DIFFERENTIAL EXPRESSION AND FEATURE SELECTION IN THE ANALYSIS OF MULTIPLE OMICS STUDIES

by

Tianzhou Ma

MS in Biostatistics, Yale University, 2013

BS in Genetics and Biotechnology (specialist), University of Toronto, Canada, 2010

Submitted to the Graduate Faculty of
the Department of Biostatistics
Graduate school of Public health in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Tianzhou Ma

It was defended on

March 2nd, 2018

and approved by

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

Zhao Ren, PhD, Assistant Professor, Department of Statistics, Dietrich School of Arts and Sciences, University of Pittsburgh

Faming Liang, PhD, Professor, Department of Statistics, Purdue University, West Lafayette, Indiana

Ying Ding, PhD, Assistant Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Robert Krafty, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

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

Zhao Ren, PhD, Assistant Professor, Department of Statistics, Dietrich School of Arts and Sciences, University of Pittsburgh

DIFFERENTIAL EXPRESSION AND FEATURE SELECTION IN THE ANALYSIS OF MULTIPLE OMICS STUDIES

Tianzhou Ma, PhD

University of Pittsburgh, 2018

ABSTRACT

With the rapid advances of high-throughput technologies in the past decades, various kinds of omics data have been generated from many labs and accumulated in the public domain. These studies have been designed for different biological purposes, including the identification of differentially expressed genes, the selection of predictive biomarkers, etc. Effective meta-analysis of omics data from multiple studies can improve statistical power, accuracy and reproducibility of single study. This dissertation covered a few methods for differential expression (Chapter 2 and 3) and feature selection (Chapter 4) in the analysis of multiple omics studies.

In Chapter 2, we proposed a full Bayesian hierarchical model for RNA-seq meta-analysis by modeling count data, integrating information across genes and across studies, and modeling differential signals across studies via latent variables. A Dirichlet process mixture prior was further applied on the latent variables to provide categorization of detected biomarkers according to their differential expression patterns across studies. We used both simulations and a real application on multiple brain region HIV-1 transgenic rats to demonstrate improved sensitivity, accuracy and biological findings of our method. In Chapter 3, we extended the previous Bayesian model to jointly integrate transcriptomic data from the two platforms: microarray and RNA-seq.

In Chapter 4, we considered a general framework for variable screening with multiple omics studies and further proposed a novel two-step screening procedure for high-dimensional regression analysis in this framework. Compared to the one-step procedure and rank-based sure independence screening procedure, our procedure greatly reduced false negative errors while keeping a low false positive rate. Theoretically, we showed that our procedure possesses the sure screening property with weaker assumptions on signal strengths and allows the number of features to grow at an exponential rate of the sample size.

Public health significance:

The proposed methods are useful in detecting important biomarkers that are either differentially expressed or predictive of clinical outcomes. This is essential for searching for potential drug targets and understanding the disease mechanism. Such findings in basic science can be translated into preventive medicine or potential treatment for disease to promote human health and improve the global healthcare system.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation covers the major methodology work during my five years' PhD studies and builds upon a foundation of excellent mentorship, generous peer support and endless love from the family. First of all, I would like to express the most genuine gratitude to Dr. George Tseng, my main advisor who really brings me to the world of "omics" and guides me how to do good research. He has been so knowledgable, patient and helpful, and serves as a role model scientist and statistician for fresh PhDs and young researchers like me. As a renowned faculty and experienced mentor, he still keeps a low profile and always stays humble, and most of time when we talk and discuss the research, he is more like a friend rather than an advisor. Having the chance to work with him and in his research group is my great honor and I will benefit from it for my entire academic career.

Next, I want to thank my co-advisor Dr. Zhao Ren for his advising on the sure screening paper. I took quite some theoretical courses he offered and became interested in theoretical statistics even if I came from a biological background in undergraduate. I really learnt a lot from him through classes and research and admired his persistent efforts in popularizing the education in statistical theory to students in biostatistics like me. I want to thank Dr. Faming Liang, for providing selfish guidance to me in the first two Bayesian papers, when the Bayesian statistics was still very new to me.

I also want to thank all my other committee members Dr. Ying Ding and Dr. Robert Krafty for providing me a great amount of help and advices on my research, presenting skills and career development.

I want to thank all my former and current lab mates for their support and help in both research and daily life. Without them, I would not have been productive during my PhD studies. Without them, I would have been lonely and may have lost my passion in research.

I would also like to thank my collaborators, Dr. Etienne Sibille from the Centre for Addiction and Mental Health (CAMH), University of Toronto, Drs. Steffi Oesterreich, Anda Vlad and Faina Linkov from Magee-Womens Research Institute (MWRI), who have helped me acquire a lot of collaborative publications and motivated me to think about new methodologies to solve statistical issues in biomedical applications of their fields.

I want to show my deepest gratitude to my parents, Xiaoyun and Aiguang, for their support both spiritually and materially. I am so lucky to have such open-minded parents like you two, you have always been hard-working, sincere and kind to others, and have set an example for me. I also want to thank my other family members, my aunt Yue and my cousin Jing for bringing the joy to life, my uncle Zhan for his positive influence on my attitude towards life and academics, my parents in law Cuifeng and Xiaomin for giving me and Zhongying the best care.

Lastly, I would like to express my greatest gratitude to my lifetime lover, my wife Zhongying. The path to an academic career is a long journey with all kinds of hardship, I am lucky to have you accompany in this journey. Thank you for your willingness to sacrifice the favorable conditions in China and stay and study in the States with me in the past six years and probably in the future. Thank you for all your contribution to the family and your understanding when I spent more time in research than being with you. Thank you and always love you and our little one.

1.0 INTRODUCTION

1.1 OVERVIEW OF HIGH-THROUGHPUT OMICS DATA AND TECHNOLOGIES

The rapid advances and prevalence of various high-throughput experimental technologies have generated abundant omics data in the public repositories in recent years and effective analytical approaches are crucial to fully understand the biological knowledge inside these data. Ending with the same suffix, these "*-omics*" data are used to study an organism's genetic material ("Genomics"), RNA transcripts ("Transcriptomics"), proteins ("Proteomics"), epigenetic modification ("Epigenomics"), etc., all of which play essential roles in the flow of biological information in the central dogma paradigm ($DNA \leftrightarrow RNA \rightarrow Protein$). This section will briefly introduce the various types of omics data, the two major platforms/technologies that generate these data and the public repository and databases of omics datasets.

1.1.1 Genomic data

Genomics is the study of the complete set of genetic material within an organism usually consisting of DNA (RNA for some viruses). Unlike genetics which studies individual genes, it usually applies high-throughput technologies such as DNA sequencing to assemble and analyze the function and structure of entire genomes including both coding and noncoding regions.

The human genome contains approximately 3.2×10^9 base pairs distributed among 22 paired chromosomes plus the two sex chromosomes and the protein-coding sequences account for only 1.5% of the whole genome (Lesk, 2017). The average proportion of nucleotide differences among different human individuals has been consistently estimated to lie between 1 in 1,000 and 1 in 1,500 (Jorde and Wooding, 2004). Considering this nucleotide diversity, personal genomics is the branch of genomics that focused on determining the genetic make-up (a.k.a. genotype) of an individual and comparing to another individual's sequence or a reference sequence.

Genetic variation among individuals can be attributed to independent assortment, cross-over and recombination during meiosis as well as various mutational events. Mutation is a permanent alteration of nucleotide sequence in the genome and the resulting change of DNA is not repairable and the errors will proceed to DNA replication and RNA transcription. It is associated with abnormal biological processes like cancer since changes in DNA can cause errors in protein sequence, creating partially or completely non-functional proteins. There are two types of major mutations: somatic mutation and germline mutation. Somatic mutation takes place in somatic cells and is usually caused by environmental factors. It is neither inherited from parents nor passed to offsprings. Germline mutation occurs in reproductive cells such as sperm or ova and is inheritable. This type of mutation can be transmitted to offspring.

Among the various types of genetic variation, single nucleotide polymorphism (SNP) is the most common one and represents difference in a single nucleotide between members that occurs in at least 1% of the population. Single-nucleotide variant (SNV) is a variation in a single nucleotide without any limitations of frequency and may arise in somatic cells. Genome-wide association study (GWAS) is known as a popular design to assess thousands to millions of common SNPs associated with a disease or a trait. Thousands of disease-susceptible variants have been discovered through the GWAS of hundreds or thousands of individuals (Hindorff et al., 2009; McCarthy et al., 2008). Recently, rare-variant association analysis has aroused more interest in the field which focuses on rare variants that might explain disease risk or trait variability in addition to common variants found in GWAS (Lee et al., 2014).

Other types of genetic variation include insertion/deletion ("indel") polymorphism in which a specific nucleotide sequence is present or absent; copy number variation (CNV), a structure variation of DNA segment due to deletion or duplication of large regions of DNA on some chromosome. CNV has been found to be related with disease phenotype and also account for regulation of genes expression and other genomic process (McCarroll and Altshuler, 2007).

1.1.2 Transcriptomic data

Transcriptomics studies the sum of all RNA molecules (a.k.a. transcripts) in an organism or in a cell, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA) and other non-coding RNA such as microRNA (miRNA), etc. Unlike the genome which is almost fixed for a given cell line, the transcriptome only reflect genes that are expressed at given time and can vary with different external conditions. Microarray and RNA sequencing (RNA-seq) are the two major platforms to quantify the transcriptome and will be introduced in the next section.

mRNA is the major family of RNA molecules that convey genetic information from DNA (known as "transcription") and produce proteins (known as "translation"). In eukaryotes, mRNA is first transcribed into precursor mRNA (pre-mRNA) and has to undertake a few processing steps including 5' cap addition, polyadenylation and splicing before it matures to generate proteins. Splicing is the editing process that removes introns (intervening sequence) from RNA and joins exons (the actual coding part of a gene) together. Since a gene contains multiple exons and mature mRNAs from the same gene can include different exons, alternative splicing can take place and produces multiple protein isoforms in the translation stage. Distinct from the stable DNA molecules, mRNA molecules have a short half life and will ultimately end in degradation.

Other RNA molecules, though not necessarily translated into protein products, play important roles in regulating and catalyzing the transcription, translation and other biological processes inside the cell. For example, rRNAs are basic components of the ribosome and catalyze the transcription; tRNAs transfers specific amino acids to a growing polypeptide to synthesize protein during translation; miRNAs function in RNA silencing and post-transcriptional regulation of gene expression.

Genetics have effects on the transcriptome. Expression quantitative trait loci (eQTLs) are genomic loci that contribute to variation in mRNA expression. Using RNA-seq samples from the 1000 Genome project, recent studies uncovered extremely widespread regulatory variation, with 3773 genes having a classical eQTL for gene expression levels (Lappalainen et al., 2013). Based on the distance to their gene-of-origin, eQTLs can be further divided into two types: cis-eQTLs (locally) and trans-eQTLs (at a distance).

1.1.3 Other omics data

There are other important omics data that are not the focus of this thesis, for example the epigenomics and the proteomics.

Epigenome is the complete set of epigenetic modifications, including DNA methylation, histone modification and chromatin structure change. It plays an indispensable role in gene expression and regulation and partially determines one's phenotype in addition to genotype and environment (Lesk, 2017). DNA methylation is the process methyl groups are added to nucleotides in DNA and is associated with a number of key processes, e.g. genomic imprinting, X-chromosome inactivation, silencing of repetitive DNA, etc. (Schübeler, 2015). Methylation takes place at the cytosine nucleotide in eukaryotes and when it is followed by a guanine nucleotide it forms a CpG site. Approximately 60% of CpG sites are methylated in somatic cells in vertebrates and those DNA regions with high frequency of CpG sites are also called CpG islands (Bird, 2002). To quantify the methylation level, scientists define the beta value for a CpG site as the percentage of methylated events out of all events which ranges between 0 and 1. The alteration of DNA methylation pattern has been outstanding in cancer, where the loss of expression of genes is about 10 times more frequently by hypermethylation of CpG islands in the promoter region than by mutations (Vogelstein et al., 2013).

Proteomics studies the entire set of protein products in an organism. Proteins are made up of long chains of amino acids with 3D configuration and perform vast array of functions inside the body. Like the transcriptome, proteome also varies with time and condition in given cell or system. To detect and quantify proteins, researchers either apply antibody-based methods (immunoassays) or mass spectrometry-based techniques. In addition to the expression profiling of proteins, computational biologists also use technologies like X-ray crystallography and NMR spectroscopy to perform structural analysis of proteins looking for e.g. potential drug binding sites.

There are multiple levels of molecular variation from different omics data that contribute to disease risk in a nonlinear, interactive and complex way and there also exist cross-talk among different types of omics data (Ritchie et al., 2015). For example, like eQTLs, researchers also characterized DNA methylation quantitative trait loci (mQTLs) and showed their important functions especially in the brain (Hannon et al., 2016). Other examples include ChIP-sequencing (ChIP-seq) which combines chromatin immunoprecipitation and DNA sequencing and is used to analyze protein (e.g. transcription factor) interaction with DNA.

1.1.4 High-throughput technologies in omics research

1.1.4.1 Microarray Before the advent of microarray techniques, biologists use northern plot or quantitative Polymerase Chain Reaction (qPCR) to study and quantify gene expression. These techniques are time consuming and expensive, since they perform gene-by-gene analysis and can only detect up to dozens of genes. As one of the earliest high-throughput technologies, the invention of DNA microarray in the early 1990s marks the start of the omics era and makes it possible to measure the expression levels of thousands of genes up to the whole genome or to genotype multiple regions of a genome simultaneously. In its application to gene expression profiling, tens of thousands of transcript-specific probes are immobilized on a solid support, such as a microscope glass slide or silicon chips, to make up the "microarray." RNA samples are reversely transcribed to cDNA, fluorescently labeled, amplified and hybridized to the microarray. The array is then washed and the expression level is quantified by measuring fluorescence intensity at each spot (Figure 1 left). In addition to gene expression profiling, the microarray technique can also be applied to detect SNPs, CNVs, DNA methylation, and protein-DNA binding.

There are a few limitations with DNA microarrays. First, microarray has detection limit at the lower end thus the intensities of low-expressed genes are un-distinguishable from background noise; Secondly, microarray only provides an indirect measure of relative concentration, at high concentrations it will become saturated and at low concentrations the equilibrium will favor no binding; Finally, DNA microarray can only detect known sequences it was designed to detect (Bumgarner, 2013; Mortazavi et al., 2008). Due to these disadvantages, microarray is now gradually being replaced by the newer RNA-seq technique for gene expression profiling.

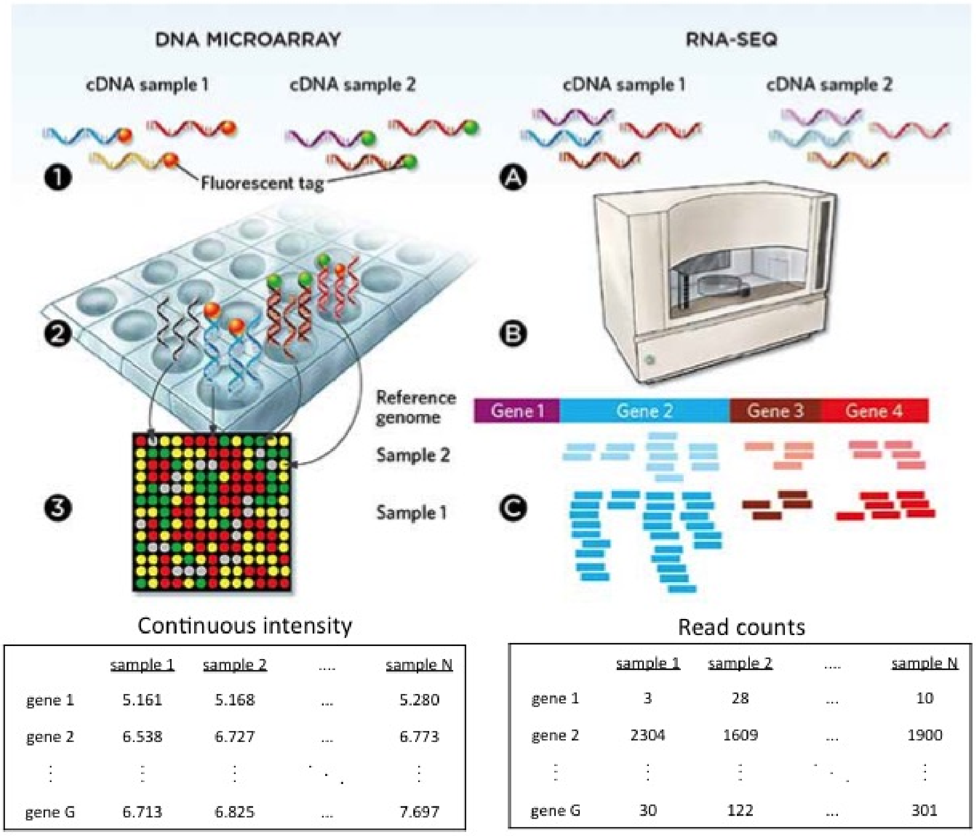
1.1.4.2 Next generation sequencing In the past decade, there has been a fundamental shift from the more traditional Sanger sequencing to the next-generation sequencing (NGS) for genomic analysis. With run time as short as a few hours to sequence the whole genome of a sample, the arrival of NGS technologies has changed the way we think about scientific approaches in basic, applied and clinical research (Metzker, 2010). NGS is also called ultra-high-throughput sequencing which can process millions of sequence reads simultaneously and have been widely applied to genome sequencing (whole genome sequencing and whole exome sequencing), transcriptome profiling (RNA-seq), DNA methylation (bisulfite sequencing), DNA-protein interaction (ChIP-sequencing), etc. In a typical RNA seq workflow, the cDNA samples are chopped into DNA fragments (called "reads") with specific adapter oligos bound to both ends and the DNA fragments are then sequenced. The reads generated by sequencers can vary by read lengths depending on user preference, technologies or platforms (e.g. Illumina, SOLiD, Roche). The sequenced reads are reassembled and aligned to a reference genome to quantify the expression levels of genes or transcripts by counting the number of mapped short reads (Figure 1 right).

Comparing to the DNA microarray, RNA-seq has quite a few advantages. First, RNA-seq has higher sensitivity and accuracy in quantifying the low-expressed genes; Secondly, RNA-seq can be used to detect novel transcripts or isoforms which is impossible in microarray with only known probes. Last but not the least, it can also be used to examine transcriptome fine structure such as allele-specific expression and splice junctions (Wang et al., 2009).

As shown in Figure 1, DNA microarray will generate a matrix of continuous intensity values while RNA-seq will generate a matrix of count data after a series of preprocessing steps for each platform. If we perform some normalization by both library size (i.e. total counts in a sample) and gene length, we can also generate continuous values such as RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million) or TPM (Transcripts Per Kilobase Million) from RNA-seq. However, such normalization introduces a length bias into the variance and is less powerful than the count data (Oshlack and Wakefield, 2009). The first paper of my thesis proposed a new method to integrate multiple RNA-seq count data and the second paper extended the method to integrate continuous data from microarray and count data from RNA-seq.

1.1.5 Public resource of omics data

With the rapid advances of high-throughput technologies and their reduction in cost in the past decades, generation of various kinds of omic data becomes affordable and prevalent in many labs. For example, large amount of transcriptomic data have been accumulated from microarray or RNA-seq experiments for different biological aims and have been stored in large data repositories such as Gene Expression Omnibus (GEO), ArrayExpress and Sequence Read Archive (SRA). Hundreds of GWAS studies have been conducted since 2000s and many datasets are stored in database of Genotypes and Phenotypes (dbGaP). In addition to these public available databases, many worldwide and nationwide consortium projects were launched in the last 10 to 15 years for different aims and generated omics data with large sample size and high quality to serve the whole scientific community. For instance, The Cancer Genome Atlas (TCGA), a "community resource project" initiated by National Cancer Institute (NCI), has profiled and analyzed a total of 33 cancer types including 10 rare cancers and generated rich amount of data at the DNA (mutation, copy number variation, etc.), RNA (gene expression, miRNA expression, etc.), protein (protein expression) and epigenetic (DNA methylation) levels. The Encyclopedia of DNA Elements (ENCODE) is a public research project aiming to identify and annotate all functional elements in the human genome including both coding and non-coding parts. The MODel organism ENCyclopedia Of DNA Elements (modENCODE) project further extends the original ENCODE project to identify the functional elements in selected model organism genomes. For omics data from in vitro cell



Continuous intensity

	sample 1	sample 2	....	sample N
gene 1	5.161	5.168	...	5.280
gene 2	6.538	6.727	...	6.773
⋮	⋮	⋮	⋱	⋮
gene G	6.713	6.825	...	7.697

Read counts

	sample 1	sample 2	....	sample N
gene 1	3	28	...	10
gene 2	2304	1609	...	1900
⋮	⋮	⋮	⋱	⋮
gene G	30	122	...	301

Figure 1: Measuring gene expression: DNA microarray vs. RNA-seq

cultures or immortal cell lines, the Cancer Cell Line Encyclopedia (CCLE) project conducted detailed genetic characterization (copy number variation, mRNA expression, mutation and more) for more than 1,000 cancer cell lines (Barretina et al., 2012).

The affluent omics datasets in the public domain provide opportunities and have motivated us to combine data from multiple cohorts (from different platforms) for different biological purposes such as differential expression analysis (paper 1 & 2) and prediction analysis with feature selection (paper 3)

1.2 OBJECTIVES OF OMICS STUDIES AND RELEVANT ANALYSIS

The various omics studies mentioned above are designed for different biological purposes. For transcriptomic studies, the most common purpose is to identify genes that are differentially expressed among predefined classes, e.g. between the diseased patients and normal controls. In addition, researchers are also interested in identifying important biomarkers (e.g. genes, SNPs, etc.) that can predict clinical outcomes or classify new patients. Some experiments are designed to identify novel subtypes based on the omics data, and other experiments aim at exploring the relationship among the genes or proteins via biological networks. In this section, I will briefly introduce these biological objectives and the types of statistical analysis involved.

1.2.1 Differential expression analysis

An important task in genomic data analysis is to identify candidate markers associated with the disease status, disease progression or environmental perturbation. In a two class comparison scenario, genomic comparative studies applied differential expression (DE) analysis methods on microarray or RNA-seq data to select genes that are differentially expressed between case and control.

The gene expression data from microarray is continuous and usually normally distributed. In the simplest scenario, we can fit the following linear model for each gene to test whether it is differentially expressed:

$$y_{gi} = \alpha_g + \beta_g X_i + \sum_{p=1}^{P} \gamma_{pg} Z_{pi},$$

where y_{gi} is the expression value for the gth gene and ith sample, X_i the indicator of the condition (e.g. 1 for case and 0 for control), α_g is the gene-specific intercept and Z_{pi} indicate the known confounding covariates you wish to adjust. The purpose is to test whether β_g is zero or not. Simple linear model via e.g. traditional t-test will underestimate variance by chance when the sample size n is small but the number of genes G is large, to overcome , Smyth (2004) proposed a specific linear model for microarray data (called "LIMMA") using empirical Bayes approach and suggested a moderated t-statistics by shrinking the estimated sample variances towards a pooled estimate for more stable inference (Smyth, 2005). SAM (short for "Significance Analysis of Microarray") is another popular tool for differential analysis in microarray that uses nonparametric statistics (Tusher et al., 2001).

For count data obtained from RNA-seq, the linear model has to be extended to a generalized linear model (GLM) setting:

$$g(E(y_{gi})) = T_i + \alpha_g + \beta_g X_i + \sum_{p=1}^{P} \gamma_{pg} Z_{pi},$$

where y_{gi} is the expression value for the gth gene and ith sample, $g(.)$ is the link function (usually use "log"), T_i is the offset of ith sample adjusting for the sequencing depth of each sample and α_g is the gene-specific intercept. Negative binomial distribution is a more popular choice than Poisson to fit y nowadays since over-dispersion is commonly seen in RNA-seq data. edgeR and DESeq are the two most widely used tools for the differential expression analysis in RNA-seq, both of which are under the GLM framework with more careful estimation of the dispersion parameter (McCarthy et al., 2012; Anders and Huber, 2010).

For DE analysis of high-dimensional genomic data, multiple comparison is one big issue always needs to be addressed. There are two general categories of methods for multiple comparison correction in the literature. The first category aims to control for the family-wise error rate (FWER) (Hochberg and Tamhane, 2009), corresponding to the probability of making at least one false discovery. Common methods falling in this category includes the Bonferroni procedure which is a popular choice in GWAS. However, such methods are usually too stringent for DE analysis in transcriptomic studies. The second less stringent category is designed to control the false discovery rate (FDR) (Benjamini and Hochberg, 1995), defined as the expected proportion of false positives among all positive "discoveries" (i.e. the type I error). Methods under this alternative category such as Benjamini-Hochberg procedure or Bayesian FDR are more popular choices in genomic studies.

The identification of important biomarkers are useful to narrow down target for further investigation, however, they may still contain little unifying biological theme for most researchers. Thus, the pathway analysis (a.k.a. gene set enrichment test) usually following the DE analysis is applied for functional annotation of the identified gene list, based on one or multiple known pathway database. Commonly used pathway databases include Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, etc.

The first and second papers in this dissertation proposed new methods for differential expression analysis when there are multiple transcriptomic datasets from multiple platforms.

1.2.2 Regression analysis with feature selection

In statistics, regression can be used to explore the relationship between independent variables and a dependent variable. When there are many independent variables present (i.e. multiple regression), we wish to identify those that are most predictive of the dependent variable. In omics studies, this might include the identification of genes or SNPs that can predict the disease status, survival or some specific quantitative measures, etc. When the outcome is binary or categorical, it becomes a classification problem in machine learning.

Consider a linear regression setting with n samples and p features (e.g. genes):

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i,$$

where y_i is the outcome for the ith sample, X_ij is the (expression) value for the ith sample and jth feature and ϵ_i is the error term. This regression model is very different from the one in the DE analysis, where y corresponds to the feature.

Very frequently in omics studies, we are facing the high-dimension data with "small n, large p" ($p >> n$). In that case, the matrix $\mathbf{X}$ is singular and most parameters are not identifiable. Conventional methods such as principal component analysis (PCA) or singular value decomposition (SVD) can be applied on $\mathbf{X}$ to reduce dimension, however, such implementation will lose the individual feature identity and interpretability. Alternatively, regularization approaches can be applied to solve such ill-posed regression problem. Over the past two decades, many regularization methods have been developed and can be summarized in the form penalized likelihood by solving the following objective function:

$$\hat{\beta} = arg \min_{\beta} \; (||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda||\beta||_q),$$

where λ is penalty parameter. When $q = 0$ (L_0 norm), it becomes the traditional model selection by AIC/BIC; when $q = 1$(L_1 norm), this is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996; Zou, 2006); when $q = 2$ (L_2 norm) this is the group version of the LASSO (Yuan and Lin, 2006); when the square of L_2 norm is used, this is the ridge regression; when both L_1 norm and L_2 norm are used, this is the elastic net (Zou and Hastie, 2005). In particular, LASSO method and its extensions (i.e. group LASSO, elastic net, adaptive LASSO, etc.) induce sparsity in the regression model and achieve the goal of feature selection. In the Bayesian school, the feature selection is achieved by putting sparsity-induced priors like Spike-and-slab prior (George and McCulloch, 1993; Ishwaran and Rao, 2005) or shrinkage priors like Laplace prior (a.k.a. Bayesian LASSO) (Park and Casella, 2008).

However, when p is very large, the computational cost can be a hurdle for most regularization methods and some theoretical assumptions may no longer hold. Sure screening methods such as sure independence screening (SIS) have been proposed as a natural way to select relevant variables based on their marginal correlation (Fan and Lv, 2008). The general idea is to first reduce a high dimensional model to a relatively lower dimensional model which still contains the true model almost surely via sure screening and then performs model selection using one of the aforementioned regularization approaches. With improvement in both speed and performance, sure screening methods have gained more popularity in various statistical fields these years.

The third paper of this dissertation proposed a new screening method for the scenario when datasets from multiple homogeneous studies are present.

1.2.3 Clustering and network analysis

In addition to those mentioned above, there are many omics studies designed for other biological purposes and applied different analysis approaches.

When the class labels are unknown, researchers will apply clustering analysis looking for novel subtypes based on e.g. gene expression profiles of the samples, which could serve as a guide to precision medicine. There are two major classes of clustering methods: distance-based clustering and model-based clustering. The former includes the most commonly used and heuristic algorithms such as K-means and hierarchical clustering, etc, and the latter is based on mixture model setting and usually require assumptions on data distributions. In addition to sample clustering, researchers may also be interested in clustering genes looking for tightly coexpressed gene modules in the transcriptomic studies. Similar clustering approaches can be applied on the other dimension.

Other experiments are designed to understand the interactions between components (e.g. genes, proteins) in a biological system. Graphical model and network analysis are the common tools to serve for this purpose. Typical networks include binary network, weighted network, directed network and undirected network, etc. Directed graphs (e.g. Bayesian network) puts directions on edges and can be used to model the causal relationship in omics data, e.g. gene regulatory mechanism. The edges of undirected graphs, on the other hand, are without direction and only represent either marginal dependence (e.g. co-expression network) or conditional dependence (e.g. gaussian graphical model) between two linked nodes.

1.3 DATA INTEGRATION AND META-ANALYSIS

In high-throughput omics studies, individual studies usually have small sample size. Combining multiple studies/cohorts using meta-analysis methods improve statistical power, estimation accuracy and reproducibility and has become popular in genomic research. Such genomic information integration of multiple transcriptomic studies is also termed "horizontal meta-analysis." On the other hand, for large cohort such as TCGA which includes multiple levels of omics data (e.g. gene expression, CNV, genotype, methylation, somatic mutation, etc.) of the same patient cohort, we are also interested in jointly analyzing these data to investigate disease subtypes, disease associated or driver genes and related regulatory network. We call such analysis "vertical integrative analysis."

1.3.1 Horizontal meta-analysis

Many "horizontal meta-analysis" methods have been developed and widely applied in the real data analysis for different biological purposes. Tseng et al. (Tseng et al., 2012) reviewed a collection of 333 microarray meta-analysis papers, in which multiple microarray studies are combined for a variety of purposes including differentially expressed gene detection, pathway analysis, inter-study prediction analysis, network and co-expression analysis, etc. More recently, methods were developed to combine multiple transcriptomic studies for other purposes including simultaneous dimension reduction (MetaPCA) (Kim et al., 2017) and robust disease subtype discovery (MetaClust) (Huo et al., 2016), etc.

There are three main categories of meta-analysis methods for transcriptomic DE analysis. The most popular one is the two-stage method, where a single summary statistics is first computed for each study by applying "state-of-the-art" methods introduced in the last section and then meta-analysis methods are used to combine the summary statistics. These methods include combining p-values (Fisher, 1925; Stouffer et al., 1949), combining effect sizes (Choi et al., 2003) or combining rank statistics (Hong et al., 2006). Among them, Fisher's method and Stouffer's method are the most popular ones to aggregate evidence from multiple studies. Adaptive-weighted Fisher's method (AW-Fisher) extends the equally weighted Fisher's method and adds binary weights to handle the study heterogeneity and categorization biomarkers (Li et al., 2011). The second category of methods merges the raw data from all studies and normalizes simultaneously (a.k.a. mega-analysis), then standard single-study analysis can be applied. These approaches have, however, been less favored in the literature since they do not guarantee to remove cross-study discrepancy and may fail to retain study-specific biomarkers. Lastly, the third category integrates DE information from all studies by using a unified and joint stochastic model. Since they are joint hierarchical models by nature, the more flexible Bayesian methods are usually applied. Depending on the hypothesis and biological questions of interest, these approaches have the potential to offer additional efficiency over the two-stage methods and, at the same time, retain the study-specific features. The meta differential analysis methods proposed in the first and second papers of this thesis applies Bayesian joint modeling and falls in the third category.

1.3.2 Vertical integrative analysis

With the large amount of omics data accumulated in public databases and depositories, vertical integrative analysis becomes appealing to explore the regulatory relationships between different levels of omics data. Omics integrative analysis has been found successful in many applications to tumor studies including ovarian cancer (Network et al., 2011), breast cancer (Network et al., 2012), stomach cancer (Network et al., 2014), to name a few.

In the field of bioinformatics, vertical integration methods have been developed for clustering and prediction analysis. Lock and Dunson (2013) fit a finite Dirichlet mixture model to perform Bayesian consensus clustering (namely "BCC") that allows both common and omic-type specific clustering patterns. Shen et al. (2009) applied a latent variable factor model (namely "iCluster") to cluster tumor samples by integrating multi-omics data. Huo and Tseng (2017) built on a sparse K-means framework to perform clustering with overlapping feature groups (Jacob et al., 2009). Wang et al. (2012) proposed an integrative Bayesian analysis of genomics data (called "iBAG") framework to identify important genes/biomarkers that can predict the clinical outcome and successfully applied their method to TCGA glioblastoma datasets.

1.4 FUNDAMENTALS OF BAYESIAN DATA ANALYSIS AND ITS APPLICATION IN OMICS STUDIES

Building upon the famous Bayes' theorem, the Bayesian statistics is characterized by its explicit use of probability for quantifying uncertainty in inferences based on statistical data analysis (Gelman et al., 2014). One major difference from the frequentist inference is that the Bayesian methods start with the assumption that the parameter is random with population or prior density while the data is fixed. In general, the process of Bayesian data analysis can be summarized into three steps according to Gelman et al. (2014):

- Setting up a full probability model. Such a probabilistic model should clearly specify the observed data/quantities and unknown parameters we wish to estimate, and take any prior knowledge into consideration.
- Conditioning on observed data to compute the posterior distribution. In Bayesian statistics, the main inference is drawn from an appropriate posterior distribution, i.e. the conditional probability distribution of unknown parameters given the observed data. According to the Bayes' theorem, there is one simple memorable form to represent the relationship among the prior, likelihood and posterior: $Posterior \propto Likelihood \times Prior$.
- Evaluating the model fit and implications of the posterior distributions. This step is similar to most frequentist approaches and involves the assessment of model fit, checking of model assumption and sensitivity analysis, etc.

Calculation of the posterior distribution in the second step usually requires elaborate and efficient Bayesian computation. There are two main categories of methods in Bayesian computation: one by obtaining samples from the posterior distribution (stochastic) and the other by computing expectations and integrals under the posterior distribution (deterministic). The most popular method in the first category is the Markov chain Monte Carlo (MCMC) approach, which draws parameter values from approximate distributions and then correct the draws to better approximate the target posterior distribution (Gelman et al., 2014). Metropolis-Hastings (MH) algorithm is one typical MCMC method which generates a random walk using a proposal density and provides procedure to either accept or reject the moves (Metropolis et al., 1953; Hastings, 1970). When the full conditional distribution of each parameter is (usually requires conjugate prior) known, Gibbs sampling algorithm can be applied instead (Geman and Geman, 1984). Sampling-based methods are usually computationally heavy, on the other hand, the second category of method tackles the problem by constructing distributional approximations and finding the posterior mode. Methods falling in this category include variational Bayes, Laplace approximation, etc.

As a general trend towards assumption-free and more robust statistics these years, the Bayesian school has also turned to more nonparametric Bayesian models with parameter space having infinite dimension. One typical example is the use of the nonparametric Dirichelet process (DP) model (a.k.a. the Chinese Restaurant Process) in clustering problems (Ferguson, 1973). Such models are assumption free and allow infinite number of clusters and have extensive application in natural language processing and bioinformatics problems. The full Bayesian model proposed in the first paper of the dissertation also includes a Dirichlet process mixture model part for biomarker categorization across studies.

There is growing body of new Bayesian approaches that are developed for application in omics studies over the years. For example, Lewin et al. (2006) proposed a full Bayesian hierarchical model to detect differentially expressed genes and accounted for the array effects in microarray. Sha et al. (2004) developed new Bayesian variable selection approach to identify genes for the classification of disease stages. Tadesse et al. (2005) developed a new method for sample clustering via finite mixture model with similar bayesian variable selection approaches using the DNA microarray data. Medvedovic and Sivaganesan (2002) developed a new procedure to cluster genes based on the Dirichlet mixture model. For the omics data other than gene expression, Morris et al. (2008) proposed Bayesian wavelet-based functional mixed models to analyze the mass spectrometry proteomic data. Zhang et al. (2010) presented a new Bayesian partition method to detect pleiotropic and epistatic eQTL modules. Li et al. (2010) proposed a two-stage hierarchical model with Bayesian lasso to model and analyze multiple SNPs in GWAS.

Comparing to frequentist approaches, Bayesian approaches have at least two major benefits which make them a popular choice especially in omics data application. First, it has the flexibility and advantage to incorporate prior biological knowledge or evidence into the statistical models (Do et al., 2006). This is very common to see in most omics data which usually involve quite some underlying biology, e.g. signaling pathway, gene regulatory mechanism, etc. Secondly, the construction of hierarchy in Bayesian model is easy and straightforward. This actually fits our perspective of the complex hierarchical biological relationships among various molecular features (the different types of omics data as measured by different platforms) inside our body.

In the first two papers in this dissertation, we developed new Bayesian hierarchical models to integrate datasets from multiple RNA-seq studies or from both RNA-seq and microarray platforms and showed the improved performance of the proposed Bayesian joint model in DE gene detection.

1.5 OVERVIEW OF THE DISSERTATION

My dissertation contains five chapters. Chapter 1 contains overall introduction of omics data, experimental techniques, high through-put analysis methods, motivation of genomic integrative analysis and fundamental knowledge of Bayesian analysis and its omics application. These contents serve as the background knowledge for the methodology development for Chapter 2, 3 and 4.

Chapter 2 introduced a full Bayesian hierarchical model for the meta-analysis of RNA-seq count data from multiple studies. We built the hierarchy based on a negative binomial regression framework in each study and allowed the sharing of information across studies ("meta-analysis"). In addition, we applied a Dirichlet process mixture (DPM) prior to the latent differential expression indicators for simultaneous biomarker detection and categorization across studies. The contents in this Chapter have been published in the Journal of the Royal Statistical Society: Series C (Ma et al., 2017c).

Chapter 3 introduced a full Bayesian hierarchical model to jointly integrate microarray continuous intensity data and RNA-seq count data from multiple transcriptomic studies. To account for the systematic bias in fold change across RNA-seq and microarray for detecting differentially expressed genes previously reported, we incorporated a normalization procedure to improve detection accuracy and power. The contents in this Chapter have been published in the Journal of Computational Biology (Ma et al., 2017b).

Chapter 4 introduced a general framework as well as a two-step screening procedure for feature selection in high-dimension regression analysis with multiple omics studies. The two-step procedure greatly reduced the false negatives errors while keeping a low false positive rate in practice and enjoyed the sure screening property with weaker assumptions.

Chapter 5 is discussion and future work. For omics data integration, we can readily propose a full Bayesian hierarchical model to meta-analyze multiple epigenomic studies from different platforms. For sure screening, we can extend our two-step screening procedure to accommodate other model settings such as generalized linear models, quantile regression, etc. by modifying the marginal measures.

2.0 RNA-SEQ META-ANALYSIS USING BAYESIAN HIERARCHICAL MODEL

2.1 INTRODUCTION

By using the next-generation sequencing technology to quantify transcriptome, RNA-seq has rapidly become a standard experimental technique in measuring RNA expression levels (Mortazavi et al., 2008; Wang et al., 2009). For RNA-seq, the abundance of transcript in each RNA sample is measured by counting the number of randomly sequenced fragments aligned to each gene. Compared to the popular microarray technology, RNA-seq has the advantage of detecting novel transcripts and quantifying a larger dynamic range of expression levels. It has been shown that RNA-seq performs better than microarray at detecting weakly expressed genes if sequencing is deep enough (Wang et al., 2014). However, new statistical challenges emerge in the differential expression analysis of RNA-seq data. First, the sequencing data are discrete counts rather than continuous intensities, so a count model is more appropriate if parametric approach is used. Secondly, since long transcripts usually have more mapped reads compared to short transcripts and the detection power of differential expression increases as the number of reads increases, short transcripts are always at a statistical disadvantage relative to long transcripts in the same dataset. Analysis of RNA-seq data needs to address such a read count bias considering the fact that many important disease markers are of short length or low expression (Oshlack and Wakefield, 2009).

Many methods have been developed to identify differentially expressed genes between two or more conditions for RNA-seq count data. Two most popular tools edgeR and DE-Seq assume a negative binomial model that takes overdispersion into account and either likelihood ratio test or exact test is used to test for differential expression (Robinson et al., 2010; Anders and Huber, 2010). Other methods such as baySeq or EBSeq applied empirical Bayes approaches to detect patterns of differential expression (Hardcastle and Kelly, 2010; Leng et al., 2013). Recently, more methods have been developed using Bayesian hierarchical model and have used either approximation methods or Markov chain Monte Carlo (MCMC) sampling schemes to estimate the parameters (Van De Wiel et al., 2012; Chung et al., 2013). No single method has been shown to outperform the other methods under all circumstances in recent comparative studies (Rapaport et al., 2013; Soneson and Delorenzi, 2013). Bayesian approaches are advantageous in handling complex models and adopting more flexible modelling of effect size and variance, and thus may increase DE detection power for lowly expressed genes (Chung et al., 2013). However, all Bayesian hierarchical models are limited to single transcriptomic study so far.

Meta-analysis in genomic research is a set of statistical tools for combining multiple "-omics" studies of a related hypothesis and can potentially increase the detection power of individual studies (Tseng et al., 2012). With the increasing availability of mRNA expression data sets, many transcriptomic meta-analysis methods for microarray data have been developed in the past decade. These methods mainly fall into three categories. The first and the most popular one is a two-stage method, where a single summary statistics is first computed for each study and then meta-analysis methods are used to combine the summary statistics. These methods include combining p-values (Fisher, 1925; Stouffer et al., 1949; Li et al., 2011), combining effect sizes (Choi et al., 2003) or combining rank statistics (Hong et al., 2006). The second category of methods merges the raw data from all microarray studies and normalize simultaneously (a.k.a. mega-analysis), then standard single-study analysis can be applied (Lee et al., 2008; Sims et al., 2008). These approaches have, however, been less favored in the literature since they do not guarantee to remove cross-study discrepancy and may fail to retain study-specific biomarkers. Instead of using two-stage approaches (i.e. DE analysis in single study + meta-analyze summary statistics in the first category, and normalization and combined DE analysis in the second category), the third category integrates differential expression information from all studies using a unified and joint stochastic model (Conlon et al., 2006; Scharpf et al., 2009). Since they are joint hierarchical models by nature, the more flexible Bayesian methods are usually applied. These approaches have the potential to offer additional efficiency over the two-stage methods and, at the same time, retain the study-specific features. This motivates us to develop a Bayesian hierarchical model for RNA-seq meta-analysis.

In the literature, almost no meta-analysis methods have been developed for RNA-seq so far. Two existing R packages claimed for RNA-seq meta-analysis – metaRNASeq (Rau et al., 2014) and metaSeq (Tsuyuzaki and Nikaido, 2013) – essentially applied naive two-stage methods by using DESeq or NOISeq methods in single study and combining p-values by Fisher's or Stouffer's method. The two-stage approach leads to loss of statistical power especially when the observed counts in a given gene are small. In this paper, we propose a Bayesian hierarchical model, BayesMetaSeq, under a unified meta-analytic framework, to jointly analyze RNA-seq data from multiple studies. Bayesian hierarchical modelling allows sharing of information across studies and genes to increases DE detection power for genes with low read counts. In addition, a Dirichlet process mixture (DPM) prior is imposed on the DE latent variables to model the homogeneous and heterogeneous differential signals across studies. Model-based clustering embedded in the full Bayesian model provides categorization of detected biomarkers according to their differential expression patterns across studies. The result facilitates better biological interpretation and hypothesis generation.

Ramasamy et al. (2008) presented seven key issues when conducting microarray meta-analysis, including identifying and extracting experimental data, preprocessing and annotating each dataset, matching genes across studies, statistical methods for meta-analysis, and final presentation and interpretation. When combining RNA-seq studies for meta-analysis, most preliminary steps and data preparation issues will similarly apply. Identification and decision to include adequate transcriptomic studies into meta-analysis greatly impacts accuracy and reproducibility of biomarker detection (Kang et al., 2012). Many useful RNA-seq preprocessing tools such as fastQC, tophat and bedtools are instrumental for alignment and preparing expression counts for downstream analysis. Genes are matched across studies using standard gene symbols or isoforms through a common reference genome (e.g. hg18 or hg19) (Oshlack et al., 2010). In the remaining of this paper, we assume that data collection and preprocessing have been carefully done and we only focus on downstream meta-analytic modeling and interpretation.

The paper is organized as follows. Section 2.2 describes the Bayesian hierarchical model and an MCMC algorithm for simulating posterior distributions of parameters. Section 2.3 explains how we perform differential expression analysis and cluster analysis based on Bayesian inference with multiple comparison addressed from a Bayesian perspective. In Section 2.4 and 2.5, we apply BayesMetaSeq to both simulation and a multi-brain-region RNA-seq dataset from HIV transgenic rat. Final conclusions and discussion are provided in Section 2.6.

2.2 BAYESIAN HIERARCHICAL MODEL

2.2.1 Notation and Assumptions

In this paper, we denote by y_{gik} the observed count for gene g and sample i in study k, $T_{ik} = \sum_{g=1}^{G} y_{gik}$ the library size (i.e. the total number of reads) for sample i in study k and $X_{ik} \in \{0, 1\}$ the phenotypic condition of sample i in study k. The observed data are:

$$D = \{(y_{gik}, T_{ik}, X_{ik}) : g = 1, \dots, G; i = 1, \dots, N_k; k = 1, \dots, K\},$$

where G is the total number of genes, N_k is the sample size of study k and K is the number of studies in the meta-analysis. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of differential expression for gene g in study k, meaning gene g is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$.

Here we assume that the raw RNA-seq count values follow a negative binomial distribution under each condition. We also assume that genes are matched across studies. Although the model could be readily extended to analyze multiple studies with similar but not completely overlapped gene sets. In the next three subsections, we will introduce the generative model within each study (Section 2.2.2), describe information integration of effect sizes across studies (Section 2.2.3) and model clusters of genes with different DE patterns across studies (Section 2.2.4). Figure 2 provides a graphical representation of the full Bayesian hierarchical model. Parameters within the rectangle form the main model and parameters outside the rectangle are hyperparameters. The gray shaded parameters δ_{gk} (latent variable of DE

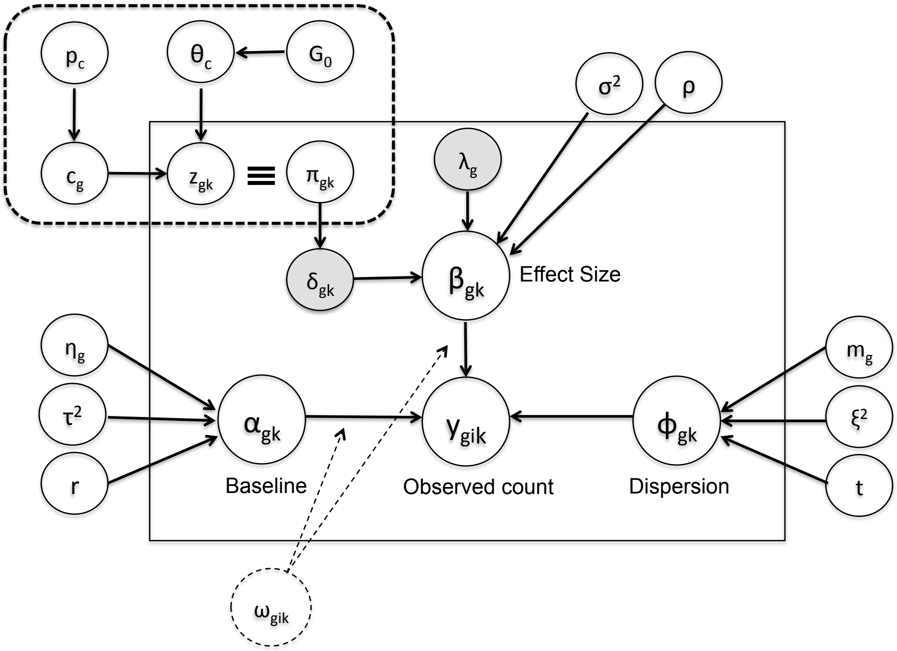


Figure 2: "BayesMetaSeq": Graphical representation of the Bayesian hierarchical model

indicator) and λ_g (DE effect size) are the parameters of interest in the model. The dashed rectangle refers to a Dirichlet process mixture (DPM) model for DE gene categorization that will be described in Section 2.2.4.

2.2.2 Generative model within each study

Below, we describe the generative model for observed data within each study. We assume the counts y_{gik}, conditioning on hyperparameters, are independent and follow a negative binomial distribution. Denote by $\mu_{gik} = E(y_{gik})$ the mean expression level and ϕ_{gk} the gene-specific dispersion parameter, we have:

$$y_{gik} \sim NB(\mu_{gik}, \phi_{gk}). \tag{2.2.1}$$

We then fit a log-linear regression model for the mean μ_{gik}, where α_{gk} denotes the baseline expression relative to the library size and β_{gk} denotes the effect size (i.e. the log fold change of expression between the two conditions):

$$\log(\mu_{gik}) = \log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik}. \tag{2.2.2}$$

Note that we set β_{gk} to depend on both g and k, allowing the existence of between study heterogeneity for the same gene. If we re-parametrize the negative binomial model in (2.2.1) in terms of proportion p ($\equiv \frac{\phi\mu}{1+\phi\mu}$) and dispersion ϕ, and let $\Psi = \text{logit}(p) = \log(\frac{\frac{\phi\mu}{1+\phi\mu}}{\frac{1}{1+\phi\mu}}) = \log(\phi\mu)$, we can re-write equation (2.2.2) as:

$$\Psi_{gik} = \log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik} + \log(\phi_{gk}). \tag{2.2.3}$$

The above equation is useful when we later use Gibbs sampling to update the parameters α_{gk} and β_{gk}. Taking equation (2.2.1) and (2.2.2) together form our basic GLM model as follows:

$$y_{gik} | \alpha_{gk}, \beta_{gk}, \phi_{gk} \sim NB(\log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik}, \phi_{gk}). \tag{2.2.4}$$

2.2.3 Information integration of effect size across studies among DE genes

Next, we select appropriate prior distributions for the model parameters in equation (2.2.4) to allow information integration across studies. We first define the following vectors:

$$\vec{\alpha}_g = (\alpha_{g1}, \ldots, \alpha_{gK})^T, \quad \vec{\beta}_g = (\beta_{g1}, \ldots, \beta_{gK})^T, \quad \log(\vec{\phi}_g) = (\log(\phi_{g1}), \ldots, \log(\phi_{gK}))^T,$$

which represent the baseline, effect size and dispersion vectors for gene g respectively. The three vectors are assumed to be a priori independent of each other. In addition, we define the vector for the differential expression indicators of gene g: $\vec{\delta}_g = (\delta_{g1}, \ldots, \delta_{gK})^T$. We assume each of the vectors $\vec{\alpha}_g, \log \vec{\phi}_g$ follows a multivariate Gaussian distribution:

$$\vec{\alpha}_g \sim N_K(\eta_g, \mathbf{\Lambda}), \quad \log \vec{\phi}_g \sim N_K(m_g, \mathbf{\Pi}), \tag{2.2.5}$$

where η_g and m_g are the gene-specific grand means for $\vec{\alpha}_g$ and $\log \vec{\phi}_g$, respectively. The covariance matrices $\mathbf{\Lambda}$ and $\mathbf{\Pi}$ are shared by all genes to be described below. For $\vec{\beta}_g$, we assume a multivariate Gaussian prior, with different means for DE and Non-DE genes:

$$\vec{\beta}_g \sim N_K(\lambda_g \vec{\delta}_g, \mathbf{\Sigma}), \tag{2.2.6}$$

where λ_g is the gene-specific grand mean for DE genes (i.e. $\delta_{gk} \neq 0$ for some k). For Non-DE genes ($\vec{\delta}_g = 0$), the grand mean is 0. We also allow a different covariance matrix of $\vec{\beta}_g$ for DE and Non-DE genes, i.e. $\mathbf{\Sigma} = \mathbf{\Sigma_1}$ for DE genes and $\mathbf{\Sigma} = \mathbf{\Sigma_0}$ for Non-DE genes.

Adopting the separation strategy on modelling covariance matrices by Barnard et al. (2000), we propose independent prior distributions on the diagonal variance components and the off-diagonal correlation matrix for all the four covariance matrices mentioned above. Let $[\boldsymbol{\rho}_{(1)kk'}]_1^K$, $[\boldsymbol{\rho}_{(0)kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$ and $[\boldsymbol{t}_{kk'}]_1^K$ denote the correlation matrices corresponding to the covariance matrices $\mathbf{\Sigma_1}$, $\mathbf{\Sigma_0}$, $\mathbf{\Lambda}$ and $\mathbf{\Pi}$ respectively, and let $[\boldsymbol{\sigma^2}_{(1),k}]_1^K$, $[\boldsymbol{\sigma^2}_{(0),k}]_1^K$, $[\boldsymbol{\tau^2}_k]_1^K$ and $[\boldsymbol{\xi^2}_k]_1^K$ denote the corresponding diagonal matrices with the variance terms on the diagonal. It is widely known that:

$$\mathbf{\Sigma_1} = \left([\boldsymbol{\sigma^2}_{(1),k}]_1^K\right)^{1/2} [\boldsymbol{\rho}_{(1)kk'}]_1^K \left([\boldsymbol{\sigma^2}_{(1),k}]_1^K\right)^{1/2},$$

$$\mathbf{\Sigma_0} = \left([\boldsymbol{\sigma^2}_{(0),k}]_1^K\right)^{1/2} [\boldsymbol{\rho}_{(0)kk'}]_1^K \left([\boldsymbol{\sigma^2}_{(0),k}]_1^K\right)^{1/2},$$

$$\mathbf{\Lambda} = \left([\boldsymbol{\tau^2}_k]_1^K\right)^{1/2} [\boldsymbol{r}_{kk'}]_1^K \left([\boldsymbol{\tau^2}_k]_1^K\right)^{1/2},$$

$$\mathbf{\Pi} = \left([\boldsymbol{\xi^2}_k]_1^K\right)^{1/2} [\boldsymbol{t}_{kk'}]_1^K \left([\boldsymbol{\xi^2}_k]_1^K\right)^{1/2}.$$

For each variance component, we propose a Jeffrey's prior, that is to say:

$$\sigma^2_{(1),k} \propto \frac{1}{\sigma^2_{(1),k}}, \qquad \sigma^2_{(0),k} \propto \frac{1}{\sigma^2_{(0),k}}, \qquad \tau^2_k \propto \frac{1}{\tau^2_k}, \qquad \xi^2_k \propto \frac{1}{\xi^2_k}.$$

For the correlation matrices, we propose an inverse-Wishart prior distribution with identity matrix as its scale matrix and $v = K + 1$ degrees of freedom, which is equivalent to putting a uniform prior on each element of the correlation matrices marginally (Gelman et al., 2014; Scharpf et al., 2009; Barnard et al., 2000), more specifically we have:

$$[\boldsymbol{\rho}_{(1)kk'}]_1^K, [\boldsymbol{\rho}_{(0)kk'}]_1^K, [\boldsymbol{r}_{kk'}]_1^K, [\boldsymbol{t}_{kk'}]_1^K \sim W^{-1}(\boldsymbol{I}, v).$$

For gene-specific grand means λ_g, η_g and m_g, we assume that they follow normal priors, e.g. $\lambda_g \sim N(\mu_\lambda, \sigma^2_\lambda)$, $\eta_g \sim N(\mu_\eta, \sigma^2_\eta)$, $m_g \sim N(\mu_m, \sigma^2_m)$ with mean $\mu_\lambda = 0$, $\mu_\eta = 0$, $\mu_m = 0$, and variance $\sigma^2_\lambda = 10^2$, $\sigma^2_\eta = 10^2$, $\sigma^2_m = 10^2$. We performed sensitivity analysis on the hyperparameter values, since the variance σ^2_λ, σ^2_η and σ^2_m are fairly large, the results show little change when the means μ_λ, μ_η and μ_m change (see Appendix for the result of a sensitivity analysis on hyperparameter μ_η).

In addition to the informative parameters listed above, we introduce one supporting parameter ω_{gik} into the model to help obtain closed-form posterior distribution for β_{gk} and α_{gk} by exploiting conditional conjugacy (Polson et al., 2013; Zhou et al., 2012b). The prior for ω_{gik} is specified as:

$$\omega_{gik} \sim PG(y_{gik} + \phi_{gk}^{-1}, 0),$$

where PG refers to the Polya-Gamma distribution, details about this distribution and how the supporting parameter facilitates conditional conjugacy are provided in the Appendix. The closed-form posterior distribution for β_{gk} and α_{gk} by conditional conjugacy speeds up MCMC simulation.

2.2.4 Model-based clustering to categorize DE genes

We next utilize the differential expression indicators δ_{gk} to cluster the DE genes and model the homogeneous and heterogeneous differential signals across studies. Since clustering based on the binary latent variable is unstable and does not take effect size into consideration, we first transform the binary vector into a standard normal vector and use Dirichlet process Gaussian mixture model to cluster the DE genes, following Medvedovic et al. (2004). Suppose $P(\delta_{gk} = 1) = \pi_{gk}$ is the prior probability that a gene g is DE in study k, the effect size is used to turn π_{gk} into a signed probability measure $\pi_{gk}^{\pm} = \pi_{gk} \times \text{sign}(\beta_{gk})$ where sign(.) is the sign function. We further rescale $\pi_{gk}^{*} = (\pi_{gk}^{\pm} + 1)/2$, so the score falls in the range $[0, 1]$. Lastly, we transform π_{gk}^{*} to a Z-score $z_{gk} = \Phi^{-1}(\pi_{gk}^{*})$ where Φ is the standard normal cumulative distribution function. Following Ferguson (1983) and Neal (2000), we construct a Dirichlet process mixture (DPM) framework to cluster the DE genes:

$$\begin{aligned}
\vec{z}_g | c_g, \boldsymbol{\theta} &\sim F(\vec{\theta}_{c_g}), \\
P(c_g = c) &= p_c, \\
\vec{\theta}_c &\sim G_0, \\
\vec{p} &\sim Dirichlet(a/C, \ldots, a/C).
\end{aligned} \tag{2.2.7}$$

where $\vec{z}_g = (z_{g1}, \ldots, z_{gK})^T$ and c_g indicates the "latent cluster" for gene g, $F(.)$ is a mixture of K-dimensional multivariate Gaussian distributions with mean $\vec{\theta}_c$ and covariance matrix being identity matrix. C is the number of clusters, which is stochastic and allowed to go to infinity under DPM. G_0 is the base distribution, in this case, $G_0 = N_K(\vec{0}, \boldsymbol{I})$ and $\vec{p} = (p_1, \ldots, p_C)$ is the mixing proportions for the clusters. a/C is the concentration parameter. In our model, we specify $a = C$ so the marginal prior distribution of each mixing proportion p_c would be Unif(0,1) under the constraint $\sum\limits_{c=1}^{C} p_c = 1$.

The above descriptions fully define the hierarchical Bayesian model proposed. The observed data are the raw counts, the library size and the phenotypic indicator $\{y_{gik}, T_{ik}, X_{ik}\}$, the parameters we need to update through sampling include δ_{gk}, β_{gk}, α_{gk}, ϕ_{gk}, λ_g, η_g, m_g, σ_k^2, τ_k^2, ξ_k^2, $\rho_{kk'}$, $r_{kk'}$, $t_{kk'}$, ω_{gik}, c_g and C. The hyperparameters we prespecify include $v = K + 1$, $\mu_\lambda = 0$, $\mu_\eta = 0$, $\mu_m = 0$, $\sigma_\lambda^2 = 10^2$, $\sigma_\eta^2 = 10^2$, $\sigma_m^2 = 10^2$ and $C_{\text{init}} = 10$.

2.2.5 Simulating posterior distribution via MCMC

We use the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) as well as the Gibbs sampling algorithm (Geman and Geman, 1984) to infer the posterior distribution of the parameters. Depending on the form of the distribution, 5 types of mechanisms are proposed to update the 16 groups of parameters.

1. The full conditional for α_{gk} and β_{gk} are bivariate normal with known $\vec{\omega}_{gk}$. The full conditional for $\vec{\omega}_{gk}$ is Polya-Gamma distribution with known α_{gk} and β_{gk} (Polson et al., 2013; Zhou et al., 2012b). We use Gibbs sampling to update them sequentially for each gene g in study k.

2. The full conditional for λ_g, η_g and m_g are multivariate Gaussian distribution for each gene g. The full conditional for each element in $[\boldsymbol{\sigma^2}_k]_1^K$, $[\boldsymbol{\tau^2}_k]_1^K$ and $[\boldsymbol{\xi^2}_k]_1^K$ is an inverse-gamma distribution. The full conditional for $[\boldsymbol{\rho}_{kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$ and $[\boldsymbol{t}_{kk'}]_1^K$ are inverse Wishart distributions. For all the above with closed form conditional distributions, we use Gibbs sampling to update them.

3. For $\vec{\phi}_g$, we propose a MH algorithm to update it for each gene g. In particular, we sample

a new value of $\vec{\phi}_g$ from a multivariate log-normal jump distribution with mean equal to the old value and covariance matrix equal to $\mathbf{\Pi}$. The acceptance ratio r is defined as the ratio of two posterior density functions, and the new value is accepted with probability $\min[1, r]$.

4. For the pair $(\beta_{gk}, \delta_{gk})$, since the support for β_{gk} depends on δ_{gk}, we jointly update them (Scharpf et al., 2009). First, a potential new value of δ_{gk} is proposed by inverting the current value, i.e. $\tilde{\delta}_{gk} = 1 - \delta_{gk}$ and a new update $\tilde{\beta}_{gk}$ is then sampled from the associated full condition given $\tilde{\delta}_{gk}$. We define the ratio of the two joint posterior distributions as r and jointly accept the new proposed values $(\tilde{\beta}_{gk}, \tilde{\delta}_{gk})$ with probability $\min[1, r]$.
5. Since our DPM model is in a conjugate context, to update the cluster assignment c_g, we follow Algorithm 3 in Neal (2000) to draw a new value from $c_g | c_{-g}, \vec{z}_g$ for $g = 1, \ldots, G$ at each iteration, where c_{-g} is the cluster assignment of all genes other than g. The number of clusters C is updated at each iteration based on c_g.

The detailed updating functions and algorithms for each group of parameters are described in the Appendix. For both simulation and real data, we ran 10,000 MCMC iterations. The selected traceplots (see Appendix) from Simulation I below showed that all parameters reached convergence after relatively small number of iterations (roughly 3,000). In light of this, the first 3,000 iterations were dropped as burn-in period in all later analysis. The remaining 7,000 of 10,000 iterations are used for inference.

2.3 BAYESIAN INFERENCE AND CLUSTERING

2.3.1 Bayesian inference and control of false discovery rate

In the Bayesian literature, Newton et al. (2004) proposed a direct approach to control FDR and defined a Bayesian false discovery rate as:

$$\mathrm{BFDR}(t) = \frac{\sum_{g=1}^{G} P_g(H_0|D) d_g(t)}{\sum_{g=1}^{G} d_g(t)},$$

where $P_g(H_0|D)$ is the posterior probability of gene g being non-DE (H_0) given data (D) and $d_g(t) = I\{P_g(H_0|D) < t\}$. The tuning parameter t can be tuned to control the BFDR at a certain α level. Throughout this paper, the Bayesian false discovery rate BFDR will be used to address the multiplicity issue for the Bayesian method so that it is comparable to the FDR control from the two-stage methods.

For fair comparison with the Fisher's method in meta-analysis, we adopt a union-intersection (UIT) hypothesis (a.k.a. conjunction hypothesis) setting following Li et al. (2011): $\mathrm{H}_0 : \bigcap\{\beta_k = 0\}$ vs $\mathrm{H}_a : \bigcup\{\beta_k \neq 0\}$, i.e. reject the null when the gene is differentially expressed in at least one study, where β_k is the effect size of study k, $1 \leq k \leq K$. Correspondingly, we define a null set $\Omega^0 = \{\vec{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) = 0\}$ and the respective DE set $\Omega^1 = \{\vec{\beta}_g : \sum_{k=1}^{K} I(\beta_{gk} \neq 0) > 0\}$. To control BFDR at the gene level, we introduce a Bayesian equivalent q-value. From the Bayesian posterior, we can calculate the probability of each gene falling in the null space: $\hat{P}_g(H_0|D) = \hat{P}(\vec{\beta}_g \in \Omega^0|D) = \frac{\sum_{t=1}^{T} I\{\vec{\delta}_g^{(t)} = \vec{0}\}}{T}$, where T is the total number of MCMC samples and $\vec{0}$ is a K-dimensional zero vector. We then define the Bayesian q-value of gene g as $q_g = \min_{t \geq \hat{P}_g(H_0|D)} \mathrm{BFDR(t)}$. This q_g will be treated similarly as q-value in the Frequentist approach by Fisher's method. Aside from detection of a DE gene list from meta-analysis, the posterior mean of δ_{gk}, $\mathrm{E}(\delta_{gk}|D)$, can be used to infer differential expression for gene g in study k.

2.3.2 Summarization of clustering posterior to categorize DE genes

Addressing the differential expression in multiple studies is more difficult than that in a single study because the gene may be concordantly or discordantly (up-regulated in some studies but not in the others) differentially expressed. The proposed Bayesian method is based on effect size, thus it would favor DE genes concordant across studies. Following Section 2.4, we use the posterior estimate of π_{gk} as an indicator of cross-study differential expression pattern to cluster the DE genes. To stabilize the estimation, we estimated π_{gk} by non-overlapping windows of every 20 MCMC simulations, i.e. $\hat{\pi}_{gk}^{(b)} = \sum_{t^{(b)}=1}^{20} \delta_{gk}^{t^{(b)}}/20$, for the bth simulation and then transformed into $\hat{z}_{gk}$ as in Section 2.4. After each chain of 20 simulations, the cluster assignment c_g is updated from the DPM model. At the end of all chains, to summarize the posterior estimates of c_g, we follow from Medvedovic et al. (2004) and Rasmussen et al. (2009) and calculate the co-occurrence probability $p_{g,h}$ for any two genes g and h as the number of times the two genes are assigned to the same cluster divided by the total number of assignments. Then we use $1-p_{g,h}$ as a dissimilarity measure to further cluster the genes using consensus clustering (Monti et al., 2003). Consensus clustering is a stable clustering method by summarizing hierarchical clustering results with Ward linkage in repeated subsampling. The default consensus clustering method does not allow scattered genes (i.e. genes not belonging to any cluster) but one can apply other methods such as tight clustering for that purpose (Tseng and Wong, 2005). As a result, genes with similar differential expression patterns over the chains are grouped together, while those with very different cross-study differential expression patterns will be separated.

2.3.3 Methods for comparison

Since other existing Bayesian methods in RNA-seq DE analysis such as "baySeq" and "EBSeq" are developed for a single study (Hardcastle and Kelly, 2010; Leng et al., 2013), they cannot be immediately extensible to meta-analysis framework and compare to our method. Thus, we will compare our method to selected two-stage approaches as frequently adopted in the literature so far. For the first stage of single study differential expression analysis of RNA-seq, we will compare two most popular tools edgeR and DESeq (Robinson et al., 2010; Anders and Huber, 2010). For meta-analysis, since no other methods have been proposed specifically for RNA-seq, Fisher's method will be applied to combine edgeR or DESeq p-values from multiple RNA-seq studies (Fisher, 1925). The meta-analysed p-values are then adjusted for multiple comparison by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). In this paper, we will compare BayesMetaSeq with the two-stage edgeR/Fisher and DESeq/Fisher approaches.

2.4 SIMULATION

We performed three types of simulation to compare BayesMetaSeq, edgeR/Fisher and DESeq/Fisher. Details are described below.

(I) Simulating homogeneous study effects to assess power and accuracy

In the first part of simulation, we assessed the performance of BayesMetaSeq for genes with low, medium and high read counts when the effects were homogeneous across all studies. We simulated expression counts of $G = 1000$ genes for $K = 2, 5$ studies, $N_k = 10$ (5 cases and 5 controls), $1 \leq k \leq K$. Library sizes for all samples were sampled from 0.3 to 0.5 millions so the average counts range roughly from 300 to 500. Baseline expressions were either high ($\alpha_{gk} \sim$ Unif$\{-5.5, -4.5\}$; mean counts $\sim$ 1500-4500), medium ($\alpha_{gk} \sim$ Unif$\{-8.5, -6.5\}$; mean counts $\sim$ 80-600) or low means ($\alpha_{gk} \sim$ Unif$\{-11, -9\}$; mean counts $\sim$ 5-50). The log-scaled dispersion were generated accordingly ($\log(\phi_{gk}) \sim$ Unif$\{-3.5, -2.5\}$ for high mean counts, $\log(\phi_{gk}) \sim$ Unif$\{-2.5, -1.5\}$) for medium mean counts and $\log(\phi_{gk}) \sim$ Unif$\{-1.5, -0.5\}$ for low mean counts), assuming genes with larger means had smaller dispersion (Anders et al., 2013). We let the first 20% genes ($N = 200$) be differentially expressed in all studies, among them, 1/2 was generated from high means and the other 1/2 from the low means. The rest of genes ($N = 800$) were non-differentially expressed, 1/4 of them were generated from high means, 1/4 from the medium means, and the other 1/2 from the low means. For differentially expressed genes, the effect size β_{gk} was drawn from Unif$\{0.8, 2\}$ or Unif$\{-2, -0.8\}$ (positive or negative log fold change, respectively). For non-differentially genes, β_{gk} was drawn from $N(0, 0.5^2)$. We repeated the above parameter sampling for all the K studies. Under the same homogeneous scenario, we also repeated the above simulations for weaker DE signals (Simulation IB), i.e. log-scaled effect size β_{gk} was drawn from Unif$\{0.7, 1.5\}$ or Unif$\{-1.5, -0.7\}$ for DE genes and from $N(0, 0.7^2)$ for Non-DE genes.

(II) Simulating heterogeneous study effects to assess power and accuracy

In the second part of simulation, we assessed the performance of BayesMetaSeq when the effects were heterogeneous in different studies. We simulated expression counts of

$G = 1000$ genes for $K = 2, 5$ studies with $N_k = 10$, $1 \leq k \leq K$. Library size, baseline expression and the corresponding log-scaled dispersion were generated in the same way as in Simulation I. We assumed the first 30% of genes ($N = 300$) be differentially expressed. For $K = 2$, 2/3 of those genes are DE only in the first study or the second study, and 1/3 are common DE; for $K = 5$, 1/3 of those genes are DE only in one study, 1/3 are DE only in two studies, and 1/3 are DE in more than two studies. Similar to the previous simulation, 1/2 of the DE genes were from high means and 1/2 from the low means. The other 70% of genes ($N = 700$) were non-differentially expressed, 1/4 of them were generated from high means, 1/4 from the medium means, and 1/2 from the low means. For differentially expressed genes, the effect size β_{gk} was drawn from Unif$\{1, 2.5\}$ or Unif$\{-2.5, -1\}$, however, no discordance was allowed. For non-differentially genes, β_{gk} was drawn from Unif$\{-0.3, 0.3\}$.

(III) Simulating cross-study differential patterns to evaluate DE gene clustering

In the fourth part of simulation, we assessed the clustering performance of our method when the DE genes were generated from varying cross-study differential patterns. We simulated expression counts of $G = 1000$ genes for $K = 3$ with $N_k = 10$, $1 \leq k \leq K$. Library size was generated in the same way as in Simulation I. The baseline expression α_{gk} was drawn from Unif$\{-8.5, -6.5\}$ (mean counts $\sim$ 80-600) and the dispersion parameter ϕ_{gk} was drawn from Unif$\{-2.5, -1.5\}$. We assumed the first 30% of genes ($N = 300$) were differentially expressed in at least 2 studies. Among them, 1/6 were up-regulated in all studies ("+++"), 1/6 were down-regulated in all studies ("- - -"), the other 2/3 were either up-regulated or down-regulated in two studies but non-DE in the third study (e.g. 50 genes with the pattern "++0", 50 genes with the pattern "- - 0", 50 genes with the pattern "+0+", 50 genes with the pattern "- 0 -"). For differentially expressed genes, the effect size β_{gk} was drawn from $N(2, 0.5^2)$ or $N(-2, 0.5^2)$ (up-regulated or down-regulated, respectively). For non-differentially expressed genes, β_{gk} was 0.

For comparison with the other methods (edgeR/Fisher and DESeq/Fisher), we assessed both power and accuracy by plotting the number of true positives against the top number of declared DE genes, as well as the ROC curves respectively for each method.

2.4.0.1 ***Simulation I, II*** The posterior means and standard errors of selected parameters were summarized and compared to their true values from Simulation IA as shown in the Appendix. The result demonstrated validity of BayesMetaSeq. In Simulation IA of homogeneous study effects, we found that BayesMetaSeq was more powerful and accurate than the edgeR/Fisher and DESeq/Fisher methods in low mean counts while performed almost equally well in high means counts (for simplicity, we combined both high mean and medium mean in this group), as shown in Figure 3(A). Comparing to the other two methods, only BayesMetaSeq had AUC above 0.9 in low mean region with both high sensitivity and specificity. As the number of study K increased, we saw more noticeable advantage of Bayesian method over the other methods in detecting DE genes with low means. Since the signals for high means were very strong, the three approaches performed almost perfectly even when $K = 2$. For Simulation IB with weaker signals, the results were similar to Simulation IA and the difference was more noticeable between BayesMetaSeq and the other two methods in low mean region, while for high mean region, the performance for all three methods were alike (Figure 3(B)).

Similarly, in Simulation II with heterogeneous study effects, though the overall signals became weaker, we found that BayesMetaSeq still performed better than the edgeR/Fisher and DESeq/Fisher methods in terms of both power and accuracy for low mean counts genes, while their performances were similar in high mean region, as shown in Figure 3(C). One thing to notice here is that, even though the Bayesian method increased the power of detecting true DE signals in low mean regions, the detection power for low mean genes was still relatively weaker than high mean genes under the same scenario, due to the inherent read count bias.

2.4.0.2 ***Simulation III*** In Simulation III, we found that BayesMetaSeq clearly identified the six clusters of DE genes with pre-specified cross-study differential patterns (Figure 4 Left). Each of the six clusters corresponded to one particular cross-study differential pattern as reflected in the heatmap of signed $\mathrm{E}(\delta_{gk}|D)$ (Figure 4 Right), for example, cluster 1 included genes up-regulated in all studies and cluster 3 included genes down-regulated in all studies.

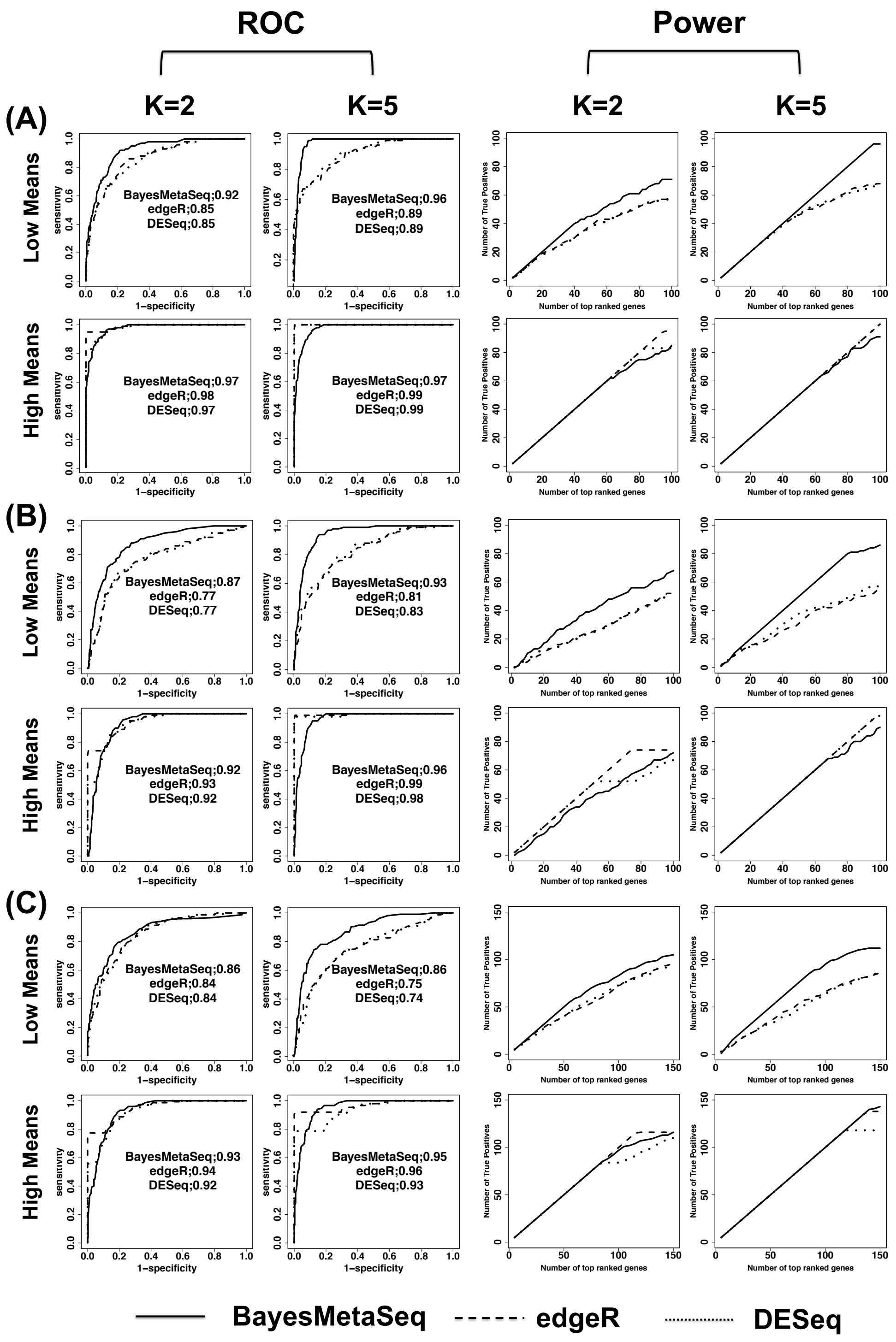


Figure 3: ROC Curve (left) and Power (right) comparison of the three methods. Panel (A) is the results from Simulation IA, panel (B) is for Simulation IB and panel (C) is for Simulation II. The solid line is for BayesMetaSeq, the dashed line is for edgeR/Fisher and dotted line is for DESeq/Fisher.

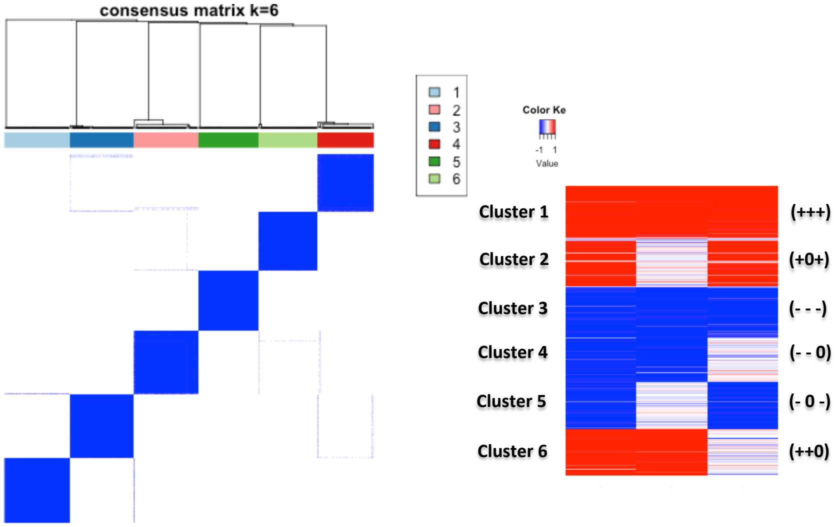


Figure 4: Clustering results from Simulation III.

Left: Correlation heatmap of DE genes based on the co-occurrence probability $p_{g,h}$ with consensus clustering. Right: The heatmap of signed posterior mean of the DE latent indicator (i.e. $\mathrm{E}(\delta_{gk}|D)$ $\times \mathrm{sign}(\beta_{gk})$) in the six clusters.

2.5 REAL DATA ANALYSIS

We applied BayesMetaSeq to a multi-brain-region HIV-1 transgenic rat experiment (GSE47474) comparing the normal F344 strain and the HIV strain (Li et al., 2013). Samples from three brain tissues (hippocampus (HIP), striatum (STR), prefrontal cortex (PFC)) were sequenced and we regarded those as 3 studies to adopt our meta-analysis framework. There were 12 samples from each brain region in each strain ($N_1 = N_2 = N_3 = 24$; $K = 3$). The experiment was designed to determine expression differences in brain regions of F344 and HIV-1 transgenic rats, in order to identify the mechanisms involved in HIV-1 neuropathology and develop efficient therapy for neuropsyhchiatric disorders associated with HIV-1 infection (Li et al., 2013). Following the guidance in edgeR (Robinson et al., 2010), we first filtered out genes with mean counts smaller than 1 in any study. After filtering, 10,280 genes remained for analysis. We applied BayesMetaSeq as well as edgeR/Fisher and DESeq/Fisher to the data. After we obtained the DE genes from each approach, we performed pathway enrichment analysis using Fisher's exact test based on the Gene Ontology (GO) database to annotate the identified genes (Khatri et al., 2012). In addition, we also analyzed the DE genes categories from BayesMetaSeq using Ingenuity Pathway Analysis (IPA) for more biological insight. IPA is a commercial curated database that contains rich functional annotation, gene-gene interaction and regulatory information (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity).

2.5.1 Differential expression analysis

Controlling FDR at 0.1, edgeR/Fisher detected 51 DE genes and DESeq/Fisher 46 DE genes respectively, while BayesMetaSeq detected 245 DE genes (Table 1). A Venn Diagram showing the number of overlapping genes indicated good agreement among the three methods (see Appendix). As shown in Figure 4(A), the DE genes detected by BayesMetaSeq have wider detection range, especially for genes with smaller read counts, smaller RPKM (reads per kilobase per million) or shorter transcript length (Mortazavi et al., 2008). Table 2 lists three DE genes detected only by BayesMetaSeq but not by the other two methods.

They typically have rare counts (a table and boxplots of normalized counts shown in the Appendix) due to short length of the transcripts (e.g. Mir212, Mir384) and/or small RPKM (e.g. Alb). microRNA-212 has been reported in previous studies to promote interleukin-17-producing T-helper cell differentiation (Nakahama et al., 2013). miRNA-384 has been found to regulate both amyloid precursor protein (APP) and β-site APP cleaving enzyme, which play an important role in the pathogenesis of Alzheimer's disease (Liu et al., 2014). Gene Alb encodes for albumin which is a primary carrier protein for steroids, fatty acids and steroid hormones in the blood, and has been used as markers of HIV disease progression in the highly active antiretroviral therapy (Shah et al., 2007).

2.5.2 Pathway enrichment analysis on detected DE genes

Detecting more DE genes does not necessarily indicate a better performance of our method. Since the underlying truth is not known in real data, we performed a pathway enrichment analysis on identified DE genes by each method. For fair comparison, we used the top 200 genes from each of the three methods and regarded them as DE genes in the pathway analysis. We tested on three pathway databases in MSigDB (`http://software.broadinstitute.org/gsea/msigdb`): GO, KEGG and Reactome, and only GO reported significant (q-value<0.05) pathways for all three methods. Controlling FDR at 0.05 by Benjamini-Hochberg correction, we found 50 GO pathways enriched with the DE genes from the BayesMetaSeq, while only 20 and 22 GO pathways were enriched for edgeR and DESeq, respectively. A cluster of enriched pathways was identified on the left of the Manhattan plot for BayesMetaSeq (circled), implying the enrichment in a major functional domain (Figure 5; pathways sorted by GO IDs). These pathways were mainly related to cell killing, leukocyte mediated cytotoxicity and T-cell mediated cytotoxicity (GO:0001906, GO:0001909, GO:0001910, GO:0001912, GO:0001913, GO:0001914, GO:0001916) and were enriched with BayesMetaSeq only (Table 3; p-values obtained from Fisher's exact test). The enrichment in these GO pathways might reflect changes in adaptive immune response against the HIV.

Table 1: Comparison of three approaches in real rat data

Method	FDR at 0.05	FDR at 0.1
BayesMetaSeq	169	245
edgeR+Fisher	36	51
DESeq+Fisher	37	46

Table 2: Three example genes that show better detection power of BayesMetaSeq to detect low expressed or short length genes.

		edgeR		DESeq		BayesMetaSeq		Ave. normalized counts			
Gene	Study	p-value	Fisher's q-value	p-value	Fisher's q-value	Posterior means	Bayesian q-value	HIV strain	Normal strain	Ave. RPKM	Transcript length(bp)
Mir212	HIP	0.02	0.21	0.05	0.51	0.89	2e-3	2.08	3.92	20.27	23
	STR	0.02		0.03		0.99		2.20	4.93	21.31	
	PFC	0.07		0.09		0.83		2.99	5.08	22.56	
Mir384	HIP	0.06	0.39	0.10	0.86	0.88	8e-3	1.59	2.84	15.53	20
	STR	0.65		0.77		0.33		2.54	2.12	14.50	
	PFC	0.004		0.01		0.98		2.02	4.59	19.28	
Alb	HIP	0.002	0.10	0.06	0.88	0.95	6e-3	8.58	2.93	1.31	2676
	STR	0.61		0.60		0.41		9.17	7.65	1.67	
	PFC	0.006		0.03		0.99		20.09	10.90	2.89	

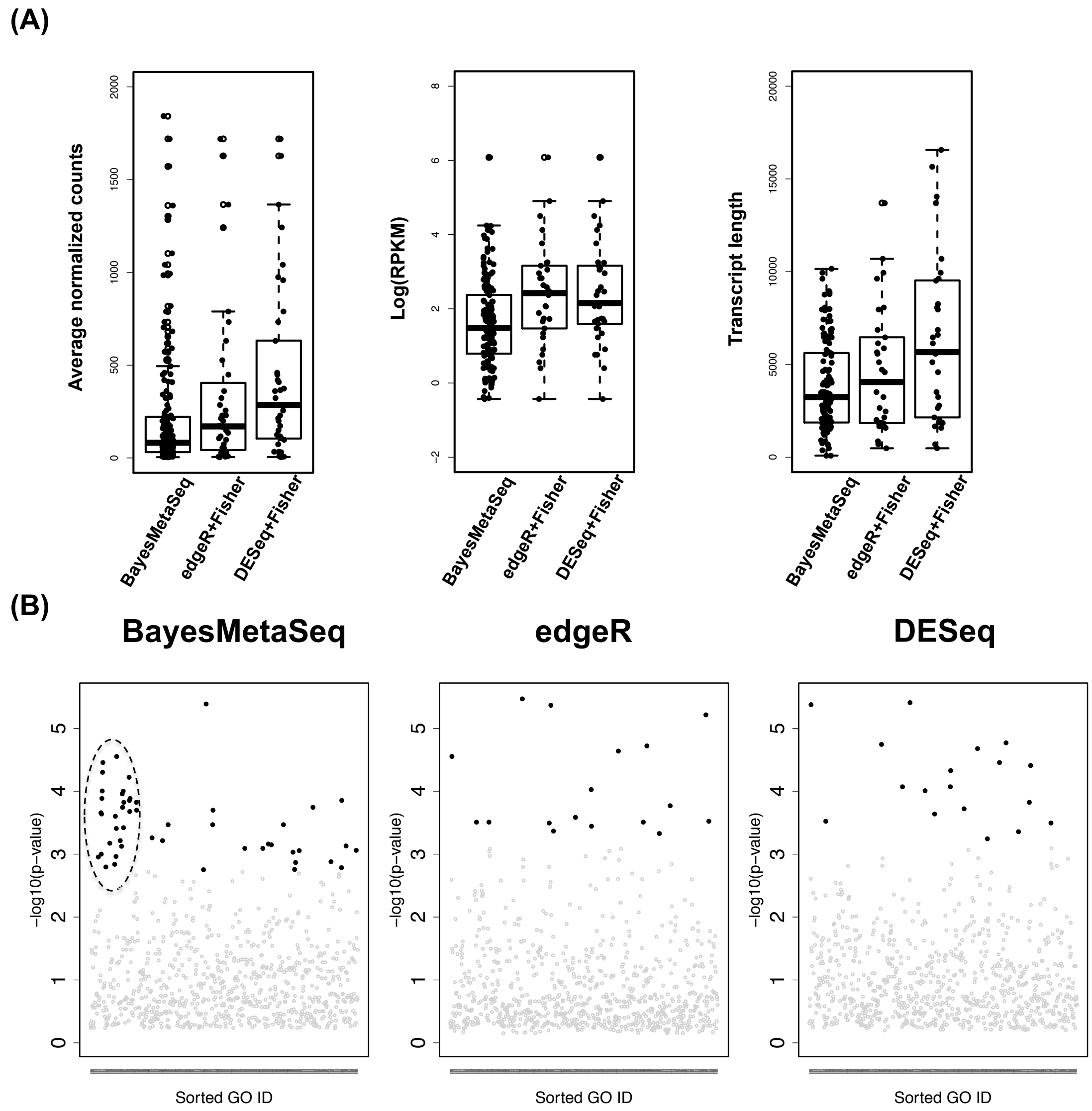


Figure 5: Comparison of three methods in real rat RNA-seq data.

(A) Boxplot of average normalized counts, log(RPKM) and transcript lengths for the declared DE genes by each method. From left to right: BayesMetaSeq, edgeR/Fisher, DESeq/Fisher.

(B) Manhattan plot of GO pathways enriched by the top 200 DE genes from each method.

Table 3: Selected GO pathways enriched only with BayesMetaSeq from Figure 4(B)

GO ID	GO Term	BayesMetaSeq p-value (logOR)	edgeR p-value (logOR)	DESeq p-value (logOR)
GO:0001906	cell killing	2.2e-4 (1.87)	0.033 (1.25)	0.12 (0.95)
GO:0001909	leukocyte mediated cytotoxicity	1e-3 (1.77)	0.105 (1.01)	0.102 (1.03)
GO:0001910	regulation of leukocyte mediated cytotoxicity	2.3e-4 (2.07)	0.056 (1.30)	0.055 (1.31)
GO:0001912	positive regulation of leukocyte mediated cytotoxicity	1.3e-4 (2.18)	0.04 (1.40)	0.043 (1.42)
GO:0001913	T cell mediated cytotoxicity	9.9e-5 (2.25)	0.039 (1.46)	0.038 (1.47)
GO:0001914	regulation of T cell mediated cytotoxicity	5e-5 (2.39)	0.029 (1.59)	0.028 (1.60)
GO:0001916	positive regulation of T cell mediated cytotoxicity	3.5E-5 (2.46)	0.024 (1.66)	0.24 (1.67)

2.5.3 Categorization of DE genes by study heterogeneity

We calculated the co-occurrence probability $p_{g,h}$ and used $1 - p_{g,h}$ as a dissimilarity measure to cluster the DE genes of BayesMetaSeq. As shown in Figure 6(A), we identified seven major clusters from the 245 DE genes. Each of the seven clusters corresponded to one particular cross-study differential patterns based on the signed $\mathrm{E}(\delta_{gk}|D)$ (Figure 6(B)). For example, genes in Cluster 1 were up-regulated in all three studies and genes in Cluster 5 were down-regulated only in STR, but not in HIP and PFC. Moreover, when we analyzed each cluster of genes separately through IPA pathway enrichment analysis, we noticed that each cluster of genes represented different functional domains that were changed in the HIV strain as compared to the normal strain in different brain regions. For example, Cluster 1 which included genes up-regulated in all three brain regions was mainly involved in antimicrobial response, while Cluster 5 which included genes down-regulated in STR region only was mainly related to nervous system development (Figure 6(C)). Cluster 7 was not shown here since it included very few DE genes and only one enriched pathway was identified. Detailed list of significant pathways in each cluster with corresponding p-values and log odds ratios can be found in the Appendix. In our analysis, we detected more region-specific DE markers (Cluster 2-7) than common DE markers (Cluster 1) which was consistent with the results reported from the original paper of this data (Li et al., 2013).

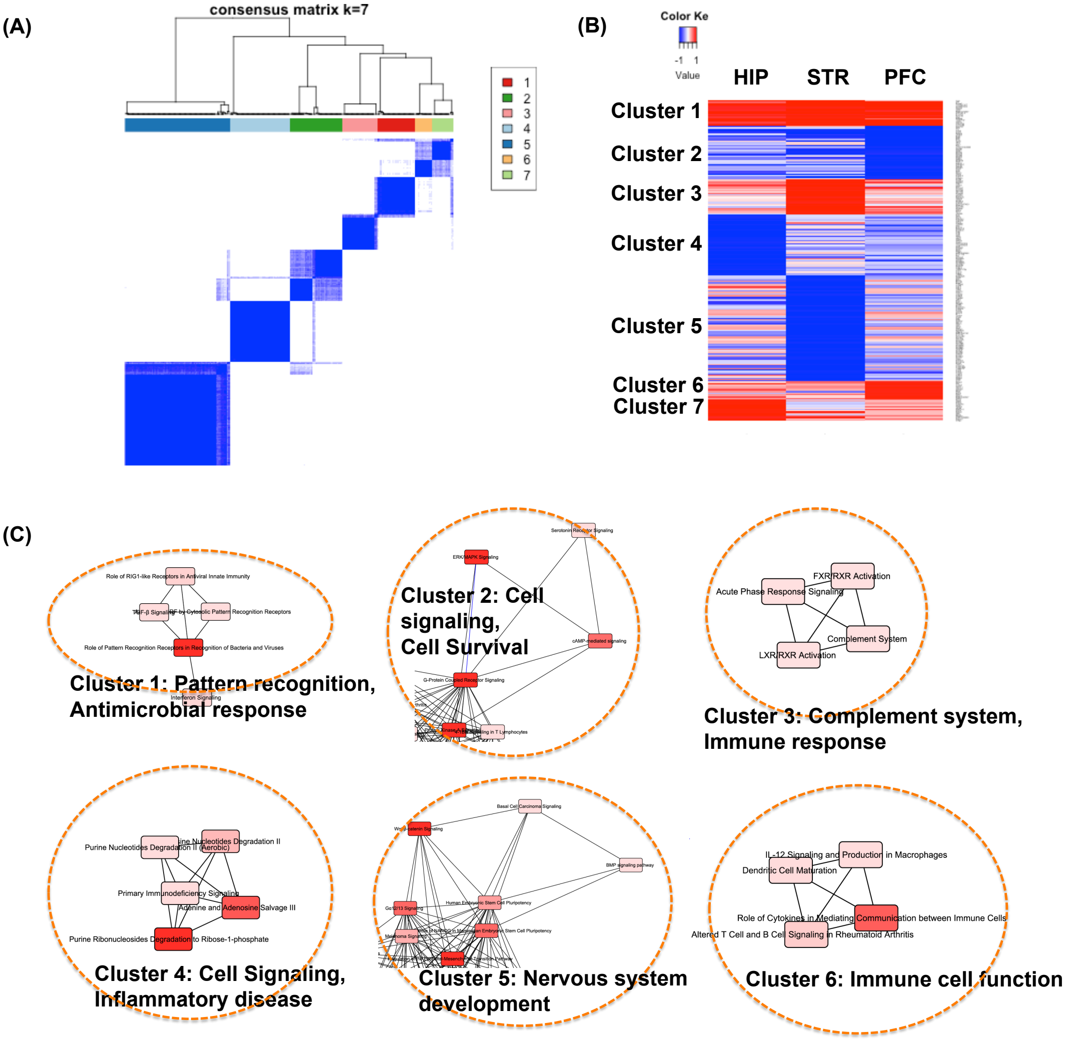


Figure 6: Real rat data clustering results.

(A) Correlation heatmap of 245 Bayesian DE genes based on the co-occurrence probability $p_{g,h}$ with consensus clustering. (B) The heatmap of signed posterior mean of DE latent indicator (i.e.E$(\delta_{gk}|D)$ ×sign(β_{gk})) in the five major clusters. (C) A collection of overlapping IPA pathways enriched with each cluster of genes (deeper color refers to more significant pathways).

2.6 DISCUSSION AND CONCLUSION

In this paper, we proposed a Bayesian hierarchical model called BayesMetaSeq to conduct meta-analysis of RNA-seq data and biomarker categorization by study heterogeneity. Based on a negative binomial framework, the model assumed study-specific differential expression pattern for each gene and allowed the shrinkage of multiple parameters. MCMC algorithm was applied to update the posterior distribution of model parameters and the multiplicity issue was addressed by global FDR from a Bayesian perspective. A Dirichlet process mixture (DPM) model embedded in the Bayesian framework automatically clustered the detected biomarkers based on cross-study differential patterns. Both the simulations and real rat data analysis showed that the Bayesian unified model was more powerful than the two-stage methods (e.g. edgeR/Fisher, DESeq/Fisher), especially in lowly expressed genes without the loss of power in highly expressed genes, and the false discovery rate was well controlled. The differentially expressed genes identified by BayesMetaSeq between HIV strains and normal strains in the real data were further validated by pathway analysis and many DE genes were enriched in pathways related to immune response. Clustering analysis of the DE genes showed that genes with unique cross-study differential patterns were involved in specific functional domains such as antimicrobial response, inflammatory response and so on.

Bayesian models have long been used in differential analysis of genomic studies such as microarray, RNA-seq and methylation (Hardcastle and Kelly, 2010; Leng et al., 2013; Van De Wiel et al., 2012; Chung et al., 2013; Park et al., 2014). Compared to other approaches, Bayesian methods can handle more complex generative mechanisms and allows the sharing of information across studies and across genes, both of which are essential for meta-analysis. Our unified Bayesian meta-analysis model increases the detection power for genes with low counts by accumulating small counts from multiple studies and encourages the sharing of information across different studies, which is not seen in the two-stage meta-analysis methods. In addition, the flexible and adaptable modelling of variance across samples in our approach also contributes to the improvement of detection power (Chung et al., 2013). Similar advantage of unified model over two-stage method has been seen in categorical analysis literature where joint modelling of count data to combine multiple sparse contingency tables was shown to be more powerful than traditional two-stage methods (Bradburn et al., 2007).

The current model relies on a fixed effects model, which assumes that differences of effect sizes are from sampling error alone. It can be readily extended to a random effects scenario, where each effect size is assumed to be drawn from a study-specific distribution (Choi et al., 2003). Model checking can be performed to determine whether the fixed effect model or the random effect model is more adequate for a given dataset. Recent statistical research on RNA-seq proposed zero-inflated negative binomial model as an alternative to the regular negative binomial model and found that it fits better to real data since excessive zeros have always been observed in the NGS data (Van De Wiel et al., 2012). Our model can be easily extended to a zero-inflation framework, and its performance and computing feasibility for applications can be assessed through simulation or real data analysis. In our current approach, only binary outcome is considered. The framework is applicable for continuous outcome or multi-class outcome, where dummy variable regression approach can be applied. Moreover, potential confounding covariates such as age, gender and other individual attributes can be included in the model.

Our real data application presents an example using the same RNA-seq platform across studies. In practice, it is possible that studies from different RNA-seq platforms are included and thus introduce significant bias. For example, the Sequencing Quality Control (SEQC) consortium performed extensive comparison on three RNA-seq platforms (Illumina HiSeq, Life Technology SOLiD and Roche 454) and determined pros and cons of different platforms (Consortium et al., 2014; Xu et al., 2013). As of 07/30/2016, more than 95% of data in GEO used Illumina sequencing systems. As a result, unless different experimental protocols (e.g. mRNA preparation kits) are used in different studies, the platform bias in RNA-seq meta-analysis is not as severe as in microarray. We, however, acknowledge that platform bias may exist or may become more serious if new competing sequencing platforms become popular in the future. Practitioners should apply batch effect diagnostic or removal tools (Leek, 2014; Liu and Markatou, 2016), or extend with random effects in our model to account for cross-platform bias.

Currently, the Bayesian hierarchical model allows study-specific DE status, but favors concordant differential expression across studies. In some applications, discordant DE genes (e.g. a biomarker is up-regulated in one brain region but down-regulated in another brain region) may be expected and another hierarchical layer will be needed to accommodate. Another limitation of our method is the relatively high computational cost. To speed up the computation, we randomly partition the whole dataset into independent gene chunks and apply explicit parallelism using "snowfall" package in R, while merging intermediate outputs for cluster analysis with all genes. It takes about 1 hour for 10,000 MCMC iterations and 10,280 genes with K=3 using 128 computing threads (8 CPUs each with Sixteen-core AMD 2.3GHz and 128GB RAM) in R code. Since the reduction of computing time is almost linear when more computing threads are used, we expect further computing time reduction when powerful computing clusters are used. Optimization of code in C++ and applying further parallel computing such as Consensus Monte Carlo Algorithm and Asynchronous Distributed Gibbs Sampling (Scott et al., 2013; Terenin et al., 2015) should further reduce computing time for general applications in the future. An R package, BayesMetaSeq, is publicly available to perform the analysis (`http://tsenglab.biostat.pitt.edu/software.htm`).

3.0 INTEGRATING MICROARRAY AND RNA-SEQ TRANSCRIPTOMIC DATA USING BAYESIAN HIERARCHICAL MODEL

3.1 INTRODUCTION

Gene expression profiling based on DNA microarray technique is a mature and powerful approach that has been widely applied in large-scale genomic analysis and biomedical research in the past two decades. More recently, with the development in next-generation sequencing technology and decreasing running cost, RNA sequencing (RNA-seq) has become a more popular tool in profiling transcriptome. Compared with the traditional probe hybridization based microarray, RNA-seq has many advantages (Mortazavi et al., 2008; Consortium et al., 2014). Firstly, RNA-seq has a wider detection range of expression levels compared to microarray. For low-expressed genes, the intensities obtained from microarray are mostly un-distinguishable from background noise. On the other hand, sequencing reads from RNA-seq can accurately quantify these genes. Secondly, RNA-seq can be used to detect novel transcripts, which is impossible in microarrray with only known probes. Thirdly, RNA-seq can also be used to examine transcriptome fine structure such as allele-specific expression and splice junctions. Despite the aforementioned benefits, there are potential biases and artefacts that needed to be appropriately addressed in the analysis of RNA-seq data as well. Due to the random RNA fragmentation and sampling nature in RNA-seq, transcript length bias is inherent to the RNA-seq studies where short transcripts with less mapped reads are usually at a statistical disadvantage relative to long transcripts in the same sample (Oshlack and Wakefield, 2009). In addition, read mapping uncertainty and sequence base composition (e.g. GC content bias) (Zheng et al., 2011) are also factors that can confound the analysis results of RNA-seq.

Many studies have been conducted to compare the two platforms in various aspects. As one of the earliest studies to introduce RNA-sequencing into the field, Marioni et al. showed that RNA-seq was comparable to microarray in differential expression analysis between human kidney and liver samples (Marioni et al., 2008). Sultan et al. further explored the performance of two platforms in the analysis of human HEK and B cells and found that RNA-seq was more sensitive than microarrays, where differentially expressed genes detected only by RNA-seq fell in the lowest range of expression levels (Sultan et al., 2008). Other studies, though restricted by small sample size, reached similar conclusions using different datasets under different scenarios (Xiong et al., 2010; Bradford et al., 2010; Su et al., 2011). As part of the third phase of large-scale MicroArray Quality Control Consortium (MAQC-III) launched by FDA (a.k.a. SEQC), Wang et al. (2014) conducted a comprehensive rat study to assess the concordance of RNA-seq and microarray using a range of chemical treatment conditions. They found that RNA-seq outperformed microarray at detecting weakly expressed genes, and the concordance between two platforms for detecting the number of differentially expressed genes (DEGs) depended on treatment effects and the abundance of genes. Furthermore, they showed a systematic difference between log fold change of RNA-seq and that of microarray for DEGs. Similar results have been reported in Robinson et al. (2015) that microarray was more systematically biased in DE analysis of low-intensity genes than RNA-seq, while the detection power of RNA-seq is more sensitive to the per-gene reading depth. In addition, they showed that the correlation between microarray and RNA-seq effect size was low for lowly expressed genes. The systematic difference in effect size between two platforms can be partially attributed to the ratio compression problem in microarray (i.e. the observed expression fold change is consistently underestimated) caused by inefficient hybridization (Draghici et al., 2006).

Meta-analysis in genomic research is a set of statistical tools for combining multiple "-omics" studies of a related hypothesis and can potentially increase the detection power of individual studies. With the increasing availability of mRNA expression data sets, many transcriptomic meta-analysis methods for microarray and some for RNA-seq have been developed in the past decade. As far as we know, no meta-analysis methods have been developed to jointly analyze the data from both microarray and RNA-seq yet. Considering the availability of both data types in the public domain, a well integration of the two platforms can potentially increase the detection power by utilizing the advantages and overcoming the disadvantages of each platform. Particularly, the cross-platform meta-analysis method needs to adjust for the systematic bias in log fold change between the two platforms, as pointed out above.

The most popular type of meta-analysis is a two-stage approach, where a summary statistics such as p-value or effect size is first computed for each study and then meta-analysis methods are used to combine the summary statistics (Tseng et al., 2012). One naive two-stage method to perform cross-platform meta-analysis is to apply some state-of-the-art tools for DE analysis in each platform individually (e.g. edgeR or DESeq2 for RNA-seq and LIMMA or SAM for microarray) (Robinson et al., 2010; Love et al., 2014; Smyth, 2005; Tusher et al., 2001), and then combine the p-values by Fisher's or Stouffer's method (Fisher, 1925; Stouffer et al., 1949). Another alternative is to integrate raw data from all studies using a joint stochastic model. These approaches have the potential to offer improved efficiency over the two-stage methods and, at the same time, retain the platform-specific features. Moreover, as one essential issue mentioned above, it is relatively simpler to adjust for the systematic bias in effect sizes between two platforms under an integrative framework than under a two-stage framework. The more flexible Bayesian methods are most adequate to fit such joint hierarchical models.

Two Bayesian hierarchical models have been developed to meta-analyze multiple microarray datasets (Conlon et al., 2006; Scharpf et al., 2009). Ma et al. (2017c) recently developed a full Bayesian hierarchical model to combine multiple RNA-seq count data. In this paper, we will combine the existing models for microarray meta-analysis (Conlon et al., 2006; Scharpf et al., 2009) and RNA-seq meta-analysis (Ma et al., 2017c) and propose a Bayesian hierarchical model to jointly analyze the data from the two platforms. To address the issue of systematic bias in effect size, we incorporated a normalization algorithm into our full model.

Ramasamy et al. (2008) presented seven key issues when conducting microarray meta-analysis, including identifying and extracting experimental data, preprocessing and annotating each dataset, matching genes across studies, statistical methods for meta-analysis, and final presentation and interpretation. When combining RNA-seq and microarray studies for meta-analysis, most preliminary steps and data preparation issues will similarly apply. In RNA-seq, preprocessing tools such as fastQC, tophat and bedtools are instrumental for alignment and preparing expression counts for downstream analysis, and lumi and affy are very popular R packages for processing microarray from different array platforms. Genes can be matched across studies using standard gene symbols from e.g. BioMart databases. In the remaining of this paper, we assume that data collection, preprocessing and gene matching have been carefully done for both platforms and we only focus on downstream meta-analytic modeling and interpretation.

In recent years, "Big data" research has rapidly become a hot topic that attracted extensive attention from academia, industry, and policy makers. In the field of genomics, the large amount of transcriptomic studies on both microarray and RNA-sequencing platforms have generated petabytes of data that constitute "Big data" from the perspective of scale and complexity. Our paper proposed one analytic method under a Bayesian framework to jointly model and analyze such high volume genomic big data and demonstrated improved biological findings. Bayesian methods have brought substantial benefits to big data research and the high-speed computation nowadays has made these methods computationally effective and scalable with the big data.

The paper is organized as follows. Section 3.2 describes the Bayesian hierarchical model as well as the embedded normalization algorithm and explains how we perform differential expression analysis based on Bayesian inference. In Section 3.3.1, we used simulation to demonstrate the benefits of our Bayesian model over two-stage methods after including the normalization algorithm. In Section 3.3.2, we apply our method to a histological subtype ("ILC") of breast cancer samples comparing early stage vs. late stage patients as the first example, and comparing PR+ vs. PR- as the second example. Final conclusion and discussion are provided in Section 3.4.

3.2 METHODS

3.2.1 Notation

Throughout the paper, we denote by Ψ_k the platform indicator where $\Psi_k = 1$ if the kth study is an RNA-seq study and $\Psi_k = 0$ if the kth study is a microarray study. y_{gik} is the observed RNA-seq count ($\Psi_k = 1$) or microarray intensity ($\Psi_k = 0$) for gene g and sample i in study k. Here we assume the intensity of microarray is already log transformed for fair comparison with the log link function used in RNA-seq count model. $T_{ik} = \sum\limits_{g=1}^{G} y_{gik}$ is the corresponding library size (i.e. the total number of reads) for sample i in study k for RNA-seq studies, and $X_{ik} \in \{0, 1\}$ the phenotypic condition of sample i in study k. The observed data are:

$$D = \{(y_{gik}, T_{ik}, X_{ik}, \Psi_k) : g = 1, \ldots, G; i = 1, \ldots, N_k; k = 1, \ldots, K\},$$

where G is the total number of genes, N_k is the sample size of study k and K is the number of studies in the meta-analysis, including both platforms. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of differential expression for gene g in study k, meaning gene g is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$.

3.2.2 Bayesian Hierarchical Model

Figure 7 provides a graphical representation of the full Bayesian hierarchical model we propose. Circles denote parameters that need to be updated, squares denote observed data or constants and dashed circles denote auxiliary parameters. Each dashed rectangle includes all parameters in a single platform model, and the parameters outside both rectangles are the parameters to be shared across two platforms in the meta-analysis.

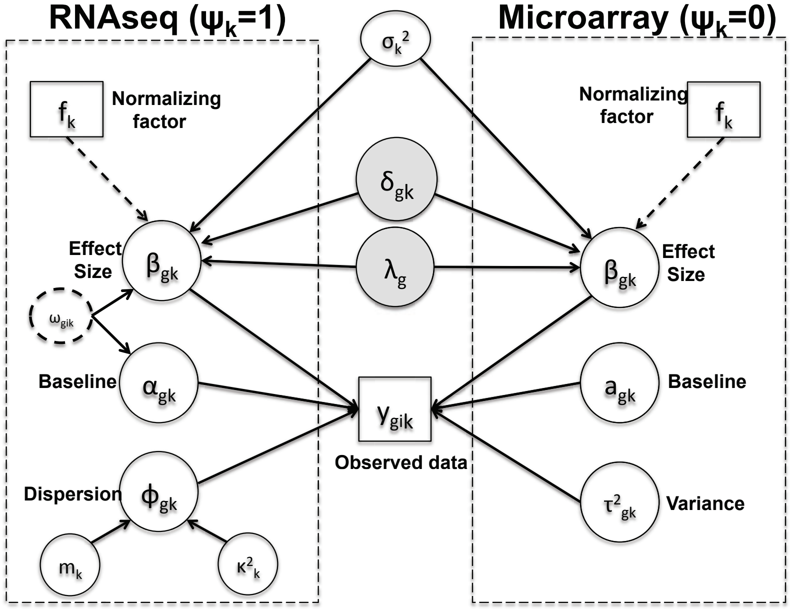


Figure 7: "CBM": Graphical representation of the Bayesian hierarchical model

For each individual study, we accommodate the widely used negative binomial regression model for RNA-seq and linear regression model for microarray respectively as follows:

$$y_{gik} \sim NB(\mu_{gik}, \phi_{gk}), \;\; \log(\mu_{gik}) = \log(T_{ik}) + \alpha_{gk} + \beta_{gk} X_{ik}, \text{ for } \Psi_k = 1,$$

$$y_{gik} \sim N(\mu_{gik}, \tau^2_{gk}), \;\; \mu_{gik} = a_{gk} + \beta_{gk} X_{ik}, \text{ for } \Psi_k = 0,$$

where $\mu_{gik} = E(y_{gik})$ is the mean expression level (mean counts in RNA-seq and mean intensity in microarray), ϕ_{gk} is the dispersion parameter for RNA-seq, and τ^2_{gk} is the variance parameter for microarray. α_{gk} denotes the baseline expression relative to the library size for RNA-seq, a_{gk} denotes the baseline intensity level for microarray, β_{gk} denotes the effect size.

We then specify the prior distributions for β_{gk}, allowing the information integration of effect size across the two platforms:

$$\vec{\beta}_g \sim N_K(\lambda_g \vec{\delta}_g, \mathbf{\Sigma}),$$

where $\vec{\beta}_g = (\beta_{g1}, \ldots, \beta_{gK}), \vec{\delta}_g = (\delta_{g1}, \ldots, \delta_{gK})$. The latent variable of interest $\delta_{gk} \in \{0, 1\}$ is the study-specific indicator of differential expression for gene g in study k, meaning gene g

is differentially expressed in study k if $\delta_{gk} = 1$ and non-differentially expressed if $\delta_{gk} = 0$. λ_g is the gene-specific grand mean across all studies for DE genes.

Here we assume the effect sizes are independent among the studies a priori (which is reasonable if no overlapping samples across studies), so $\mathbf{\Sigma}$ is a diagonal matrix with the kth diagonal component being the variance σ_k^2. We give different variance $\sigma^2_{(1),k}$ and $\sigma^2_{(0),k}$ for DE and Non-DE genes, respectively. Each variance component is assumed to follow a non-informative Jeffrey's prior, i.e. $\sigma^2_{(1),k} \sim \frac{1}{\sigma^2_{(1),k}}$, $\sigma^2_{(0),k} \sim \frac{1}{\sigma^2_{(0),k}}$.

For the prior of dispersion parameter ϕ_{gk}, we follow from Wu et al. (2013) and assume a log normal prior with a study-specific mean and variance common to all genes:

$$\log \phi_{gk} \sim N(m_k, \kappa_k^2),$$

where m_k is assumed to follow normal prior $N(\mu_m, \sigma_m^2)$ with pre-specified mean $\mu_m = 0$ and variance $\sigma_m^2 = 5^2$. κ_k^2 is assumed to follow a non-informative Jeffrey's prior, i.e. $\kappa_k^2 \sim \frac{1}{\kappa_k^2}$. Similarly, the variance of the linear model τ^2_{gk} is assumed to follow a non-informative Jeffrey's prior, i.e. $\tau^2_{gk} \sim \frac{1}{\tau^2_{gk}}$.

For the baseline expression α_{gk}, a_{gk} as well as the grand mean effect size λ_g, we assume a normal prior with pre-specified mean and variance:

$$\alpha_{gk} \sim N(\mu_\alpha, \sigma_\alpha^2), a_{gk} \sim N(\mu_a, \sigma_a^2), \lambda_g \sim N(\mu_\lambda, \sigma_\lambda^2),$$

where $\mu_\alpha = 0$, $\sigma_\alpha^2 = 5^2$, $\mu_a = 0$, $\sigma_a^2 = 5^2$, $\mu_\lambda = 0$, $\sigma_\lambda^2 = 5^2$. To complete the hierarchy, we also specify the prior for the DE indicator δ_{gk}: $P(\delta_{gk} = 1) = \pi_k, \pi_k \sim Unif(0, 1)$.

In addition to the informative parameters listed above, we introduce one auxiliary parameter ω_{gik} (dashed circle in Fig 7) into the negative binomial model to help obtain closed-form posterior distribution for β_{gk} and α_{gk} by exploiting conditional conjugacy (Polson et al., 2013; Zhou et al., 2012b). The prior for ω_{gik} is specified as:

$$\omega_{gik} \sim PG(y_{gik} + \phi_{gk}^{-1}, 0),$$

where PG refers to the Polya-Gamma distribution. The description above fully defines the proposed Bayesian hierarchical model. The observed data are RNA-seq count or microarray intensity, the library size for RNA-seq samples, the phenotypic indicator and the platform indicator $\{y_{gik}, T_{ik}, X_{ik}, \Phi_k\}$. We use Markov chain Monte Carlo (MCMC) sampling algorithm to sample the posterior distribution of unknown parameters that need to be updated, including δ_{gk}, β_{gk}, α_{gk}, a_{gk}, ϕ_{gk}, τ^2_{gk}, λ_g, σ^2_k, m_k, κ^2_k and ω_{gik}. A brief summary of updating functions and algorithms for each parameter is described in the Appendix.

3.2.3 Normalization Algorithm

Previous comparative studies on RNA-seq and microarray data found a systematic difference in log fold change (logFC) between the two platforms (Wang et al., 2014; Robinson et al., 2015), where RNA-seq always has a larger absolute logFC than microarray. To adjust for this inherent cross-platform bias in our full model, we hereby introduce a simple normalization algorithm:

1. The logFCs are first computed from each study. We then choose genes with absolute logFC greater than a pre-specified threshold in at least half of the studies as our candidate gene list for calculating the normalization factors. The threshold can be based on quantiles or values of biological significance (e.g. 2-fold change), and in the examples we show below, the selection of threshold based on effect size is quite robust. The selected set is denoted as $\mathscr{G}$.

2. Using one RNA-seq data as the reference, a simple linear model is used to test for the difference in absolute logFC between a test study k ($k = 1, 2, \ldots, K-1$) and the reference study:

$$abs(\log FC)_{gk} = p_k + \epsilon_{gk},$$

   where $g \in \mathscr{G}$, $abs(\log FC)_{gk}$ is the observed absolute log fold change of gene g in kth study, p_k denotes the platform effect of the kth study.

3. If the difference between platforms is significant (i.e. p-value for the coefficient p_k is smaller than $\frac{0.05}{(K-1)}$ after Bonferroni correction), the normalization factor f_k is calculated as the median difference of logFC between two platforms in the gene set $\mathscr{G}$; otherwise, no normalization is required (i.e. $f_k = 0$).

4. Lastly, the normalization factor is incorporated into the Bayesian model while updating the grand mean effect size parameter λ_g. More specifically, the new study-specific effect size becomes $\beta'_{gk} = \beta_{gk} + f_k$ and then λ_g is sampled using the new β'_{gk}. Details of this modification in MCMC algorithm can be referred to the Appendix.

   Remarks:

- Normalization works by adding constant normalization factor to the effect size estimates of microarray which is usually underestimated due to inefficient hybridization. The new estimates become more commensurate to that of RNA-seq while updating the grand mean.
- A normalization algorithm can be potentially incorporated into a two-stage effect size model. The effectiveness of normalization in such scenario needs to be further explored and is beyond the scope of this paper. Note that the normalization is infeasible for two-stage Fisher's method since it involves the combination of p-values.
- For the ILC example in our application, there is only one RNA-seq study so we will just use that study as the reference. In the case when there are multiple RNA-seq studies present, we will choose the study with the largest sample size, whose logFC estimates are more reliable (with smaller variability).

3.2.4 Evidence for necessity of normalization

We give three examples to show the necessity of performing normalization and demonstrate our normalization algorithm, using three publicly available datasets (GSE11045, GSE5350, GSE65365) from previous studies (Marioni et al., 2008; Su et al., 2011; Robinson et al., 2015). Each study consists of same samples measured by both RNA-seq and microarray from human, rat, and yeast respectively. We first selected a list of candidate genes using absolute logFC threshold of 0.5 in all three studies. In Figure 8, we showed the boxplots of logFC in the two platforms separately for up-regulated and down-regulated genes selected. As we can see, RNA-seq has a significantly larger absolute logFC than microarray in Marioni and Su's data for both up-regulated and down-regulated genes ($p < 0.05$), while no significant difference is found between the two platforms in Storey's data ($p > 0.05$). Thus, in this case, we will need to perform normalization for Marioni and Su's data, but not for Storey's data.

3.2.5 Inference for Differential Expression

In the Bayesian literature, Newton et al. (2004) proposed a direct approach to control FDR and defined a Bayesian false discovery rate as:

$$\mathrm{BFDR}(t) = \frac{\sum_{g=1}^{G} P_g(H_0|D) d_g(t)}{\sum_{g=1}^{G} d_g(t)},$$

where $P_g(H_0|D)$ is the posterior probability of gene g being non-DE (H_0) given data (D) and $d_g(t) = I\{P_g(H_0|D) < t\}$ as the indicator of claiming DE genes. t is a tuning parameter to control the BFDR at a certain α level. The Bayesian false discovery rate BFDR will be used to address the multiplicity issue for the Bayesian method throughout this paper so that it is comparable to the FDR control from the frequentist two-stage methods.

For fair comparison with the other frequentist meta-analysis methods (e.g. Fisher's method), we adopt an union-intersection (UIT) hypothesis (a.k.a. conjunction hypothesis) setting following Li et al. (2011): $\mathrm{H}_0 : \bigcap\{\beta_k = 0\}$ vs $\mathrm{H}_a : \bigcup\{\beta_k \neq 0\}$, i.e. reject the null when the gene is differentially expressed in at least one study, where β_k is the effect size of study k, $1 \leq k \leq K$. Correspondingly, we define a null set $\Omega^0 = \{\vec{\beta}_g : \sum\limits_{k=1}^{K} I(\beta_{gk} \neq 0) = 0\}$ and the respective DE set $\Omega^1 = \{\vec{\beta}_g : \sum\limits_{k=1}^{K} I(\beta_{gk} \neq 0) > 0\}$. To control BFDR at the gene level, we introduce a Bayesian equivalent q-value. From the Bayesian posterior, we can calculate the probability of each gene falling in the null space: $\hat{P}_g(H_0|D) = \hat{P}(\vec{\beta}_g \in \Omega^0|D) = \frac{\sum\limits_{t=1}^{T} I\{\vec{\delta}_g^{(t)} = \vec{0}\}}{T}$, where $\delta_g^{(t)} = (\delta_{g1}^{(t)}, \ldots, \delta_{gK}^{(t)})$ is the vector of DE indicators at the tth MCMC iteration, T is the total number of MCMC samples and $\vec{0}$ is a K-dimensional zero vector. We then define the Bayesian q-value of gene g as $q_g = \min\limits_{t \geq \hat{P}_g(H_0|D)} \mathrm{BFDR(t)}$. This q_g will be treated similarly as q-value in the frequentist approach.

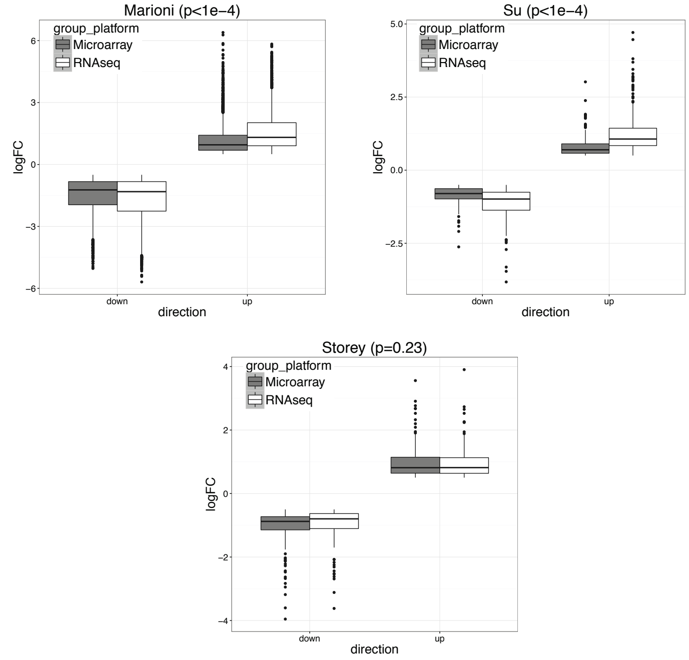


Figure 8: Boxplot of logFC from either microarray or RNA-seq in three public studies. The up-regulated or down-regulated genes are separately plotted. p-values from the linear model are attached to each plot.

3.2.6 Methods for comparison

Since no other cross-platform meta-analysis methods for integrating microarray and RNA-seq have been proposed, we will compare our method to three widely used two-stage methods in this paper: Fisher's method with edgeR (for RNA-seq) and limma (for microarray) used in single study DE analysis, the Fixed Effect Model (FEM) and the Random Effect Model (REM) (Fisher, 1925; Choi et al., 2003) with single study log fold change and variance estimated by DESeq2 (for RNA-seq) and limma (for microarray). The meta-analysed p-values are then adjusted for multiple comparison by Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

3.3 RESULTS

3.3.1 Simulation

Simulation setting

In this section, we provide one simulation example to show the benefits of the Bayesian integrative method over the other two-stage methods, especially after the inclusion of normalization algorithm. To mimic the real data, we randomly picked up 2000 genes from the TCGA-BRCA study (which includes both RNA-seq and microarray data) and used the estimated baseline expression (i.e. α and a) of these genes to simulate four studies, including two RNA-seq studies and two microarray studies. For RNA-seq, the library sizes for all samples were sampled from 0.4 to 0.8 million reads so the average counts range roughly from 200 to 400. The average intensity after log transformation is around 5 for microarray. We assumed the first 400 genes as DE genes, and the rest 1600 genes as Non-DE. For DE genes, we fixed the effect size of RNA-seq studies to be ± 1.25, the effect size of microarray studies to be ± 1 considering the systematic fold change difference between the two platforms. For Non-DE genes, the effect size is 0. The variance of microarray τ^2 is assumed to be 1, and the log dispersion $\log \phi$ is sampled from $Unif(-2, -1)$.

Simulation results

We compared our Bayesian method with and without normalization scheme (BayesNorm & Bayes, respectively) to the three two-stage meta-analysis methods: Fisher's method, FEM and REM. For a fair comparison, we assessed the power by plotting the number of detected true positives against the top number of declared DE genes in each method. As we can see from Figure 9, the full Bayesian model with normalization algorithm detected more true DE genes than any of the other four methods among the declared DE genes. In addition, the BayesNorm method was also more accurate than the other methods (ROC and PR curves shown in the Appendix). Note that even though both our method and the FEM/REM methods were effect size based, the integrative model was more powerful than the two-stage methods since two-stage approaches involve data reduction and theoretically lose efficiency.

3.3.2 Application

Data description

We applied the proposed model to two real datasets of invasive lobular carcinoma (ILC) breast cancer. ILC is the secondly most frequently diagnosed histological subtype of invasive breast cancer, consisting of $\sim 10\% - 15\%$ of all cases. As opposed to the most frequent invasive ductal carcinoma (IDC), ILC is less studied in its molecular mechanism, thus provides limited insight into the biological characteristics of the disease. In general, ILC cases usually express estrogen receptors (ER) but show no over-expression for HER2 protein (Ciriello et al., 2015). Here we collected one RNA-seq data from TCGA-BRCA study (Network et al., 2012), one microarray data from METABRIC (Curtis et al., 2012), one microarray data from Sotiriou study (Metzger-Filho et al., 2013), and a combination of 4 microarray datasets from GEO repository (GSE2109, GSE21653, GSE5460, GSE5764). Here, the four GEO studies contained four microarray datasets using Affymetrix U133 Plus 2.0, all are of small sample size. As a result, we obtained the raw data (CEL files) for simultaneous preprocessing and

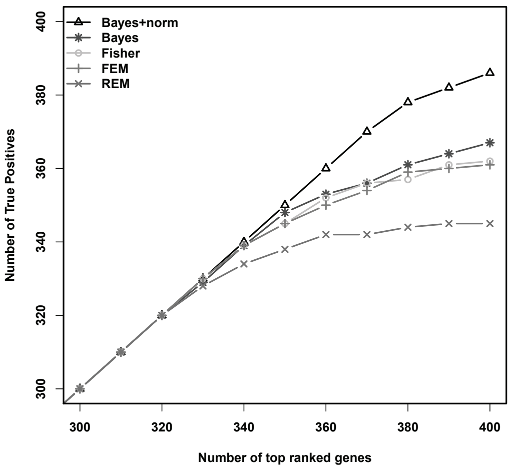


Figure 9: Power comparison of different methods in simulation.

All genes are ordered by the significance levels, the number of true positives among the top declared DE genes are compared. Red is the Bayesian method with normalization algorithm added, blue is without normalization algorithm, green is Fisher's method, brown and gray are fixed effect model and random effect model, respectively.

directly merged all qualified samples as the fourth study. All ILC samples used in the analysis are restricted to ER+ only. A summary of the ILC studies used in this paper can be found in the Appendix.

In the first example, we aim to identify biomarkers differentially expressed between early vs. late stage ILC breast cancer. To avoid confusing or erroneous tumor staging, we regarded pathological stage 0 and 1 as early stage and stage 3 and 4 as late stage and exclude the intermediate stage 2. Taking the stage information into account, we collected 69 ($N_{early} = 16, N_{late} = 53$), 57 ($N_{early} = 50, N_{late} = 7$), 57 ($N_{early} = 29, N_{late} = 28$) and 15 ($N_{early} = 5, N_{late} = 10$) samples from the four ILC studies, respectively. We first preprocessed the TCGA RNA-seq study by filtering out genes with mean counts less than 1. After merging and gene matching, 14621 genes were retained for ILC stage analysis.

In the second example, we aim to identify biomarkers differentially expressed between progesterone-receptor-positive (PR+) vs. progesterone-receptor-negative (PR-) ILC breast cancer. Taking the PR information into account, we collected 162 ($N_{PR+} = 144, N_{PR-} = 18$), 130 ($N_{PR+} = 80, N_{PR-} = 50$), 130 ($N_{PR+} = 93, N_{PR-} = 37$) and 43 ($N_{PR+} = 33, N_{PR-} = 10$) samples from the four studies, respectively. We similarly preprocessed the TCGA RNA-seq study by filtering out genes with mean counts less than 1. After merging and gene matching, 14636 genes were used for ILC PR analysis.

ILC stage example

As described in Section 3.2.3, for stage data, we first took genes with absolute logFC greater than 0.2 in at least 3 studies and used them to calculate the normalization factor. In Figure 10 (A), we noticed a significant difference in logFC between the TCGA RNA-seq study and the first two microarray studies ($p < \frac{0.05}{3}$) for ILC stage data. On the other hand, there was no significant difference in logFC between the RNA-seq study and the third microarray study ($p > \frac{0.05}{3}$). As a result, we performed embedded normalization on the first two microarray studies but not the third one while applying BayesNorm. The normalization factor was calculated as the median absolute difference of logFC (where RNA-seq always has a larger absolute logFC than microarray) for those selected genes.

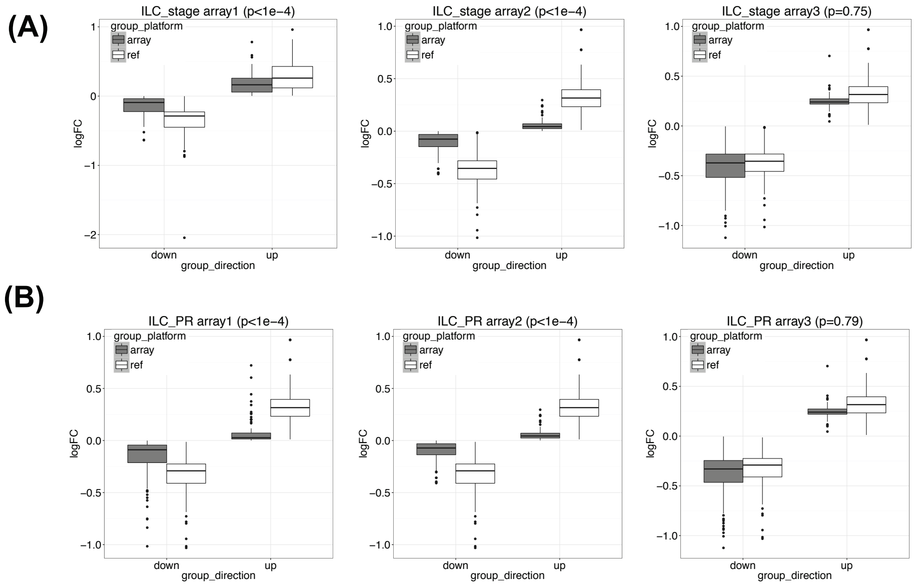


Figure 10: Cross-platform logFC comparison.

Panel (A) is from ILC-stage and panel (B) is from ILC-PR. White is the reference RNA-seq study from TCGA, boxplots of logFC from different platforms stratified by the directionality were included. p-values from the linear model are attached to each plot.

We applied five approaches (Bayes, BayesNorm, Fisher, FEM and REM) to the ILC stage example. As shown in Table 4, the Bayesian method without normalization detected 267 DE genes at q<0.05. With normalization, there were 279 DE genes detected. Both Bayesian models were more powerful than the two stage methods. We selected 3 representative genes that benefitted from normalization (Table 5). The log fold change and standard error (in the parenthesis) obtained from DESeq2 or limma are shown for all four studies. Without normalization, these genes are only marginally significant. After normalization, the significance level has been increased showing the necessity of normalization. "GLYATL2" is a gene coding for transferase that produces N-acyl glycines in humans and has been found to be differentially expressed across different breast cancer subtypes (Milioli et al., 2015). "FOSB" is an oncogene belonging to the FOS family and has been implicated as regulators of cell proliferation, differentiation, and transformation. Previous studies found that this gene was down-regulated in poorly differentiated breast carcinomas (Milde-Langosch et al., 2003). "KCNQ5" gene is a member of the KCNQ potassium channel gene family that yields currents which activate slowly with depolarization and recent review papers have regarded them as potential biomarkers for various types of cancer including breast cancer, glioblastoma and colorectal cancer (Lastraioli et al., 2015).

For the 279 DE genes detected by BayesNorm at q<0.05, we further performed a single-platform DE analysis using Bayesian model and compared the significance levels of the two platforms. Overall, RNA-seq is more significant than microarray in this dataset as shown in Figure 11 (A). Further, we found that for genes with lower RPKM (i.e. lowly expressed genes), RNA-seq is even more significant than microarray. This is consistent with the features of the two technologies: RNA-seq has a wider detection range and delivers low background signal, while microarray has a detection limit in the lower end.

To associate the detected biomarkers with the biological functions, we further performed pathway enrichment analysis using Fisher's exact test. For fair comparison, we used the top 500 genes identified from BayesNorm and Fisher's method for ILC data (FEM and REM are excluded due to too weak signals). For Fisher's method, this roughly corresponded to a q-value cutoff at 0.15. In Figure 6 (A), controlling FDR at 0.05, we identified 37 significant GO pathways from the BayesNorm method, while no significant pathways were identified

Table 4: Number of DE genes detected by five approaches at varying cutoff

Example	Method	q<0.01	q<0.05	q<0.1
ILC-stage	Bayes	167	267	365
	BayesNorm	161	279	400
	Fisher	19	57	195
	FEM	0	18	45
	REM	0	0	0
ILC-PR	Bayes	283	543	822
	BayesNorm	286	549	825
	Fisher	45	262	890
	FEM	0	1	176
	REM	0	1	44

Table 5: ILC stage: Three example genes that show the necessity of applying normalization

Gene	logFCseq1 (SE)	logFCarray1 (SE)	logFCarray2 (SE)	logFCarray3 (SE)	qBayes	qBayesNorm
GLYATL2	0.78 (0.28)	0.13 (0.17)	0.14 (0.24)	-0.68 (0.60)	0.02	0.002
FOSB	-0.85 (0.28)	-0.63 (0.44)	-0.08 (0.25)	-0.64 (0.52)	0.07	0.02
KCNQ5	0.82 (0.28)	0.05 (0.05)	0.56 (0.24)	0.05 (0.18)	0.07	0.04
		Normalized	Normalized	No Norm.		

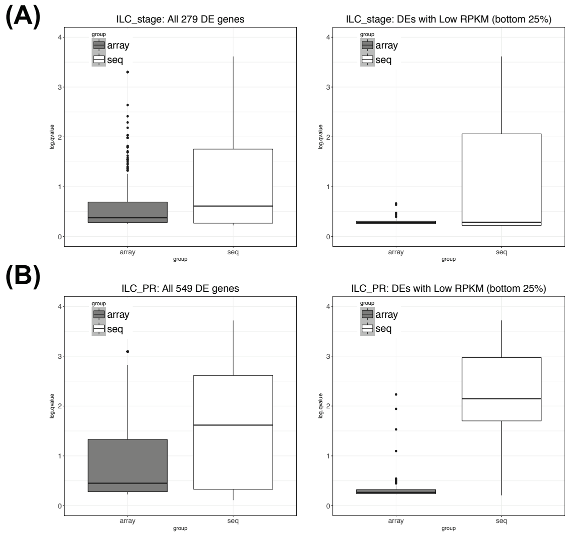


Figure 11: Comparison of significance of RNA-seq vs. microarray in BayesNorm detected DE genes.

Panel (A) is for ILC stage and panel (B) is for ILC PR. Y-axis is the negative log q-value, i.e. -log10(q), from the single platform DE analysis. White is for RNA-seq and black is for microarray. In the figure on the left, we included all DE genes, while on the right, we focused only on the genes with lower RPKM (bottom 25%).

from the Fisher's method. Intriguingly, we identified many cell fate and lineage pathways to be differentially activated comparing early and late stage ILC tumors (see pathway enrichment q-values and odds ratios in Table 6). Genes include members of the HOX and NKX gene family, SOX genes, EYA1 and others. This finding implies that early and late stage ILC tumors might have different precursors, or that significant changes in differentiation pathways contribute to progression of the disease.

ILC PR example

For ILC PR data, we took genes with absolute logFC greater than 0.2 in at least 3 studies and used them to calculate the normalization factor. In Figure 10 (B), we noticed a significant difference in logFC between the TCGA RNA-seq study and the first two microarray studies ($p < \frac{0.05}{3}$) for ILC stage data. On the other hand, there was no significant difference in logFC between the RNA-seq study and the third microarray study ($p > \frac{0.05}{3}$). As a result, we only performed normalization for the first two microarray studies with normalization factor calculated from the median absolute difference of logFC for those selected genes.

As shown in Table 4, the Bayesian method with normalization detected 549 DE genes at q<0.05 while Fisher's method only detected 262. We also selected 2 representative genes with increased significance level after normalization (Table 7) and among them for example, PTPRD is a tumor suppressor that is frequently inactivated in human cancers and has been identified to predict for poor prognosis in breast cancer (Veeriah et al., 2009). For the 549 DE genes detected by BayesNorm at q<0.05, we further performed a single-platform DE analysis using Bayesian model and compared the significance levels of the two platforms. As shown in Figure 11 (B), similar to that in the stage example, RNA-seq is more significant than microarray for genes with lower RPKM. For the PR example, there are 30 GO pathways identified by Bayesian method and 6 GO pathways identified by Fisher's method at FDR cutoff of 0.05 (Figure 3.3.2 (B)). As shown in Table 8, our pathway analysis showed a significant enrichment of genes involved in proteolysis and regulation of peptidase activity. These include many members of the serpin family, such as SERPINB5, SERPINA3, and SERPINA1. These proteins are inhibitors of serine proteases, and are known to mediate

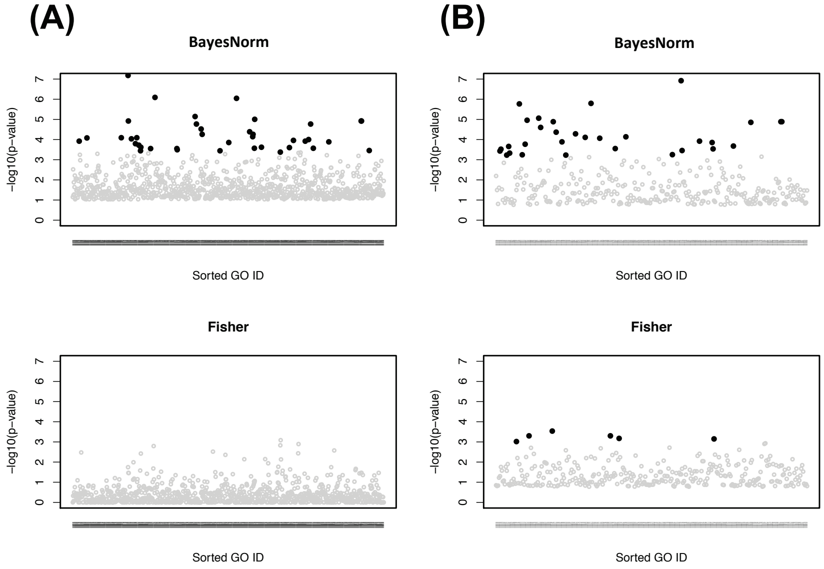


Figure 12: GO enrichment analysis results using the top 500 genes from the two methods. Panel (A) is for ILC stage and panel (B) is for ILC PR. Manhattan plot of GO pathways enriched by the top 500 DE genes from each method. X axis refers to the GO pathways sorted by GO IDs, Y axis refers to the -log10(p-values) from the Fisher's exact test, the highlighted points are the GO pathways with FDR < 0.05.

Table 6: ILC stage: Selected top pathways enriched with BayesNorm

Pathway Name	BayesNorm q-value	Fisher q-value
	(Odds Ratio)	(Odds Ratio)
GO:0007267: cell-cell signaling	6e-5 (2.21)	1 (1.43)
GO:0010817: regulation of hormone levels	3e-4 (2.74)	1 (1.28)
GO:0048665: neuron fate specification	0.02 (8.72)	1 (0)
GO:0048663: neuron fate commitment	0.03 (3.87)	1 (1.16)
KEGG Neuroactive ligand-receptor interaction	0.01 (2.91)	0.68 (1.62)
KEGG Steroid hormone biosynthesis	0.05 (4.21)	1 (0.94)
Reactome GPCR ligand binding	1e-3 (2.76)	1 (1.06)

Table 7: ILC PR: Three example genes that show the necessity of applying normalization

Gene	logFC.seq1	logFC.array1	logFC.array2	logFC.array3	qBayes	qBayesNorm
	(SE)	(SE)	(SE)	(SE)		
PTPRD	-0.44 (0.21)	-0.07 (0.06)	-0.02 (0.14)	-0.10 (0.20)	0.05	0.02
SULF2	-0.49 (0.17)	-0.25 (0.10)	-0.01 (0.10)	-0.51 (0.29)	0.05	0.02
		Normalized	Normalized	No Norm.		

breast cancer cell invasion and metastases, and some of the genes have been shown to be strong predictive biomarkers (Duffy et al., 2014).

3.4 DISCUSSION AND CONCLUSION

In this paper, we proposed a Bayesian hierarchical model to meta-analyze gene expression data generated from two popular transcriptome profiling platforms: microarray and RNA-seq. Within each platform, we adopted a negative binomial model for RNA-seq and linear model for microarray and we allowed the information integration of effect sizes across platforms among DE genes. An additional normalization algorithm was embedded in the Bayesian model to correct for the systematic cross-platform bias in effect sizes, as shown in previous studies and in the examples provided in our paper. To the best of our knowledge, the proposed model is the first cross-platform joint model for integrating microarray and RNA-seq transcroptomic data. Through simulation, we found that normalization was necessary and had increased the detection power of biomarkers. The application to ILC breast cancer data showed the advantage of our model in identifying DE genes comparing to the two-stage methods such as Fisher's method or the fixed/random effect models, and identified DE genes were validated by functional annotation (pathway) analysis.

During the analysis, we found that RNA sequencing was more powerful than microarray for lowly expressed genes. Similar findings have been shown in previous comparative studies (Sultan et al., 2008; Su et al., 2011; Wang et al., 2014). Using a comprehensive study design with 15 chemical treatments, Wang et al. (2014) showed that the concordance between two

Table 8: ILC PR: Selected top pathways enriched with BayesNorm

Pathway Name	BayesNorm q-value (Odds Ratio)	Fisher q-value (Odds Ratio)
GO:0010466: negative regulation of peptidase activity	1.e-4 (3.78)	1 (1.40)
GO:0010951: negative regulation of endopeptidase activity	5e-4 (3.53)	1 (1.46)
GO:0052547: regulation of peptidase activity	5e-4 (2.80)	1 (1.09)
GO:0045861: negative regulation of proteolysis	7e-4 (2.82)	1 (1.36)
GO:0052548: regulation of endopeptidase activity	8e-4 (2.73)	1 (1.16)
KEGG Drug metabolism - other enzymes	0.04 (7.14)	1 (2.64)

platforms dropped to below 40% for genes with below median expression, and a direct comparison to qPCR results indicated a better performance of RNA-seq in detecting differential gene expression at low expression levels than microarray. These results are consistent with the pros and cons of the two technologies where RNA-seq has a wider detection range and delivers low background signal, while on the other hand, microarray has a detection limit in the lower end. Though no advantage of microarray has been found in our application examples, we expect that microarray would be more powerful than RNA-seq in detecting DE genes with short lengths, considering the transcript length bias of RNA-seq.

Many studies have previously reported the systematic difference in log fold change between the two platforms (Wang et al., 2014; Robinson et al., 2015). In this paper, we reproduced the results using the same datasets and suggested that this difference was quite universal. More specifically, RNA-seq tend to have consistently larger absolute value of logFC than microarray under the same set of DE genes. Thus, to adjust for the difference, we introduced a simple normalization algorithm into the Bayesian model by taking the median difference of absolute logFC of representative genes between the two platforms as a constant normalization factor. Other normalization algorithm such as using adaptive normalization factor (e.g. varies according to expression levels, etc.) can also apply. In the ILC data application, the normalization algorithm increased the significance levels of some DE genes which were otherwise underpowered due to the log fold change difference.

Comparing to other methods, the Bayesian method has a few benefits. Firstly, it is relatively flexible to incorporate the normalization algorithm under Bayesian framework. Since the Bayesian estimation is sampling based (MCMC), the normalization factor can be directly put into the updating functions; Secondly, our Bayesian model includes a latent DE indicator, an individual effect size parameter and the overall effect size parameter. With this setting, the underpowered study/platform will be down-weighted automatically for some genes in a sense that its individual effect size will less likely contribute to the overall effect size. Such analysis to allow heterogeneity is relatively hard to achieve in a two-stage scenario. Thirdly, under Bayesian, we can allow the information of dispersion parameter to be shared across genes, which is fairly important in the entity of dispersion estimation.

There exist different "platforms" for both microarray and RNA-seq technologies. For example in microarray, data can be generated from Illumina platform, Affymetrix platform, etc; while in RNA-seq, the most popular platform is Ilumina which generates 95% of all sequencing data stored in GEO repository. Each platform has its own technical characteristics and protocol for handling and processing data. While combining microarray and RNA-seq, our Bayesian model only considers single platform from each technology. It can be readily extended to accommodate the multi-platform scenarios by including random effects or one more layer to explain for the account for the cross-platform difference within each technology.

Since the advent of next generation sequencing technology, RNA-seq has gradually become a standard experimental technique in measuring RNA expression levels while taking the place of traditional microarray technology. However, the large availability of historical microarray datasets in the GEO repository gives us a good reason of utilizing microarray in addition to RNA-seq in the DE analysis. Some of our findings in comparing the two platforms were consistent with the results reported from the third phase of MicroArray Quality Control (MAQC) project (a.k.a. SEQC) initiated by FDA (Consortium et al., 2014; Wang et al., 2014).

One limitation of our current method is that the normalization factors were estimated a priori and inserted into the Bayesian full model. Joint estimation of these parameters inside the model could be a potential extension in the future. Secondly, our model failed to take gene lengths into account, which could be one potential factor that will affect the detection power of different platforms. Our core MCMC updating algorithms were written in C++ and Rcpp was used to integrate the C++ codes into R. An R package, CBM ("Cross-platform Bayesian Model"), is publicly available to perform the analysis on the author's website (http://tsenglab.biostat.pitt.edu/software.htm).

4.0 VARIABLE SCREENING WITH MULTIPLE STUDIES

4.1 INTRODUCTION

In many areas of scientific disciplines nowadays such as omics studies (including genomics, transcriptomics, etc.), biomedical imaging and signal processing, high dimensional data with much greater number of features than the sample size (i.e. $p >> n$) have become rule rather than exception. For example, biologists may be interested in predicting certain clinical outcome (e.g. survival) using the gene expression data where we have far more genes than the number of samples. With the advancement of technologies and affordable prices in recent biomedical research, more and more experiments have been performed on a related hypothesis or to explore the same scientific question. Since the data from one study often have small sample size with limited statistical power, effective information integration of multiple studies can improve statistical power, estimation accuracy and reproducibility. Direct merging of the data (a.k.a. "mega-analysis") is usually less favored due to the inherent discrepancy among the studies (Tseng et al., 2012). New statistical methodologies and theories are required to solve issues in high-dimensional problem when integrating multiple related studies.

Various regularization methods have been developed in the past two decades and frequently used for feature selection in high-dimensional regression problems. Popular methods include, but are not limited to, Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005) and adaptive Lasso (Zou, 2006). When group structure exists among the variables (for example, a set of gene features belonging to a pre-specified pathway), group version of regularization methods can be applied (Yuan and Lin, 2006; Meier et al., 2008; Nardi et al., 2008). One can refer to Fan and Lv (2010) and Huang et al. (2012) for a detailed overview of variable selection and group selection in high-dimensional models.

When the number of features grows significantly larger than the sample size, most regularization methods perform poorly due to the simultaneous challenges of computation expediency, statistical accuracy and algorithmic stability (Fan et al., 2009). Variable screening methods become a natural way to consider by first reducing to a lower or moderate dimensional problem and then performing variable regularization. Fan and Lv (2008) first proposed a sure independent screening (SIS) method to select features based on their marginal correlations with the response in the context of linear regression models and showed such fast selection procedure enjoyed a "sure screening property". Since the development of SIS, many screening methods have been proposed for generalized linear models (Fan et al., 2009, 2010; Chang et al., 2013), nonparametric additive models or semiparametric models (Fan et al., 2011; Chang et al., 2016), quantile linear regression (Ma et al., 2017a), Gaussian graphical models (Luo et al., 2014; Liang et al., 2015) or exploit more robust measures for sure screening (Zhu et al., 2011; Li et al., 2012, 2017). However, all these screening methods are limited to single study so far.

In this paper, we first propose a general framework for simultaneous variable screening with multiple related studies. Compared to single study scenario, inclusion of multiple studies gives us more evidence to reduce dimension and thus increases the accuracy and efficiency of removing unimportant features during screening. To our knowledge, our paper is the first to utilize multiple studies to help variable screening in high-dimensional linear regression model. Such a framework provides a novel perspective to the screening problem and opens a door to the development of methods using multiple studies to perform screening under different types of models or with different marginal utilities. In this framework, it is natural to apply a selected screening procedure to each individual study, respectively. However, important features with weak signals in some studies may be incorrectly screened out if only such a one-step screening is performed. To avoid such false negative errors and fully take advantage of multiple studies, we further propose a two-step screening procedure, where one additional step of combining studies with potential zero correlation is added to the one-step procedure for a second check. This procedure has the potential to save those important features with weak signals in individual studies but strong aggregate effect across studies during the screening stage. Compared to the naive multiple study extension of SIS method, our procedure greatly reduces the false negative errors while keeping a low false positive rate. These merits are confirmed by our theoretical analysis. Specifically, we show that our procedure possesses the sure screening property with weaker assumptions on signals and allows the number of features to grow at an exponential rate of the sample size. Furthermore, we only require the data to have sub-Gaussian distribution via using novel self-normalized statistics. Thus our procedure can be applied to more general distribution family other than Gaussian distribution, which is considered in Fan and Lv (2008) and Bühlmann et al. (2010) for a related screening procedure under single study scenarios.

After screening, we further apply two general and applicable variable selection algorithms: the multiple study extension of PC-simple algorithm proposed by Bühlmann et al. (2010) as well as a two-stage feature selection method to choose the final model in a lower dimension.

The rest of the paper is organized as follows. In Section 4.2, we present a framework for variable screening with multiple related studies as well as notations. Then we propose our two-step screening procedure in Section 4.3. Section 4.4 provides the theoretical properties of our procedure, and demonstrates the benefits of multiple related studies as well as the advantages of our procedure. General algorithms for variable selection that can follow from our screening procedure are discussed in Section 4.5. Section 4.6 and 4.7 include the simulation studies and a real data application on three breast cancer transcriptomic studies, which illustrate the advantage of our method in reducing false negative errors and retaining important features as compared to the rank-based SIS method. We conclude and discuss possible extensions of our procedure in Section 4.8. The technical proofs to the major theorems are provided in the Appendix.

4.2 MODEL AND NOTATION

Suppose we have data from K related studies, each has n observations. Consider a random design linear model in each study $k \in [K]$ ($[K] = 1, \dots, K$):

$$Y^{(k)} = \sum_{j=1}^{p} \beta_j^{(k)} X_j^{(k)} + \epsilon^{(k)}, \tag{4.2.1}$$

where each $Y^{(k)} \in \mathbb{R}$, each $X^{(k)} = (X_1^{(k)}, \ldots, X_p^{(k)})^T \in \mathbb{R}^p$ with $E(X^{(k)}) = \mu_X^{(k)}$ and $\text{cov}(X^{(k)}) = \Sigma_X^{(k)}$, each $\epsilon^{(k)} \in \mathbb{R}$ with $E(\epsilon^{(k)}) = 0$ and $\text{var}(\epsilon^{(k)}) = \sigma^2$ such that $\epsilon^{(k)}$ is uncorrelated with $X_1^{(k)}, \ldots, X_p^{(k)}$, and $\beta^{(k)} = (\beta_1^{(k)}, \ldots, \beta_p^{(k)})^T \in \mathbb{R}^p$. We assume implicitly with $E(Y^{(k)2}) < \infty$ and $E\{(X_j^{(k)})^2\} < \infty$ for $j \in [p]$ ($[p] = 1, \ldots, p$).

When p is very large, we usually assume that only a small set of covariates are true predictors that contribute to the response. In other words, we assume most of $\beta_j = (\beta_j^{(1)}, \ldots, \beta_j^{(K)})^T$, where $j \in [p]$, are equal to a zero vector. In addition, in this paper, we assume $\beta_j^{(k)}$'s are either zero or non-zero in all K studies. This framework is partially motivated by a high-dimensional linear random effect model considered in literature (e.g. Jiang et al. (2016)). More specifically, we can have $\beta = (\beta_{(1)}^T, 0^T)^T$, where $\beta_{(1)}$ is the vector of the first s_0 non-zero components of β ($1 \leq s_0 \leq p$). Consider a random effect model where only the true predictors of each study are treated as the random effect, that is, $\beta^{(k)} = (\beta_{(1)}^{(k)}, 0)^T$ and $\beta_{(1)}^{(k)}$ is distributed as $N(\beta_{(1)}, \tau^2 I_{s_0})$, where τ^2 is independent of ϵ and X. Consequently, $\beta_j^{(k)}$'s are either zero or non-zero in all K studies with probability one. Such assumption fits the reality well, for example, in a typical GWAS study, a very small pool of SNPs are reported to be associated with a complex trait or disease among millions (Jiang et al., 2016).

With n i.i.d. observations from model (4.2.1), our purpose is to identify the non-zero $\beta_{(1)}$, thus we define the following index sets for active and inactive predictors:

$$\begin{aligned} \mathcal{A} &= \{j \in [p]; \beta_j \neq 0\} = \{j \in [p]; \beta_j^{(k)} \neq 0 \text{ for all } k\}; \\ \mathcal{A}^C &= \{j \in [p]; \beta_j = 0\} = \{j \in [p]; \beta_j^{(k)} = 0 \text{ for all } k\}, \end{aligned} \tag{4.2.2}$$

where $\mathcal{A}$ is our target. Clearly, under our setting, $\mathcal{A}$ and $\mathcal{A}^C$ are complementary to each other so that the identification of $\mathcal{A}^C$ is equivalent to the identification of $\mathcal{A}$. Let $|\mathcal{A}| = s_0$, where $|\cdot|$ denotes the cardinality.

4.3 SCREENING PROCEDURE WITH MULTIPLE STUDIES

4.3.1 Sure independence screening

For a single study ($K = 1$), Fan and Lv (2008) first proposed the variable screening method called sure independence screening (SIS) which ranked the importance of variables according to their marginal correlation with the response and showed its great power in preliminary screening and dimension reduction for high-dimensional regression problems. Bühlmann et al. (2010) later introduced the partial faithfulness condition that a zero partial correlation for some separating set S implied a zero regression coefficient and showed that it held almost surely for joint normal distribution. In the extreme case when $S = \emptyset$, it is equivalent to the SIS method.

The purpose of sure screening is to identify a set of moderate size d (with $d << p$) that will still contain the true set $\mathcal{A}$. Equivalently, we can try to identify $\mathcal{A}^C$ or subsets of $\mathcal{A}^C$ which contain unimportant features that need to be screened out. There are two potential errors that may occur in any sure screening methods (Fan and Lv, 2010):

1. **False Negative (FN):** Important predictors that are marginally uncorrelated but jointly correlated with the response fail to be selected.
2. **False Positive (FP):** Unimportant predictors that are highly correlated with the important predictors can have higher priority to be selected than other relatively weaker important predictors.

The current framework for variable screening with multiple studies is able to relieve us from the FP errors significantly. Indeed, we have multiple studies in our model setting thus we have more evidence to exclude noises and reduce FP errors than single study. In addition, sure screening is used to reduce dimension at a first stage, so we can always include a second stage variable selection methods such as Lasso or Dantzig selection to further refine the set and reduce FP errors.

The FN errors occur when signals are falsely excluded after screening. Suppose ρ_j is the marginal correlation of the jth feature with the response, with which we try to find the set $\{j : \rho_j = 0\}$ to screen out. Under the assumption of partial faithfulness (for explicit definition, see Section 4.4.3), these variables have zero coefficients for sure so the FN errors are guaranteed to be excluded. However, this might not be true for the empirical version of marginal correlation. For a single study ($K = 1$), to rule out the FN errors in empirical case, it is well-known that the signal-to-noise ratio has to be large (at least of an order of $(\log p/n)^{1/2}$ after Bonferroni adjustment). In the current setting with multiple studies, the requirement on strong signals remains the same if we naively perform one-step screening in each individual study. As we will see next, we propose a novel two-step screening procedure which allows weak signals in individual studies as long as the aggregate effect is strong enough. Therefore our procedure is able to reduces FN errors in the framework with multiple studies.

Before closing this section, it is worthwhile to mention that, to perform a screening test, one usually applies Fisher's z-transformation on the sample correlation (Bühlmann et al., 2010). However, this will require the bivariate normality assumption. Alternatively, in this paper, we propose to use the self-normalized estimator of correlation that works generally well even for non-Gaussian data (Shao, 1999). Similar ideas have been applied in the estimation of large covariance matrix (Cai and Liu, 2016).

4.3.2 Two-step screening procedure with multiple studies

In the presence of multiple studies, we have more evidence to reduce dimension and $\rho_j^{(k)} = 0$ for any k will imply a zero coefficient for that feature. On one hand, it is possible for features with zero β_j to have multiple non-zero $\rho_j^{(k)}$'s. On the other hand, a non-zero β_j will have non-zero $\rho_j^{(k)}$'s in all studies. Thus, we aim to identify the following two complementary sets while performing screening with multiple studies:

$$\begin{aligned} \mathcal{A}^{[0]} &= \{j \in [p]; \quad \min_k |\rho_j^{(k)}| = 0\}, \\ \mathcal{A}^{[1]} &= \{j \in [p]; \quad \min_k |\rho_j^{(k)}| \neq 0\}. \end{aligned} \tag{4.3.1}$$

We know for sure that $\mathcal{A}^{[0]} \subseteq \mathcal{A}^C$ and $\mathcal{A} \subseteq \mathcal{A}^{[1]}$ with the partial faithfulness assumption. For $j \in \mathcal{A}^{[0]}$, the chance of detecting a zero marginal correlation in at least one study has been greatly increased with increasing K, thus unimportant features will more likely be screened out as compared to single study scenario.

One way to estimate $\mathcal{A}^{[1]}$ is to test $H_0 : \rho_j^{(k)} = 0$ of each k for each feature j. When any of the K tests is not rejected for a feature, we will exclude this feature from $\hat{\mathcal{A}}^{[1]}$ (we call it the "One-Step Sure Independence Screening" procedure, or "OneStep-SIS" for short). This can be viewed as an extension of the screening test to multiple study scenario. However, in reality, it is possible for important features to have weak signals thus small $|\rho_j^{(k)}|$'s in at least one study. These features might be incorrectly classified into $\hat{\mathcal{A}}^{[0]}$ since weak signals can be indistinguishable from null signals in individual testing. It will lead to the serious problem of false exclusion of important features (FN) from the final set during screening.

This can be significantly improved by adding a second step to combine those studies with potential zero correlation (i.e., fail to reject the null $H_0 : \rho_j^{(k)} = 0$) identified in the first step and perform another aggregate test. For the features with weak signals in multiple studies, as long as their aggregate test statistics is large enough, they will be retained. Such procedure will be more conservative in screening features as to the first step alone, but will guarantee to reduce false negative errors.

For simplicity, we assume n i.i.d. observations $(X_i^{(k)}, Y_i^{(k)})$, $i \in [n]$, are obtained from all K studies. It is straightforward to extend the current procedure and analysis to the scenarios with different sample sizes across multiple studies, and thus omitted. Our proposed "Two-Step Aggregation Sure Independence Screening" procedure ("TSA-SIS" for short) is formally described below:

Step 1. Screening in each study

In the first step, we perform screening test in each study $k \in [K]$ and obtain the estimate of study set with potential zero correlations $\hat{l}_j$ for each $j \in [p]$ as:

$$\hat{l}_j = \{k; |\hat{T}_j^{(k)}| \leq \Phi^{-1}(1 - \alpha_1/2)\} \quad \text{and} \quad \hat{T}_j^{(k)} = \frac{\sqrt{n}\hat{\sigma}_j^{(k)}}{\sqrt{\hat{\theta}_j^{(k)}}}, \tag{4.3.2}$$

where $\hat{\sigma}_j^{(k)} = \frac{1}{n}\sum_{i=1}^n (X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)})$ is the sample covariance and $\hat{\theta}_j^{(k)} = \frac{1}{n}\sum_{i=1}^n [(X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)}) - \hat{\sigma}_j^{(k)}]^2$. $\hat{T}_j^{(k)}$ is the self-normalized estimator of covariance between $X_j^{(k)}$ and $Y^{(k)}$. Φ is the CDF of standard normal distribution and α_1 the pre-specified significance level.

In each study, we test if $|\hat{T}_j^{(k)}| > \Phi^{-1}(1-\alpha_1/2)$, if not, we will include study k into $\hat{l}_j$. This step does not screen out any variables, but instead separates potential zero and non-zero study-specific correlations for preparation of the next step. Define the cardinality of $\hat{l}_j$ as $\hat{\kappa}_j = |\hat{l}_j|$. If $\hat{\kappa}_j = 0$ (i.e., no potential zero correlation), we will for sure retain feature j and not consider it in step 2; Otherwise, we move on to step 2.

Remark 1. By the scaling property of $\hat{T}_j^{(k)}$, it is sufficient to impose assumptions on the standardized variables: $W^{(k)} = \frac{Y^{(k)}-E(Y^{(k)})}{\sqrt{\text{var}(Y^{(k)})}}, Z_j^{(k)} = \frac{X_j^{(k)}-E(X_j^{(k)})}{\sqrt{\text{var}(X_j^{(k)})}}$. Thus $\hat{T}_j^{(k)}$ can also be treated as the self-normalized estimator of correlation. We thus can define $\theta_j^{(k)} = \text{var}(Z_j^{(k)}W^{(k)})$ and $\sigma_j^{(k)} = \text{cov}(Z_j^{(k)}, W^{(k)}) = \rho_j^{(k)}$.

Remark 2. In our analysis, the index set in (4.3.2) is shown to coincide with $l_j(j \in \mathcal{A}^{[0]})$ and $l_j(j \in \mathcal{A}^{[1]})$ which will be introduced in more details in Section 4.4.

Step 2. Aggregate screening

In the second step, we wish to test whether the aggregate effect of potential zero correlations in $\hat{l}_j$ identified in step 1 is strong enough to be retained. Define the statistics $\hat{L}_j = \sum_{k\in\hat{l}_j} (\hat{T}_j^{(k)})^2$ and this statistics will approximately follow a $\chi^2_{\hat{\kappa}_j}$ distribution with degree of freedom $\hat{\kappa}_j$ under null. Thus we can estimate $\hat{\mathcal{A}}^{[0]}$ by:

$$\hat{\mathcal{A}}^{[0]} = \{j \in [p]; \hat{L}_j \le \varphi^{-1}_{\hat{\kappa}_j}(1-\alpha_2) \text{ and } \hat{\kappa}_j \neq 0\}, \tag{4.3.3}$$

or equivalently estimate $\hat{\mathcal{A}}^{[1]}$ by:

$$\hat{\mathcal{A}}^{[1]} = \{j \in [p]; \hat{L}_j > \varphi^{-1}_{\hat{\kappa}_j}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\}, \tag{4.3.4}$$

where $\varphi_{\hat{\kappa}_j}$ is the CDF of chi-square distribution with degree of freedom equal to $\hat{\kappa}_j$ and α_2 the pre-specified significance level.

The second step takes the sum of squares of $\hat{T}_j^{(k)}$ from studies with potential zero correlation as the test statistics. For each feature j, we test if $\sum_{k\in\hat{l}_j} (\hat{T}_j^{(k)})^2 > \varphi^{-1}_{\hat{\kappa}_j}(1-\alpha_2)$. If rejected, we conclude that the aggregate effect is strong and the feature needs to be retained,

Table 9: Toy example to demonstrate the strength of two-step screening procedure.

		S1 (signal)	S2 (signal)	N1 (noise)
	k=1	$\lvert\hat{T}_1^{(1)}\rvert = 3.71$	$\lvert\hat{T}_2^{(1)}\rvert = 3.70$	$\lvert\hat{T}_3^{(1)}\rvert = 0.42$
	k=2	$\lvert\hat{T}_1^{(2)}\rvert = 3.16$	$\lvert\hat{T}_2^{(2)}\rvert = 2.71$	$\lvert\hat{T}_3^{(2)}\rvert = 0.54$
	k=3	$\lvert\hat{T}_1^{(3)}\rvert = 3.46$	$\lvert\hat{T}_2^{(3)}\rvert = 2.65$	$\lvert\hat{T}_3^{(3)}\rvert = 0.56$
	k=4	$\lvert\hat{T}_1^{(4)}\rvert = 3.63$	$\lvert\hat{T}_2^{(4)}\rvert = 2.68$	$\lvert\hat{T}_3^{(4)}\rvert = 0.12$
	k=5	$\lvert\hat{T}_1^{(5)}\rvert = 3.24$	$\lvert\hat{T}_2^{(5)}\rvert = 1.94$	$\lvert\hat{T}_3^{(5)}\rvert = 0.69$
TSA-SIS	$\hat{l}_j$	$\emptyset$	$\{2, 3, 4, 5\}$	$\{1, 2, 3, 4, 5\}$
	$\hat{\kappa}_j$	0	4	5
	$\hat{L}_j$	-	$25.31 > \varphi_4(0.95)$	$1.27 < \varphi_5(0.95)$
	$\hat{\mathcal{A}}^{[0]}$	N	N	Y
	$\hat{\mathcal{A}}^{[1]}$	Y	Y	N
OneStep-SIS	$\hat{\mathcal{A}}^{[0]}$	N	Y	Y
	$\hat{\mathcal{A}}^{[1]}$	Y	N (FN)	N

otherwise, we will screen it out. This step performs a second check in addition to the individual testing in step 1 and potentially saves those important features with weak signals in individual studies but strong aggregate effect.

In Table 9, we use a toy example to demonstrate our idea and compare the two approaches ("OneStep-SIS" vs. "TSA-SIS"). In this example, suppose we have five studies ($K = 5$) and three features (two signals and one noise). "S1" is a strong signal with $\beta = 0.8$ in all studies, "S2" is a weak signal with $\beta = 0.4$ in all studies and "N1" is a noise with $\beta = 0$. In hypothesis testing, both small β and zero β can give small marginal correlation and are sometimes indistinguishable. Suppose $T = 3.09$ is used as the threshold (corresponding to $\alpha_1 = 0.001$). For the strong signal "S1", all studies have large marginal correlations, so both "OneStep-SIS" and "TSA-SIS" procedures include it correctly. For the weak signal "S2", since in many studies it has small correlations, it is incorrectly screened out by "OneStep-

SIS" procedure (False Negative). However, the "TSA-SIS" procedure saves it in the second step (with $\alpha_2 = 0.05$). For the noise "N1", both methods tend to remove it after screening.

4.4 THEORETICAL PROPERTIES

4.4.1 Assumptions and conditions

We impose the following conditions to establish the model selection consistency of our procedure:

(C1) (Sub-Gaussian Condition) There exist some constants $M_1 > 0$ and $\eta > 0$ such that for all $|t| \leq \eta$, $j \in [p]$, $k \in [K]$:

$$E\{\exp(tZ_j^{(k)2})\} \leq M_1, \quad E\{\exp(tW^{(k)2})\} \leq M_1.$$

In addition, there exist some $\tau_0 > 0$ such that $\min\limits_{j,k} \theta_j^{(k)} \geq \tau_0$.

(C2) The number of studies $K = O(p^b)$ for some constant $b \geq 0$. The dimension satisfies: $\log^3(p) = o(n)$ and $\kappa_j \log^2 p = o(n)$, where κ_j is defined next.

(C3) For $j \in \mathcal{A}^{[0]}$, $l_j(j \in \mathcal{A}^{[0]}) = \{k; \rho_j^{(k)} = 0\}$ and $\kappa_j = |l_j|$. If $k \notin l_j$, then $|\rho_j^{(k)}| \geq C_3\sqrt{\frac{\log p}{n}}\sqrt{1.01\theta_j^{(k)}}$, where $C_3 = 3(L+1+b)$.

(C4) For $j \in \mathcal{A}^{[1]}$, $l_j(j \in \mathcal{A}^{[1]}) = \{k; |\rho_j^{(k)}| < C_1\sqrt{\frac{\log p}{n}}\sqrt{0.99\theta_j^{(k)}}\}$ and $\kappa_j = |l_j|$, where $C_1 = L+1+b$. If $k \notin l_j$, then $|\rho_j^{(k)}| \geq C_3\sqrt{\frac{\log p}{n}}\sqrt{1.01\theta_j^{(k)}}$. In addition, we require $\sum\limits_{k \in l_j} |\rho_j^{(k)}|^2 \geq \frac{C_2(\log^2 p + \sqrt{\kappa_j \log p})}{n}$, where C_2 is some large positive constant.

The first condition (C1) assumes that each standardized variable $Z_j^{(k)}$ or $W^{(k)}$, $j \in [p]$, $k \in [K]$, marginally follow a sub-Gaussian distribution in each study. This condition relaxes the normality assumption in (Fan and Lv, 2008; Bühlmann et al., 2010). The second part of (C1) assumes there always exist some positive τ_0 not greater than the minimum variance of $Z_j^{(k)}W^{(k)}$. In particular, if $(X_j^{(k)}, Y^{(k)})$ jointly follows a multivariate normal distribution, then $\theta_j^{(k)} = 1 + \rho_j^{(k)2} \geq 1$, so we can always pick $\tau_0 = 1$.

The second condition (C2) allows the dimension p to grow at an exponential rate of sample size n, which is a fairly standard assumption in high-dimensional analysis. Many sure screening methods like "SIS", "DC-SIS" and "TPC" have used this assumption (Fan and Lv, 2008; Li et al., 2012, 2017). Though the PC-simple algorithm (Bühlmann et al., 2010) assumes a polynomial growth of p_n as a function of n, we notice that it can be readily relaxed to an exponential of n level. Further, we require the product $\kappa_j \log^2 p$ to be small, which is used to control the errors in the second step of our screening procedure. It is always true if $K \log^2 p = o(n)$.

Conditions (C3) assumes a lower bound on non-zero correlation (i.e. $k \notin l_j$) for features from $\mathcal{A}^{[0]}$. In other words, if the marginal correlation $|\rho_j^{(k)}|$ is not zero, then it must have a large enough marginal correlation to be detected. While this has been a key assumption for a single study in many sure screening methods (Fan and Lv, 2008; Bühlmann et al., 2010; Li et al., 2012, 2017), we only impose this assumption for $j \in \mathcal{A}^{[0]}$ rather than all $j \in [p]$. This condition is used to control for type II error in step 1 for features from $\mathcal{A}^{[0]}$.

Condition (C4) gives assumptions on features from $\mathcal{A}^{[1]}$. We assume the correlations to be small for those $k \in l_j$ and large for those $k \notin l_j$ so that studies with strong or weak signals can be well separated in the first step. This helps control the type II error in step 1 for features from $\mathcal{A}^{[1]}$. For those studies in l_j, we further require their sum of squares of correlations to be greater than a threshold, so that type II error can be controlled in step 2. This condition is different from other methods with single study scenario, where they usually assume a lower bound on each marginal correlation for features from $\mathcal{A}^{[1]}$ just like (C3). We relax this condition and only put restriction on their L_2 norm. This allows features from $\mathcal{A}^{[1]}$ to have weak signals in each study but combined strong signal. To appreciate this relaxation, we compare the minimal requirements with and without step 2. For each $j \in \mathcal{A}^{[1]}$, in order to detect this feature, we need $|\rho_j^{(k)}| \geq C(\log p/n)^{1/2}$ with some large constant C for all $k \in l_j$, and thus at least $\sum\limits_{k \in l_j} |\rho_j^{(k)}|^2 \geq C^2 \kappa_j \log p/n$. In comparison, the assumption in (C4) is much weaker in reasonable settings $\kappa_j >> \log p$.

4.4.2 Consistency of the two-step screening procedure

We state the first theorem involving the consistency of screening in our step 1:

Theorem 1. Consider a sequence of linear models as in (4.2.1) which satisfy assumptions and conditions (C1)-(C4), define the event $A = \{\hat{l}_j = l_j \text{ for all } j \in [p]\}$, there exists a sequence $\alpha_1 = \alpha_1(n,p) \to 0$ as $(n,p) \to \infty$ where $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ with $\gamma = 2(L+1+b)$ such that:

$$P(A) = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.4.1}$$

The proof of Theorem 1 can be found in the Appendix. This theorem states that the screening in our first step correctly identifies the set l_j for features in both $\mathcal{A}^{[0]}$ and $\mathcal{A}^{[1]}$ (in which strong and weak signals are well separated) and the chance of incorrect assignment is low. Given the results in Theorem 1, we can now show the main theorem for the consistency of the two-step screening procedure:

Theorem 2. Consider a sequence of linear models as in (4.2.1) which satisfy assumptions and conditions (C1)-(C4), we know there exists a sequence $\alpha_1 = \alpha_1(n,p) \to 0$ and $\alpha_2 = \alpha_2(n,p) \to 0$ as $(n,p) \to \infty$ where $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ with $\gamma = 2(L+1+b)$ and $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$ with $\gamma_{\kappa_j} = \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})$ and some constant $C_4 > 0$ such that:

$$P\{\hat{\mathcal{A}}^{[1]}(\alpha_1, \alpha_2) = \mathcal{A}^{[1]}\} = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.4.2}$$

The proof of Theorem 2 can be found in the Appendix. The result shows that the two-step screening procedure enjoys the model selection consistency and identifies the model specified in (4.3.1) with high probability. The choice of significance level that yields consistency is $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ and $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$.

4.4.3 Partial faithfulness and Sure screening property

Bühlmann et al. (2010) first came up with the partial faithfulness assumption which theoretically justified the use of marginal correlation or partial correlation in screening as follows:

$$\rho_{j|S} = 0 \text{ for some } S \subseteq \{j\}^C \text{ implies } \beta_j = 0, \tag{4.4.3}$$

where S is the set of variables conditioned on. For independence screening, $S = \emptyset$.

Under the two conditions: the positive definiteness of Σ_X and non-zero regression coefficients being realization from some common absolutely continuous distribution, they showed that partial faithfulness held almost surely (Theorem 1 in Bühlmann et al. (2010)). Since the random effect model described in Section 4.2 also satisfies the two conditions, the partial faithfulness holds almost surely in each study.

Thus, we can readily extend their Theorem 1 to a scenario with multiple studies:

Corollary 1. Consider a sequence of linear models as in (4.2.1) satisfying the partial faithfulness condition in each study and true active and inactive set defined in (4.2.2), then the following holds for every $j \in [p]$:

$$\rho_{j|S}^{(k)} = 0 \text{ for some } k \text{ for some } S \subseteq \{j\}^C \text{ implies } \beta_j = 0. \tag{4.4.4}$$

The proof is straightforward and thus omitted: if $\rho_{j|S}^{(k)} = 0$ for some study k, then with partial faithfulness, we will have $\beta_j^{(k)} = 0$ for that particular k. Since we only consider features with zero or non-zero $\beta_j^{(k)}$'s in all studies in (4.2.2), we will have $\beta_j = 0$. In the case of independence screening (i.e. $S = \emptyset$), $\rho_j^{(k)} = 0$ for some k will imply a zero β_j.

With the model selection consistency in Theorem 2 and the extended partial faithfulness condition in Corollary 1, the sure screening property of our two-step screening procedure immediately follows:

Corollary 2. Consider a sequence of linear models as in (4.2.1) which satisfy assumptions and conditions (C1)-(C4) as well as the extended partial faithfulness condition in Corollary 1, there exists a sequence $\alpha_1 = \alpha_1(n,p) \to 0$ and $\alpha_2 = \alpha_2(n,p) \to 0$ as $(n,p) \to \infty$ where $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ with $\gamma = 2(L+1+b)$ and $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$ with $\gamma_{\kappa_j} = \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})$ such that:

$$P\{\mathcal{A} \subseteq \hat{\mathcal{A}}^{[1]}(\alpha_1, \alpha_2)\} = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.4.5}$$

The proof of this Corollary simply combines the results of Theorem 2 and the extended partial faithfulness and is skipped here.

4.5 ALGORITHMS FOR VARIABLE SELECTION WITH MULTIPLE STUDIES

Usually, performing sure screening once may not remove enough unimportant features. In our case since there are multiple studies, we expect our two-step screening procedure to remove many more unimportant features than in single study. If the dimension is still high after applying our screening procedure, we can readily extend the two-step screening procedure to an iterative variable selection algorithm by testing the partial correlation with gradually increasing size of the conditional set S. Since such method is a multiple study extension of the PC simple algorithm in Bühlmann et al. (2010), we call it "Multi-PC" algorithm (Section 4.5.1).

On the other hand, if the dimension has already been greatly reduced with the two-step screening, we can simply add a second stage group-based feature selection techniques to select the final set of variables (Section 4.5.2).

4.5.1 Multi-PC algorithm

We start from $S = \emptyset$, i.e., our two-step screening procedure and build a first set of candidate active variables:

$$\hat{\mathcal{A}}^{[1,1]} = \hat{\mathcal{A}}^{[1]} = \{j \in [p]; \hat{L}_j > \varphi^{-1}_{\hat{\kappa}_j}(1 - \alpha_2) \text{ or } \hat{\kappa}_j = 0\}. \tag{4.5.1}$$

We call this set $stage_1$ active set, where the first index in $[,]$ corresponds to the stage of our algorithm and the second index corresponds to whether the set is for active variables ($[,1]$) or inactive variables ($[,0]$). If the dimensionality has already been decreased by a large amount, we can directly apply group-based feature selection methods such as group lasso to the remaining variables (to be introduced in Section 4.5.2).

However, if the dimension is still very high, we can further reduce dimension by increasing the size of S and considering partial correlations given variables in $\hat{\mathcal{A}}^{[1,1]}$. We follow the similar two-step procedure but now using partial correlation of order one instead of marginal correlation and yield a smaller $stage_2$ active set:

$$\hat{\mathcal{A}}^{[2,1]} = \{j \in \hat{\mathcal{A}}^{[1,1]}; \hat{L}_{j|q} > \varphi^{-1}_{\hat{\kappa}_{j|q}}(1-\alpha_2) \text{ or } \hat{\kappa}_{j|q} = 0, \text{ for all } q \in \hat{\mathcal{A}}^{[1,1]} \backslash \{j\}\}, \tag{4.5.2}$$

where each self-normalized estimator of partial correlation can be computed by taking the residuals from regressing over the variables in the conditional set.

We can continue screening high-order partial correlations, resulting in a nested sequence of m active sets:

$$\hat{\mathcal{A}}^{[m,1]} \subseteq \ldots \subseteq \hat{\mathcal{A}}^{[2,1]} \subseteq \hat{\mathcal{A}}^{[1,1]}. \tag{4.5.3}$$

Note that the active and inactive sets at each stage are non-overlapping and the union of active and inactive sets at a stage m will be the active set in a previous stage $m-1$, i.e., $\hat{\mathcal{A}}^{[m,1]} \cup \hat{\mathcal{A}}^{[m,0]} = \hat{\mathcal{A}}^{[m-1,1]}$. This is very similar to the original PC-simple algorithm, but now at each order-level, we perform the two-step procedure. The algorithm can stop at any stage m when the dimension of $\hat{\mathcal{A}}^{[m,1]}$ already drops to low to moderate level and other common group-based feature selection techniques can be used to select the final set. Alternatively, we can continue the algorithm until the candidate active set does not change anymore. The algorithm can be summarized as follows:

Algorithm 1. Multi-PC algorithm for variable selection.

1. Set $m = 1$, perform the two-step screening procedure to construct $stage_1$ active set:

$$\hat{\mathcal{A}}^{[1,1]} = \{j \in [p]; \hat{L}_j > \varphi^{-1}_{\hat{\kappa}_j}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\}.$$

2. Set $m = m+1$. Construct the $stage_m$ active set:

$$\hat{\mathcal{A}}^{[m,1]} = \{j \in \hat{\mathcal{A}}^{[m-1,1]}; \hat{L}_{j|S} > \varphi^{-1}_{\hat{\kappa}_{j|S}}(1-\alpha_2) \text{ or } \hat{\kappa}_{j|S} = 0,$$
$$\text{for all } S \subseteq \hat{\mathcal{A}}^{[m-1,1]} \backslash \{j\} \text{ with } |S| = m-1\}.$$

3. **Repeat** Step 2 until $m = \hat{m}_{reach}$, where $\hat{m}_{reach} = \min\{m : |\hat{\mathcal{A}}^{[m,1]}| \leq m\}$.

4.5.2 Two-stage feature selection

As an alternative to "Multi-PC" algorithm for variable selection, we also introduce here a two-stage feature selection algorithm by combining our two-step screening procedure and other regular feature selection methods together. In single study, for example, Fan & Lv (2008) performed sure independence screening in the first stage followed by model selection techniques including Adaptive Lasso, Dantzig Selector and SCAD, etc., and named those procedures as "SIS-AdaLasso","SIS-DS", "SIS-SCAD" , accordingly.

In our case, since the feature selection is group-based, we adopt a model selection technique using group Lasso penalty in the second stage:

$$\min_{\beta} \sum_{k=1}^{K} ||y^{(k)} - X^{(k)}_{\hat{\mathcal{A}}^{[1]}} \beta^{(k)}_{\hat{\mathcal{A}}^{[1]}}||_2^2 + \lambda \sum_{j \in \hat{\mathcal{A}}^{[1]}} ||\beta_j||_2 \quad , \tag{4.5.4}$$

where $\hat{\mathcal{A}}^{[1]}$ is the active set identified from our two-step screening procedure and the tuning parameter λ can be chosen by cross-validation or BIC in practice just like for a regular group Lasso problem. We call such two-stage feature selection algorithm as "TSA-SIS-groupLasso".

In addition, at any stages of the "Multi-PC" algorithm when the dimension has already been dropped to a moderate level, the group Lasso-based feature selection techniques can always take over to select the final set of variables.

4.6 NUMERICAL EVIDENCE

In this section, we demonstrate the advantage of TSA-SIS procedure in comparing to the multiple study extension of SIS (named "Min-SIS"), which ranks the features by the minimum absolute correlation among all studies. We simulated data according to the linear model in (4.2.1) including p covariates with zero mean and covariance matrix $\Sigma^{(k)}_{i,j} = r^{|i-j|}$ where $\Sigma^{(k)}_{i,j}$ denotes the (i, j)th entry of $\Sigma^{(k)}_X$.

In the first part of simulation, we fixed the sample size $n = 100$, $p = 1000$, the number of studies $K = 5$ and performed $B = 1000$ replications in each setting. We assumed that the true active set consisted of only ten variables and all the other variables had zero coefficients (i.e., $s_0 = 10$). The indices of non-zero coefficients were evenly spaced between 1 and p. The variance of the random error term in linear model was fixed to be 0.5^2. We randomly drew r from $\{0, 0.2, 0.4, 0.6\}$ and allowed different r's in different studies. We considered the following four settings:

1. Homogeneous weak signals across all studies: nonzero β_j generated from Unif$(0.1, 0.3)$ and $\beta_j^{(1)} = \beta_j^{(2)} = \ldots = \beta_j^{(K)} = \beta_j$.
2. Homogeneous strong signals across all studies: nonzero β_j generated from Unif$(0.7, 1)$ and $\beta_j^{(1)} = \beta_j^{(2)} = \ldots = \beta_j^{(K)} = \beta_j$.
3. Heterogeneous weak signals across all studies: nonzero β_j generated from Unif$(0.1, 0.3)$ and $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.
4. Heterogeneous strong signals across all studies: nonzero β_j generated from Unif$(0.7, 1)$ and $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.

We evaluated the performance of Min-SIS using receiver operating characteristic (ROC) curves which measured the accuracy of variable selection independently from the issue of choosing good tuning parameters (for Min-SIS, the tuning parameter is the top number of features d). The OneStep-SIS procedure we mentioned above was actually one special case of the Min-SIS procedure (by thresholding at α_1). In presenting our TSA-SIS procedure, we fixed $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$ so the result was just one point on the sensitivity vs. 1-specificity plot. We also performed some sensitivity analysis on the two cutoffs based on the first simulation (see Table 10) and found the two values to be optimal since they had both high sensitivity and high specificity. Thus we suggested fixing these two values in all the simulations.

Figure 4.6 showed the results of simulation 1-4. When the signals were homogeneously weak in all studies as in (1), TSA-SIS clearly outperformed the Min-SIS procedure (above its ROC curve). It reached about 90% sensitivity with controlled false positive errors (specificity $\sim$ 95%). In order to reduce false negatives, Min-SIS had to sacrifice the specificity and increased the false positives, which in the end lost the benefits of performing screening (i.e. end up keeping too many features). When the signals became strong as in (2), both procedures performed equally well. This fit our motivation and theory and showed the

Table 10: Sensitivity analysis on the choice of α_1 and α_2 in simulation

Sensitivity/Specificity	$\alpha_2 = 0.15$	0.05	0.01	0.001
α_1=0.01	0.793/0.901	0.525/0.984	0.210/0.999	0.142/1.000
0.001	0.947/0.826	0.864/0.943	0.691/0.990	0.373/0.999
0.0001	0.966/0.816	0.922/0.932	0.840/0.985	0.681/0.998

Note: All value are based on average results from $B = 1000$ replications.

strength of our two-step procedure in saving weak signals without much increase in false positive rates. When the signals became heterogeneous as in (3) and (4), both procedures performed worse than before. But the Min-SIS procedure never outperformed the TSA-SIS procedure since it only examined the minimum correlation among all studies while the two-step procedure additionally considered the aggregate statistics.

4.7 REAL DATA APPLICATION

We next demonstrated our method in three microarray datasets of triple-negative breast cancer (TNBC, sometimes a.k.a. basal-like), an aggressive subtype of breast cancer usually with poor prognosis. Previous studies have shown that the tumor suppressor protein “p53” played an important role in breast cancer prognosis and its expression was associated with both disease-free survival and overall survival in TNBC (Yadav et al., 2015). Our purpose was to identify the genes most relevant and predictive to the response - the expression level of *TP53* gene, which encodes p53 protein. The three datasets are publicly available on authors’ website or at GEO repository including METABRIC (a large cohort consisting of roughly 2000 primary breast tumours), GSE25066 and GSE76250 (Curtis et al., 2012; Itoh et al., 2014; Liu et al., 2016). We subset the data to focus on the TNBC cases only and ended up with 275, 178 and 165 TNBC samples in each dataset, respectively. After

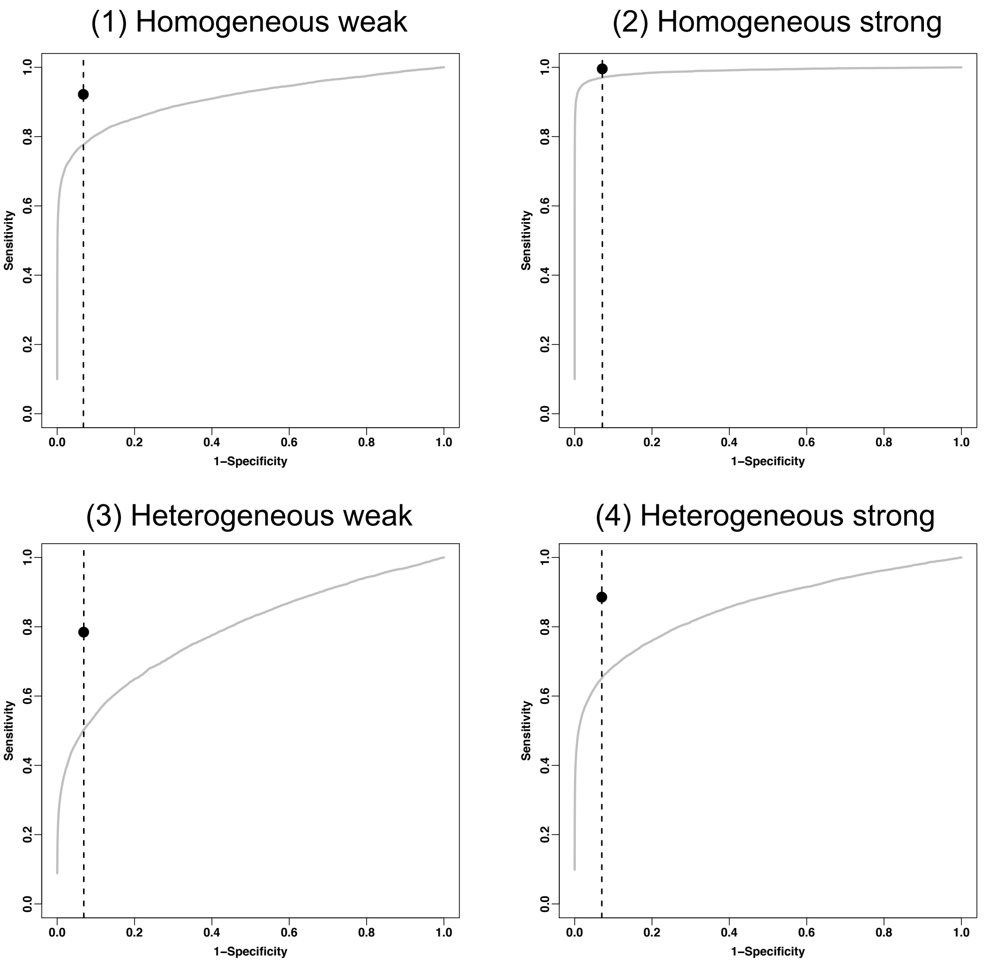


Figure 13: Simulation results 1-4.
The ROC curve is for Min-SIS, the black point is for our TSA-SIS using $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$.

routine preprocessing and filtering by including genes sufficiently expressed and with enough variation, a total of 3377 genes remained in common for the analysis.

We applied our Multi-PC algorithm and compared to the OneStep-SIS procedure as well as the Min-SIS method by using $d = n/\log(n) = 49$ (as suggested by their paper). We used $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$ (as determined by sensitivity analysis in simulation) and the "Multi-PC" algorithm only ran up to the first order (i.e. $m = 2$) and stopped

Table 11: The six genes selected by our TSA-SIS procedure.

Gene	METABRIC	GSE25066	GSE76250	Min-SIS	Rank in	OneStep-SIS
	Est (SE)	Est (SE)	Est (SE)	d=49	Min-SIS	$\lvert S \rvert$=25
Intercept	7.600 (1.502)	0.213 (0.553)	-1.783 (0.971)	-	-	-
EXOC1	0.251 (0.081)**	0.278 (0.157)$^{\cdot}$	0.293 (0.167)$^{\cdot}$	N	164	N
ITGB1BP1	-0.134 (0.045)**	0.003 (0.111)	-0.178 (0.194)	N	123	N
RBM23	0.168 (0.078)*	0.144 (0.167)	0.367 (0.168)*	N	152	N
SETD3	-0.166 (0.081)*	0.366 (0.184)*	-0.080 (0.175)	N	101	N
SQSTM1	-0.114 (0.050)*	0.029 (0.099)	0.245 (0.183)	N	98	N
TRIOBP	-0.126 (0.062)*	0.084 (0.118)	0.628 (0.261)*	N	91	N
Adjusted-R^2	0.151	0.522	0.359			

Note: "." indicates significant level of 0.1, "$*$" for level of 0.05, "$**$" for level of 0.01.

with six features. This again showed the power of screening with multiple studies. After feature selection, we fit the linear model in each study to obtain the coefficient estimates and adjusted R^2. Table 11 showed the coefficient estimates and standard errors of the final set of six genes selected by our procedure. We added three columns to indicate whether they were also retained by the Min-SIS (and their relative rank) or OneStep-SIS procedures. As we can see from the table, all the six genes selected by our procedure were missed by the other methods. Those genes typically had weak signals in one or more studies thus were very likely to be incorrectly excluded if only one step screening is performed. Since the METABRIC study had a larger sample size, all the coefficients appeared to be more significant than the other two studies.

The gene *EXOC1* and p53 are both components of the Ras signaling pathway which is responsible for cell growth and division and can ultimately lead to cancer (Rajalingam et al., 2007). *RBM23* encodes for an RNA-binding protein implicated in the regulation of estrogen-mediated transcription and has been found to be associated with p53 indirectly via a heat shock factor (Asano et al., 2016). *ITGB1BP1* encodes for an integrin protein which is essential for cell adhesion and other downstream signaling pathways that are also modulated by p53 (Brakebusch et al., 2002).

4.8 DISCUSSION

In this paper, we proposed a two-step screening procedure for high-dimensional regression analysis with multiple related studies. In a fairly general framework with weaker assumptions on the signal strength, we showed that our procedure possessed the sure screening property for exponentially growing dimensionality without requiring the normality assumption. We have shown through simulations that our procedure consistently outperformed the rank-based SIS procedure independent of their tuning parameter d. As far as we know, our paper is the first proposed procedure to perform variable screening in high-dimensional regression when there are multiple related studies. In addition, we also introduced two applicable variable selection algorithms following the two-step screening procedure.

Variable selection in regression with multiple studies have been studied in a subfield of machine learning called multi-task learning (MTL) before and the general procedure is to apply regularization methods by putting group Lasso penalty, fused Lasso penalty or trace norm penalty, etc. (Argyriou et al., 2007; Zhou et al., 2012a; Ji and Ye, 2009). However, at ultra-high dimension, such regularization methods usually fail due to challenges in computation expediency, statistical accuracy and algorithmic stability. Instead, sure screening can be used as a fast algorithm for preliminary feature selection, and as long as it exhibits comparable statistical performance both theoretically and empirically, its computational advantages make it a good choice in application (Genovese et al., 2012). Our method has provided an alternative to target the high-dimensional multi-task learning problems.

The current two-step screening procedure is based on the linear models but relaxes the Gaussian assumption to sub-Gaussian distribution. One can apply a modified Fisher's z-transformation estimator rather than our self-normalized estimator to readily accommodate general elliptical distribution families (Li et al., 2017). In biomedical applications, noncontinuous outcomes such as categorical, count or survival outcomes are more commonly seen. Fan et al. (2010) extended SIS and proposed a more general independent learning approach for generalized linear models by ranking the maximum marginal likelihood estimates. Fan et al. (2011) further extended the correlation learning to marginal nonparametric learning for screening in ultra-high dimensional additive models. Other researchers exploited more robust measure for the correlation screening (Zhu et al., 2011; Li et al., 2012; Balasubramanian et al., 2013). All these measures can be our potential extension by modifying the marginal utility used in the screening procedure. Besides, the idea of performing screening with multiple studies is quite general and is applicable to relevant statistical models other than the regression model, for example, Gaussian graphical model with multiple studies. We leave these interesting problems in future study.

5.0 DISCUSSION AND FUTURE WORKS

5.1 DISCUSSION

The first and second papers proposed Bayesian hierarchical models for the meta-analysis of transcriptomic data to identify important differentially expressed genes. The Bayesian approach is preferred for its flexibility in constructing hierarchical model to share the information (in our case, multiple transcriptomic studies and multiple genes) and incorporating prior knowledge and its easiness in computation and parameter estimation (via MCMC sampling). We expect increasing Bayesian applications in omics data integration (combine experimental data, prior biological knowledge, external biological databases and clinical data) in the near future. In addition, our cross-platform Bayesian model can be readily modified to meta-analyze epigenomic studies using methyl-seq and methylation array platforms as well.

The third paper proposed a general framework and a novel two-step screening procedure for the feature selection in high dimension regression analysis with multiple omics studies. The regression problem when we have multiple studies has already been considered in “multi-task learning”, a subfield of machine learning, however, our approach is the first to apply screening method to such setting. Moreover, the two-step procedure proposed was more beneficial than the naive one-step procedure or the rank-based SIS for its capability in reducing the serious false negative errors.

5.2 EXTENSION OF THE SCREENING PROCEDURE TO NON-LINEAR CASE

(Zhu et al., 2011) and (Li et al., 2012) proposed more robust and model-free marginal measures in place of the Pearson correlation for feature screening and ranking when the linearity assumption is not met. (Li et al., 2017) extended the PC simple algorithm (Bühlmann et al., 2010) to a wider family of elliptical linear regression models for robustness. (Li et al., 2017) proposed several marginal measures for sure screening in quantile regression. We can also generalize our procedure to accommodate a wide variety of commonly used parametric and semiparametric models (e.g. survival model) by modifying our marginal measure. Such methods will be likely to have more applications in the biomedical field.

APPENDIX A

APPENDIX FOR "BAYESMETASEQ"

A.1 *PARAMETER ESTIMATION BY GIBBS SAMPLING AND THE METROPOLIS-HASTINGS ALGORITHM*

In this section, we described the detailed updating conditional distributions or algorithms if there were no closed form conditional distributions for some parameters. The full conditional posterior is as follows:

$$\begin{aligned} P(-|Y_{gik},T_{ik},X_{ik}) \propto\ & P(Y_{gik}|\alpha_{gk},\beta_{gk},\phi_{gk})\times \\ & f(\alpha_{gk}|\eta_g,\tau_k^2,r)f(\beta_{gk}|\lambda_g,\delta_{gk},\sigma_k^2,\rho)f(\phi_{gk}|m_g,\xi_k^2,t) \\ & f(\eta_g|N(\mu_\eta,\sigma_\eta^2))(1/\tau_k^2)f(r|InvWishart(I,K+1)) \\ & f(\lambda_g|N(\mu_\lambda,\sigma_\lambda^2))(1/\sigma_k^2)f(\rho|InvWishart(I,K+1)) \\ & f(m_g|N(\mu_m,\sigma_m^2))(1/\xi_k^2)f(t|InvWishart(I,K+1)) \\ & f(\delta_{gk}|\pi_{gk})f(\pi_{gk}|\theta,c_g)f(\theta|G_0)f(c_g|p)f(p|a,C). \end{aligned} \quad \text{(A.1.1)}$$

To update each parameter, we simply integrate out the rest from the above.

Step 1

Gibbs sampling is used to update α_{gk},β_{gk}. The two sets of parameters would be updated for each gene in each study, for simplicity, I will drop the suffix g and k here. The posterior distributions of these two parameters have closed form conditioning on the supporting parameter

ω from the Polya-Gamma (PG) distribution. Following Polson et al. (2013), $\omega \sim PG(b, c)$ is an infinite convolution of gamma distributions defined as:

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{k - 1/2^2 + c^2/(4\pi^2)}.$$

where each $g_k \sim Gamma(b, 1)$ is an independent gamma random variable with $b > 0$, $c \in \Re$, and $\overset{D}{=}$ denotes equality in distribution.

The PG distribution has two important properties. Firstly, if $\omega \sim PG(b, 0)$, then by Laplace transform, we would have $E\{\exp(-\omega t)\} = cosh^{-b}(\sqrt{t/2})$, where $cosh(x) = \frac{e^x + e^{-x}}{2}$. Let $\omega \sim PG(y + \phi^{-1}, 0)$, the negative binomial likelihood in terms of proportion p and dispersion ϕ can thus be expressed as:

$$L(p, \phi) \propto p^y (1-p)^{\phi^{-1}} = \frac{[\exp(\Psi)]^y}{[1 + \exp(\Psi)]^{y+\phi^{-1}}} = \frac{2^{-(y+\phi^{-1})} \exp(\frac{(y-\phi^{-1})\Psi}{2})}{cosh^{y+\phi^{-1}}(\frac{\Psi}{2})}$$
$$\propto \exp(\frac{(y-\phi^{-1})\Psi}{2}) E_\omega\{\exp(-\frac{\omega \Psi^2}{2})\}. \quad \text{(A.1.2)}$$

In other words, conditioning on ω, the above will end up with some negative quadratic form of Ψ (see equation 2.2.3 in the main text) within the exponential. Thus, the normal prior on Ψ would be a conjugate prior conditioning on ω. Let's go back to equation 2.2.3 in the main text, assume $\boldsymbol{B} = (\alpha, \beta)^T$ and $\boldsymbol{Z_i} = (1, X_i)^T$, then conditioning on known ω_i's, we know the likelihood of $\boldsymbol{B}$ is equal to:

$$L(\boldsymbol{B}) \propto \prod_{i=1}^{N} \exp\{-\frac{\omega_i}{2}(\boldsymbol{Z_i^T B} - (\frac{y_i - \phi^{-1}}{2\omega_i} - \log T_i))^2\}. \quad \text{(A.1.3)}$$

Let $\boldsymbol{\Omega} = diag(\omega_1, \ldots, \omega_n)$ and $u_i = \frac{y_i - \phi^{-1}}{2\omega_i} - \log T_i$, $\boldsymbol{U} = (u_1, \ldots, u_n)^T$, and we have the prior $\boldsymbol{B} \sim N(\boldsymbol{c}, \boldsymbol{C})$, where $\boldsymbol{c} = (\eta, \lambda\delta)^T, \boldsymbol{C} = diag(\tau^2, \sigma^2)$, so the conditional posterior we used to update $\boldsymbol{B}$ would be:

$$(\boldsymbol{B}|-) \sim N(\boldsymbol{m}, \boldsymbol{V}), \text{ where } \boldsymbol{V = (Z\Omega Z^T + C^{-1})^{-1}, m = V(Z\Omega U + C^{-1}c)}. \quad \text{(A.1.4)}$$

Another important property of PG distribution is that any $PG(b, c)$ random variable ω has the following pdf (where the expectation in the denominator is taken w.r.t. $PG(b, 0)$):

$$p(\omega|b, c) = \frac{\exp(-\frac{c^2}{2}\omega) p(\omega|b, 0)}{E_\omega\{\exp(-\frac{c^2}{2}\omega)\}}$$

In other words, the posterior distribution of $\omega \sim PG(b, 0)$ given c still belongs to the PG class (in our case, $b = y + \phi^{-1}, c = \Psi$):

$$P(\omega|\Psi) \propto \exp(-\frac{\omega\Psi^2}{2})PG(y + \phi^{-1}, 0) \propto PG(y + \phi^{-1}, \Psi). \tag{A.1.5}$$

We can update each ω_i based on the above distribution using Gibbs sampling.

Step 2

For ϕ_g, since no closed form posterior distribution is available, we used Metropolis Hasting (MH) algorithm to update ϕ_g for all studies together. For each g, we proposed a new vector $\log(\vec{\phi}_{new}) = (\log(\phi_1), \ldots, \log(\phi_K))^T$ from some jump distribution $N_K(\log(\vec{\phi}_{old}), \mathbf{\Pi})$. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N_K(\log\vec{\phi}_{g_{new}}; m_g, \Pi)\prod_{k=1}^{K}\prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}X_{ik}, \phi_{gk_{new}})}{N_K(\log\vec{\phi}_{g_{old}}; m_g, \Pi)\prod_{k=1}^{K}\prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}X_{ik}, \phi_{gk_{old}})}. \tag{A.1.6}$$

If the proposal is accepted, we replace the old $\log(\vec{\phi})$ with the new one, otherwise, we keep the current value of $\log(\vec{\phi})$.

Step 3

We used Gibbs sampling to update λ_g, η_g, m_g based on their full conditional Gaussian distributions as follows:

$$\begin{aligned}
(\lambda_g|-) &\sim N_K(\lambda_\mu, \Sigma_\lambda),\ where\, \Sigma_\lambda = (diag(1/(\sigma_\lambda^2)) + K\Sigma^{-1})^{-1}, \\
&\lambda_\mu = \Sigma_\lambda(diag(1/(\sigma_\lambda^2))\vec{\mu}_\lambda + K\Sigma^{-1}\vec{\beta}_g) \\
(\eta_g|-) &\sim N_K(\eta_\mu, \Sigma_\eta),\ where\, \Sigma_\eta = (diag(1/(\sigma_\eta^2)) + K\Lambda^{-1})^{-1}, \\
&\eta_\mu = \Sigma_\eta(diag(1/(\sigma_\eta^2))\vec{\mu}_\eta + K\Lambda^{-1}\vec{\alpha}_g) \\
(m_g|-) &\sim N_K(m_\mu, \Sigma_m), \Sigma_m = (diag(1/(\sigma_m^2)) + K\Pi^{-1})^{-1}, \\
&m_\mu = \Sigma_m(diag(1/(\sigma_m^2))\vec{\mu}_m + K\Pi^{-1}\log\vec{\phi}_g)
\end{aligned} \tag{A.1.7}$$

To update λ_g, we will only use those β_{gk} for which $\delta_{gk} = 1$, if $\vec{\delta}_g = \vec{0}$, we would redraw from its prior $N(\mu_\lambda, \sigma_\lambda^2)$. Since we only need one value for each of the above parameters in every iteration, we took the average of each result.

Step 4

The full conditional for $[\boldsymbol{\sigma^2}_{(1),k}]_1^K$, $[\boldsymbol{\sigma^2}_{(0),k}]_1^K$, $[\boldsymbol{\tau^2}_k]_1^K$ and $[\boldsymbol{\xi^2}_k]_1^K$ have closed forms and are updated using Gibbs sampling for each k:

$$\begin{aligned}
\sigma^2_{(1),k} &\sim InvGamma(\frac{\sum_{g=1}^G \delta_{gk}}{2}, \frac{1}{2}\sum_{g=1}^G \delta_{gk}(\beta_{gk} - \lambda_g)^2) \\
\sigma^2_{(0),k} &\sim InvGamma(\frac{\sum_{g=1}^G (1-\delta_{gk})}{2}, \frac{1}{2}\sum_{g=1}^G (1-\delta_{gk})(\beta^2_{gk})) \\
\tau^2_k &\sim InvGamma(\frac{G}{2}, \frac{1}{2}\sum_{g=1}^G (\alpha_{gk} - \eta_g)^2) \\
\xi^2_k &\sim InvGamma(\frac{G}{2}, \frac{1}{2}\sum_{g=1}^G (\log \phi_{gk} - m_g)^2)
\end{aligned} \tag{A.1.8}$$

Step 5

The full conditional for $[\boldsymbol{\rho}_{(1)kk'}]_1^K$, $[\boldsymbol{\rho}_{(0)kk'}]_1^K$, $[\boldsymbol{r}_{kk'}]_1^K$, $[\boldsymbol{t}_{kk'}]_1^K$ have closed forms and are updated using Gibbs sampling:

$$\begin{aligned}
\text{For } \vec{\delta}_g \neq 0 \text{ , } [\boldsymbol{\rho}_{(1)kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^K (\bar{\beta}_k - \bar{\lambda})(\bar{\beta}_k - \bar{\lambda})^T, v = 2K+1) \\
\text{For } \vec{\delta}_g = 0 \text{ , } [\boldsymbol{\rho}_{(0)kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^K (\bar{\beta}_k)(\bar{\beta}_k)^T, v = 2K+1) \\
[\boldsymbol{r}_{kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^K (\bar{\alpha}_k - \bar{\eta})(\bar{\alpha}_k - \bar{\eta})^T, v = 2K+1) \\
[\boldsymbol{t}_{kk'}]_1^K &\sim InvWishart(\Psi = I + \sum_{k=1}^K (\log\bar{\phi}_k - \bar{m})(\log\bar{\phi}_k - \bar{m})^T, v = 2K+1)
\end{aligned} \tag{A.1.9}$$

where the average is taken over all genes for β_k, λ, α_k, η, $\log\phi_k$ and m. After drawing a new covariance matrix from the above posterior, the actual correlation matrix can be obtained by integrating out the variance components.

Step 6

Since the support for β_{gk} depends on the choice of δ_{gk}, we update $(\delta_{gk}, \beta_{gk})$ together for each g and k. Specifically, a new value $\delta^{new}_{gk} = 1 - \delta^{old}_{gk}$ is proposed, and we then generate

β_{gk}^{new} from the posterior in Step 1 based on δ_{gk}^{new}. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N(\beta_{gk}^{new}; \delta_{gk}^{new}\lambda_g, \sigma^2) \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}^{new} X_{ik}, \phi_{gk})}{N(\beta_{gk}^{old}; \delta_{gk}^{old}\lambda_g, \sigma^2) \prod_{i=1}^{I(k)} NB(y_{gik}; \log T_{ik} + \alpha_{gk} + \beta_{gk}^{old} X_{ik}, \phi_{gk})} \tag{A.1.10}$$

We accept or reject the proposed values jointly from the above.

Step 7

Lastly, upon obtaining the updates of δ_{gk}, we can estimate π_{gk} for every 20 chains, and we transform it into z_{gk} through the steps described in Section 2.3.2. Based on the vector $\vec{z}_g$, we can update the cluster assignment c_g for each gene by Gibbs sampling using the following conditional probabilities:

$$\text{If } c = c_h \text{ for some } h \neq g : P(c_g = c | c_{-g}, \vec{z}_g) = b\frac{n_c}{G-1+a} \int F(\vec{z}_g, \theta_c) dH_{-g,c}(\theta_c)$$

$$P(c_g \neq c_h \text{ for all } h \neq g | c_{-g}, \vec{z}_g) = b\frac{a}{G-1+a} \int F(\vec{z}_g, \theta) dG_0(\theta)$$

where $H_{-g,c}$ is the posterior distribution of θ_c based on the prior G_0 and all observations $\vec{z}_h$ for which $h \neq g$ and $c_h = c$, n_c is the cluster size of cluster c, b is the normalizing constant to make the probability sum to 1. More specifically, $\int F(\vec{z}_g, \theta_c) dH_{-g,c}(\theta_c) = f(\vec{z}_g; N_K(\mu_K = \frac{n_c}{n_c+1}\vec{\bar{z}}_h, \Sigma = diag(\frac{n_c+2}{n_c+1}, K))$, $\int F(\boldsymbol{z}_g, \theta) dG_0(\theta) = f(\vec{z}_g; N_K(\mu_K = \mathbf{0}_K, \Sigma = diag(2, K))$.

A.2 *SUPPLEMENTAL FIGURES AND TABLES*

Table 12: Comparison of parameters estimates by BayesMetaSeq with their true values from Simulation IA, K=2

Parameters	True values	Posterior mean (SE)
β_0	0	-0.01 (0.42)
β_1^+	(0.8,2)	1.21 (0.50)
β_1^-	(-2,-0.8)	-1.25 (0.59)
α^{high}	(-8.5,-4.5)	-7.06 (1.32)
α^{low}	(-11,-9)	-11.17 (0.84)

Table 13: Sensitivity analysis on hyperparameter μ_η

Value of μ_η	α^{high} Posterior mean (SE)	α^{low} Posterior mean (SE)
0	-7.06 (1.32)	-11.17 (0.84)
-3	-7.05 (1.31)	-11.11 (0.79)
-5	-7.03 (1.31)	-11.10 (0.78)
-7	-7.04 (1.31)	-11.11 (0.78)

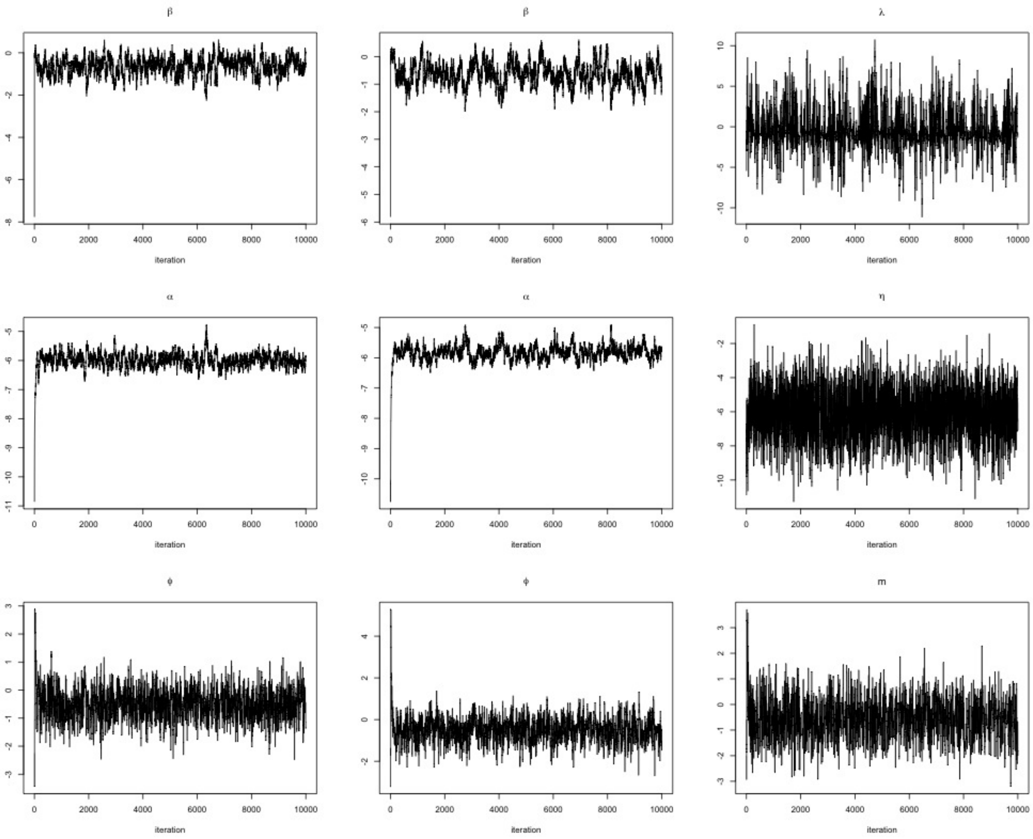


Figure 14: Traceplots of selected parameters from Simulation IA.

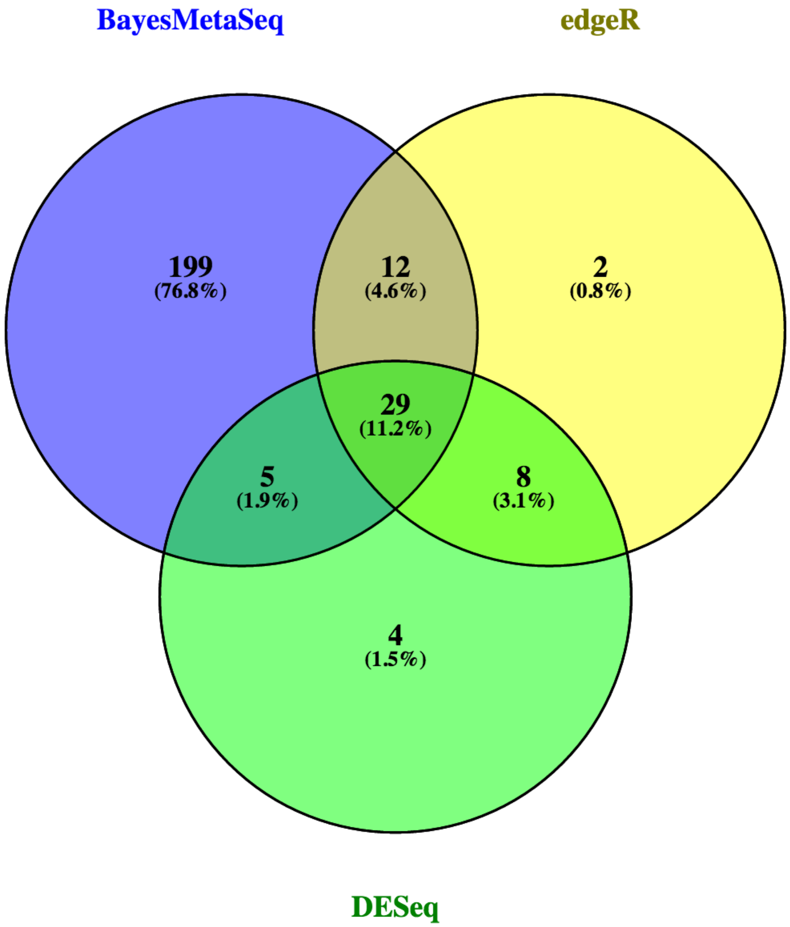


Figure 15: Venn Diagram of number of overlapping DE genes (FDR < 0.1) among the three methods applied in real data

Table 14: Normalized counts (rounded) for the three genes shown in Table 2.

Gene	Study	HIV strain	Normal strain
Mir212	HIP	(0,0,1,1,1,1,2,2,3,4,4,6)	(2,2,3,3,3,3,4,4,4,5,6,7)
	STR	(1,1,1,1,2,2,3,3,3,3,3,4)	(0,1,1,2,2,4,4,4,7,9,11,13)
	PFC	(0,1,1,1,2,2,3,3,4,5,6,7)	(0,1,2,4,5,6,6,6,7,7,8,9)
Mir384	HIP	(0,1,1,1,1,1,1,2,2,3,3,3)	(0,1,1,2,2,2,3,4,4,5,5,5)
	STR	(0,0,0,1,1,2,2,2,3,5,6,8)	(0,0,1,1,1,2,2,2,3,3,5,5)
	PFC	(1,1,1,1,1,1,2,2,3,3,4,4)	(2,2,3,3,4,4,5,5,7,7,7,7)
Alb	HIP	(3,4,4,4,5,5,5,6,6,8,13,41)	(1,2,2,2,2,2,3,3,3,3,4,8)
	STR	(0,3,3,4,4,4,8,11,14,18,20,21)	(1,1,2,3,3,4,7,8,11,12,13,26)
	PFC	(10,13,14,14,14,14,15,15,16,19,37,60)	(5,8,8,9,9,9,10,11,14,14,15,19)

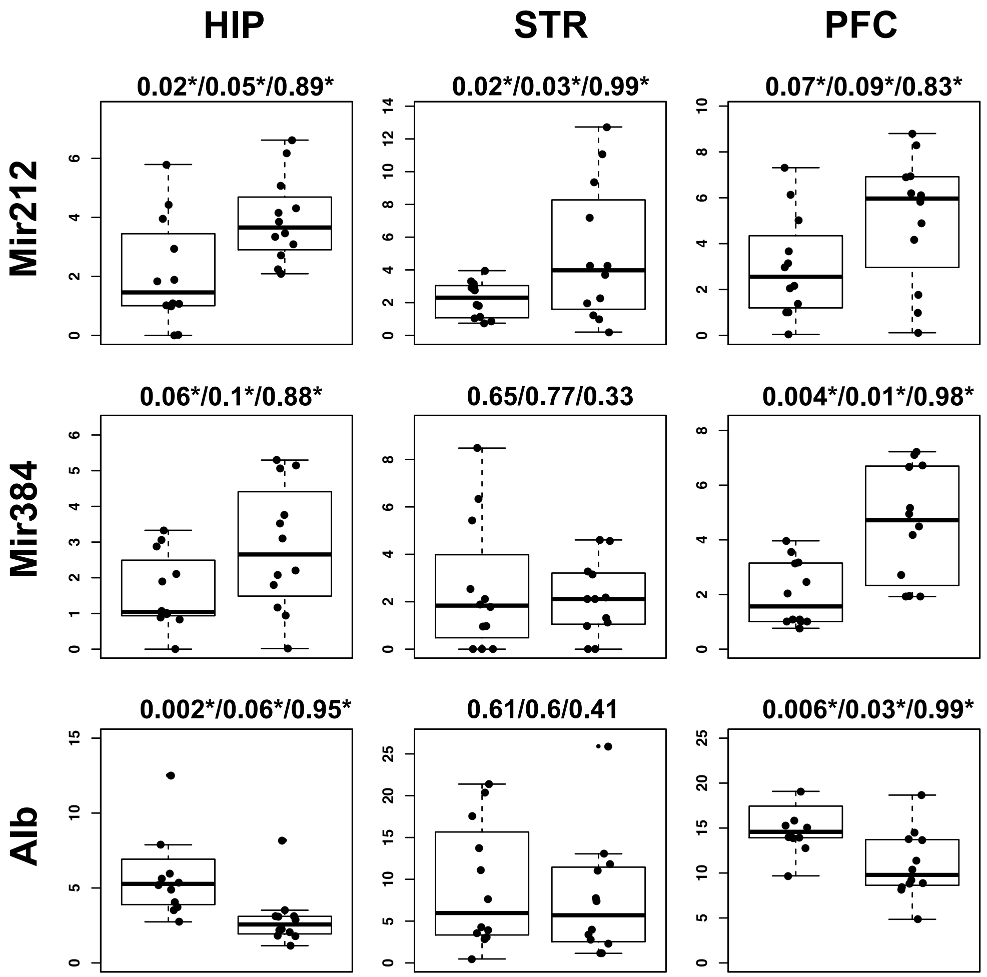


Figure 16: Distribution of normalized counts for the three genes shown in Table 2. Left: HIV strain; Right: Normal strain. The values above the boxplots correspond to the respective p-values or posterior means from edgeR/DESeq/BayesMetaSeq, with stars indicating the significance (e.g. p-value ≤ 0.1 or $E(\delta_{gk}|D) \geq 0.8$).

Table 15: List of significant IPA pathways (p-value < 0.05) from Cluster 1-4 in Figure 6.

Cluster	**Pathway Name**	**p-value**	**logOR**
Cluster 1	Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	0.003	3.79
	Role of JAK1, JAK2 and TYK2 in Interferon Signaling	0.017	4.67
	Allograft Rejection Signaling	0.023	4.33
	Autoimmune Thyroid Disease Signaling	0.023	4.33
	OX40 Signaling Pathway	0.030	4.08
	Role of RIG1-like Receptors in Antiviral Innate Immunity	0.031	4.03
	Interferon Signaling	0.033	3.98
	Activation of IRF by Cytosolic Pattern Recognition Receptors	0.046	3.60
Cluster 2	PI3K Signaling in B Lymphocytes	0.002	3.11
	G-Protein Coupled Receptor Signaling	0.002	2.55
	Protein Kinase A Signaling	0.003	2.45
	ERK/MAPK Signaling	0.005	2.72
	cAMP-mediated signaling	0.010	2.44
	Acute Phase Response Signaling	0.043	2.31
Cluster 3	Maturity Onset Diabetes of Young (MODY) Signaling	0.014	4.85
	Acyl-CoA Hydrolysis	0.016	4.67
	Stearate Biosynthesis I (Animals)	0.039	3.69
	Complement System	0.042	3.63
Cluster 4	Catecholamine Biosynthesis	0.008	5.99
	Adenine and Adenosine Salvage III	0.008	5.99
	Serotonin and Melatonin Biosynthesis	0.020	4.60
	Sphingomyelin Metabolism	0.020	4.60
	Adenine and Adenosine Salvage II	0.023	4.38
	Purine Nucleotides Degradation II (Aerobic)	0.035	3.91
	Tryptophan Degradation X (Mammalian, via Tryptamine)	0.046	3.59
	Primary Immunodeficiency Signaling	0.050	3.50

Table 16: List of significant IPA pathways (p-value < 0.05) from Cluster 5-7 in Figure 6.

Cluster	Pathway Name	p-value	logOR
Cluster 5	Regulation of the Epithelial-Mesenchymal Transition Pathway	7e-4	2.51
	Dendritic Cell Maturation	0.003	2.44
	Intrinsic Prothrombin Activation Pathway	0.003	3.77
	IL-4 Signaling	0.004	2.81
	Wnt/β-catenin Signaling	0.004	2.34
	Fc Epsilon RI Signaling	0.006	2.63
	Atherosclerosis Signaling	0.006	2.63
	Role of NANOG in Mammalian Embryonic Stem Cell Pluripotency	0.010	2.44
	Gα12/13 Signaling	0.010	2.44
	Human Embryonic Stem Cell Pluripotency	0.018	2.19
	Docosahexaenoic Acid (DHA) Signaling	0.018	2.78
	CTLA4 Signaling in Cytotoxic T Lymphocytes	0.020	2.74
	Melanoma Signaling	0.021	2.71
	Role of JAK1 and JAK3 in Cytokine Signaling	0.024	2.64
	Virus Entry via Endocytic Pathways	0.026	2.58
	IL-15 Signaling	0.028	2.55
	Endometrial Cancer Signaling	0.029	2.52
Cluster 6	Role of Cytokines in Mediating Communication between Immune Cells	0.008	5.48
	Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	0.029	4.16
Cluster 7	Serotonin Receptor Signaling	0.034	3.71

APPENDIX B

APPENDIX FOR “CBM”

B.1 *SAMPLE THE POSTERIOR DISTRIBUTION BY MCMC*

In this section, we described the detailed updating functions and algorithms in MCMC for the 12 groups of parameters in our model. We use both the Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953; Hastings, 1970) as well as the Gibbs sampling algorithm (Geman and Geman, 1984) to infer the posterior distribution of the parameters depending on whether closed form conditional distributions exist:

1. The full conditional of α_{gk} and β_{gk} for $\Psi_k = 1$ (i.e. RNA-seq) are bivariate normal with known $\vec{\omega}_{gk}$. We use Gibbs sampling to update them sequentially for each gene g in study k (subscript omitted for simplicity):

   $$(\boldsymbol{B}|-) \sim N(\boldsymbol{m}, \boldsymbol{V}), \text{ with } \boldsymbol{V} = (\boldsymbol{Z\Omega Z^T} + \boldsymbol{C^{-1}})^{-1}, \boldsymbol{m} = \boldsymbol{V}(\boldsymbol{Z\Omega U} + \boldsymbol{C^{-1}c}).$$

   where $\boldsymbol{Z_i} = (1, X_i)^T$, $\boldsymbol{\Omega} = diag(\omega_1, \ldots, \omega_n)$, $\boldsymbol{c} = (\mu_\alpha, \lambda\delta)^T$, $\boldsymbol{C} = diag(\sigma_\alpha^2, \sigma^2)$, $\boldsymbol{U} = (u_1, \ldots, u_n)^T$ and $u_i = \frac{y_i - \phi^{-1}}{2\omega_i} - \log T_i$.

2. The full conditional of $\vec{\omega}_{gk}$ is Polya-Gamma distribution with known α_{gk}, β_{gk} and $\log(\phi_{gk})$ (Polson et al., 2013; Zhou et al., 2012b). We use Gibbs sampling to update each ω_i:

   $$P(\omega|\Phi) \propto \exp(-\frac{\omega\Phi^2}{2})PG(y + \phi^{-1}, 0) \propto PG(y + \phi^{-1}, \Phi).$$

   where $\Phi_i = \log(T_i) + \alpha + \beta X_i + \log(\phi)$.

3. The full conditional of β_{gk} for $\Psi_k = 0$ (i.e. microarray) is Gaussian distribution for each gene g in each study k:

$$(\beta_{gk}|-) \sim N(\frac{(\lambda_g \delta_{gk})/\sigma_k^2 + \sum_{i=1}^{N_k}(X_{ik}y_{gik} - a_{gk})/\tau_{gk}^2}{1/\sigma_k^2 + \sum_{i=1}^{N_k} X_{ik}/\tau_{gk}^2}; (1/\sigma_k^2 + \sum_{i=1}^{N_k} X_{ik}/\tau_{gk}^2)^{-1})$$

4. The full conditional of a_{gk} is Gaussian distribution for each gene g in each study k:

$$(a_{gk}|-) \sim N(\frac{\mu_a/\sigma_a^2 + \sum_{i=1}^{N_k}(y_{gik} - X_{ik}\beta_{gk})/\tau_{gk}^2}{1/\sigma_a^2 + N_k/\tau_{gk}^2}; (1/\sigma_a^2 + N_k/\tau_{gk}^2)^{-1})$$

5. If $\vec{\delta_g} \neq \vec{0}$, we use Metropolis Hasting (MH) algorithm to update λ_g. For each g, we proposed a new λ_{new} from a jump distribution $N(\lambda_{old}, \sigma_\lambda^2)$. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N(\lambda_{new}; \mu_\lambda, \sigma_\lambda^2) N_K(\beta \vec{+} f_k; \lambda_{new}\vec{\delta}_k, diag(\sigma_{(1),k}^2))}{N(\lambda_{old}; \mu_\lambda, \sigma_\lambda^2) N_K(\beta \vec{+} f_k; \lambda_{old}\vec{\delta}_k, diag(\sigma_{(1),k}^2))}.$$

where the normalization factor f_k is included in this step to adjust for the difference in effect size across studies. If the proposal is accepted, we replace λ_{old} with λ_{new}, otherwise, we keep the current value of λ_{old}. If $\vec{\delta_g} = \vec{0}$, we will redraw from its prior $N(\mu_\lambda, \sigma_\lambda^2)$.

6. For ϕ_{gk}, since no closed form posterior distribution is available, we use Metropolis Hasting (MH) algorithm to update. For each g and k, we proposed a new $\log(\phi_{new})$ from a jump distribution $N(\log(\phi_{old}), \sigma_m^2)$. The proposal is accepted with probability $\min(1, r)$, where r is the acceptance ratio:

$$r = \frac{N(\log(\phi_{new}); m, \kappa^2) \prod_{i=1}^{N} NB(y_i; \log T_i + \alpha + \beta X_i, \phi_{new})}{N(\log(\phi_{old}); m, \kappa^2) \prod_{i=1}^{N} NB(y_i; \log T_i + \alpha + \beta X_i, \phi_{old})}.$$

If the proposal is accepted, we replace $\log(\phi_{old})$ with $\log(\phi_{new})$, otherwise, we keep the current value of $\log(\phi_{old})$.

7. We update m_k and κ_k^2 for each k from the following distributions:

$$(m_k|-) \sim N(\frac{\mu_m/\sigma_m^2 + \sum\limits_{g=1}^{G} \log(\phi_g)/\kappa_k^2}{1/\sigma_m^2 + G/\kappa_k^2}; (1/\sigma_m^2 + G/\kappa_k^2)^{-1});$$
$$(\kappa_k^2|-) \sim InvGamma(\frac{G}{2}; \frac{1}{2}\sum_{g=1}^{G}(\log(\phi_g) - m_k)^2)$$

8. We update $\sigma_{(1),k}^2$ (for those genes with $\vec{\delta}_g \neq \vec{0}$) and $\sigma_{(0),k}^2$ (for those genes with $\vec{\delta}_g = \vec{0}$) for each k from the following distributions:

$$\sigma_{(1),k}^2 \sim InvGamma(\frac{\sum_{g=1}^{G} \delta_{gk}}{2}, \frac{1}{2}\sum_{g=1}^{G} \delta_{gk}(\beta_{gk} - \lambda_g)^2)$$
$$\sigma_{(0),k}^2 \sim InvGamma(\frac{\sum_{g=1}^{G} (1 - \delta_{gk})}{2}, \frac{1}{2}\sum_{g=1}^{G}(1 - \delta_{gk})(\beta_{gk}^2))$$

9. We update τ_{gk}^2 for each g and k (for $\Psi_k = 0$) from the following distribution:

$$\tau_{gk}^2 \sim InvGamma(\frac{N_k}{2}, \frac{1}{2}\sum_{i=1}^{N_k}(y_{gik} - a_{gk} - b_{gk}X_{ik})^2)$$

10. To update δ_{gk}, we apply the MH algorithm. We propose a new value of δ_{gk} from the Bernoulli distribution with $P(\delta_{gk}^{new} = 1) = \pi_k$. If $\delta_{gk}^{new} = \delta_{gk}^{old}$, we just keep the same value. If $\delta_{gk}^{new} \neq \delta_{gk}^{old}$, we define the ratio of the two posterior density functions as r and accept the new proposed value δ_{gk}^{new} with probability $\min[1, r]$ (here we suppose $\delta_{gk}^{new} = 1$, similar r can be derived for $\delta_{gk}^{new} = 0$):

$$r = \frac{N(\beta_{gk}; \lambda_g, \sigma_{(1),k}^2)}{N(\beta_{gk}; 0, \sigma_{(0),k}^2)}$$

If the proposal is accepted, we replace δ_{gk}^{old} with δ_{gk}^{new}, otherwise, we keep the current value of δ_{gk}^{old}.

11. Lastly, we update π_k from $Dir(1 + \sum\limits_{g} \delta_{gk}, 1 + \sum\limits_{g}(1 - \delta_{gk}))$ and take the first element.

For both simulation and real data, we ran 10,000 MCMC iterations, the first 3,000 iterations were dropped as burn-in period in all analysis. The remaining 7,000 of 10,000 iterations are used for inference.

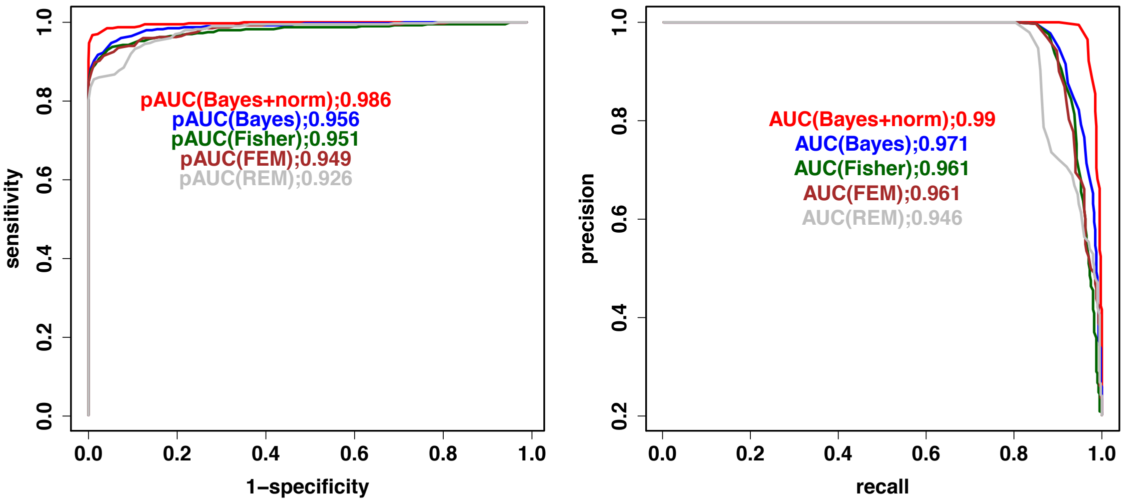


Figure 17: ROC Curve (left) and PR Curve (right) comparison of different methods. The AUC or partial AUC values are attached to each plot.

B.3 *DESCRIPTION OF ILC DATASETS*

Table 17: Overview of ILC datasets used in the application

Study	Platform	Stage sample size (early/late)	PR sample size (PR+/PR-)	Reference
TCGA BRCA	RNA-seq (Illumina)	69 (16/53)	162 (144/18)	Network et al. (2012)
METABRIC	microarray (Illumina)	57 (50/7)	130 (80/50)	Curtis et al. (2012)
Sotiriou	microarray (Affymetrix)	57 (29/28)	130 (93/37)	Metzger-Filho et al. (2013)
GSE2109, GSE21653, GSE5460, GSE5764	microarray (Affymetrix)	15 (5/10)	43 (33/10)	Sabatier et al. (2011); Lu et al. (2008); Turashvili et al. (2007)

APPENDIX C

APPENDIX FOR "TSA-SIS"

C.1 PROOFS

We start by introducing three technical lemmas that are essential for the proofs of the main results. By the scaling property of $\hat{T}_j^{(k)}$ and Remark 1, without loss of generality, we can assume $E(X_j^{(k)}) = E(Y^{(k)}) = 0$ and $\mathrm{var}(X_j^{(k)}) = \mathrm{var}(Y^{(k)}) = 1$ for all $k \in [K]$, $j \in [p]$. Therefore in the proof we do not distinguish between $\sigma_j^{(k)}$ and $\rho_j^{(k)}$. The first lemma is on the concentration inequalities of the self-normalized covariance and $\hat{\theta}_j^{(k)}$.

Lemma 1. Under the assumptions (C1) and (C2), for any $\delta \geq 2$ and $M > 0$, we have:

(i) $P(\max\limits_{j,k} |\frac{\hat{\sigma}_j^{(k)} - \sigma_j^{(k)}}{(\hat{\theta}_j^{(k)})^{1/2}}| \geq \delta\sqrt{\frac{\log p}{n}}) = O((\log p)^{-1/2} p^{-\delta+1+b})$,

(ii) $P(\max\limits_{j,k} |\hat{\theta}_j^{(k)} - \theta_j^{(k)}| \geq C_\theta\sqrt{\frac{\log p}{n}}) = O(p^{-M})$,

where C_θ is a positive constant depending on M_1, η and M only.

The second and third lemmas, which will be used in the proof of Theorem 2, describe the concentration behaviors of $\hat{H}_j^{(k)} := \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)}) - \rho_j^{(k)}]}{\sqrt{\theta_j^{(k)}}} = \hat{T}_j^{(k)}\sqrt{\frac{\hat{\theta}_j^{(k)}}{\theta_j^{(k)}}} - \frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}}$ and $\check{H}_j^{(k)} := \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{ij}^{(k)} Y_i^{(k)} - \rho_j^{(k)})}{\sqrt{\theta_j^{(k)}}}$.

Lemma 2. There exists some constant $c > 0$ such that,

$$P(|\sum_{k \in l_j}[\check{H}_j^{(k)2} - 1]| > t) \leq 2\exp(-c\min[\frac{t^2}{\kappa_j}, t^{1/2}]),$$

where c depends on M_1 and η only.

Lemma 3. There exists some constant $C_H > 0$ such that,

$$\begin{aligned} P(\max_{j,k} |\check{H}_j^{(k)} - \hat{H}_j^{(k)}| > C_H \sqrt{\frac{\log^2 p}{n}}) &= O(p^{-M}), \\ P(\max_{j,k} |\check{H}_j^{(k)2} - \hat{H}_j^{(k)2}| > C_H \sqrt{\frac{\log^3 p}{n}}) &= O(p^{-M}), \end{aligned}$$

where C_H depends on M_1, η, M and τ_0 only.

The proofs of the three lemmas are provided in the supplemental materials.

Proof of Theorem 1. We first define the following error events:

$$E_{j,k}^{I,\mathcal{A}^{[0]}} = \{|\hat{T}_j^{(k)}| > \Phi^{-1}(1-\alpha_1/2) \text{ and } j \in \mathcal{A}^{[0]}, k \in l_j\},$$

$$\begin{aligned} E_{j,k}^{II,\mathcal{A}^{[0]}} &= \{|\hat{T}_j^{(k)}| \leq \Phi^{-1}(1-\alpha_1/2) \text{ and } j \in \mathcal{A}^{[0]}, k \notin l_j\}, \\ E_{j,k}^{I,\mathcal{A}^{[1]}} &= \{|\hat{T}_j^{(k)}| > \Phi^{-1}(1-\alpha_1/2) \text{ and } j \in \mathcal{A}^{[1]}, k \in l_j\}, \\ E_{j,k}^{II,\mathcal{A}^{[1]}} &= \{|\hat{T}_j^{(k)}| \leq \Phi^{-1}(1-\alpha_1/2) \text{ and } j \in \mathcal{A}^{[1]}, k \notin l_j\}. \end{aligned}$$

To show Theorem 1 that $P(A) = 1 - O(p^{-L})$, it suffices to show that,

$$P\{\bigcup_{j,k}(E_{j,k}^{I,\mathcal{A}^{[0]}} \cup E_{j,k}^{II,\mathcal{A}^{[0]}})\} = O(p^{-L}), \tag{C.1.1}$$

and

$$P\{\bigcup_{j,k}(E_{j,k}^{I,\mathcal{A}^{[1]}} \cup E_{j,k}^{II,\mathcal{A}^{[1]}})\} = O(p^{-L}). \tag{C.1.2}$$

One can apply Lemma 1 to bound each component in (C.1.1) and (C.1.2) with $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ and $\gamma = 2(L+1+b)$. Specifically, we obtain that,

$$P(\bigcup_{j,k} E_{j,k}^{I,\mathcal{A}^{[0]}}) = P(\max_{j \in \mathcal{A}^{[0]}, k \in l_j} |\hat{T}_j^{(k)}| \geq \gamma\sqrt{\log p}) = O(\frac{1}{\sqrt{\log p}} p^{-\gamma+1+b}) = o(p^{-L}), \tag{C.1.3}$$

where the second equality is due to Lemma 1 (i) with $\delta = \gamma$, noting that $\sigma_j^{(k)} = 0$ and $\hat{T}_j^{(k)} = \sqrt{n}\hat{\sigma}_j^{(k)}/\sqrt{\hat{\theta}_j^{(k)}}$. In addition, we have that,

$$
\begin{aligned}
P(\bigcup_{j,k} E_{j,k}^{I,\mathcal{A}^{[1]}}) =& P\{\max_{j\in\mathcal{A}^{[1]},k\in l_j} |\hat{T}_j^{(k)}| \geq \gamma\sqrt{\log p}\} \\
\leq& P(\max_{j\in\mathcal{A}^{[1]},k\in l_j} |\frac{\hat{\sigma}_j^{(k)} - \rho_j^{(k)}}{(\hat{\theta}_j^{(k)})^{1/2}}| \geq (\gamma - C_1)\sqrt{\frac{\log p}{n}}) + O(p^{-L}) \\
=& O(\frac{1}{\sqrt{\log p}} p^{-(\gamma - C_1)+1+b}) + O(p^{-L}) \\
=& O(p^{-L}),
\end{aligned}
\tag{C.1.4}
$$

where the inequality on the second line is due to assumption (C4) on l_j for $j \in \mathcal{A}^{[1]}$, Lemma 1 (ii) with $M = L$, and assumption (C1) $\min_{j,k}\theta_j^{(k)} \geq \tau_0$, i.e., $\hat{\theta}_j^{(k)} \geq \theta_j^{(k)} - C_\theta(\log p/n)^{1/2} \geq 0.99\theta_j^{(k)}$. The equality on the third line follows from Lemma 1 (i) where $\delta = \gamma - C_1 = L+1+b$. In the end, we obtain that,

$$
\begin{aligned}
P\{\bigcup_{j,k}(E_{j,k}^{II,\mathcal{A}^{[0]}} \cup E_{j,k}^{II,\mathcal{A}^{[1]}})\} =& P(\max_{j,k\notin l_j} |\hat{T}_j^{(k)}| < \gamma\sqrt{\log p}) \\
\leq& P(\max_{j,k\notin l_j} |\frac{\hat{\sigma}_j^{(k)} - \rho_j^{(k)}}{(\hat{\theta}_j^{(k)})^{1/2}}| \geq (C_3 - \gamma)\sqrt{\frac{\log p}{n}}) + O(p^{-L}) \\
=& O(\frac{1}{\sqrt{\log p}} p^{-(C_3 - \gamma)+1+b}) + O(p^{-L}) \\
=& O(p^{-L}),
\end{aligned}
\tag{C.1.5}
$$

where the inequality on the second line is due to assumptions (C3) and (C4) on l_j, Lemma 1 (ii) with $M = L$ and assumption (C1) on sub-Gaussian distributions, i.e., $\hat{\theta}_j^{(k)} \leq \theta_j^{(k)} + C_\theta(\log p/n)^{1/2} \leq 1.01\theta_j^{(k)}$. In particular, we have implicitly used the fact that $\max_{j,l}\theta_j^{(k)}$ is upper bounded by a constant depending on M_1 and η only. The equality on the third line follows from Lemma 1 (i) where $\delta = C_3 - \gamma = L + 1 + b$.

Finally, we complete the proof by combining (C.1.3)-(C.1.5) to show (C.1.1)-(C.1.2).

□

Proof of Theorem 2. We first define the following error events:

$$E_j^{\mathcal{A}^{[0]},2} = \{|\hat{L}_j| > \varphi^{-1}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\} \text{ for } j \in \mathcal{A}^{[0]},$$

$$E_j^{\mathcal{A}^{[1]},2} = \{|\hat{L}_j| < \varphi^{-1}(1-\alpha_2) \text{ and } \hat{\kappa}_j \neq 0\} \text{ for } j \in \mathcal{A}^{[1]}.$$

To prove Theorem 2, we only need to show that,

$$P(\bigcup_{j\in\mathcal{A}^{[0]}} E_j^{\mathcal{A}^{[0]},2}) = O(p^{-L}) \quad \text{and} \quad P(\bigcup_{j\in\mathcal{A}^{[1]}} E_j^{\mathcal{A}^{[1]},2}) = O(p^{-L}), \tag{C.1.6}$$

with $\alpha_{2,\kappa_j} := 1 - \varphi_{\kappa_j}[\kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})] := 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$.

Recall the event A defined in Theorem 1. Thus we have that,

$$\begin{aligned} &P\{(\cup_{j\in\mathcal{A}^{[0]}} E_j^{\mathcal{A}^{[0]},2})\bigcup(\cup_{j\in\mathcal{A}^{[1]}} E_j^{\mathcal{A}^{[1]},2})\} \\ \leq &P(A^C) + p\max_{j\in\mathcal{A}^{[0]}} P(\sum_{k\in l_j}\hat{T}_j^{(k)2} > \gamma_{\kappa_j}) + p\max_{j\in\mathcal{A}^{[1]},\kappa_j\neq 0} P(\sum_{k\in l_j}\hat{T}^{(k)2} < \gamma_{\kappa_j}). \end{aligned}$$

Therefore, given the results in Theorem 1, it suffices to show,

$$P(\sum_{k\in l_j}\hat{T}^{(k)2} > \gamma_{\kappa_j}) = O(p^{-L-1}) \text{ for any } j \in \mathcal{A}^{[0]}, \tag{C.1.7}$$

and

$$P(\sum_{k\in l_j}\hat{T}^{(k)2} < \gamma_{\kappa_j}) = O(p^{-L-1}) \text{ for any } j \in \mathcal{A}^{[1]} \text{ and } \kappa_j > 0. \tag{C.1.8}$$

We first prove equation (C.1.7). Since $j \in \mathcal{A}^{[0]}$, we have $\hat{H}_j^{(k)} = \hat{T}_j^{(k)}\sqrt{\frac{\hat{\theta}_j^{(k)}}{\theta_j^{(k)}}}$. We are ready to bound the probability of $\sum_{k\in l_j} \hat{T}_j^{(k)2} > \gamma_{\kappa_j}$ below.

$$\begin{aligned}
&P(\sum_{k\in l_j} \hat{T}_j^{(k)2} > \gamma_{\kappa_j})\\
\leq &P(\sum_{k\in l_j} \hat{H}_j^{(k)2} > (1-\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}})\gamma_{\kappa_j}) + O(p^{-L-1})\\
\leq &P(\sum_{k\in l_j} (\check{H}_j^{(k)2}-1) > (1-\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}})\gamma_{\kappa_j} - \kappa_j - \kappa_j C_H\sqrt{\frac{\log^3 p}{n}}) + O(p^{-L-1})\\
= &P(\sum_{k\in l_j} (\check{H}_j^{(k)2}-1) > \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p}) - \frac{C_\theta}{\tau_0}\sqrt{\frac{\kappa_j^2 \log p}{n}}\\
&-\frac{C_\theta C_4}{\tau_0}(\sqrt{\frac{\log^5 p}{n}} + \sqrt{\frac{\kappa_j \log^2 p}{n}}) - \kappa_j - \kappa_j C_H\sqrt{\frac{\log^3 p}{n}}) + O(p^{-L-1})\\
\leq &P(\sum_{k\in l_j} (\check{H}_j^{(k)2}-1) > C_2'(\log^2 p + \sqrt{\kappa_j \log p})) + O(p^{-L-1})\\
= &O(p^{-L-1}).
\end{aligned}$$

The inequality on the second line is due to assumption (C1) that $\min_{j,k} \theta_j^{(k)} \geq \tau_0 > 0$ and Lemma 1 (ii) with $M = L+1$. The inequality on the third line follows from Lemma 3 with $M = L+1$. The inequality on the fifth line is by the choice of γ_{κ_j} with a sufficiently large $C_4 > 0$ and the assumption (C2) that $\log^3 p = o(n)$ and $\kappa_j \log^2 p = o(n)$. The last equality follows from Lemma 2.

Lastly, we prove (C.1.8) as follows,

$$\begin{aligned}
&P(\sum_{k\in l_j}\hat{T}_j^{(k)2}<\gamma_{\kappa_j})\\
=&P(\sum_{k\in l_j}(\hat{H}_j^{(k)}+\frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}})^2\frac{\theta_j^{(k)}}{\hat{\theta}_j^{(k)}}<\gamma_{\kappa_j})\\
\leq&P(\sum_{k\in l_j}(\hat{H}_j^{(k)}+\frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}})^2\leq(1+\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}})\gamma_{\kappa_j})+O(p^{-L-1})\\
\leq&P(\sum_{k\in l_j}(\check{H}_j^{(k)2}-1)\leq\kappa_jC_H\sqrt{\frac{\log^3 p}{n}}-\kappa_j+(1+\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}})\gamma_{\kappa_j}-C_mn\sum_{k\in l_j}\rho_j^{(k)2}\\
&-2\sum_{k\in l_j}\check{H}_j^{(k)}\frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}}+2C_H\sqrt{\frac{\log^2 p}{n}}\sum_{k\in l_j}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}})+O(p^{-L-1}).
\end{aligned}\tag{C.1.9}$$

The inequality on the third line is due to assumption (C1) that $\min_{j,k}\theta_j^{(k)}\geq\tau_0>0$ and Lemma 1 (ii) with $M=L+1$. The inequality on the fourth line follows from Lemma 3 (both equations) and $\min_{j,k}(\theta_j^{(k)})^{-1}:=C_m>0$, guaranteed by the sub-Gaussian assumption in assumption (C1).

We can upper bound the term $2C_H\sqrt{\frac{\log^2 p}{n}}\sum_{k\in l_j}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}}$ in (C.1.9) as follow,

$$2C_H\sqrt{\frac{\log^2 p}{n}}\sum_{k\in l_j}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}}\leq 2C_H\sqrt{\frac{\log^2 p}{n}}\frac{\sqrt{n}}{\sqrt{\tau_0}}\sqrt{\kappa_j}\sqrt{\sum_{k\in l_j}\rho_j^{(k)2}}=o(\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}}).\tag{C.1.10}$$

The first inequality is by the Cauchy-Schwarz inequality and assumption (C1), and the second equality by the assumption (C2) that $\kappa_j\log^2 p=o(n)$.

We next upper bound the term $-2\sum_{k\in l_j}\check{H}_j^{(k)}\frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}}$ with high probability. Note that $\theta_j^{(k)}$ is bounded below and above, i.e., $\tau_0\leq\theta_j^{(k)}\leq C_m^{-1}$ by assumption (C1). In addition, $\check{H}_j^{(k)}$ has zero mean and is sub-exponential with bounded constants by assumption (C1). By Bernstein inequality (Proposition 5.16 in Vershynin (2010)), we have with some constant $c'>0$,

$$P(|2\sum_{k\in l_j}|\check{H}_j^{(k)}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}}|>t)\leq 2\exp(-c'\min[\frac{t^2}{n\sum_{k\in l_j}\rho_j^{(k)2}}],\frac{t}{\max_{k\in l_j}\sqrt{n}|\rho_j^{(k)}|}).$$

We pick $t = C_B\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}\log^2 p}$ with a large constant C_B in the inequality above and apply (C.1.10) to reduce (C.1.9) as follows,

$$\begin{aligned}
&P(\sum_{k\in l_j}\hat{T}_j^{(k)2} < \gamma_{\kappa_j})\\
\le& P(\sum_{k\in l_j}(\check{H}_j^{(k)2}-1) \le -C_m n\sum_{k\in l_j}\rho_j^{(k)2} + 2C_B\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}\log^2 p}\\
&+2C_4\sqrt{\kappa_j\log p} + 2C_4\log^2 p) + O(p^{-L-1})\\
\le& P(\sum_{k\in l_j}(\check{H}_j^{(k)2}-1) \le -C_m C_2(\log^2 p + \sqrt{\kappa_j\log p}) + 2C_B\sqrt{C_2\log^2 p(\log^2 p + \sqrt{\kappa_j\log p})}\\
&+2C_4\sqrt{\kappa_j\log p} + 2C_4\log^2 p) + O(p^{-L-1})\\
\le& P(\sum_{k\in l_j}(\check{H}_j^{(k)2}-1) \le -C_2'(\log^2 p + \sqrt{\kappa_j\log p})) + O(p^{-L-1})\\
=&O(p^{-L-1}).
\end{aligned}$$

The inequality on the first line is obtained by the choice of γ_{κ_j} with the chosen $C_4 > 0$ and the assumption (C2) that $\kappa_j\log^2 p = o(n)$. The inequalities on the second line and third line are by the assumption (C4) that $\sum_{k\in l_j}|\rho_j^{(k)}|^2 \ge \frac{C_2(\log^2 p+\sqrt{\kappa_j\log p})}{n}$ for a sufficiently large $C_2 > 0$. The last equality is by Lemma 2.

This completes the proof of (C.1.7) and (C.1.8), which further yields to

$$P\{(\cup_{j\in\mathcal{A}^{[0]}}E_j^{\mathcal{A}^{[0]},2})\bigcup(\cup_{j\in\mathcal{A}^{[1]}}E_j^{\mathcal{A}^{[1]},2})\} = O(p^{-L}),$$

with the results from Theorem 1. Therefore we complete the proof of Theorem 2.

□

BIBLIOGRAPHY

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of rna sequencing data using r and bioconductor. *Nature protocols*, 8(9):1765–1786.

Argyriou, A., Evgeniou, T., and Pontil, M. (2007). Multi-task feature learning. In *Advances in neural information processing systems*, pages 41–48.

Asano, Y., Kawase, T., Okabe, A., Tsutsumi, S., Ichikawa, H., Tatebe, S., Kitabayashi, I., Tashiro, F., Namiki, H., and Kondo, T. (2016). Ier5 generates a novel hypo-phosphorylated active form of hsf1 and contributes to tumorigenesis. *Scientific reports*, 6.

Balasubramanian, K., Sriperumbudur, B., and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In *Artificial Intelligence and Statistics*, pages 126–134.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607.

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

Bird, A. (2002). Dna methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.

Bradburn, M. J., Deeks, J. J., Berlin, J. A., and Russell Localio, A. (2007). Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Statistics in medicine*, 26(1):53–77.

Bradford, J. R., Hey, Y., Yates, T., Li, Y., Pepper, S. D., and Miller, C. J. (2010). A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1):1.

Brakebusch, C., Bouvard, D., Stanchi, F., Sakai, T., and Fassler, R. (2002). Integrins in invasive growth. *The Journal of clinical investigation*, 109(8):999.

Bühlmann, P., Kalisch, M., and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the pc-simple algorithm. *Biometrika*, 97(2):261–278.

Bumgarner, R. (2013). Overview of dna microarrays: types, applications, and their future. *Current protocols in molecular biology*, pages 22–1.

Cai, T. T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association*, 111(513):229–240.

Chang, J., Tang, C. Y., and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Annals of statistics*, 41(4).

Chang, J., Tang, C. Y., and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Annals of statistics*, 44(2):515.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(suppl 1):i84–i90.

Chung, L. M., Ferguson, J. P., Zheng, W., Qian, F., Bruno, V., Montgomery, R. R., and Zhao, H. (2013). Differential expression analysis for paired rna-seq data. *BMC bioinformatics*, 14(1):110.

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519.

Conlon, E. M., Song, J. J., and Liu, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics*, 7(1):247.

Consortium, S.-I. et al. (2014). A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903–914.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352.

Do, K.-A., Müller, P., and Vannucci, M. (2006). *Bayesian inference for gene expression and proteomics*. Cambridge University Press.

Draghici, S., Khatri, P., Eklund, A. C., and Szallasi, Z. (2006). Reliability and reproducibility issues in dna microarray measurements. *TRENDS in Genetics*, 22(2):101–109.

Duffy, M. J., McGowan, P. M., Harbeck, N., Thomssen, C., and Schmitt, M. (2014). upa and pai-1 as biomarkers in breast cancer: validated for clinical use in level-of-evidence-1 studies. *Breast Cancer Research*, 16(4):1.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101.

Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10(Sep):2013–2038.

Fan, J., Song, R., et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24(1983):287–302.

Fisher, R. A. (1925). *Statistical methods for research workers.* Genesis Publishing Pvt Ltd.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741.

Genovese, C. R., Jin, J., Wasserman, L., and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13(Jun):2107–2143.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., Troakes, C., Turecki, G., ODonovan, M. C., Schalkwyk, L. C., et al. (2016). Methylation quantitative trait loci in the developing brain and their enrichment in schizophrenia-associated genomic regions. *Nature neuroscience*, 19(1):48.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):422.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.

Hochberg, Y. and Tamhane, A. C. (2009). Multiple comparison procedures.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). Rankprod: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4).

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

Huo, Z. and Tseng, G. C. (2017). Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery. *The Annals of Applied Statistics*, In press.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, pages 730–773.

Itoh, M., Iwamoto, T., Matsuoka, J., Nogami, T., Motoki, T., Shien, T., Taira, N., Niikura, N., Hayashi, N., Ohtani, S., et al. (2014). Estrogen receptor (er) mrna expression and molecular subtype distribution in er-negative/progesterone receptor-positive breast cancers. *Breast cancer research and treatment*, 143(2):403–409.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th annual international conference on machine learning*, pages 457–464. ACM.

Jiang, J., Li, C., Paul, D., Yang, C., Zhao, H., et al. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics*, 44(5):2127–2160.

Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and'race'. *Nature genetics*, 36:S28–S33.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). Metaqc: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic acids research*, 40(2):e15–e15.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375.

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2017). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics*.

Lappalainen, T., Sammeth, M., Friedländer, M. R., ACt Hoen, P., Monlong, J., Rivas, M. A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506.

Lastraioli, E., Iorio, J., and Arcangeli, A. (2015). Ion channel expression as promising cancer biomarker. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(10):2685–2702.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., Liau, L. M., Horvath, S., Mischel, P. S., and Nelson, S. F. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC medical genomics*, 1(1):52.

Leek, J. T. (2014). svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic acids research*, page gku864.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Lesk, A. (2017). *Introduction to genomics*. Oxford University Press.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. (2006). Bayesian modeling of differential gene expression. *Biometrics*, 62(1):10–18.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2010). The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523.

Li, J., Tseng, G. C., et al. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.

Li, M. D., Cao, J., Wang, S., Wang, J., Sarkar, S., Vigorito, M., Ma, J. Z., and Chang, S. L. (2013). Transcriptome sequencing of gene expression in the brain of the hiv-1 transgenic rat. *PloS one*, 8(3):e59582.

Li, R., Liu, J., and Lou, L. (2017). Variable selection via partial correlation. *Statistica Sinica*, 27(3):983.

Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139.

Liang, F., Song, Q., and Qiu, P. (2015). An equivalent measure of partial correlation coefficients for high-dimensional gaussian graphical models. *Journal of the American Statistical Association*, 110(511):1248–1265.

Liu, C.-G., Wang, J.-L., Li, L., and Wang, P.-C. (2014). Microrna-384 regulates both amyloid precursor protein and β-secretase expression and is a potential biomarker for alzheimer's disease. *International journal of molecular medicine*, 34(1):160–166.

Liu, Q. and Markatou, M. (2016). Evaluation of methods in removing batch effects on rna-seq data. *Infectious Diseases and Translational Medicine*, 2(1):3–9.

Liu, Y.-R., Jiang, Y.-Z., Xu, X.-E., Hu, X., Yu, K.-D., and Shao, Z.-M. (2016). Comprehensive transcriptome profiling reveals multigene signatures in triple-negative breast cancer. *Clinical cancer research*, 22(7):1653–1662.

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

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1.

Lu, X., Lu, X., Wang, Z. C., Iglehart, J. D., Zhang, X., and Richardson, A. L. (2008). Predicting features of breast cancer with gene expression patterns. *Breast cancer research and treatment*, 108(2):191–201.

Luo, S., Song, R., and Witten, D. (2014). Sure screening for gaussian graphical models. *arXiv preprint arXiv:1407.7819*.

Ma, S., Li, R., and Tsai, C.-L. (2017a). Variable screening via quantile partial correlation. *Journal of the American Statistical Association*, pages 1–14.

Ma, T., Liang, F., Oesterreich, S., and Tseng, G. C. (2017b). A joint bayesian model for integrating microarray and rna sequencing transcriptomic data. *Journal of Computational Biology*.

Ma, T., Liang, F., and Tseng, G. C. (2017c). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(4):847–867.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517.

McCarroll, S. A. and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics*, 39(7s):S37.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9(5):356.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, 18(9):1194–1206.

Medvedovic, M., Yeung, K. Y., and Bumgarner, R. E. (2004). Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Metzger-Filho, O., Michiels, S., Bertucci, F., Catteau, A., Salgado, R., Galant, C., Fumagalli, D., Singhal, S., Desmedt, C., Ignatiadis, M., et al. (2013). Genomic grade adds prognostic value in invasive lobular carcinoma. *Annals of oncology*, 24(2):377–384.

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nature reviews. Genetics*, 11(1):31.

Milde-Langosch, K., Kappes, H., Riethdorf, S., Löning, T., and Bamberger, A.-M. (2003). Fosb is highly expressed in normal mammary epithelia, but down-regulated in poorly differentiated breast carcinomas. *Breast cancer research and treatment*, 77(3):265–275.

Milioli, H. H., Vimieiro, R., Riveros, C., Tishchenko, I., Berretta, R., and Moscato, P. (2015). The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabric data set. *PloS one*, 10(7):e0129711.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–489.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Nakahama, T., Hanieh, H., Nguyen, N. T., Chinen, I., Ripley, B., Millrine, D., Lee, S., Nyati, K. K., Dubey, P. K., Chowdhury, K., et al. (2013). Aryl hydrocarbon receptor-mediated induction of the microrna-132/212 cluster promotes interleukin-17–producing t-helper cell differentiation. *Proceedings of the National Academy of Sciences*, 110(29):11964–11969.

Nardi, Y., Rinaldo, A., et al. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Network, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumors. *Nature*, 490(7418):61.

Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609.

Network, C. G. A. R. et al. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From rna-seq reads to differential expression results. *Genome biology*, 11(12):1.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4(1):14.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). Methylsig: a whole genome dna methylation analysis pipeline. *Bioinformatics*, page btu339.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Rajalingam, K., Schreck, R., Rapp, U. R., and Albert, S. (2007). Ras oncogenes and their downstream targets. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1773(8):1177–1195.

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

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biol*, 14(9):R95.

Rasmussen, C. E., De la Cruz, B. J., Ghahramani, Z., and Wild, D. L. (2009). Modeling and visualizing uncertainty in gene expression clusters using dirichlet process mixtures. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 6(4):615–628.

Rau, A., Marot, G., and Jaffrézic, F. (2014). Differential meta-analysis of rna-seq data from multiple studies. *BMC bioinformatics*, 15(1):91.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature reviews. Genetics*, 16(2):85.

Robinson, D. G., Wang, J. Y., and Storey, J. D. (2015). A nested parallel experiment demonstrates differences in intensity-dependence between rna-seq and microarrays. *Nucleic acids research*, 43(20):e131–e131.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Sabatier, R., Finetti, P., Cervera, N., Lambaudie, E., Esterni, B., Mamessier, E., Tallet, A., Chabannon, C., Extra, J.-M., Jacquemier, J., et al. (2011). A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast cancer research and treatment*, 126(2):407–420.

Scharpf, R. B., Tjelmeland, H., Parmigiani, G., and Nobel, A. B. (2009). A bayesian model for cross-study differential gene expression. *Journal of the American Statistical Association*, 104(488).

Schübeler, D. (2015). Function and information content of dna methylation. *Nature*, 517(7534):321.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H., George, E., and McCulloch, R. (2013). Bayes and big data: The consensus monte carlo algorithm. In *EFaBBayes 250 conference*, volume 16.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., et al. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3):812–819.

Shah, S., Smith, C., Lampe, F., Youle, M., Johnson, M., Phillips, A., and Sabin, C. (2007). Haemoglobin and albumin as markers of hiv disease progression in the highly active antiretrovial therapy era: relationships with gender*. *HIV medicine*, 8(1):38–45.

Shao, Q.-M. (1999). A cramér type large deviation result for student's t-statistic. *Journal of Theoretical Probability*, 12(2):385–398.

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

Sims, A. H., Smethurst, G. J., Hey, Y., Okoniewski, M. J., Pepper, S. D., Howell, A., Miller, C. J., and Clarke, R. B. (2008). The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. *BMC medical genomics*, 1(1):42.

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

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):91.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The american soldier: adjustment during army life.(studies in social psychology in world war ii, vol. 1.).

Su, Z., Li, Z., Chen, T., Li, Q.-Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R. G., et al. (2011). Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chemical research in toxicology*, 24(9):1486–1493.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., et al. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960.

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

Terenin, A., Simpson, D., and Draper, D. (2015). Asynchronous distributed gibbs sampling. *arXiv preprint arXiv:1509.08999*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785–3799.

Tseng, G. C. and Wong, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16.

Tsuyuzaki, K. and Nikaido, I. (2013). metaseq: Meta-analysis of rna-seq count data.

Turashvili, G., Bouchal, J., Baumforth, K., Wei, W., Dziechciarkova, M., Ehrmann, J., Klein, J., Fridman, E., Skarda, J., Srovnal, J., et al. (2007). Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC cancer*, 7(1):1.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

Van De Wiel, M. A., Leday, G. G., Pardo, L., Rue, H., Van Der Vaart, A. W., and Van Wieringen, W. N. (2012). Bayesian analysis of rna sequencing data by estimating multiple shrinkage priors. *Biostatistics*, page kxs031.

Veeriah, S., Brennan, C., Meng, S., Singh, B., Fagin, J. A., Solit, D. B., Paty, P. B., Rohle, D., Vivanco, I., Chmielecki, J., et al. (2009). The tyrosine phosphatase ptprd is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proceedings of the National Academy of Sciences*, 106(23):9435–9440.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

Wang, C., Gong, B., Bushel, P. R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., et al. (2014). A comprehensive study design reveals treatment-and transcript abundance–dependent concordance between rna-seq and microarray data. *Nature biotechnology*, 32(9):926.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G., and Do, K.-A. (2012). ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159.

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

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–243.

Xiong, Y., Chen, X., Chen, Z., Wang, X., Shi, S., Wang, X., Zhang, J., and He, X. (2010). Rna sequencing shows no dosage compensation of the active x-chromosome. *Nature genetics*, 42(12):1043–1047.

Xu, J., Su, Z., Hong, H., Thierry-Mieg, J., Thierry-Mieg, D., Kreil, D. P., Mason, C. E., Tong, W., and Shi, L. (2013). Cross-platform ultradeep transcriptomic profiling of human reference rna samples by rna-seq. *Scientific data*, 1:140020–140020.

Yadav, B. S., Chanana, P., and Jhamb, S. (2015). Biomarkers in triple negative breast cancer: A review. *World journal of clinical oncology*, 6(6):252.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, W., Zhu, J., Schadt, E. E., and Liu, J. S. (2010). A bayesian partition method for detecting pleiotropic and epistatic eqtl modules. *PLoS computational biology*, 6(1):e1000642.

Zheng, W., Chung, L. M., and Zhao, H. (2011). Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, 12(1):1.

Zhou, J., Liu, J., Narayan, V. A., and Ye, J. (2012a). Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM.

Zhou, M., Li, L., Dunson, D., and Carin, L. (2012b). Lognormal and gamma mixed negative binomial regression. In *Machine learning: proceedings of the International Conference. International Conference on Machine Learning*, volume 2012, page 1343. NIH Public Access.

Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.