. Liu, N. Rahman, C. Turnbull, A. M. Dunning, P. Pharoah, Q. Waisfisz, H. Meijers-Heijboer, A. G. Uitterlinden, F. Rivadeneira, D. Nicolae, D. F. Easton, N. J. Cox, and A. S. Whittemore. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer Epidemiology Biomarkers and Prevention*, 23(4):658–669, 2014.

P. S. Albert and D. Follmann. Shared-parameter models. In *Longitudinal Data Analysis*, pages 433–452. Chapman and Hall/CRC, 2008.

American Cancer Society. Breast cancer facts & figures 2013-2014. 2013.

V. Baladandayuthapani, R. Talluri, Y. Ji, K. R. Coombes, Y. Lu, B. T. Hennessy, M. A. Davies, and B. K. Mallick. Bayesian sparse graphical models for classification with application to protein expression data. *The Annals of Applied Statistics*, 8(3):1443, 2014.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002.

G. N. Brock, J. R. Shaffer, R. E. Blakesley, M. J. Lotz, and G. C. Tseng. Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. *BMC Bioinformatics*, 9:12, 2008.

M. Celton, A. Malpertuy, G. Lelandais, and A. G. de Brevern. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. *BMC Genomics*, 11(1):15, 2010.

N. Chatterjee, Z. Kalaylioglu, R. Moslehi, U. Peters, and S. Wacholder. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *The American Journal of Human Genetics*, 79(6):1002–1016, 2006.

B. Chen, J. Wang, D. Dai, Q. Zhou, X. Guo, Z. Tian, X. Huang, L. Yang, H. Tang, and X. Xie. AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. *Journal of Experimental & Clinical Cancer Research*, 36(1):65, 2017a.

L. S. Chen, L. Hsu, E. R. Gamazon, N. J. Cox, and D. L. Nicolae. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91:997–986, 2012.

L. S. Chen, R. L. Prentice, and P. Wang. A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics*, 70(2):312–322, 2014.

L. S. Chen, J. Wang, X. Wang, and P. Wang. A mixed-effects model for incomplete data from labeling-based quantitative proteomics experiments. *Annals of Applied Statistics*, 11 (1):114–138, 2017b.

L. S. Chen, P. Wang, and J. Wang. *mixEMM: A Mixed-Effects Model for Analyzing Cluster-Level Non-Ignorable Missing Data*, 2017c. URL `https://CRAN.R-project.org/package=mixEMM`. R package version 1.0.

A. Chhibber, C. E. French, S. W. Yee, E. R. Gamazon, X. Qin, E. Theusch, A. Webb, S. Weiss, M. W. Medina, R. M. Krauss, S. E. Scherer, N. J. Cox, K. M. Giacomini, and S. E. Brenner. Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. *The Pharmacogenomics Journal*, 1:9, 2016.

T. Clough, M. Key, I. Ott, S. Ragg, G. Schadow, and O. Vitek. Protein quantification in label-free LC-MS experiments. *Journal of Proteome Research*, 8(11):5275–5284, 2009.

F. S. Collins and H. Varmus. A new initiative on precision medicine. *The New England Journal of Medicine*, 372(9):793–795, 2015.

P. J. Colvin and R. J. Mermelstein. Adolescents smoking outcome expectancies and acute emotional responses following smoking. *Nicotine & Tobacco Research*, 12(12):1203–1210, 2010.

J. Y. Dai, C. Kooperberg, M. Leblanc, and R. L. Prentice. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika*, 99 (4):929–944, 2012.

P. Danaher, P. Wang, and D. M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

R. B. Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3): 323–333, 1980.

P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49, 1994.

A. S. Dimas, A. C. Nica, S. B. Montgomery, B. E. Stranger, T. Raj, A. Buil, T. Giger, T. Lappalainen, M. Gutierrez-Arcelus, MuTHER Consortium, M. I. McCarthy, and E. T. Dermitzakis. Sex-biased genetic effects on gene regulation in humans. *Genome Research*, 22:2368–2375, 2012.

Y. Donner, T. Feng, C. Benoist, and D. Koller. Imputing gene expression from selectively reduced probe sets. *Nature Methods*, 9(11):1120–1125, 2012.

S. C. Elbein, E. R. Gamazon, S. K. Das, N. Rasouli, P. A. Kern, and N. J. Cox. Genetic risk factors for type 2 diabetes: A trans-regulatory genetic architecture? *The American Journal of Human Genetics*, 91(3):466–477, 2012.

M. Ellis, M. Gillette, S. Carr, A. Paulovich, R. Smith, K. Rodland, R. Townsend, C. Kinsinger, M. Mesri, H. Rodriguez, D. Liebler, and CPTAC. Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery*, 3(10):1108–1112, 2013a.

M. J. Ellis, M. Gillette, S. A. Carr, A. G. Paulovich, R. D. Smith, K. K. Rodland, R. R. Townsend, C. Kinsinger, M. Mesri, H. Rodriguez, and D. C. Liebler. Connecting genomic alterations to cancer biology with proteomics: {The NCI Clinical Proteomic Tumor Analysis Consortium}. *Cancer Discovery*, 3:1108–1112, 2013b.

S. C. Fears, S. K. Service, B. Kremeyer, C. Araya, X. Araya, J. Bejarano, M. Ramirez, G. Castrillón, J. Gomez-Franco, M. C. Lopez, G. Montoya, P. Montoya, I. Aldana, T. M. Teshiba, Z. Abaryan, N. B. Al-Sharif, M. Ericson, M. Jalbrzikowski, J. J. Luykx, L. Navarro, T. A. Tishler, L. Altshuler, G. Bartzokis, J. Escobar, D. C. Glahn, J. Ospina-duque, N. Risch, A. Ruiz-Linares, P. M. Thompson, R. M. Cantor, C. Lopez-jaramillo, G. Macaya, J. Molina, V. I. Reus, C. Sabatti, N. B. Freimer, and C. E. Bearden. Multisystem component phenotypes of bipolar disorder for genetic investigations of extended pedigrees. *JAMA Psychiatry*, 71(4):375–387, 2014.

T. Flutre, X. Wen, J. Pritchard, and M. Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, 9, 2013.

J. Friedman. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, 4:404–408, 1977.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

E. R. Gamazon, H. E. Wheeler, K. P. Shah, S. V. Mozaffari, K. Aquino-Michaels, R. J. Carroll, A. E. Eyler, J. C. Denny, GTEx Consortium, D. L. Nicolae, N. J. Cox, and H. K. Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47:1091–1098, 2015.

J. M. Giltnane and J. M. Balko. Rationale for targeting the Ras/MAPK pathway in triple-negative breast cancer. *Discovery Medicine*, 17(95):275–283, 2014.

H. Glanz and L. Carvalho. An expectation-maximization algorithm for the matrix normal distribution. *arXiv preprint arXiv:1309.6609*, 2013.

E. Grundberg, K. S. Small, A. s. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, J. Nisbett, M. Sekowska, A. Wilk, S.-Y. Shin, D. Glass, M. Travers, J. L. Min, S. Ring, K. Ho, G. Thorleifsson, A. Kong, U. Thorsteindottir, C. Ainali, A. S. Dimas, N. Hassanali, C. Ingle, D. Knowles, M. Krestyaninova, C. E. Lowe, P. Di Meglio, S. B. Montgomery, L. Parts, S. Potter, G. Surdulescu, L. Tsaprouni, S. Tsoka, V. Bataille, R. Durbin, F. O. Nestle, S. O'Rahilly, N. Soranzo, C. M. Lindgren, K. T. Zondervan, K. R. Ahmadi, E. E. Schadt, K. Stefansson, G. D. Smith, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, and T. D. Spector. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10):1084–1089, 2012.

A. Hajjem, F. Bellavance, and D. Larocque. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6):1313–1328, 2014.

A. Hajjem, D. Larocque, and F. Bellavance. Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126:114–118, 2017.

D. Hedeker and R. D. Gibbons. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2(1):64, 1997.

Y. Hu, S. I. Berndt, S. Gustafsson, A. Ganna, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, J. Hirschhorn, K. E. North, E. Ingelsson, and D. Y. Lin. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *American Journal of Human Genetics*, 93(2):236–248, 2013.

J. E. Hunter, F. L. Schmidt, and G. B. Jackson. *Meta-analysis*. Sage Publications, 1982.

D. Huo, Y. Feng, S. Haddad, Y. Zheng, S. Yao, Y.-J. Han, T. O. Ogundiran, C. Adebamowo, O. Ojengbede, A. G. Falusi, W. Zheng, W. Blot, Q. Cai, L. Signorello, E. M. John, L. Bernstein, J. J. Hu, R. G. Ziegler, S. Nyante, E. V. Bandera, S. A. Ingles, M. F. Press, S. L. Deming, J. L. Rodriguez-Gil, K. L. Nathanson, S. M. Domchek, T. R. Rebbeck, E. A. Ruiz-Narváez, L. E. Sucheston-Campbell, J. T. Bensen, M. S. Simon, A. Hennis, B. Nemesure, M. C. Leske, S. Ambs, L. S. Chen, F. Qian, E. R. Gamazon, K. L. Lunetta, N. J. Cox, S. J. Chanock, L. N. Kolonel, A. F. Olshan, C. B. Ambrosone, O. I. Olopade, J. R. Palmer, and C. A. Haiman. Genome-wide association studies in women of african ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. *Human Molecular Genetics*, 25(21):4835, 2016.

J. G. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009.

S. Jiao, L. Hsu, S. Bézieau, H. Brenner, A. T. Chan, J. Chang-Claude, L. Le Marchand, M. Lemire, P. A. Newcomb, M. L. Slattery, and U. Peters. SBERIA: Set-based gene-environment interaction test for rare and common variants in complex diseases. *Genetic Epidemiology*, 37(5):452–464, 2013.

M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457, 2016.

N. A. Karp, W. Huber, P. G. Sadowski, P. D. Charles, S. V. Hester, and K. S. Lilley. Addressing accuracy and precision issues in iTRAQ quantitation. *Molecular & Cellular Proteomics*, 9:1885–1897, 2010.

J. Keen and H. Moore. The Genotype-Tissue Expression (GTEx) project: linking clinical data with molecular analysis to advance personalized medicine. *Journal of Personalized Medicine*, 5(1):22–29, 2015.

C. Kooperberg and M. LeBlanc. Increasing the power of identifying gene×gene interactions in genome-wide association studies. *Genetic Epidemiology*, 32(3):255–263, 2008.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38 (4):963–974, 1982.

S. Lee, T. M. Teslovich, M. Boehnke, and X. Lin. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics*, 93(1):42–53, 2013.

G. Li, A. A. Shabalin, I. Rusyn, F. A. Wright, and A. B. Nobel. An empirical Bayes approach for multiple tissue eQTL analysis. *arXiv Prepr.*, 2013.

S. G. Liao, Y. Lin, D. D. Kang, D. Chandra, J. Bon, N. Kaminski, F. C. Sciurba, and G. C. Tseng. Missing value imputation in high-dimensional phenomic data: imputable or not, and how? *BMC Bioinformatics*, 15(1):346, 2014.

A. Liaw and M. Wiener. Classification and Regression by randomForest. *R News*, 2:18–22, 2002.

A. W.-C. Liew, N.-F. Law, and H. Yan. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, 12(5):498–513, 2011.

R. Little. Selection and pattern-mixture models. In *Longitudinal Data Analysis*, pages 409–431. Chapman and Hall/CRC, 2008.

R. J. A. Little. A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483, 1994.

R. J. A. Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90:1112–1121, 1995.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002.

J. Liu, W. Liu, L. Wu, and G. Yan. A flexible approach for multivariate mixed-effects models with non-ignorable missing values. *Journal of Statistical Computation and Simulation*, 85 (18):3727–3743, 2015.

L. C. Liu and D. Hedeker. A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, 62(1):261–268, 2006.

Q. Liu, L. S. Chen, D. L. Nicolae, and B. L. Pierce. A unified set-based test with adaptive filtering for gene-environment interaction analyses. *Biometrics*, 72(2):629–638, 2016.

J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, A. Sturcke, J. Paschal, J. M. Anderson, E. L. Wilder, L. K. Derr, E. D. Green, J. P. Struewing, G. Temple, S. Volpi, J. T. Boyer, E. J. Thomson, M. S. Guyer, C. Ng, A. Abdallah, D. Colantuoni, T. R. Insel, S. E. Koester, A. R. Little, P. K. Bender, T. Lehner, Y. Yao, C. C. Compton, J. B. Vaught, S. Sawyer, N. C. Lockhart, J. Demchok, and H. F. Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

T. Lumley, R. Kronmal, and S. Ma. Relative risk regression in medical research: models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series. Working Paper 293*, 2006.

R. Luo, C. M. Colangelo, W. C. Sessa, and H. Zhao. Bayesian analysis of iTRAQ data with nonrandom missingness: identification of differentially expressed proteins. *Statistics in Biosciences*, 1:228–245, 2009.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher,

A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 2009.

M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Review Genetics*, 9:356–369, 2008.

P. Mertins, N. D. Udeshi, K. R. Clauser, D. R. Mani, J. Patel, S.-e. Ong, J. D. Jaffe, and S. A. Carr. iTRAQ Labeling is Superior to mTRAQ for Quantitative Global Proteomics and Phosphoproteomics. *Molecular & Cellular Proteomics*, 11(6), 2012.

P. Mertins, D. R. Mani, K. V. Ruggles, M. A. Gillette, K. R. Clauser, P. Wang, X. Wang, J. W. Qiao, S. Cao, F. Petralia, E. Kawaler, F. Mundt, K. Krug, Z. Tu, J. T. Lei, M. L. Gatza, M. Wilkerson, C. M. Perou, V. Yellapantula, K.-l. Huang, C. Lin, M. D. McLellan, P. Yan, S. R. Davies, R. R. Townsend, S. J. Skates, J. Wang, B. Zhang, C. R. Kinsinger, M. Mesri, H. Rodriguez, L. Ding, A. G. Paulovich, D. Fenyö, M. J. Ellis, S. A. Carr, and NCI CPTAC. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*, 534(7605):55–62, 2016.

C. E. Murcray, J. P. Lewinger, and W. J. Gauderman. Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, 169(2):219–226, 2009.

B. Neale, M. Rivas, B. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. Purcell, K. Roeder, and M. Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.

L. Pagani, P. A. St. Clair, T. M. Teshiba, S. K. Service, S. C. Fears, C. Araya, X. Araya, J. Bejarano, M. Ramirez, G. Castrillón, J. Gomez-Makhinson, M. C. Lopez, G. Montoya, C. P. Montoya, I. Aldana, L. Navarro, D. G. Freimer, B. Safaie, L.-W. Keung, K. Greenspan, K. Chou, J. I. Escobar, J. Ospina-Duque, B. Kremeyer, A. Ruiz-Linares, R. M. Cantor, C. Lopez-Jaramillo, G. Macaya, J. Molina, V. I. Reus, C. Sabatti, C. E. Bearden, J. S. Takahashi, and N. B. Freimer. Genetic contributions to circadian activity rhythm and sleep pattern phenotypes in pedigrees segregating for severe bipolar disorder. *Proceedings of the National Academy of Sciences*, 113(6):E754–E761, 2016.

G. Paré, N. R. Cook, P. M. Ridker, and D. I. Chasman. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. *PLoS Genetics*, 6(6):e1000981, 2010.

A. G. Paulovich, D. Billheimer, A.-J. L. Ham, L. Vega-Montoto, P. A. Rudnick, D. L. Tabb, P. Wang, R. K. Blackman, D. M. Bunk, H. L. Cardasis, K. R. Clauser, C. R. Kinsinger, B. Schilling, T. J. Tegeler, A. M. Variyath, M. Wang, J. R. Whiteaker, L. J. Zimmerman, D. Fenyo, S. A. Carr, S. J. Fisher, B. W. Gibson, M. Mesri, T. A. Neubert, F. E. Regnier,

H. Rodriguez, C. Spiegelman, S. E. Stein, P. Tempst, and D. C. Liebler. Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Molecular & Cellular Proteomics*, 9(2):242–254, 2010.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

C. Peterson, S. K. Service, A. Jasinska, F. Gao, I. Zelaya, T. Teshiba, C. Bearden, V. Reus, G. Macaya, C. López-Jaramillo, M. Bogomolov, Y. Benjamini, E. Eskin, G. Coppola, N. B. Freimer, and C. Sabatti. Characterization of expression quantitative trait loci in pedigrees from colombia and costa rica ascertained for bipolar disorder. *PLoS Genetics*, 12(5):e1006046, 2016.

B. L. Pierce, L. Tong, L. S. Chen, R. Rahaman, M. Argos, F. Jasmine, S. Roy, R. Paul-Brutus, H.-J. Westra, L. Franke, T. Esko, R. Zaman, T. Islam, M. Rahman, J. A. Baron, M. G. Kibriya, and H. Ahsan. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS Genetics*, 10(12):e1004818, 2014.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81:559–575, 2007.

T. Raj, K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, M. Lee, N. Asinovski, I. Frohlich, S. Imboywa, A. Von Korff, Y. Okada, N. A. Patsopoulos, S. Davis, C. McCabe, H.-i. Paik, G. P. Srivastava, S. Raychaudhuri, D. A. Hafler, D. Koller, A. Regev, N. Hacohen, D. Mathis, C. Benoist, B. E. Stranger, and P. L. De Jager. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, 344(6183):519–23, May 2014.

M. V. Rockman and L. Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, 2006.

P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3:1154–1169, 2004.

N. Rothman, M. Garcia-Closas, N. Chatterjee, N. Malats, X. Wu, J. D. Figueroa, F. X. Real, D. Van Den Berg, G. Matullo, D. Baris, M. Thun, L. A. Kiemeney, P. Vineis, I. De Vivo, D. Albanes, M. P. Purdue, T. Rafnar, M. A. T. Hildebrandt, A. E. Kiltie, O. Cussenot, K. Golka, R. Kumar, J. A. Taylor, J. I. Mayordomo, K. B. Jacobs, M. Kogevinas, A. Hutchinson, Z. Wang, Y.-P. Fu, L. Prokunina-Olsson, L. Burdett, M. Yeager,

W. Wheeler, A. Tardón, C. Serra, A. Carrato, R. García-Closas, J. Lloreta, A. Johnson, M. Schwenn, M. R. Karagas, A. Schned, G. Andriole, R. Grubb, A. Black, E. J. Jacobs, W. R. Diver, S. M. Gapstur, S. J. Weinstein, J. Virtamo, V. K. Cortessis, M. Gago-Dominguez, M. C. Pike, M. C. Stern, J.-M. Yuan, D. J. Hunter, M. McGrath, C. P. Dinney, B. Czerniak, M. Chen, H. Yang, S. H. Vermeulen, K. K. Aben, J. A. Witjes, R. R. Makkinje, P. Sulem, S. Besenbacher, K. Stefansson, E. Riboli, P. Brennan, S. Panico, C. Navarro, N. E. Allen, H. B. Bueno-de Mesquita, D. Trichopoulos, N. Caporaso, M. T. Landi, F. Canzian, B. Ljungberg, A. Tjonneland, F. Clavel-Chapelon, D. T. Bishop, M. T. W. Teo, M. A. Knowles, S. Guarrera, S. Polidoro, F. Ricceri, C. Sacerdote, A. Allione, G. Cancel-Tassin, S. Selinski, J. G. Hengstler, H. Dietrich, T. Fletcher, P. Rudnai, E. Gurzau, K. Koppova, S. C. E. Bolick, A. Godfrey, Z. Xu, J. I. Sanz-Velez, M. D García-Prats, M. Sanchez, G. Valdivia, S. Porru, S. Benhamou, R. N. Hoover, J. F. Fraumeni, D. T. Silverman, and S. J. Chanock. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature Genetics*, 42(11):978–984, 2010.

J. Roy and X. Lin. Analysis of multivariate longitudinal outcomes with nonignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association*, 97(457):40–52, 2002.

C. Saha and M. P. Jones. Asymptotic bias in the linear mixed effects model under nonignorable missing data mechanisms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):167–182, 2005.

E. E. Schadt, S. Woo, and K. Hao. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics*, 44:603–608, 2012.

J. S. Schildcrout, S. P. Garbett, and P. J. Heagerty. Outcome vector dependent sampling with longitudinal continuous response data: stratified sampling based on summary statistics. *Biometrics*, 69(2):405–416, 2013.

R. J. Sela and J. S. Simonoff. Re-em trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2):169–207, 2012.

A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

A. Shah, N. Laird, and D. Schoenfeld. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*, 92(438): 775–779, 1997.

P. G. Smith and N. E. Day. The design of case-control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology*, 13:356–365, 1984.

D. Soave, H. Corvol, N. Panjwani, J. Gong, W. Li, P. Y. Boëlle, P. R. Durie, A. D. Paterson, J. M. Rommens, L. J. Strug, and L. Sun. A Joint Location-Scale Test Improves Power to

Detect Associated SNPs, Gene Sets, and Pathways. *American Journal of Human Genetics*, 97(1):125–138, 2015.

D. J. Stekhoven and P. Bühlmann. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

J. Stephan, O. Stegle, and A. Beyer. A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6:7432, 2015.

S. A. Stouffer, E. A. Suchman, L. C. Devinney, S. A. Star, and R. M. Williams Jr. *The American soldier: adjustment during army life. (Studies in social psychology in World War II, Vol. 1.)*. Princeton University Press, 1949.

B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis. Patterns of Cis regulatory variation in diverse human populations. *PLoS Genetics*, 8(4), 2012.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

D. Thomas. Gene–environment-wide association studies: emerging approaches. *Nature Reviews Genetics*, 11(4):259–272, 2010.

R. Tibshirani. Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1994.

J. M. Torres, E. R. Gamazon, E. J. Parra, J. E. Below, A. Valladares-Salgado, N. Wacher, M. Cruz, C. L. Hanis, and N. J. Cox. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. *The American Journal of Human Genetics*, 95(5): 521–534, 2014.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.

J. W. Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

S. Van Buuren and K. Groothuis-Oudshoorn. Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, 45:1–67, 2011.

C. F. Van Loan. The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics*, 123(1):85–100, 2000.

A. Vazquez, D. Bates, G. Rosa, D. Gianola, and K. Weigel. Technical note: an R package for fitting generalized linear mixed models in animal breeding. *Journal of Animal Science*, 88(2):497–504, 2010.

J. Wang and L. S. Chen. *MixRF: A Random-Forest-Based Approach for Imputing Clustered Incomplete Data*, 2016. URL `https://CRAN.R-project.org/package=MixRF`. R package version 1.0.

J. Wang, E. R. Gamazon, B. L. Pierce, B. E. Stranger, H. K. Im, R. D. Gibbons, N. J. Cox, D. L. Nicolae, and L. S. Chen. Imputing gene expression in uncollected tissues within and beyond GTEx. *The American Journal of Human Genetics*, 98(4):697–708, 2016.

P. Wang, H. Tang, H. Zhang, J. Whiteaker, A. G. Paulovich, and M. Mcintosh. Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing*, pages 315–326, 2006.

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42:D1001–D1006, 2014.

S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid. Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics*, 7(3): 340–350, 2007.

F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y.-H. Zhou, A. Abdellaoui, S. Batista, C. Butler, G. Chen, T.-H. Chen, D. D'Ambrosio, P. Gallins, M. J. Ha, J. J. Hottenga, S. Huang, M. Kattenberg, J. Kochar, C. M. Middeldorp, A. Qu, A. Shabalin, J. Tischfield, L. Todd, J.-Y. Tzeng, G. van Grootheest, J. M. Vink, Q. Wang, W. Wang, W. Wang, G. Willemsen, J. H. Smit, E. J. de Geus, Z. Yin, B. W. J. H. Penninx, and D. I. Boomsma. Heritability and genomics of gene expression in peripheral blood. *Nature Genetics*, 46(5):430–7, 2014.

M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44(4): 369–375, 2012.

Y. Yuan and R. J. A. Little. Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, 65:478–486, 2009.