Computational Modeling for Differential Analysis of RNA-seq and Methylation Data

Xiao Wang

Dissertation submitted to the faculty of Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Electrical and Computer Engineering

Jianhua J. Xuan
Yue J. Wang
Amos L. Abbott
Dong S. Ha
Wenjing Lou

August 2nd, 2016
Arlington, VA

Keywords: Differential Analysis, Bayesian Modeling, Markov Random Field, RNA-seq Data Analysis, Markov Chain Monte Carlo (MCMC)

Computational Modeling for Differential Analysis of RNA-seq and Methylation Data

Xiao Wang

ABSTRACT

Computational systems biology is an inter-disciplinary field that aims to develop computational approaches for a system-level understanding of biological systems. Advances in high-throughput biotechnology offer broad scope and high resolution in multiple disciplines. However, it is still a major challenge to extract biologically meaningful information from the overwhelming amount of data generated from biological systems. Effective computational approaches are of pressing need to reveal the functional components. Thus, in this dissertation work, we aim to develop computational approaches for differential analysis of RNA-seq and methylation data to detect aberrant events associated with cancers.

We develop a novel Bayesian approach, BayesIso, to identify differentially expressed isoforms from RNA-seq data. BayesIso features a joint model of the variability of RNA-seq data and the differential state of isoforms. BayesIso can not only account for the variability of RNA-seq data but also combines the differential states of isoforms as hidden variables for differential analysis. The differential states of isoforms are estimated jointly with other model parameters through a sampling process, providing an improved performance in detecting isoforms of less differentially expressed.

We propose to develop a novel probabilistic approach, DM-BLD, in a Bayesian framework to identify differentially methylated genes. The DM-BLD approach features a hierarchical model, built upon Markov random field models, to capture both the local dependency of measured loci and the dependency of methylation change. A Gibbs sampling procedure is designed to estimate the posterior distribution of the methylation change of CpG sites. Then, the differential methylation score of a gene is calculated from the estimated methylation changes of the involved CpG sites and the significance of genes is assessed by permutation-based statistical tests.

We have demonstrated the advantage of the proposed Bayesian approaches over conventional methods for differential analysis of RNA-seq data and methylation data. The joint estimation of the posterior distributions of the variables and model parameters using sampling procedure has demonstrated the advantage in detecting isoforms or methylated genes of less differential. The applications to breast cancer data shed light on understanding the molecular mechanisms underlying breast cancer recurrence, aiming to identify new molecular targets for breast cancer treatment.

Computational Modeling for Differential Analysis of RNA-seq and Methylation Data

Xiao Wang

GENERAL AUDIENCE ABSTRACT

Computational systems biology is an inter-disciplinary field that aims to develop computational approaches for a system-level understanding of biological systems. Advances in high-throughput biotechnology offer broad scope and high resolution in multiple disciplines. With the accumulation of various kinds of omics data, differential analysis is a very powerful and widely used approach, particularly in cancer research, to identify biomarkers by comparing molecular datasets of different phenotypes. The identification of aberrant events from differential analysis of the omics data can help diagnosis and make prognostic decisions to prevent the development of disease. In this dissertation research, we develop two novel Bayesian approaches to detect differentially expressed transcripts and differentially methylated genes. Variables and model parameters are jointly estimated by MCMC sampling procedures. We demonstrate the advantage of the proposed methods by comprehensive simulation studies and real data studies with benchmarks. We apply the proposed method to breast cancer recurrence studies The identified differential genes shed light on understanding the molecular mechanisms underlying breast cancer recurrence, aiming to identify new molecular targets for breast cancer treatment.

Acknowledgement

I would like to take this opportunity to acknowledge all of the people who give me their kind help and support in completing this dissertation.

First and foremost, I want to thank my advisor, Dr. Jason J. Xuan. He brought me to the new field of computational biology and provided me with professional and insightful guidance on my research. He encouraged me to be positive when I encountered difficulties in research and life; and helped me greatly in my paper writing and the dissertation. He is the mentor in both my research and life.

I would like to thank my committee members, Dr. Yue J. Wang, Dr. A. Lynn Abbott, Dr. Dong S. Ha, and Dr. Wenjing Lou, for their valuable comments on my preliminary examination and suggestion in my presentation. I would also thank my collaborators, Dr. Leena Hilakivi-Clarke, Dr. Ayesha N. Shajahan-Haq, and Dr. Robert Clarke at Georgetown University, Dr. Tian-Li Wang, Dr. Ie-Ming Shih at John Hopkins Medical Institute. They helped me identifying the biological problems and provided me support on biological interpretation for the publications. I thank Dr. Jinghua Gu for his help in my research and life. I also enjoyed the time working with all of my colleagues and lab mates in CBIL.

Finally, special thanks to my parents, my younger sister and my best friends, Dr. Xin Lin, Dr. Xun Yan, Dr. Shuping Jiang, and Dr. Qian Li. I would like to dedicate this dissertation to them for their love, encouragement, and support in my research and life.

Table of Contents

List of figures

List of tables

List of Abbreviations

AUC	Area-under-the-curve
AUCpr	area-under-precision-recall curve
BRCA	breast cancer
CAR	conditional autoregressive
cDNA	complementary DNA
ChIP-seq	Chromatin immunoprecipitation (ChIP) with massively parallel sequencing
CpG	-C-phosphate-G-
DAVID	the Database for Annotation, Visualization and Integrated Discovery
DDI	protein Domain-Domain interaction
DM	differential methylation
DMR	differentially methylated region
DMRF	discrete Markov random field
E2	estrogen
ENC	external_normalized_count
ER+	estrogen receptor positive
ERCC	External RNA Control Consortium
FPKM	fragments-per-kilobase-per-million
GMRF	Gaussian Markov random field
hbr	human brain reference RNA
HPRD	the Human Protein Reference Database
ICAR	intrinsic conditional autoregressive
Illumina 450k	the Illumina Infinium HumanMethylation450 BeadChip Kit
IPA	Ingenuity Pathway Analysis
K-S test	Kolmogorov–Smirnov test
$\|\log_2 FC\|$	the absolute log2 ratio of fold change
LS	least squares
MAQC	the MicroArray Quality Control Project

MCMC	Markov Chain Monte Carlo
M-H sampling	Metropolis-Hasting sampling
MRF	Markov random field
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
PPI	Protein-Protein interaction
qRT-PCR	quantitative real-time reverse transcription polymerase chain reaction
RNA-seq	RNA sequencing
ROC curve	receiver operating characteristic curve
RPKM	reads-per-kilobase-per-million
RPPA	reverse phase protein array
SEQC	Sequencing Quality Control
SNP	single-nucleotide polymorphism
SNR	signal-to-noise-ratio
SRA	sequence read archive
TCGA	The Cancer Genome Atlas project
TMM	trimmed mean of M-values
uhr	human reference RNA
UQUA	upper quantile normalization

1 Introduction

1.1 Motivation and background

As an inter-disciplinary field, systems biology [1-3] focuses on studying complex interactions in biological systems, aiming to address the pluralism of causes and effects in biological networks. Systems biology can be interpreted as the process to obtain quantitative measurements of various components of biological systems, and then, analyze and integrate the complex datasets for the identification of molecular modules, networks, and pathways. The development of biotechnologies offers unprecedented scope and resolution for systems biology research. Advanced high-throughput experiments provide large quantities of high-quality data from multiple study fields. However, due to the complexity of biological systems and the overwhelming amount of data, how to excavate biologically meaningful information from these multiple data sets is still a huge challenge. Data analysis and integration require more realistic and advanced mathematical and computational models, which fall under the remit of computational systems biology. Computational systems biology [4] aims to develop and use efficient algorithms, data structure, visualization, and communication tools for a system-level understanding of the underlying biological mechanisms. Effective and efficient computational and machine learning techniques are in high demand to reveal the function of the components of biological systems as well as their interactions.

Biological systems are complex systems, in which a large number of functionally diverse sets of components interact selectively and nonlinearly to produce coherent behaviors [4]. Different types of data that are acquired from multiple studies can reveal different aspects of a biology system. In the field of molecular biology, studies mainly focus on three levels: DNA, RNA, and protein. There is an exponential increase in data from multiple disciplines, such as genomics, epigenomics, transcriptomics, interactomics, proteomics, etc., as shown in Figure 1.1. Genomics and epigenomics are on the DNA level. Genomics focuses on the complete set of DNA of an organism [5]. The three main components of genomic study are sequencing the entire

DNA, assembling the sequence to reconstruct the original genome, and generating biological information by analyzing the function and structure of the genome [6, 7]. DNA sequencing makes it possible to identify single-nucleotide polymorphism (SNP) [8, 9], copy number variations [10, 11], or other structure variations [12, 13] associated with diseases in a high-throughput manner. Epigenomics [14] is the study of epigenetic modifications on the epigenome, such as DNA methylation [15] and histone modification [16], which plays an essential role in mRNA transcription (gene expression) and regulation and has been revealed to be associated with numerous cellular processes such as differentiation/development [17, 18] and tumorigenesis [19, 20]. Transcriptomics focuses on RNA molecules, where gene expression can be measured by microarray assays [21] or more advanced sequencing techniques [22]. Gene expression

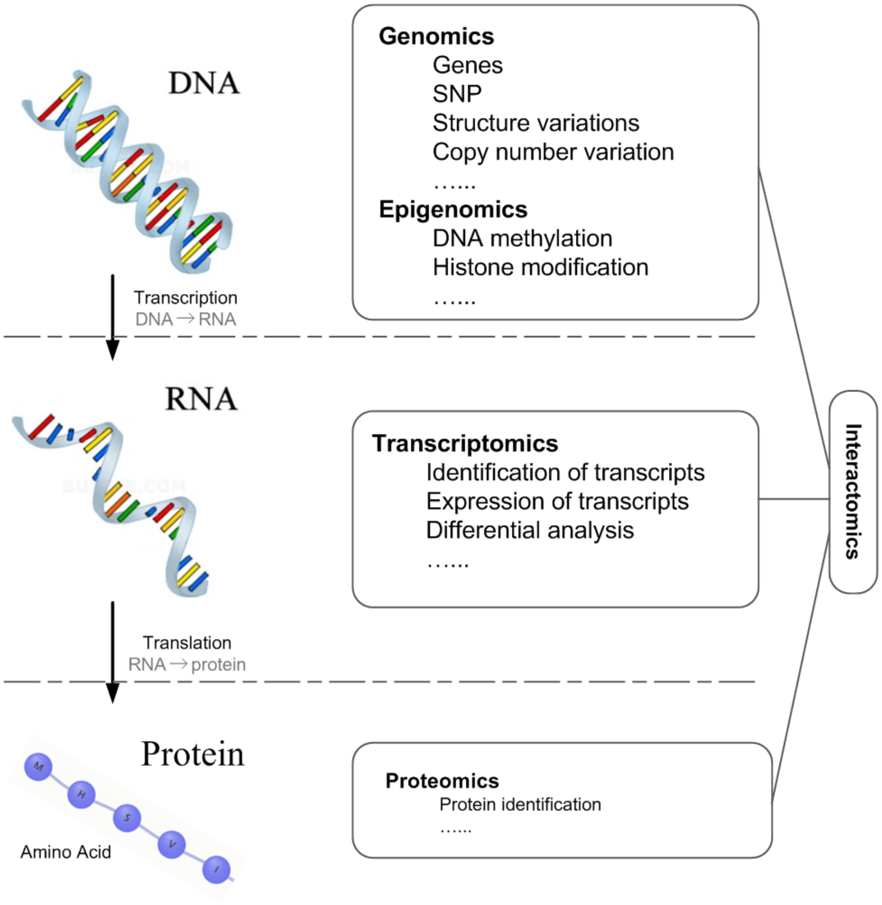


Figure 1.1 Multiple study fields for molecular biology study.

profiling can be used to explore how transcript patterns are affected by various developmental aspects of tissues, diseases, or environmental factors such as drugs, pollutants, etc. Proteomics [23] is the next step of genomics and transcriptomics, which studies the functionality and structure of proteins. It focuses on post-translational modification such as phosphorylation, ubiquitination, etc. Interactomics is a study of the interactions between molecules, where protein-protein interaction (PPI) is currently the authoritative molecular discipline in the field. The PPI data provide the information of the cause/effect/binding relationship between proteins, which facilitates our understanding on how the proteins function and how they interact with others. PPI data for human can be downloaded from multiple databases [24, 25]. Comprehensive and integrative analysis of the various omics data can deepen our understanding of the underlying mechanisms of biological systems.

With the accumulation of various kinds of omics data, differential analysis is a very powerful and widely used approach, particularly in cancer research, which identifies biomarkers by comparing molecular datasets of different phenotypes. Disease biomarkers identified from differential analysis comparing healthy subjects with diseased subjects or comparing different phenotypes of diseased subjects can be used to understand the pathogenic process. Thus, the identification of aberrant events from differential analysis of the omics data can help diagnosis and make prognostic decisions to prevent the development of disease. Among all these disciplines associated with computational systems biology, transcriptomics and epigenomics are two study fields of critical importance for biomarker identification. RNA transcriptions and DNA methylation play crucial roles in biological processes, and the abnormality of these two molecular processes may result in the disorder of a biological system. The anomalous expression of many transcripts has been demonstrated to be implicated in disease development; the abnormality of DNA methylation has been revealed to play an important role in regulating gene expression. Thus, the abnormality of DNA methylation can be involved in the molecular mechanism of disease development through inducing the anomaly of transcript expression. Figure 1.2 presents an illustration of DNA methylation, transcription, as well as their relation.

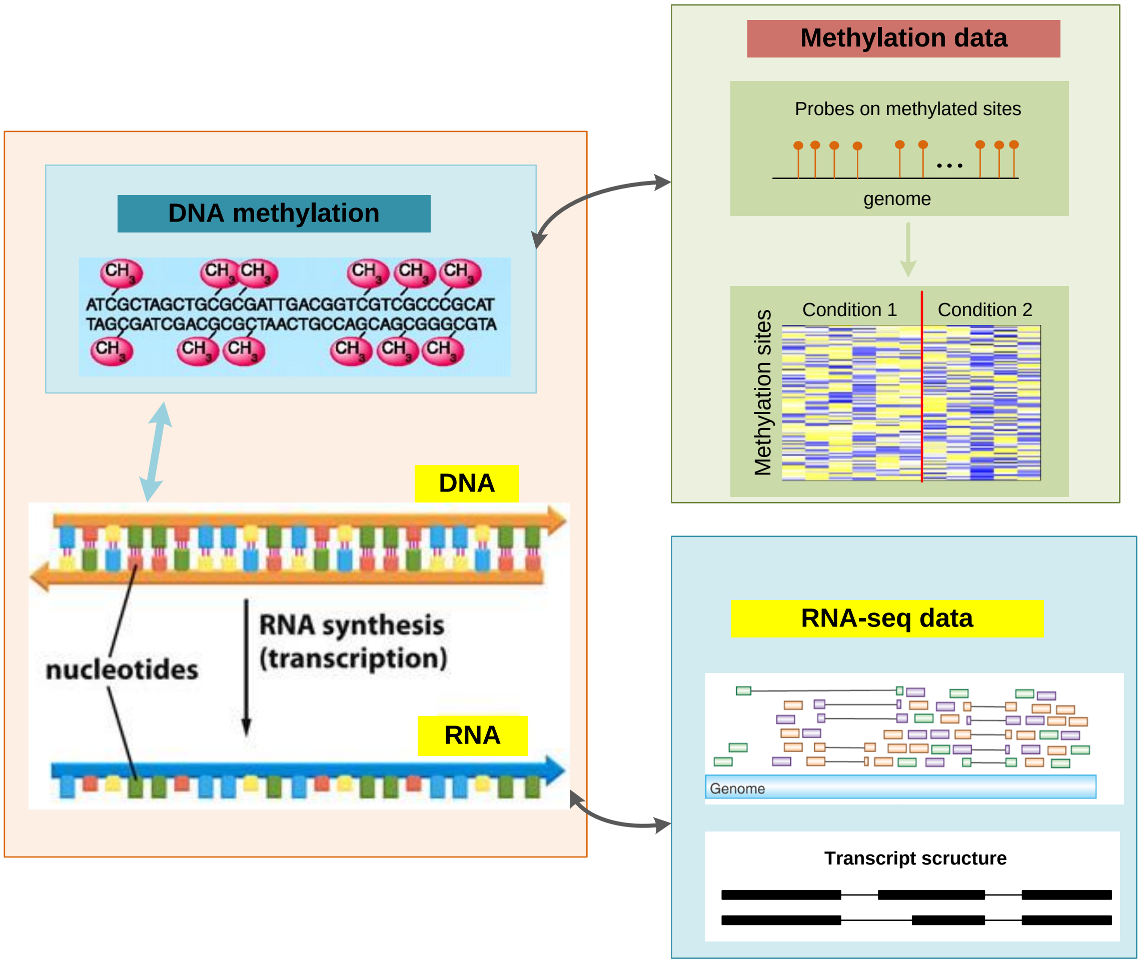


Figure 1.2 An illustration of the relation between DNA methylation and transcription.

Understanding the transcriptome is essential for revealing how the functional elements of the genome affect cell development and cause disease. In the past decades, DNA microarray was a dominant technique to measure the relative abundance of gene expression. With the gene expression profiling, researchers focused on studying gene expression pattern changes associated with different phenotypes such as clinical disease subtypes. The development of next-generation sequencing (NGS) opens a new era of transcriptomics study. RNA sequencing (RNA-seq) is a revolutionary approach to transcriptome profiling using deep-sequencing technique [22]. An illustration of RNA-seq data is shown in Figure 1.2, where the existence and expression levels of the transcripts can be estimated from millions of sequencing reads. Compared with DNA microarrays, RNA-seq provides the measurements in much larger dynamic range with much higher accuracy and much lower background noise. In addition to studying gene expression changes, RNA-seq, by virtue of its single-base resolution, makes it capable of investigating new biological problems, such as alternative splicing, differential isoforms, gene fusion, etc. While the high-throughput sequencing data offers wide coverage and high resolution, there are many nontrivial challenges for transcriptome analysis including the high variability of RNA-seq data produced between different runs as well as complicate biases introduced by library preparation protocols, sequencing platforms, nucleotide compositions and so on.

DNA methylation [15], a molecular modification of DNA, is a stable, heritable, and also reversible process that plays a crucial role in epigenetic regulation of gene expression without altering DNA primary structure. As an epigenetic mark, DNA methylation is an important component in various biological processes [17], such as cell division, stem cell differentiation, etc. Besides its crucial function in normal cell development, DNA methylation has also been demonstrated to be associated with many diseases [26] including cancer [19, 20, 27]. Aberrant DNA methylation of specific genes may lead to aberrant activation of growth-promoting genes and aberrant silencing of tumor-suppressor genes [28]. It has been firmly established that hypermethylation of tumor suppressor genes is one of the most common mechanisms for gene regulation in cancer [29, 30]. The development of high-throughput technologies provides the opportunity to obtain genome-wide DNA methylation mapping with a very high resolution. Illumina’s methylation arrays are widely used for measuring the status of methylation sites due to they're high-quality, relatively low-cost techniques and require a small number of samples. The functional regions of most of the genes are covered by multiple probes, as shown in Figure 1.2. The high coverage of the Illumina’s methylation array [31, 32] makes it a very powerful platform for exploring genome-wide DNA methylation landscape. Despite the advantage of high-throughput profiling, the high resolution poses challenges to computational analysis for differentially methylated gene detection from the huge amount of measured CpG (shorted for ‘-C-phosphate-G-’) sites. The characteristic and variability of the methylation level of the measured CpG sites need to be well modeled for accurate differential analysis.

Differential analysis of transcript expression and DNA methylation can help reveal the underlying aberrant events of biological systems, especially in the field of cancer research. There is strong evidence suggesting that DNA methylation affects biological systems via altering gene expression [33]. Considering the complexity of high-throughput RNA-seq and methylation data, it is an important yet challenging problem to detect underlying genetic and epigenetic events and uncover their relationship.

1.2 Objectives and problem statement

In this dissertation research, we mainly focus on differential analysis of RNA-seq data and methylation data in the disciplines of transcriptomics and epigenomics. We propose two novel computational methods to model these high-throughput biological datasets for differential analysis, in order to uncover the underlying molecular mechanisms associated with cancer. Specifically, the major focuses of this dissertation work are: (1) to develop a novel joint model that accounts for the variability of RNA-seq data to identify differentially expressed genes/isoforms in a Bayesian framework; (2) to develop a hierarchical Bayesian model that exploits local dependency for the detection of differentially methylated gene from high-throughput methylation data.

1.2.1 Identification of differentially expressed isoforms from RNA-seq data

Next generation sequencing technology has opened a new era for transcriptome analysis. The advent of rapid sequencing technologies along with reduced costs makes RNA-seq become the standard method for measuring RNA expression levels, particularly for cancer research. RNA-seq has clear advantages over traditional array-based techniques. As illustrated in Figure 1.2, by piling up millions of sequencing reads along the reference genome, the expression level of RNAs can be obtained in a much larger dynamic range with much higher accuracy. Moreover, RNA-seq technology makes it possible to identify the expressed isoforms of the genes, to detect differentially expressed isoforms, etc.

One of the main goals in RNA-seq experiments is to detect differentially expressed transcripts by comparing replicates of different phenotypes. As a high-throughput technology with large dynamic range and high accuracy, RNA-seq technology facilitates the detection of differentially expressed transcripts between different phenotypes, yet posing nontrivial challenges due to the high variability of sequencing data. Several computational tools for differential analysis of RNA-seq data have been developed in the past few years, which fall into two categories: count-based approaches and isoform-level approaches. Count-based differential analysis methods are initially developed for gene-level differential analysis, in which the input is a count matrix where each element is the number of reads assigned to a gene in a sample. Initial count-based approaches model the observed read counts using Poisson distribution based on the uniform assumption of read distribution. However, it is observed that the biological variability across samples cannot be well approximated by Poisson distribution, i.e., the variance is much larger than the mean, termed over-dispersion. EdgeR [34] and DESeq [35] use negative binomial distribution instead of Poisson distribution to address the problem of over-dispersion among samples in a phenotype group. The dispersion parameters are estimated by borrowing information across genes. DSS [36] also uses a negative binomial model but exploits an empirical Bayes shrinkage estimate of the dispersion parameters. EBSeq [37] is an empirical Bayesian approach that models the variability of read counts of genes or isoform expression, aiming to improve overall fitting of expression data. Those count-based methods cannot be directly applied to the aligned sequencing data for isoform-level differential analysis because the number of reads on each isoform (or isoform expression level) cannot be directly counted from the aligned sequencing reads due to the uncertainty of read alignment. Cuffdiff 2 [38] is one of the most popular tools for isoform-level different analysis, in which BAM files (the binary version of sequence aligned data) are used as input. A beta negative binomial distribution is used to account for the variability and read-mapping ambiguity. Ballgown [39], which works together with Cufflinks [40], has improved the detection performance because of its flexible selection of several statistical models.

These existing approaches have demonstrated their initial success in differential analysis of RNA-seq data. However, a systematic effort to address the variability in RNA-seq data is lacking. Count-based methods, with read counts of genes or estimated isoform expression levels as the input, mainly focus on modeling the variability among biological samples in the same phenotype group [41], yet miss the variability of RNA-seq data along genomic loci. To account for the variability along genomic loci, Cuffdiff 2 incorporates a fragment bias model [42] for isoform expression estimation, where positional and sequence-specific biases are estimated. However, the parameters of the bias model are estimated in a global perspective, which is insufficient to account for the complex patterns observed from data. Moreover, in the existing methods, such as Cuffdiff 2 and Ballgown, a statistic test is used as a second step (after isoform expression estimation) to detect differentially expressed isoforms. Thus, a joint model that takes into account both the variability of RNA-seq data and differential states of isoforms is needed for differential isoform identification.

In this dissertation, we propose to develop a novel Bayesian approach, namely BayesIso, to identify differentially expressed isoforms from RNA-seq data. The BayesIso approach features a novel joint model that accounts for both the variability of RNA-seq data and the differential states of isoforms. A Poisson-Lognormal regression model [43] is used to account for the variability of sequencing reads along the genomic loci of the transcripts, which is capable of estimating the over-dispersion pattern of each transcript, instead of dealing with multiples sources of technical biases and variation separately. A Gamma-Gamma model [44] is used to analyze the differential expression of isoforms while accounting for the variability of the replicates. A Markov Chain Monte Carlo (MCMC) procedure [45, 46], which is a combination of Metropolis-Hasting (M-H) sampling [47] and Gibbs sampling [48], is designed for a joint estimation of the model parameters and the differential states of isoforms.

1.2.2 Detection of differential methylation genes from methylation data

The advent of high-throughput DNA methylation profiling techniques allows for whole genome-wide epigenetic study with high resolution. As the most stable epigenetic mark, DNA methylation is widely regarded as a major mechanism for influencing patterns of gene expression, cell differentiation, and cell phenotype. DNA methylation pattern changes are pivotal marks contributing to the complexity of organisms’ cellular subtypes. In recent decades, there is strong evidence that aberrant DNA methylation can give rise to various diseases including cancer [49] and can be used for clinical outcome prediction [50]. Given the essential roles of methylation, it is of increasing interest to detect biologically meaningful methylation pattern changes that may alter gene expression and eventually cause diseases.

By virtue of the high-throughput biotechnologies, the methylation level of each gene is measured at multiple CpG sites across the genomic location of the gene, providing a more comprehensive measurement for a methylation event. However, the high resolution also poses challenges to computational analysis for the detection of differentially methylated genes from the huge number of measured CpG sites. Initial site-level differential analysis approaches by statistical tests lack statistical power due to the problem of multiple testing; moreover, the methylation change of an individual CpG site is of limited value without considering the methylation status of its neighbors. Thus, combining information from multiple neighboring CpG sites to detect differentially methylated regions (DMRs) is of prime interest, and several methods have been proposed. IMA [84] first generates an index of the methylation value of predefined regions (such as genes, promoter regions, etc.), and then uses statistical tests to identify differentially methylated regions. Bumphunter [86] first estimates the association between the methylation level and the phenotypes for each site and then identifies DMRs after a smoothing operation in a *de novo* manner without relying on the predefined regions. DMRcate [87] is another *de novo* approach, which first calculates a statistic from differential test for each site, and then detects DMRs incorporating the neighboring information via a Gaussian kernel. Comb-P [88] combines the spatially assigned p-values of each site calculated from statistical test to find regions of enrichment. Probe Lasso [89] is a window-based approach that detects DMRs using neighboring significant-signals. These region-based methods have demonstrated their advantage in detecting methylation pattern changes. However, most of the existing DMR detection methods are based on statistic tests, and the neighboring information is not jointly considered when estimating the methylation change of CpG sites.

In this dissertation, we propose a novel probabilistic approach, DM-BLD, to detect differentially methylated genes based on a Bayesian framework. The DM-BLD approach features a joint model to capture both the local dependency of measured loci and the dependency of methylation changes in samples. Specifically, the local dependency is modeled by a Gaussian Markov random field (GMRF), i.e., Leroux conditional autoregressive structure; the dependency of methylation changes is modeled by a discrete Markov random field (MRF). A hierarchical Bayesian model is developed to fully take into account the local dependency for differential analysis, in which differential states are embedded as hidden variables. The differential methylation scores of the genes are calculated from the estimated methylation changes of the involved CpG sites. Permutation-based statistical tests are designed to assess the significance of the detected differentially methylated genes.

1.3 Summary of contributions

In this dissertation study, we focus on developing novel computational approaches for differential analysis of RNA-seq data and methylation data via investigating and modeling the characteristics of the data of interest. We summarize the major contributions as follows:

(1) We develop a novel Bayesian approach (BayesIso) for differential analysis of RNA-seq data at isoform-level by joint modeling the variability of RNA-seq data with the differential state of isoforms as hidden variables. BayesIso uses a Poisson-lognormal model to model the variability of sequencing data along the genomic locus of each transcript. A unique feature is that the variability is modeled at the isoform level; the dispersion of read count on each exon is modeled using a parameter specific to each isoform. A Gamma-Gamma model is used to model the expression level of transcripts of the replicates of different phenotypes, which is capable of capturing the variability across replicates. Moreover, differential states of the transcripts are embedded into the Gamma-Gamma model as hidden variables, affecting the distribution of transcript expressions in each group or condition. An MCMC sampling procedure is used to jointly estimate the posterior probability of differential state with other model parameters capturing expression variability. BayesIso aims to improve the performance of differentially expressed isoforms detection, particularly on isoforms with moderate expression change, so as to obtain a complete understanding of the difference between different phenotypes, such as different stages of tumor samples.

(2) We develop a hierarchical Bayesian approach (DM-BLD) for differential methylation detection using a hierarchical Bayesian model exploiting the local dependency of CpG sites measured in methylation data. DM-BLD uses MRF models to account for the local dependency of methylation levels of the CpG sites and the dependency of methylation changes of nearby CpG sites. To be specific, the Leroux conditional autoregressive structure (a Gaussian MRF) is used to model the local dependency of the CpG sites; a discrete MRF is used to account for the dependency of methylation changes. The Leroux model is capable of accounting for different levels of local dependency. A hierarchical Bayesian model is built upon the two MRF models, and a Gibbs sampling procedure is developed to estimate all of the variables and model parameters jointly. With the estimates of the methylation levels of CpG sites in two phenotypes, the differential methylation scores of the genes are calculated from the estimated methylation changes of CpG sites, and permutation-based statistical tests are performed to assess the significance of the identified genes. DM-BLD is proposed to identify differentially methylated gene that involves a sequence of CpG sites with methylation change, particularly when the methylation change of the CpG sites is moderate or the variability of methylation in samples is high.

1.4 Organization of the dissertation

The major objective of this dissertation work is to develop computational methods for differential analysis of RNA-seq data and methylation data to understand aberrant transcript expression and methylation change associated with cancer development. The remainder of the dissertation is organized as follows.

In Chapter 2, BayesIso is proposed to detect differentially expressed isoforms from RNA-seq data. First, we conduct a comprehensive analysis of the variability of read count data using multiple real RNA-seq datasets and dissect the variability into two dimensions: within-sample variability and between sample variability. Then, we introduce the details of the unified Bayesian model BayesIso. To systematically assess the performance of BayesIso, we simulate multiple RNA-seq datasets with different scenarios and compare the performance of BayesIso to several existing methods. We also apply BayesIso and the competing methods onto real RNA-seq datasets with benchmarks. Simulation studies and real data studies with benchmarks demonstrate that BayesIso outperforms the other competing methods, especially on detecting isoforms that are moderately differentially expressed. Finally, we apply the BayesIso approach to breast cancer RNA-seq data to identify isoforms associated with breast cancer recurrence.

In Chapter 3, we describe our DM-BLD approach for the detection of differentially methylated genes. Based on the observation of the intrinsic local dependency among methylation sites from real methylation data sets, we propose the hierarchical Bayesian model of DM-BLD and describe the Gibbs sampling procedure for model parameter estimation. As a next step of estimating the methylation change of CpG sites, we introduce the calculation of differential methylation score of the genes as well as the significant tests used to assess the significance of the genes. To demonstrate the advantage of DM-BLD over the existing methods, we apply all of the competing methods to multiple simulation data sets and a breast cancer data set for performance comparison. Finally, we use our approach to identify a set of functional differentially methylated genes that are also differentially expressed, which may help reveal the underlying mechanism of breast cancer recurrence.

In Chapter 4, we summarize the contributions of this dissertation work, lay out future tasks for further analysis and integration of RNA-seq data and methylation data, and finally draw conclusions of this dissertation research.

2 A novel Bayesian model for differential analysis of RNA-seq data

2.1 Introduction

With the advent of next-generation sequencing technologies, RNA sequencing (RNA-seq) has become a major molecular profiling technique in the field of cancer research for transcriptome analysis [22, 51, 52]. The procedure of a typical RNA-seq experiment is as follows. First, a population of RNA is converted to a library of complementary DNA (cDNA) fragments. Then, the fragments are sequenced in a high-throughput manner to obtain short sequencing reads. Finally, the reads are either aligned to a reference genome or assembled *de novo* without the genome information to construct a whole-genome transcription map. The transcriptional structure and expression level of the transcripts can be derived from the transcription map. Compared to conventional hybridization-based microarray and Sanger sequencing-based methods, RNA-seq provides a more comprehensive understanding of transcriptomes. It allows for quantification of the transcripts in a much larger dynamic range with much higher accuracy. As the cost of sequencing techniques becomes lower, more tumor samples will likely be profiled by RNA-seq than with other current technologies.

In recent RNA-seq profiling studies, the detection of differentially expressed transcripts (or isoforms) between different types of cancers (or sub-types of cancer) has become a major task in the field of cancer research [38, 53, 54]. While RNA-seq has the advantage of wide coverage and high resolution, there are many nontrivial challenges for transcriptome analysis including the variability of RNA-seq data and the uncertainty of read assignment. Variability in RNA-seq data can arise from transcript length bias, library size bias (the total number of sequencing reads in each sample), sequencing biases (GC-content bias, random hexamer priming bias) and other sources [40, 55-57]. We have investigated the variability of real RNA-seq datasets, and dissected the variability along two dimensions: variability along the genomic region of a gene in a sample, termed within-sample variability, and variability across samples from the same biological group, termed between-sample variability. Our investigation into the variability reveals that bias patterns exist but cannot yet be fully explained by known sources; moreover, different genes/transcripts may exhibit different and complex bias patterns [58]. As many genes have multiple transcripts (isoforms), many of which share exons, some reads cannot be assigned unequivocally to a specific isoform. Thus, the uncertainty of read mapping to each transcript is an inherent problem for RNA-seq data analysis.

Current efforts on differential analysis of RNA-seq data can be divided into two categories: count-based differential analysis methods and isoform-level differential analysis methods. Count-based approaches are initially developed for differential analysis of RNA-seq data at the gene level, in which the input is a count matrix where each element is the number of reads assigned to a gene in a sample. Count-based methods mainly focus on modeling the variability among biological samples in the same phenotype group: the between-sample variability [41]. Initial count-based approaches model the observed reads using Poisson distribution based on the uniform assumption of read distribution. However, it is observed that the biological variability across samples cannot be well approximated by Poisson distribution. EdgeR [34] is the first method that models the between-sample variability by replacing Poisson model with negative binomial model, which can account for over-dispersion among samples in a phenotype group. DESeq [35] also uses negative binomial distribution and models the variance as a non-linear function of the mean/medium counts. DSS [36] presents an empirical Bayes shrinkage estimate of the dispersion parameters in the negative binomial model. EBSeq [37] is developed for the detection of differentially expressed genes/isoforms with read counts of genes or estimates of isoform expression as the input. It is an empirical Bayes approach that models the variability of the read counts of genes or the estimated isoform expression levels, aiming to improve overall fitting of count data. Those count-based methods cannot be directly applied to the aligned sequencing data for isoform-level differential analysis because the number of reads on each isoform (or isoform expression level) cannot be directly counted from the aligned sequence due to the uncertainty of read alignment. Cuffdiff 2 [38] is one of the most popular tools for differential analysis of RNA-seq data at isoform (or transcript) level. BAM files (the binary version of sequence aligned data) are used as input and a beta negative binomial distribution accounts for the between-sample variability and read-mapping ambiguity. Cuffdiff 2 is a two-step approach, which first estimates isoform expression and then detects differentially expressed isoforms with a statistical test. Cuffdiff 2 is an overly conservative method for detecting differentially expressed isoforms since it misses many differential isoforms [39]. Ballgown, a newly developed method that works together with Cufflinks [40], has improved the detection performance because of its flexible selection of several statistical models [39].

The above-mentioned approaches have demonstrated their initial success in differential analysis of RNA-seq data. However, a systematic effort to address the variability in RNA-seq data is lacking. Specifically, the within-sample variability of RNA-seq data, the variability along genomic loci, is not well modeled for the identification of differentially expressed genes/isoforms. In the count-based methods for the identification of differentially expressed genes, the overall read count of a gene is used to assess its expression level, without considering large variance of read counts among genomic loci, i.e. within-sample variability. The within-sample variability is more critical for differential analysis of RNA-seq data at isoform level. The expression of a gene consists of the expressions of multiple isoforms, increasing the complexity of bias at different genomic locations along the gene. To deal with the within-sample variability, Cuffdiff 2 incorporates a fragment bias model [42] for isoform expression estimation, where positional and sequence-specific biases are estimated. However, the parameters of the bias model are estimated in a global perspective, assuming that the positional bias of transcripts of lengths within a range is the same. This assumption, however, is insufficient to account for the complex patterns of within-sample variability observed from data. Moreover, Cuffdiff 2 uses a statistic test as its second step (after isoform expression estimation) to detect differentially expressed isoforms. Many differentially expressed isoforms may not reach statistical significance due to the huge amount of transcripts in consideration. Therefore, a joint model that takes into account both the variability of RNA-seq data and the differential states of isoforms is needed for differential isoform identification.

In this study, we develop a novel Bayesian approach, namely BayesIso, for differential analysis of RNA-seq data at the isoform level. The BayesIso approach is built upon a novel joint model that accounts for both the variability of RNA-seq data and the differential states of isoforms. Specifically, a Poisson-Lognormal regression model [43] is used to account for the within-sample variability. Instead of dealing with multiples sources of technical biases and variation separately, the proposed method can estimate the over-dispersion pattern of each transcript. A Gamma-Gamma model [44] is used to analyze the differential expression of isoforms while accounting for the between-sample variability. Importantly, for genes with multiple isoforms, the within-sample variability is modeled at the isoform level; the dispersion of read count on each exon is modeled using a parameter specific to each isoform. A Markov Chain Monte Carlo (MCMC) procedure [45, 46] is designed for a joint estimation of the model parameters and the differential states of isoforms. Simulation studies and real data studies with benchmarks demonstrate that the BayesIso approach has significantly improved the performance in identifying differentially expressed isoforms, especially on isoforms that are moderately differentially expressed. We have applied the BayesIso approach to breast cancer RNA-seq data to identify isoforms associated with breast cancer recurrence. The identified differentially expressed isoforms are enriched in cell death, cell survival, and signaling pathways (such as PI3K/AKT/mTOR signaling and PTEN signaling pathways), shedding light on the underlying mechanisms of isoforms in driving breast cancer recurrence.

2.2 Methods

2.2.1 Variability observed from real RNA-seq datasets

The variability of RNA-seq count data has been announced from various sources. To forward our understanding of the complexity of the variance of RNA-seq data, we looked into several real RNA-seq datasets. Figure 2.1 presents the complex variability observed from three data sets: basal breast cancer samples from The Cancer Genome Atlas (TCGA) project [59], human B cell datasets from Cheung et al. [60], and a mouse dataset [61]. We dissected the variance in sequencing counts along two dimensions: within-sample variation and between-sample variation. Within-sample variation typically leads to large variance of read counts among genomic loci (e.g., nucleotides or exons) which have similar expression level in the same sample. It is typically caused by technical artifacts such as uncorrected systematic bias and gene-specific random effects. On the other hand, between-sample variation is mostly due to biological differences among samples under the same condition.

The variability of RNA-seq data is assessed in Figure 2.1, where Figure 2.1(a) presents between-sample variability, and Figure 2.1(b) and (c) present within-sample variability from different perspectives. To avoid the ambiguity caused by the uncertainty of read assignment, genes with only one isoform are investigated. Figure 2.1(a) shows the scatter plots of the mean versus the variance of the number of reads that are mapped to the same gene across multiple samples under the same biological condition. The slopes of the least-squares (LS) fit lines for all scatter plots are apparently larger than those for Poisson model, which implies severe between-sample over-dispersion in all three RNA-seq datasets. Figure 2.1(b) shows the scatter plots of counts that fall in 100nt bins along the same gene within the same sample. One TCGA breast cancer sample and one MCF7 breast cancer cell line sample [62] are used as examples. Figure 2.1(b) indicates strong within-sample over-dispersion of read counts in both RNA-Seq samples. Figure 2.1(c), which presents the within-sample variance from another perspective, shows the variation of sequencing bias within one sample. Despite the overall tendency where read coverage is biased towards the 3'-end of transcript, subgroups of genes exhibit diverse patterns: a bias towards the 5'-end, or having depleted coverage on both ends. Figure 2.1(c) indicates that sequencing bias should be caused by multiple complex factors besides the location in transcript. In Figure 2.1(d), we further show an example of the read coverage for gene S100A9 (exon 2) across four samples from TCGA basal breast cancer dataset. The base level coverage has two distinct patterns, indicating the large variation of unknown read bias in the same biological group. Moreover, the ambiguity in the coverage patterns cannot be explained by deterministic systematic bias.

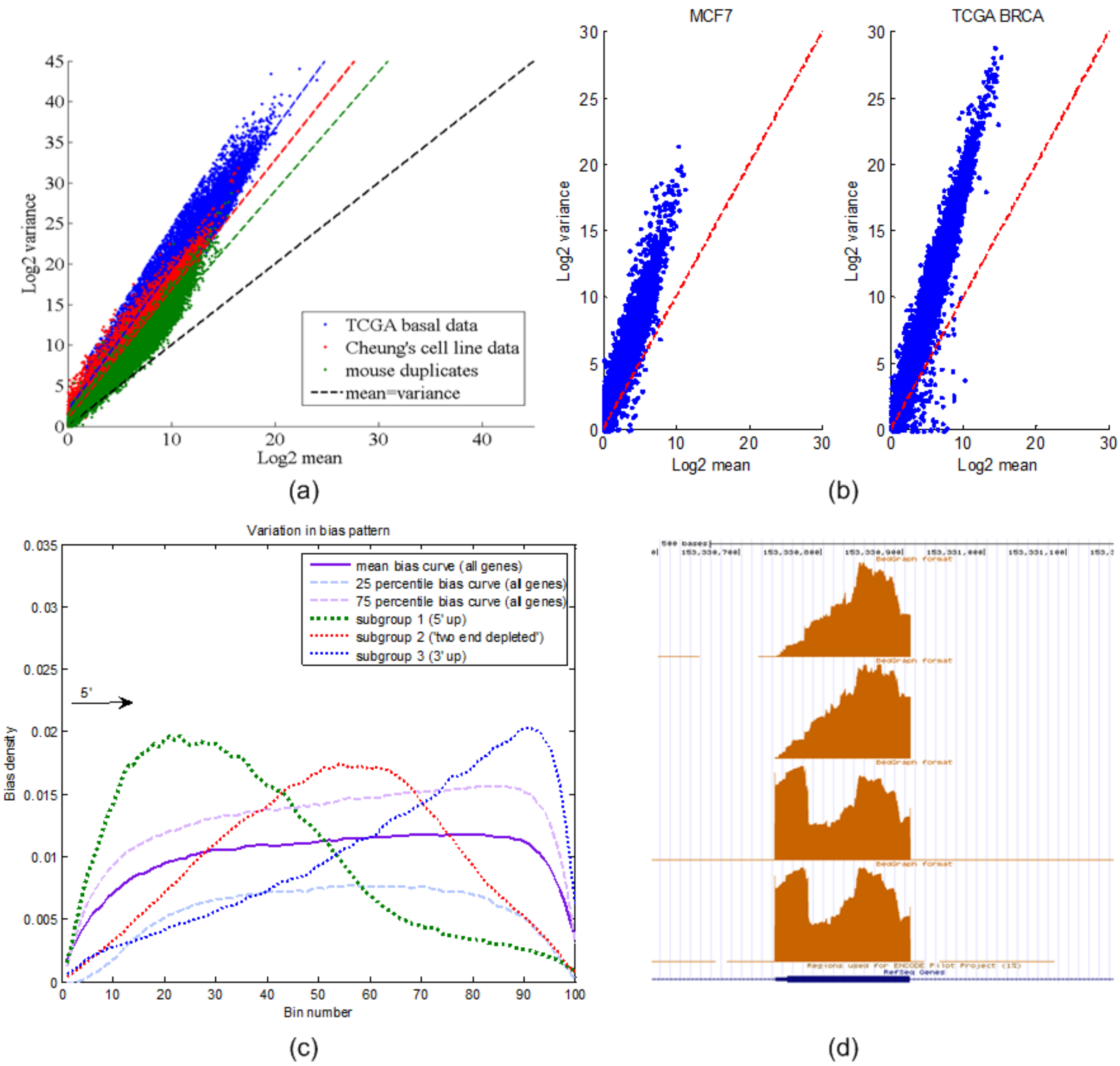


Figure 2.1 Between-sample variation and within-sample variation in RNA-seq data. (a) Between-sample over-dispersion observed from three real RNA-seq datasets. Scatter plots are in log 2 scale. (b) Scatter plots of the variance and mean of read counts in 100nt bins from an MCF7 breast cancer sample and a TCGA breast cancer tumor sample. (c) Bias patterns of genes in the same sample: purple solid line shows the overall bias pattern towards the right-tailed (biased towards the 3'-end of transcript); however, subgroups of genes have diverse bias patterns, either biased expression towards the 5'-end, or depleted expression in both ends shown by the green dotted line and red dotted line, respectively. (d) The coverage of an example region in four samples in the same biological group.

2.2.2 Framework of the BayesIso approach

An overview of the BayesIso approach for the identification of differentially expressed isoforms from RNA-seq data is shown in Figure 2.2. In our proposed joint model, a Poisson-Lognormal model can account for the within-sample variability. As noted before, different loci may have different bias patterns, which cannot be well explained by known sources. The Poisson-Lognormal model is capable of modeling different bias patterns along genomic loci at the isoform level. A Gamma-Gamma model is then used to model both the isoform abundance of multiple samples and the differential isoform abundance between two phenotypes. Specifically, the differential states of isoforms are introduced in the Gamma-Gamma model as hidden variables that control the differential isoform abundance of the samples between two phenotypes. The joint model, in which the Poisson-Lognormal model and the Gamma-Gamma model work together, can account for the between-sample variability in addition to the within-sample variability.

Based on the joint model, a Bayesian approach is used to estimate the posterior probability of the differential state of isoforms (the hidden variable). Since the joint model is defined by a set of parameters, a Markov Chain Monte Carlo (MCMC) sampling algorithm is used to estimate the parameters and the posterior probability of hidden variable. The MCMC sampling process consists of Gibbs sampling [48] and Metropolis-Hasting (M-H) sampling [47], generating samples from the conditional distributions. By virtue of the sampling process, the (marginal) posterior distributions of the parameters and the hidden variable can be estimated (or approximated) by the samples drawn from the MCMC sampling procedure.

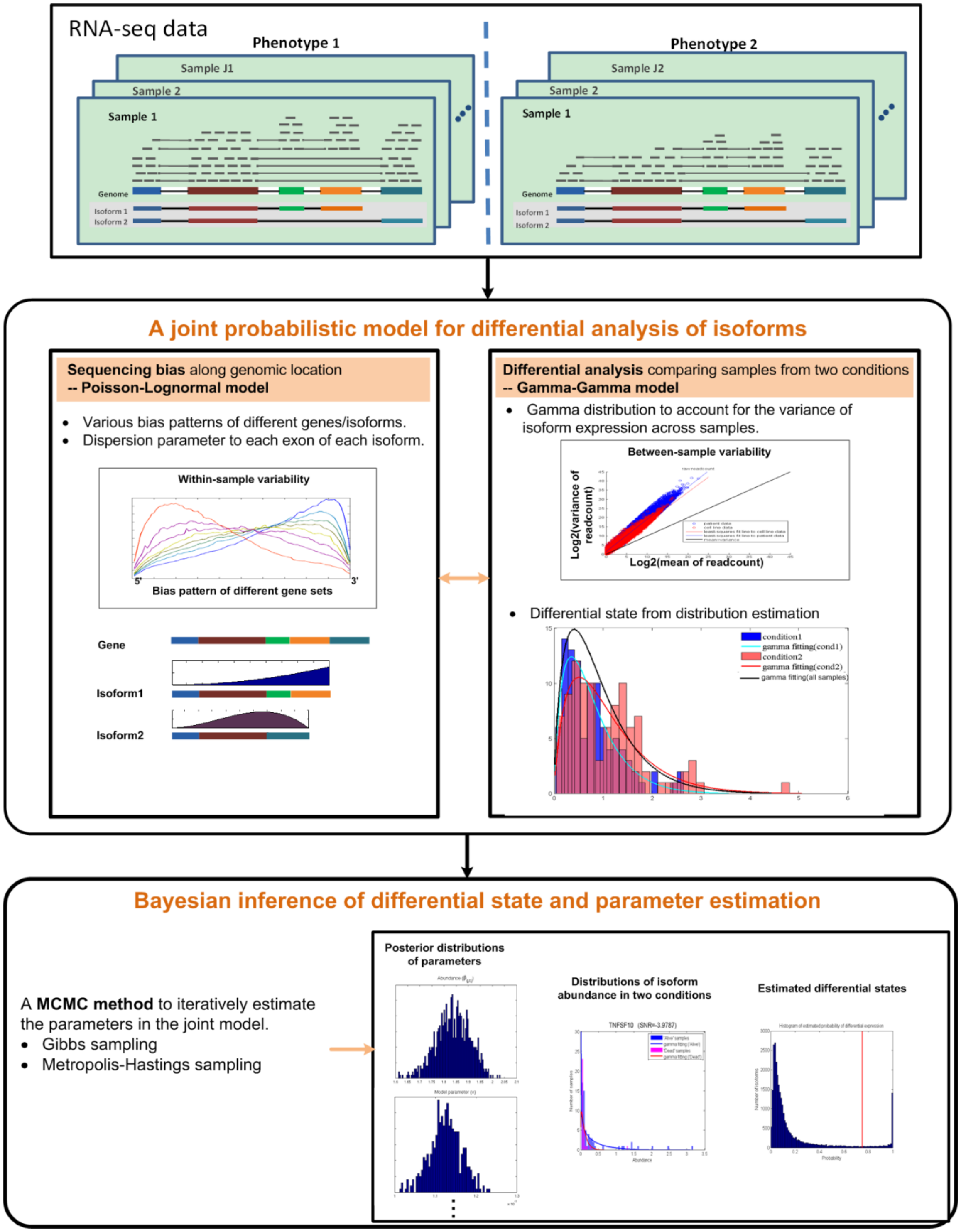

RNA-seq data
Phenotype 1
Phenotype 2
Sample J1
Sample J2
Sample 2
Sample 1
Genome
Isoform 1
Isoform 2
A joint probabilistic model for differential analysis of isoforms
Sequencing bias along genomic location
-- Poisson-Lognormal model
Various bias patterns of different genes/isoforms.
Dispersion parameter to each exon of each isoform.
Within-sample variability
5'
3'
Bias pattern of different gene sets
Gene
Isoform1
Isoform2
Differential analysis comparing samples from two conditions
-- Gamma-Gamma model
Gamma distribution to account for the variance of isoform expression across samples.
Between-sample variability
Log2(variance of readcount)
Log2(mean of readcount)
Differential state from distribution estimation
condition1
gamma fitting(cond1)
condition2
gamma fitting(cond2)
gamma fitting(all samples)
Bayesian inference of differential state and parameter estimation
A MCMC method to iteratively estimate the parameters in the joint model.
Gibbs sampling
Metropolis-Hastings sampling
Posterior distributions of parameters
Distributions of isoform abundance in two conditions
Estimated differential states
TNFSF10 (SNR=-3.9787)

Figure 2.2 Framework of BayesIso.

2.2.3 Bayesian model for differential analysis of RNA-seq data

Let $y_{t,g,i,j}$ represent the read counts that fall into the i^{th} $\left(1 \le i \le I_g\right)$ exon region of isoform $t\left(1 \le t \le T\right)$ of gene g $\left(1 \le g \le G\right)$ in sample $j\left(1 \le j \le J\right)$. T is the number of isoforms of gene g given by the annotation information. I_g is the number of exons in gene g. G is the total number of genes. $J = J_1 + J_2$ is the total number of samples, where J_1 and J_2 denote the number of samples in phenotype 1 and 2, respectively. Let $y_{g,i,j}$ represent the observed counts in the i^{th} exon region of gene g. For genes with only one isoform, i.e., $t = T = 1$, there is no uncertainty in read assignment caused by multiple isoforms, and thus,

$$y_{g,i,j} = y_{t,g,i,j}\,;$$

for genes with multiple isoforms, $y_{g,i,j}$ is the combination of all potential isoforms, as defined by

$$y_{g,i,j} = \sum_t \left(s_{t,g,i}\, y_{t,g,i,j}\right),$$

where $s_{t,g,i}$ is a binary value indicating whether exon i is included in isoform t of gene g. For each isoform, we use a Poisson-Lognormal regression model to account for the within-sample variability of RNA-seq data. Specifically, $y_{t,g,i,j}$ follows a Poisson distribution with mean $\gamma_{t,g,i,j}$:

$$y_{t,g,i,j} \sim Poiss(\gamma_{t,g,i,j})\,. \tag{2-1}$$

According to the Poisson-Lognormal model [43],

$$\gamma_{t,g,i,j} = x_{g,i}\beta_{t,g,\mathrm{j}} \exp\left(U_{g,t,i}\right), \tag{2-2}$$

$$U_{g,t,i} \sim N\left(0,\tau\right), \text{ s.t. } \sum_i U_{g,t,i} = 0\,, \tag{2-3}$$

$$\tau \sim Gamma\left(a,b\right), \tag{2-4}$$

where $\beta_{t,g,\mathrm{j}}$ is the true expression level of isoform *t* of gene *g* in sample *j*. $x_{g,i}$ is the length of the i^{th} exon weighted by the library size of sample *j*. $U_{g,t,i}$ is a model parameter representing the within-sample variability (or dispersion) for exon *i* of isoform *t* of gene *g*. Thus, the dispersion of different loci, exons of the isoforms, is modeled by different parameters. Precision parameter τ, which further follows a Gamma distribution with fixed hyperparameters *a* and *b*, controls the overall degree of within-sample variability.

Instead of being constant across samples in the same phenotype, the expression level $\beta_{t,g,j}$ carries between-sample variation. We adopt the Gamma-Gamma model [63] to model $\beta_{t,g,j}$ across samples collected from two phenotypes. The differential state, as a hidden variable in this Bayesian model, affects the distribution of $\beta_{t,g,j}$ among samples in each of the two phenotypes. $d_{t,g}$, a binary value, indicates the differential state of isoform *t* of gene *g*, where $d_{t,g}=1$ means isoform *t* of gene *g* is differentially expressed; $d_{t,g}=0$, otherwise. Note that the between-sample variability is captured by the Gamma distribution. From the Gamma-Gamma model, the isoform expression level $\beta_{t,g,\mathrm{j}}$ is given by:

if $d_{t,g}=0$,

$$\beta_{t,g,j} \sim Gamma\left(\alpha, \lambda_{t,g}\right), \tag{2-5}$$

$$\lambda_{t,g} \sim Gamma\left(\alpha_0, v\right); \tag{2-6}$$

if $d_{t,g}=1$

$$\beta_{t,g,\mathrm{j}_1} \sim Gamma\left(\alpha, \lambda_{t,g}^{(1)}\right),\ \beta_{t,g,\mathrm{j}_2} \sim Gamma\left(\alpha, \lambda_{t,g}^{(2)}\right), \tag{2-7}$$

$$\lambda_{t,g}^{(1)},\ \lambda_{t,g}^{(2)} \sim Gamma\left(\alpha_0, v\right), \tag{2-8}$$

and

$$v \sim Gamma\left(a_0, b_0\right), \tag{2-9}$$

where α is the shape parameter; $\lambda_{t,g}$ is the rate parameter that depends on differential state $d_{t,g}$. If $d_{t,g} = 0$, $\lambda_{t,g}^{(1)} = \lambda_{t,g}^{(2)} = \lambda_{t,g}$; if $d_{t,g} = 1$, $\lambda_{t,g}^{(1)} \neq \lambda_{t,g}^{(2)}$. $\lambda_{t,g}$ is further assumed to follow a Gamma distribution with shape parameter α_0 and rate parameter v.

Figure 2.3 presents the Bayesian hierarchical dependency graph for all the parameters involved in the proposed model. Observation **y**, in which each element is the read count on each exon of each gene in each sample, is represented by a shaded circle. Observation **y** depends on mRNA abundance **β**, design matrix **x**, **s**, and within-sample variance parameter **U**. **β** depends on Gamma parameters **λ** and α, and **λ** further depends on differential state **d**, Gamma parameters v and α_0. Thus, the parameter set $\{\alpha, \alpha_0, v\}$ controls the expression level of the isoforms of all samples, and thus controls the between-sample variance. On the other hand, **U** depends on τ which controls the overall within-sample variance. a_0, b_0, π, a, and b, shown in shaded square, are the hyper-parameters. In marked contrast to existing methods like edgeR, DESeq, Cuffdiff 2 that use statistical tests to identify differentially expressed genes/isoforms, the differential states of isoforms are introduced and modeled in the proposed joint model of BayesIso. A joint estimation of the differential states with other model parameters is accomplished by a Markov Chain Monte Carlo (MCMC) sampling method as described in detail in the next section.

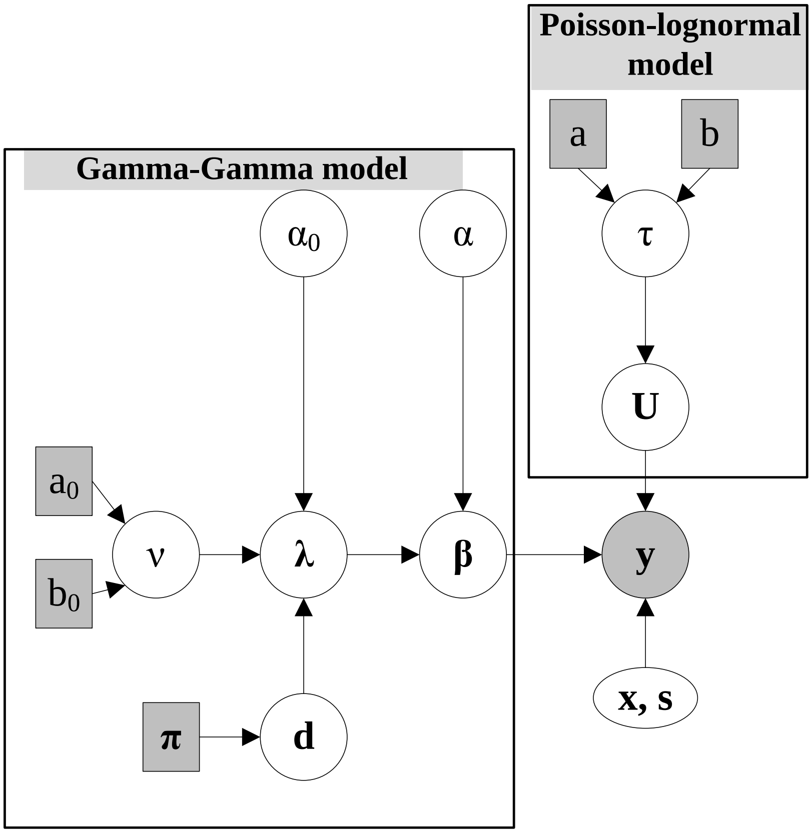


Figure 2.3 Dependency graph of model parameters. Observation (read count on each exon) **y** is shaded, while the random variables are denoted as circles. Fixed parameters are denoted as shaded squares. Exon length and library size information **x** and **s** is denoted as oval.

2.2.4 Model parameter estimation via an MCMC scheme

Due to the complexity of the joint model, it is challenging to estimate directly the model parameters and the hidden variables (i.e., the differential states, d = [$d_{t,g}$]). We have designed an MCMC method to estimate the parameters and the hidden variables (**d**). The MCMC sampling process is a combination of Gibbs sampling and M-H sampling, with which as many samples as possible can be generated or drawn from the conditional distributions. By virtue of the sampling process, the marginal posterior distributions of the parameters and the hidden variables can be approximated by the samples drawn from the MCMC sampling procedure. Next, we will describe the MCMC algorithm and the associated conditional distributions.

For genes with one isoform and genes with multiple isoforms, the conditional posterior distribution of parameters $\boldsymbol{\beta}$, $\mathbf{U}$ and τ for the Poisson-Lognormal regression model can be derived differently.

For genes with one isoform, isoform-level differential analysis is the same as gene-level analysis. The number of reads on exon *i* of gene *g* in sample *j* can be directly obtained as $y_{g,i,j}$. Thus, the conditional posterior distributions of the parameters $\mathbf{U}$ and τ for the Poisson-Lognormal regression model are

$$\begin{aligned} P(U_{g,i} \left| \mathbf{y}_{g,i}, \beta_{g,j}, \tau) \right. &\sim \prod_{j} P\left(y_{g,i,j} \left| U_{g,i}, \beta_{g,j}\right.\right) \times P\left(U_{g,i} \left|\tau\right.\right) \\ &\sim \prod_{j} \left(\beta_{g,j} x_{g,i} e^{U_{g,i}}\right)^{y_{g,i,j}} \times \exp\left(-\beta_{g,j} x_{g,i} e^{U_{g,i}}\right) \times \exp\left(-\frac{\tau U_{g,i}^2}{2}\right), \end{aligned} \tag{2-10}$$

$$\begin{aligned} P\left(\tau \left| \mathbf{U}\right.\right) \sim P\left(\mathbf{U} \left|\tau\right.\right) \times P(\tau) &\sim \prod_{g} \prod_{i} P(U_{g,i} \left|\tau) \right. \times P(\tau) \\ &\sim Gamma\left(a + \frac{\sum_{g} I_g}{2}, b + \sum_{g,i} \frac{U_{g,i}^2}{2} \right). \end{aligned} \tag{2-11}$$

The conditional posterior distributions of $\beta_{g,j}$ under the two conditions are

$$\begin{aligned} P\left(\beta_{g,j_1} \left| \mathbf{y}_{g,j_1}, U_{g,i}, \lambda_g^{(1)}, \alpha\right.\right) &\sim P\left(\mathbf{y}_{g,j_1} \left| \beta_{g,j_1}, U_{g,i}\right.\right) \times P\left(\beta_{g,j_1} \left| \lambda_g^{(1)}, \alpha\right.\right) \\ &\sim Gamma\left(\sum_{i} y_{g,i,j_1} + \alpha, \sum_{i} x_{g,i} \exp\left(U_{g,i}\right) + \lambda_g^{(1)} \right), \end{aligned} \tag{2-12}$$

$$\begin{aligned} P\left(\beta_{g,j_2} \left| \mathbf{y}_{g,j_2}, U_{g,i}, \lambda_g^{(2)}, \alpha\right.\right) &\sim P\left(\mathbf{y}_{g,j_2} \left| \beta_{g,j_2}, U_{g,i}\right.\right) \times P\left(\beta_{g,j_2} \left| \lambda_g^{(2)}, \alpha\right.\right) \\ &\sim Gamma\left(\sum_{i} y_{g,i,j_2} + \alpha, \sum_{i} x_{g,i} \exp\left(U_{g,i}\right) + \lambda_g^{(2)} \right). \end{aligned} \tag{2-13}$$

For genes with multiple isoforms, it is more complicated. Based on the assumption that the expression levels of the transcripts are independent, observation $\mathbf{y}$ follows:

$$y_{g,i,j} \sim \text{Poiss}\left(\sum_t \left(s_{t,g,i} \gamma_{t,g,i,j} \right) \right). \tag{2-14}$$

Thus, the likelihood of observation $y_{g,i,j}$ is

$$P\left(y_{g,i,j} \middle| \boldsymbol{\beta}_{g,j}, U_{g,t,i} \right) \sim \text{Poiss}\left(\sum_t \left(s_{t,g,i} \beta_{t,g,j} x_{g,i} \exp\left(U_{g,t,i} \right) \right) \right). \tag{2-15}$$

The conditional posterior distributions of the parameters $\mathbf{U}$ and τ of the Poisson-Lognormal model can be derived as follows:

$$\begin{aligned} P(U_{g,t,i} \left| \mathbf{y}_{g,i}, \boldsymbol{\beta}_{g,j}, \tau \right.) &\sim \prod_j P\left(\mathbf{y}_{g,i,j} \middle| U_{g,t,i}, \boldsymbol{\beta}_{g,j} \right) \times P\left(U_{g,t,i} \middle| \tau \right) \\ &\sim \prod_j \left(\left(\sum_t \left(s_{t,g,i} \beta_{t,g,j} x_{g,i} \exp\left(U_{g,t,i} \right) \right) \right)^{y_{g,i,j}} \times \exp\left(-\sum_t \left(s_{t,g,i} \beta_{t,g,j} x_{g,i} \exp\left(U_{g,t,i} \right) \right) \right) \right), \\ &\quad \times \exp\left(-\frac{\tau \left(U_{g,t,i} \right)^2}{2} \right) \end{aligned} \tag{2-16}$$

$$\begin{aligned} P\left(\tau \middle| \mathbf{U} \right) &\sim \prod_{g,t,i} P\left(U_{g,t,i} \middle| \tau \right) \times P(\tau) \\ &\sim Gamma\left(a + \frac{\sum_{t,g,i} s_{t,g,i}}{2}, b + \sum_{g,t,i} \frac{\left(U_{g,t,i} \right)^2}{2} \right). \end{aligned} \tag{2-17}$$

The conditional posterior distributions of $\beta_{t,g,j}$ under the two conditions are

$$\begin{aligned} P\left(\beta_{g,t,j_1} \middle| \mathbf{y}_{g,j_1}, U_{g,t,i}, \lambda_{g,t}^{(1)}, \alpha \right) &\sim P\left(\mathbf{y}_{g,j_1} \middle| \beta_{g,t,j_1}, U_{g,t,i} \right) \times P\left(\beta_{g,t,j_1} \middle| \lambda_{g,t}^{(1)}, \alpha \right) \\ &\sim \prod_i \left(\left(\sum_t \left(s_{g,t,i} \beta_{g,t,j_1} x_{g,i} \exp\left(U_{g,t,i} \right) \right) \right)^{y_{g,i,j_1}} \times \exp\left(-\sum_t \left(s_{g,t,i} \beta_{g,t,j_1} x_{g,i} \exp\left(U_{g,t,i} \right) \right) \right) \right), \\ &\quad \times \beta_{g,t,j_1}^{\alpha} \times \exp(-\lambda_{g,t}^{(1)} \beta_{g,t,j_1}) \end{aligned} \tag{2-18}$$

$$\begin{aligned} &P\left(\beta_{g,t,j_2}\left|\mathbf{y}_{g,j_2},U_{g,t,i},\lambda_{g,t}^{(2)},\alpha\right.\right)\sim P\left(\mathbf{y}_{g,j_2}\left|\beta_{g,t,j_2},U_{g,t,i}\right.\right)\times P\left(\beta_{g,t,j_2}\left|\lambda_{g,t}^{(2)},\alpha\right.\right)\\ &\sim\prod_i\left(\left(\sum_t\left(s_{g,t,i}\beta_{g,t,j_2}x_{g,i}\exp\left(U_{g,t,i}\right)\right)\right)^{y_{g,i,j_2}}\times\exp\left(-\sum_t\left(s_{g,t,i}\beta_{g,t,j_2}x_{g,i}\exp\left(U_{g,t,i}\right)\right)\right)\right).\\ &\quad\times\beta_{g,t,j_2}^{\alpha}\times\exp(-\lambda_{g,t}^{(2)}\beta_{g,t,j_2}) \end{aligned} \tag{2-19}$$

Given isoform expression level $\boldsymbol{\beta}$, the conditional posterior distributions of parameters $\boldsymbol{\lambda},\alpha,\alpha_0,\nu$, and $\mathbf{d}$ for the Gamma-Gamma model are the same for isoforms from genes with one isoform and from genes with multiple isoforms, which can be derived as follows. $\lambda_{g,t}$ depends on differential state $d_{g,t}$. Thus, if $d_{g,t}=1$,

$$\begin{aligned} P\left(\lambda_{g,t}^{(1)}\left|\boldsymbol{\beta}_{g,t}^{(1)},\nu,\alpha,\alpha_0\right.\right)&\sim P\left(\boldsymbol{\beta}_{g,t}^{(1)}\left|\lambda_{g,t}^{(1)},\alpha\right.\right)\times P\left(\lambda_{g,t}^{(1)}\left|\nu,\alpha_0\right.\right)\\ &\sim Gamma\left(J_1\alpha+\alpha_0,\sum_{j_1}\beta_{g,t,j_1}+\nu\right), \end{aligned} \tag{2-20}$$

$$\begin{aligned} P\left(\lambda_{g,t}^{(2)}\left|\boldsymbol{\beta}_{g,t}^{(2)},\nu,\alpha,\alpha_0\right.\right)&\sim P\left(\boldsymbol{\beta}_{g,t}^{(2)}\left|\lambda_{g,t}^{(2)},\alpha\right.\right)\times P\left(\lambda_{g,t}^{(2)}\left|\nu,\alpha_0\right.\right)\\ &\sim Gamma\left(J_2\alpha+\alpha_0,\sum_{j_2}\beta_{g,t,j_2}+\nu\right); \end{aligned} \tag{2-21}$$

if $d_{g,t}=0$,

$$\begin{aligned} P\left(\lambda_{g,t}\left|\boldsymbol{\beta}_{g,t},\nu,\alpha,\alpha_0\right.\right)&\sim P\left(\boldsymbol{\beta}_{g,t}\left|\lambda_{g,t},\alpha\right.\right)\times P\left(\lambda_{g,t}\left|\nu,\alpha_0\right.\right)\\ &\sim Gamma\left(J\alpha+\alpha_0,\sum_{j}\beta_{g,t,j}+\nu\right). \end{aligned} \tag{2-22}$$

The conditional posterior distribution of υ is

$$\begin{aligned} P\left(\nu\left|\boldsymbol{\lambda},\mathbf{d},\alpha,\alpha_0\right.\right)&\sim P\left(\boldsymbol{\lambda}\left|\mathbf{d},\nu,\alpha,\alpha_0\right.\right)\times P\left(\nu\right)\\ &\sim Gamma\left(\left(\sum_{g,t}\left(1+d_{g,t}\right)\right)\alpha_0+a_0,\sum_{g,t}\left(\lambda_{g,t}^{(1)}+\lambda_{g,t}^{(2)}\times d_{g,t}\right)+b_0\right). \end{aligned} \tag{2-23}$$

According to [44], the posterior distribution of $\mathbf{d}$ given $\boldsymbol{\beta}, \alpha, \alpha_0$ and ν can be derived as:

$$
\begin{aligned}
P\left(d_{g,t} \middle| \boldsymbol{\beta}, \lambda_{g,t}, \alpha, \alpha_0, \nu\right) &\sim P\left(\boldsymbol{\beta} \middle| d_{g,t}, \alpha, \alpha_0, v\right) P\left(d_{g,t}\right) \\
&\sim \left(K_1 K_2 \frac{\left(\prod_j \beta_{g,t,j}\right)^{\alpha-1}}{\left(\nu + \sum_{j_1} \beta_{g,t,j_1}\right)^{J_1\alpha+\alpha_0} \left(\nu + \sum_{j_2} \beta_{g,t,j_2}\right)^{J_2\alpha+\alpha_0}} \right)^{d_{g,t}} \\
&\quad \times \left(K \frac{\left(\prod_j \beta_{g,t,j}\right)^{\alpha-1}}{\left(\nu + \sum_{j_1} \beta_{g,t,j_1} + \sum_{j_2} \beta_{g,t,j_2}\right)^{(J_1+J_2)\alpha+\alpha_0}} \right)^{1-d_{g,t}} \times \pi_{g,t}
\end{aligned}
\quad , \qquad \text{(2-24)}
$$

where

$$
K_1 = \frac{\nu^{\alpha_0} \Gamma\left(J_1\alpha + \alpha_0\right)}{\Gamma^{J_1}(\alpha)\Gamma(\alpha_0)}, \; K_2 = \frac{\nu^{\alpha_0} \Gamma\left(J_2\alpha + \alpha_0\right)}{\Gamma^{J_2}(\alpha)\Gamma(\alpha_0)}, \text{ and } K = \frac{\nu^{\alpha_0} \Gamma\left(\left(J_1 + J_2\right)\alpha + \alpha_0\right)}{\Gamma^{J_1+J_2}(\alpha)\Gamma(\alpha_0)}.
$$

The posterior distribution of α_0 and α are given by:

$$
P\left(\alpha_0 \middle| \boldsymbol{\lambda}, \mathbf{d}, \alpha_0, \nu\right) \sim \prod_{g,t} \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha_0)} \left(\lambda_{g,t}^{(1)}\right)^{\alpha_0-1} \times \left(\frac{\nu^{\alpha_0}}{\Gamma(\alpha_0)} \left(\lambda_{g,t}^{(2)}\right)^{\alpha_0-1} \right)^{d_{g,t}} \right), \qquad \text{(2-25)}
$$

$$
P\left(\alpha \middle| \boldsymbol{\beta}, \boldsymbol{\lambda}\right) \sim \prod_{g,t} \left(\prod_{j_1} \frac{\left(\lambda_{g,t}^{(1)}\right)^{\alpha}}{\Gamma(\alpha)} \beta_{g,t,j_1}^{\alpha-1} \times \prod_{j_2} \frac{\left(\lambda_{g,t}^{(2)}\right)^{\alpha}}{\Gamma(\alpha)} \beta_{g,t,j_2}^{\alpha-1} \right). \qquad \text{(2-26)}
$$

For genes with only one isoform, the subscription *t* in Eq.(2-20) ~ Eq.(2-26) is can be deleted.

With the derived conditional posterior distributions, the MCMC algorithm is designed with the steps for Gibbs sampling and M-H sampling. Note that M-H sampling is used to sample the parameters without conjugate priors, while Gibbs sampling is used to sample the parameters with conjugate priors.

INPUT: Observed read counts **y**, library size weighted isoform structure **x**, design matrix **s**, number of iterations N

OUTPUT: Estimates of all of the parameters and the differential state **d** in the joint Bayesian model

Algorithm:

Step 1. Initialization: each parameter is set an arbitrary value and non-informative prior knowledge is used for the parameters.

Step 2. Draw samples iteratively from the conditional distributions of parameters $\boldsymbol{\beta}, \mathbf{U}, \tau$ (in the Poisson-Lognormal model) and parameters $\boldsymbol{\lambda}, \alpha, \alpha_0, \nu,$ and **d** (in the Gamma-Gamma model). Perform the following sampling steps for N iterations:

Use Gibbs sampling to draw samples of $\boldsymbol{\beta}, \tau, \boldsymbol{\lambda}, \nu$ from their conditional distributions that follow standard probability distributions;

Use Metropolis-Hasting (M-H) sampling to draw samples of $\mathbf{U}$, **d**, α, α_0 from their conditional distributions in sequence. Since these parameters do not have conjugate priors, M-H sampling is used to approximate their posterior distributions.

Step 3. Estimate differential state **d** as well as other parameters $\boldsymbol{\beta}, \mathbf{U}, \tau, \boldsymbol{\lambda}, \alpha, \alpha_0, \nu$ from the samples, after the burn-in period, generated from the MCMC procedure.

2.3 Simulation studies

To systematically assess the performance of BayesIso, we simulated multiple RNA-seq datasets with different scenarios. In each experiment, a gene set was randomly selected from the annotation file from the UCSC genome browser database (version: GRCh37/hg19; http://genome.ucsc.edu/). 30% of the genes/isoforms were randomly selected as differentially expressed. We generated simulation data sets using two simulators: (1) our simulator that produced aligned reads following the proposed model; (2) RNAseqReadSimulator [64] that generates raw sequencing reads. We conducted simulation studies in different scenarios regarding different variance of RNA-seq data and differential levels of true differentially expressed isoforms.

2.3.1 Simulation data at varying model parameters

To mimic real RNA-seq data, we adopted a simulation strategy proposed by Wu *et al*. [36]. We first estimated model parameters from real datasets and then used them as the baseline to generate sequencing data sets based on human annotation file (version: GRCh37/hg19). Two RNA-seq datasets were investigated in the study: 1) a mouse dataset with 10 C57BL/6J (B6) mouse samples and 11 DBA/2J (D2) mouse samples [61]; 2) 23 basal type breast cancer samples which received chemotherapy treatment from the TCGA project [59]. For the TCGA dataset, we divided the patients into two groups with 13 samples and 10 samples, respectively. Without losing generality, we used all 15137 genes with one isoform to estimate the model parameters.

We assigned non-informative priors for hyper-parameters τ $(\mathrm{a}=1,\mathrm{b}=0)$, α, α_0, d_g ($P\left(d_g=1\right)=\pi_g=0.5$) and v ($a_0=1$, $b_0=0$). We used thin=10 for the sampling process to record every 10th sample. Figure 2.4 shows the estimates of model parameters from the real data sets. We explored the variability of RNA-seq data by close examination of estimated parameters. Precision parameter τ (inverse of Gaussian variance σ^2) in the Poisson-Lognormal model controls the overall degree of within-sample variance. Smaller τ indicates larger variation of read counts within a sample. Estimated from the sampling process, $\tau=0.44$ in the mouse dataset and the $\tau=1.78$ in TCGA dataset. Between-sample variance is jointly determined by α, α_0, and v. Parameter α and α_0 jointly affect the expression level of isoforms and differential level of the differentially expressed isoforms. In general, as α increases, the expression level of isoforms increases, and the differential level comparing samples from two phenotypes also increases. In contrast, for larger α_0 values, the expression level and the difference between two phenotypes become smaller.

With parameters estimated from the two real datasets as the baseline, we varied model parameter τ to simulate data sets with different within-sample variabilities, and varied model parameter α and α_0 to simulate data sets with different expression levels and differential levels. In the simulation studies regarding varying within-sample variabilities, we used the average correlation of estimated expression with ground truth expression to measure the accuracy of abundance quantification across multiple samples. In the simulation studies regarding varying abundance and differential levels of isoforms, we used Area-under-the-curve (AUC) of the receiver operating characteristic (ROC) curve, precision, recall, and F-score as the metrics to evaluate the performance of the competing methods. F-score is calculated by the following definition:

$$\text{F-score} = \frac{2 \times precision \times recall}{precision + recall} .$$

In the experiments, α varied among (1.4, 1, 0.6) with $a_0 = 0.5$, $\nu = 0.1$, $\tau = 1.78$; α_0 varied among (0.2, 0.6, 1) with $a_0 = 1.5$, $\nu = 0.1$, $\tau = 1.78$. To show the differential levels of isoforms in the simulation data sets, signal-to-noise-ratio (SNR) was calculated from the ground truth expression of isoforms of samples in two conditions by:

$$SNR = 20\log_{10}\left(\frac{\left| mean(\text{cond}\,1) - mean(\text{cond}\,2) \right|}{sqrt(var(\text{cond}\,1) + var(\text{cond}\,2))} \right).$$

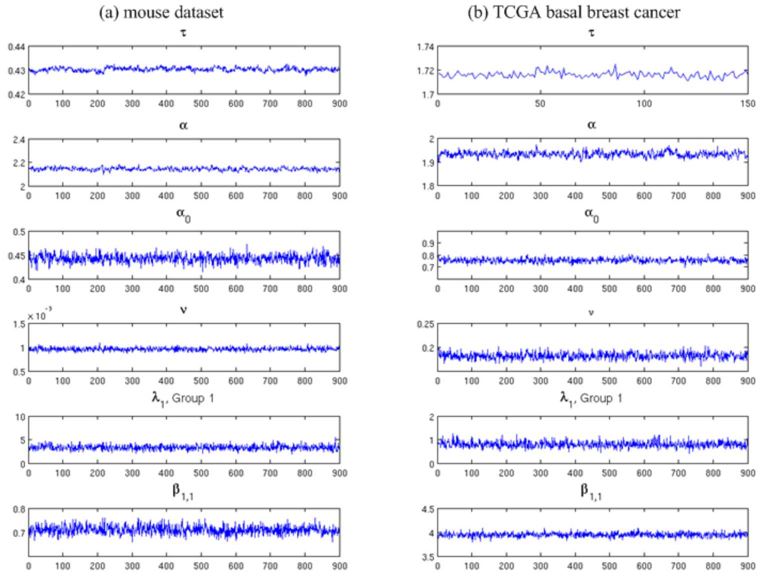
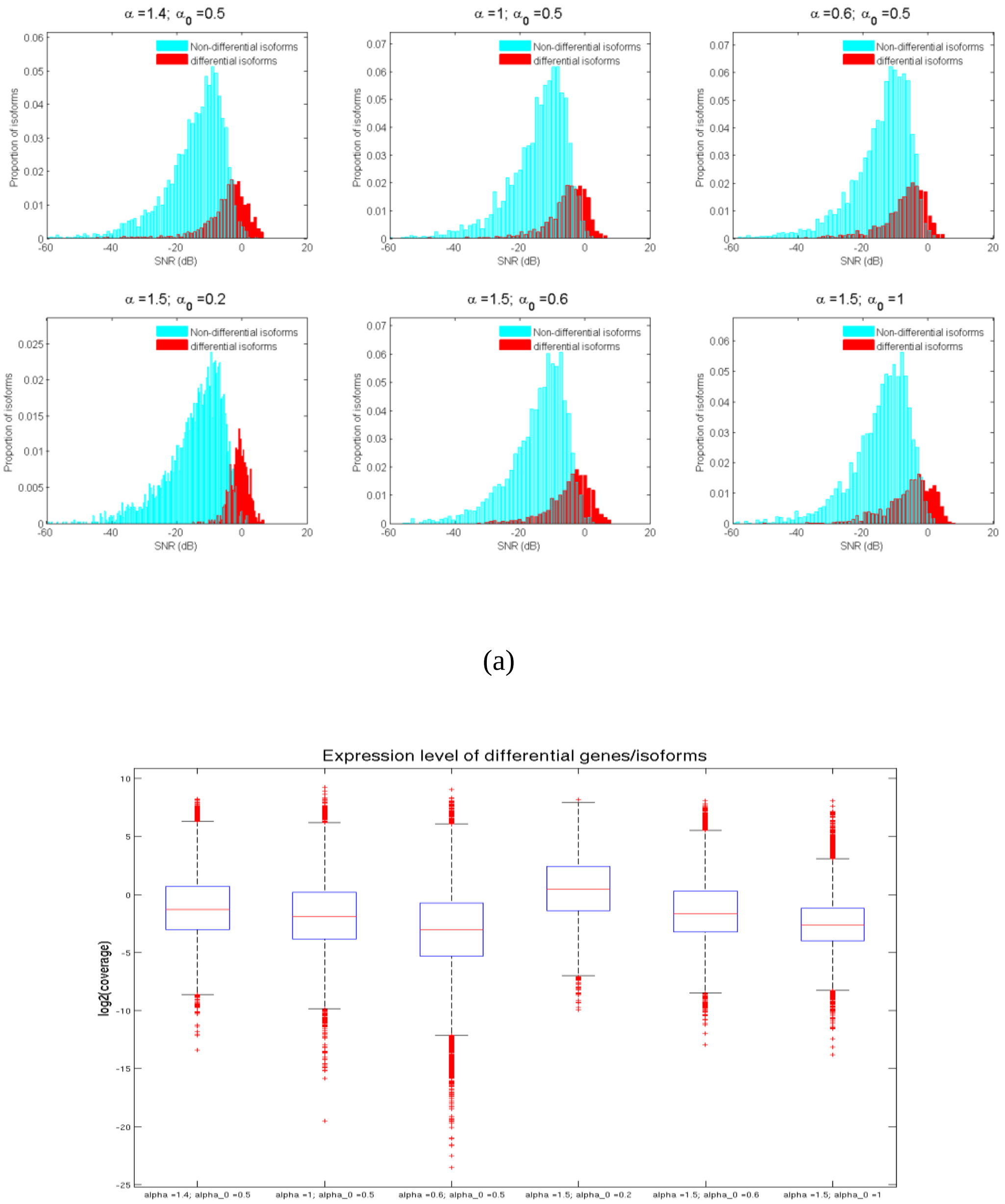


Figure 2.4 Estimate of model parameters from real datasets. (a) Model parameters estimated from the mouse dataset. (b) Model parameters estimated from the TCGA basal breast cancer dataset.



(a)

(b)

Figure 2.5 Differential level and abundance of differentially expressed isoforms in simulation data sets at varying parameters α, α_0: (a) histograms of SNRs of differentially expressed isoforms and non-differentially expressed isoforms; (b) boxplot of the expression level (log2(coverage)) of differentially expressed isoforms.

Figure 2.5(a) shows the histograms of SNRs of differentially expressed isoforms and non-differentially expressed isoforms at varying parameters α , α_0 , and Figure 2.5(b) shows the boxplot of the expression levels of differentially expressed isoforms. We can see that decreasing α and increasing α_0 lead to lower expression and lower differential level. In all of the simulation experiments, q-value < 0.05 was used as the criteria for differentially expressed isoforms detected by DESeq, edgeR, DSS, Ballgown and Cuffdiff 2, and Prob(DE)>0.95 was used for BayesIso and EBSeq.

2.3.2 Performance comparison at varying model parameters

We first used our simulator to generate simulation data at varying model parameters, and evaluated the performance of BayesIso on genes with one isoform and genes with multiple isoforms, separately. As introduced in Section 2.2.2, isoform-level differential analysis of RNA-seq data on genes with single isoform is the same as gene-level differential analysis. Thus, the count-based methods, such as DESeq, edgeR, EBSeq, and DSS were applicable. Therefore, we compared the proposed method BayesIso to the count-based methods on genes with single isoform. For genes with multiple isoforms, we compared the proposed method BayesIso to isoform-level differential analysis approaches, i.e., Cuffdiff 2 and Ballgown.

A. Performance comparison on genes with only one isoform

As an intermediate step for differential analysis, we first evaluated the performance on abundance quantification at different within-sample variability. We set $\tau = 1.78, 1, 0.44$, with $\alpha = 2$, $\alpha_0 = 0.5$, $\nu = 0.1$. In general, the smaller the precision parameter τ , the higher the overall within-sample variability. We compared our method with four widely used methods for RNA-seq normalization, which are: reads-per-kilobase-per-million (RPKM), DESeq, trimmed mean of M-values (TMM), and upper quantile normalization (UQUA). RPKM [65] is calculated by normalizing reads by the length of genomic features (genes and exons) and total library size. DESeq normalization is implemented by DESeq (1.14.1) [35], a differential gene identification method based on a negative binomial model. TMM is first developed in edgeR [34] and later included into BioConductor package NOISeq [66]. NOISeq (2.0.0) also has separate implementations of RPKM and UQUA methods in addition to TMM, all of which were used for performance comparison in this study.

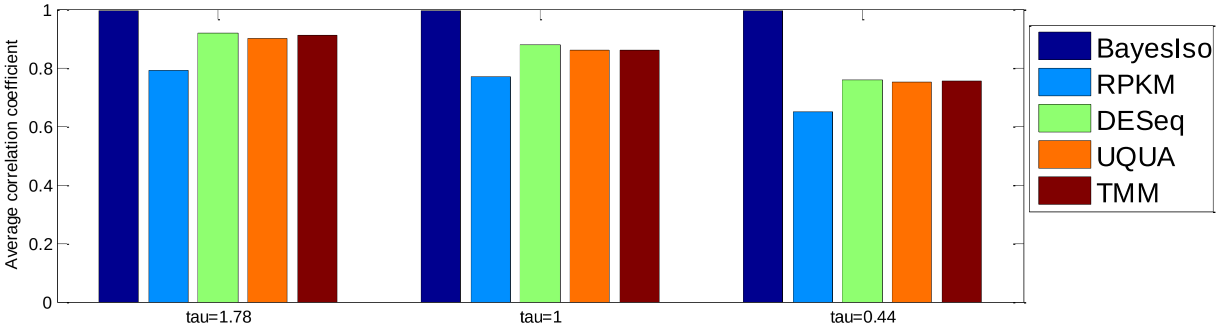


Figure 2.6 Performance comparison for abundance estimation on genes with single isoform.

Figure 2.6 gives the average correlation for all competing methods under different parameter settings. We see that BayesIso achieved robust performance under different over-dispersion settings; the abundance estimated by BayesIso was highly correlated with ground truth. DESeq, TMM and UQUA achieved comparable performance across multiple parameter settings, while RPKM had the least favorable performance in all scenarios. Our computational results were quite consistent with the observation by Dillies *et al*. [67] that DESeq and TMM (edgeR) are much better normalization methods than RPKM.

We further evaluated the performance on differential analysis in multiple scenarios regarding varying differential levels controlled by parameters α , α_0 , as aforementioned. We compared with four existing count-based methods: DESeq (1.14.1, fitType=local), edgeR (3.4.2, default), DSS (2.0.0, default), EBSeq (1.3.1, default). From Figure 2.7, we can see that, BayesIso outperformed the other methods in all scenarios with varying differential levels and abundance of differentially expressed isoforms (genes). To further assess the ability of the competing methods regarding different differential levels of isoforms, we divided the true differentially expressed isoforms into three groups according to their SNRs. Performance on different groups of truly differential isoforms and non-differential isoforms were evaluated separately, as shown in Figure 2.8. We can see that BayesIso achieved a much higher recall when the isoforms were less differentially expressed (‘diff-group2’ with -4 dB< SNR <-1 dB). This improvement can be attributed in part to the more accurate abundance quantification of BayesIso.

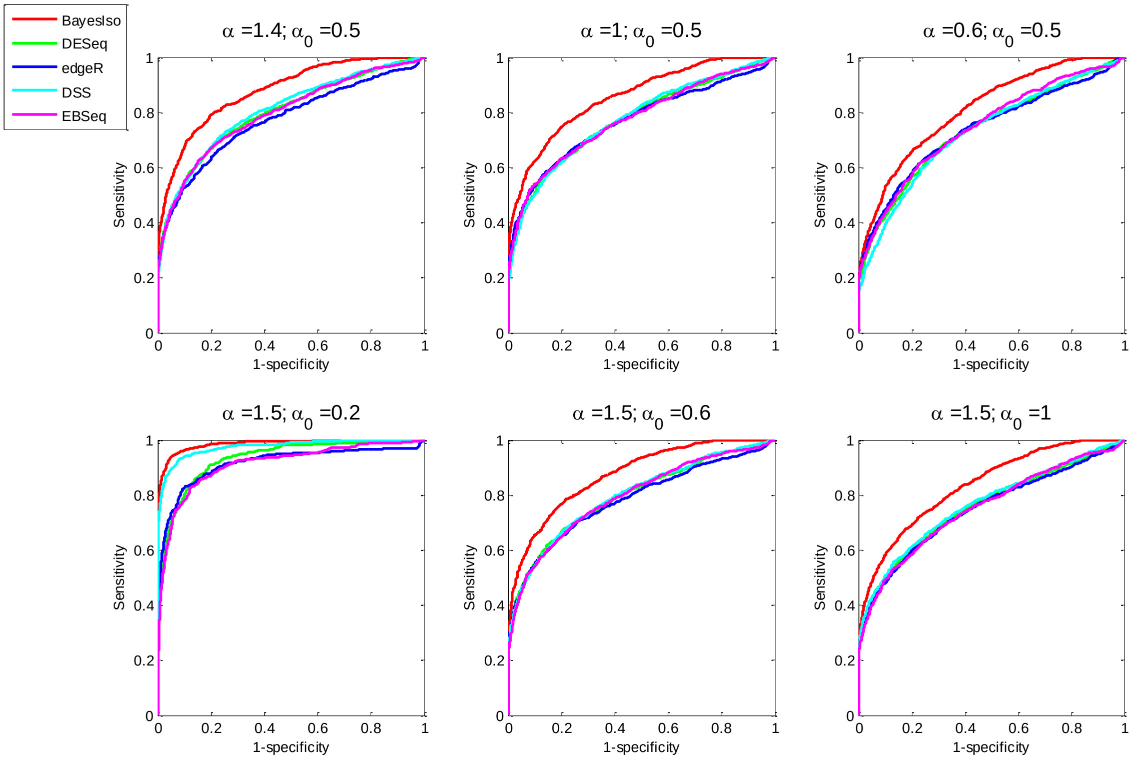


Figure 2.7 ROC curves for performance comparison on differential analysis on genes with single isoform at varying model parameters α, α_0.

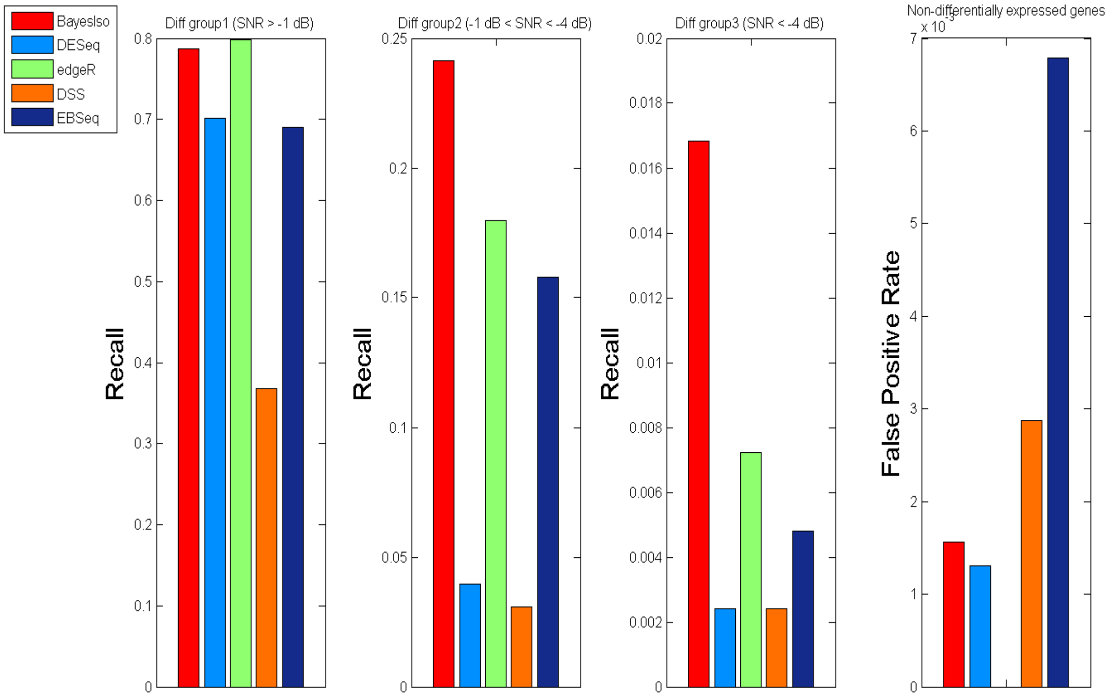


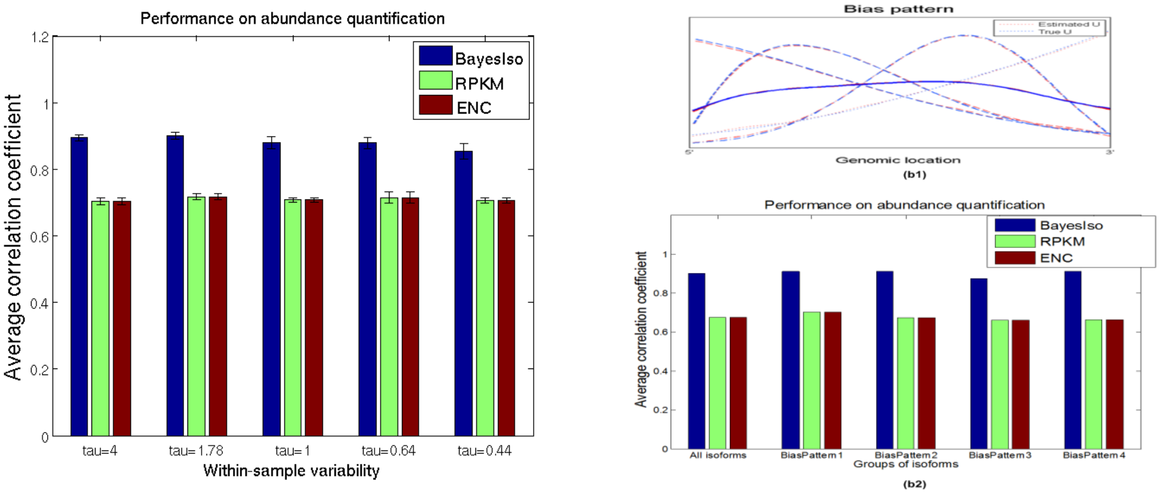
Figure 2.8 Precision, Recall, and F-score for performance comparison on genes with single isoform with different SNRs.

B. Performance comparison on genes with multiple isoforms

Similar to the performance comparison on genes with single isoform, we first evaluated the performance on abundance quantification on simulation data sets with different within-sample variance. We compared the abundance (β in Eq. (2.5)) estimated by our method with two measurements of abundance estimated by Cufflinks and Cuffdiff 2: RPKM and 'external_normalized_count' (ENC). ENC is an alternative measurement estimated by Cuffdiff 2 to take into account the phenotype information and variance of the samples.

First, we varied parameter τ that affects the overall within-sample variability, with other model parameters set as $\alpha = 1$, $\alpha_0 = 0.5$, $\upsilon = 0.1$. From Figure 2.9(a), we can see that, at different τ , the average correlation coefficient between our estimation and the true value was consistently higher than that of RPKM and ENC. Note that although RPKM and ENC were calculated by different methods, their correlations to the true expression level across samples were about the same.

Moreover, we tested the performance on simulation data with different sequencing bias patterns (along genomic location). Rather than randomly drawn from τ , parameter $\mathbf{U}$ was designed to follow different bias patterns for different sets of genes. We simulated four patterns, as shown by the blue curves in Figure 2.9(b), to mimic the observed patterns from real RNA-seq data in our studies. We can see from the estimated patterns shown in Figure 2.9(b) (the red curves) that our proposed model was able to capture the bias patterns accurately. The correct estimation of bias patterns consequently contributes to more accurate abundance estimation (Figure 2.9(b)). As shown by the barplot in Figure 2.9(b), the performances on groups of genes with different bias patterns were comparable, indicating that our model can deal with various bias patterns.



(a) (b)

Figure 2.9 Performance comparison on abundance quantification: (a) different overall within-sample variability; (b) different bias patterns along genomic location. Average correlation coefficient between the estimated abundance and true abundance of isoforms is used to evaluate the performance.

We further compared the performance on differentially expressed isoform identification to isoform-level differential analysis approaches Cuffdiff 2 and Ballgown. . We ran the experiments using genes with an increasing number of isoforms, starting from genes with two isoforms. Table 2-1 lists the results of performance comparison on varying parameter settings. As we can see, our method consistently outperformed the other methods on simulation data generated at various parameters. Cuffdiff 2 achieved a high precision; however, it missed many differentially expressed isoforms. Ballgown could identify more differentially expressed isoforms but with less precision. From the table, we can see that our method was more effective when the isoform abundance was generally lower or the isoforms were less differentially expressed. The improved performance of BayesIso in differential analysis is gained from: (1) using a model to account for both within-sample variability and between-sample variability; (2) estimating differential states of isoforms by a Bayesian method using a joint model of the variability of RNA-seq data and the differential state of isoforms.

Table 2-1. Performance comparison on differential analysis at varying parameters α or α_0 (while other parameters are fixed).

	BayesIso			Cuffdiff 2			Ballgown		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
$\alpha = 1.4$	0.948	0.455	**0.612**	0.95	0.113	0.194	0.74	0.4	0.515
$\alpha = 1$	0.932	0.431	**0.587**	0.667	0.011	0.021	0.747	0.373	0.49
$\alpha = 0.6$	0.921	0.312	**0.464**	0.364	0.046	0.072	0.721	0.246	0.361
$\alpha_0 = 0.2$	0.898	0.622	**0.732**	0.828	0.119	0.208	0.753	0.311	0.419
$\alpha_0 = 0.6$	0.942	0.38	**0.54**	0.854	0.116	0.198	0.826	0.158	0.265
$\alpha_0 = 1$	0.939	0.271	**0.42**	0.867	0.044	0.082	0.564	0.052	0.09

Similar to the performance evaluation on genes with single isoforms, we also divided ground truth differentially expressed isoforms into three groups according to SNR. The performance was shown by Table 2-2. When the performance on all of the isoforms was compared, the recall performance of Cuffdiff 2 was the lowest, which affected its overall performance measured by F-score. Consistent with the performance on genes with single isoform, BayesIso achieved a much higher recall when the isoforms were less differentially expressed (‘diff-group2’ with -4 dB< SNR <-1 dB). We can also see that the false positive rate (defined by the portion of non-differential isoforms included) of our method was lower than Ballgown. Note that the false positive rate of Cuffdiff 2 was low because fewer differential isoforms were identified.

Table 2-2. Performance comparison on differential analysis at different SNR levels.

		BayesIso	**Cuffdiff 2**	**Ballgown**
All isoforms	Precision	**0.926**	0.907	0.808
	Recall	**0.468**	0.117	0.297
	F-score	**0.622**	0.207	0.435
‘Diff-group 1’ (SNR>-1dB)	Recall	**0.859**	0.449	0.539
‘Diff-group 2’ (-4dB<SNR<-1dB)	Recall	**0.457**	0.173	0.231
‘Diff-group 3’ (SNR<-4dB)	Recall	**0.055**	0.051	0.084
‘Non-diff-group’	FPR*	**0.034**	0.047	0.065

*FPR: False Positive Rate

Table 2-3. Performance comparison on differential analysis of genes with more than 2 isoforms.

		K=3	K=4	K=5
BayesIso	Precision	0.883	0.839	0.842
	Recall	0.485	0.429	0.441
	F-score	**0.626**	**0.561**	**0.571**
Cuffdiff 2	Precision	0.777	0.69	0.706
	Recall	0.345	0.261	0.271
	F-score	0.476	0.374	0.379
Ballgown	Precision	0.738	0.482	0.606
	Recall	0.394	0.177	0.272
	F-score	0.508	0.244	0.361

We also evaluated the performance of the competing methods on genes with 3, 4, and 5 isoforms. The experimental results demonstrated that our proposed method outperformed the other methods in multiple scenarios (Table 2-3). The performance comparison was consistent with experiments on genes with two isoforms.

2.3.3 Simulation study using RNAseqReadSimulator

We also generated synthetic data using another RNA-seq simulator (RNAseqReadSimulator [64]) to test the performance of all competing methods: DESeq, edgeR, DSS, EBSeq, Cuffdiff 2 and Ballgown. A set of 4000 genes, which includes 7810 isoforms, was randomly selected from the around 23000 RefSeq genes for the experiment, where 20% of the isoforms were differentially expressed. RNAseqReadSimulator was used to generate raw RNA-seq reads, and TopHat 2 was used to align the reads to reference genome hg19. In order to implement the count-based methods for isoform-level differential analysis, the read count of each transcript was first calculated from RPKM estimated by Cufflinks for a fair comparison.

Figure 2.10 is the histogram of the SNRs of differentially expressed isoforms and non-differentially expressed isoforms, to assess the differential level of simulation data. ROC curves of all competing methods were shown in Figure 2.11. We can see that, consistent with simulation studies using our simulator, BayesIso outperformed the other methods. By closely investigating the ‘left’ ROC curves, we can see that Cuffdiff 2 (the gray curve) was efficient in detecting a subset of differentially expressed isoforms; however, its performance was less favorable due to high false negatives. The superiority of BayesIso may be gained from comprehensive modeling of the variabilities of RNA-seq data as well as joint estimation of all model parameters together with differential state.

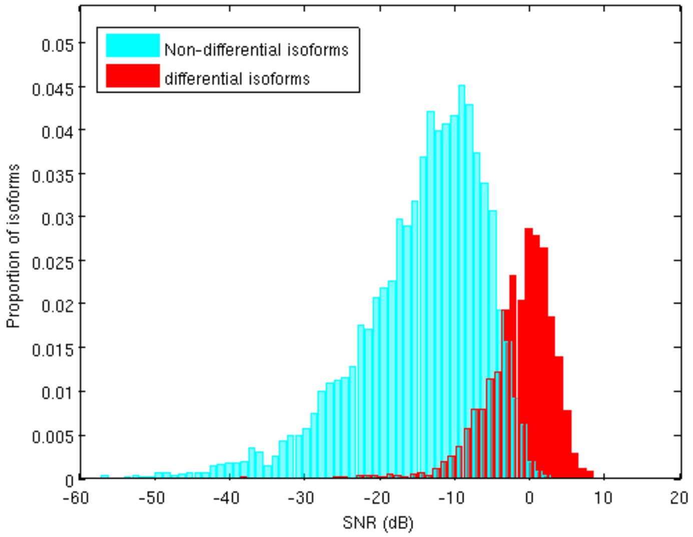


Figure 2.10 Histogram of SNRs of the differentially expressed isoforms and non-differentially expressed isoforms.

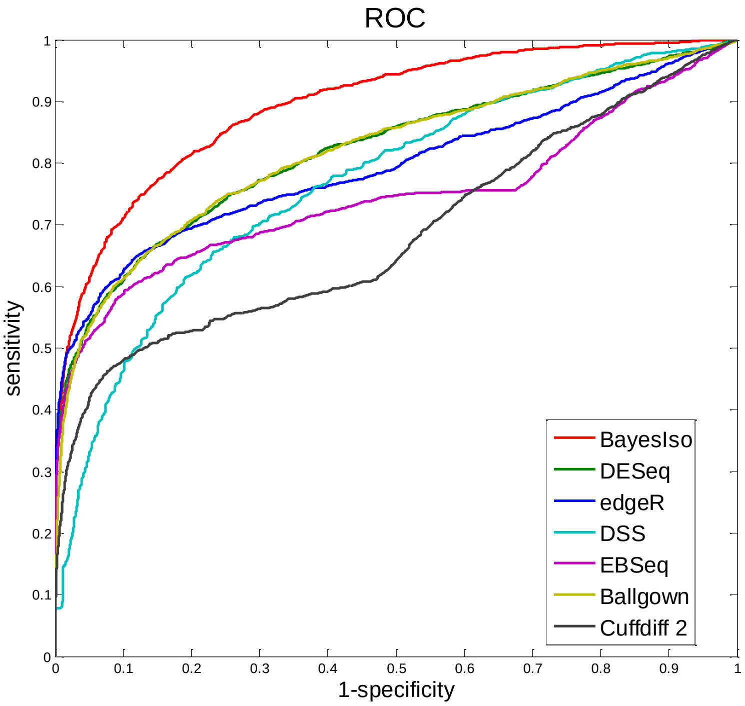


Figure 2.11 ROC curves of all competing methods on simulation data generated by RNAseqReadSimulator.

2.4 Real data analysis

We first used two benchmark datasets to evaluate the performance of all competing methods of differential analysis of RNA-seq data. Both datasets are part of the MicroArray Quality Control Project (MAQC) for benchmarking microarray technology [68, 69] as well as to characterize RNA-seq technology. The MAQC project includes replicated samples of human brain reference RNA (hbr) and universal human reference RNA (uhr). The first data set is the Sequencing Quality Control (SEQC) dataset [70], which includes RNA spike-in controls along with the true RNA library. The RNA spike-in controls is a set of synthetic RNAs from External RNA Control Consortium (ERCC), which can be used as benchmark. The second data set is another set of hbr and uhr samples. The expression of close to 1000 genes in uhr and hbr were validated by TaqMan qRT-PCR, which can be used as benchmark. Finally, we applied our proposed method to breast cancer tumor samples to identify differentially expressed isoforms associated breast cancer recurrence.

2.4.1 Performance comparison benchmarked by ERCC RNAs

We first compared the performance of BayesIso with four existing methods (DESeq, edgeR, DSS and EBSeq) on the SEQC dataset with ERCC spike-in controls. 92 artificial transcripts were mixed into a real RNA-Seq library with different ratios (1:1 for none differentially expressed genes, and 4:1, 2:3 and 1:2 for differentially expressed genes), which were used as ground truth for differential analysis. Gene level counts were downloaded from http://bitbucket.org/soccin/seqc, with 5 replicates in each group.

Figure 2.12 shows the ROC curves of the five competing methods. We can see that BayesIso had the best performance among all five methods by achieving an AUC very close to 0.9. The second best method was DSS with an AUC about 0.85. DESeq (AUC=0.7624) and edgeR (AUC=0.7675) had very close performance, which was consistent with the previous results reported by Rapaport *et al.* [70]. EBSeq, on the other hand, had the least favorable performance (AUC=0.71) on this specific dataset and it failed to detect the most strongly differentially expressed genes: its sensitivity was less than 0.1 when its specificity was about 0.9.

By close examination of the ‘left’ ROC curves of the five methods, we can further infer that BayesIso should have significantly better precision than the other competing methods, as the sensitivity of BayesIso went up to 0.7 before any sacrifice in specificity.

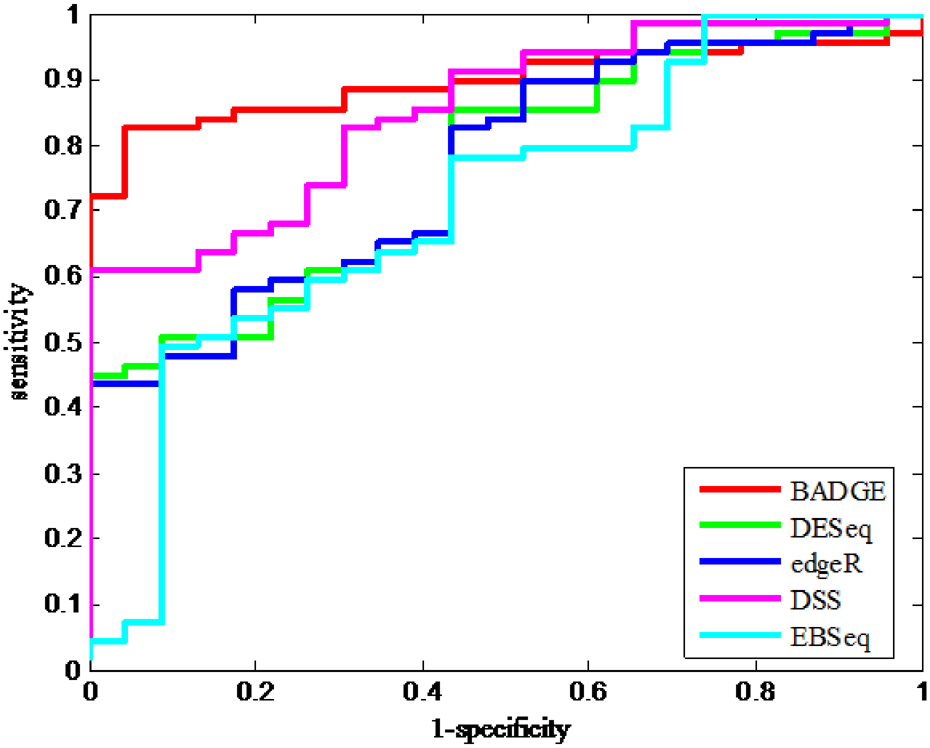


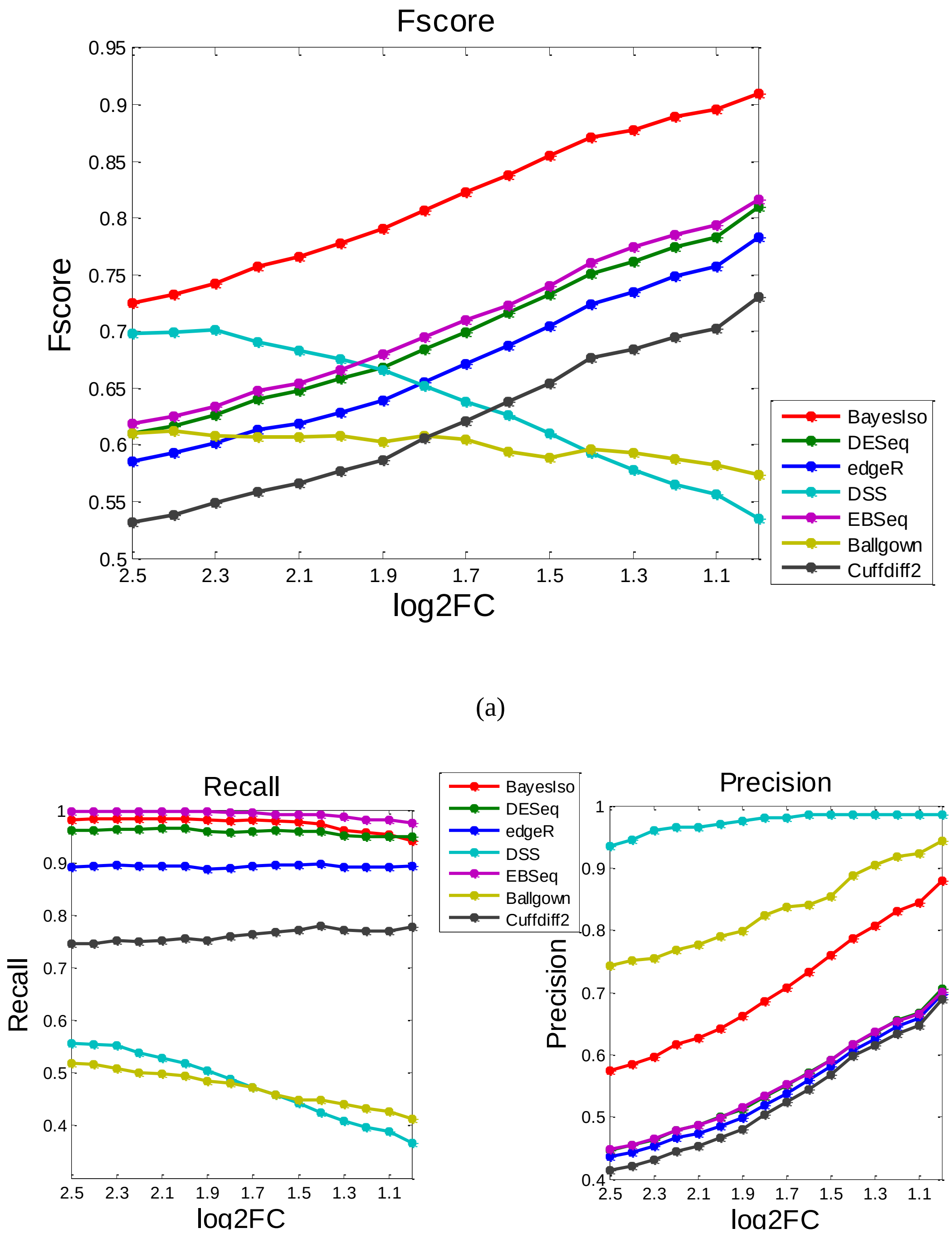
Figure 2.12 Performance comparison on differential analysis using the SEQC dataset benchmarked by ERCC RNAs.

2.4.2 Performance comparison benchmarked by TaqMan data

The set of roughly 1000 genes of hbr and uhr measured by TaqMan qRT-PCR is a set of more comprehensive benchmarks as it spans a wider range of expression ratios and represents a sampling of true human transcripts [71]. We evaluated the ability of all competing methods to detect differentially expressed genes on another RNA-seq data set with replicated hbr and uhr samples. These samples were generated by Dr. Dudoit from University of California at Berkeley

[72] from NCBI sequence read archive (SRA) with ID SRZ016359 and SRX016367, with 7 samples in each phenotype. To apply all competing methods to this data set for differential analysis, we first performed alignment using 'TopHat 2 (TopHat v2.0.12)' with UCSC hg19 as the reference genome. Then, Cuffdiff 2 and Ballgown were applied to the aligned reads for differential analysis of the RefSeq genes. For the other count-based methods, i.e., DESeq, edgeR, EBSeq, and DSS, the required input, a matrix of read count, was calculated from RPKM that was estimated by Cufflinks for a fair comparison. All of the methods were implemented under default setting, and q-value < 0.05 was used as the criteria for differential genes detection for Cuffdiff 2, Ballgown, DESeq, edgeR, and DSS, while Prob(DE) > 0.95 was used for BayesIso and EBSeq.

TaqMan qRT-PCR measurements of roughly 1000 genes were downloaded from https://bitbucket.org/soccin/seqc. Among the roughly 1000 genes, 844 genes were overlapped with RefSeq genes, which were used to benchmark the performance on differential analysis. We used the absolute log2 ratio of fold change (| $\log_2$FC |) of the expression between the two phenotypes to determine differentially expressed genes and non-differentially expressed genes. Genes with | $\log_2$FC | larger than a threshold T were considered as differentially expressed, while the other genes were non-differentially expressed. We varied the threshold T from 2.5 to 1 (5.6 × expression change to 1.5 × expression change measured by qRT-PCR) to evaluate the performance of the competing methods at decreasing cutoff values of qRT-PCR expression change, which defined sets of differentially expressed genes with decreasing differential level. We used recall, precision, and F-score as the metrics, as shown in Figure 2.13. From Figure 2.13(a), we can see that the overall performance, evaluated by F-score, of BayesIso outperformed all other methods in various scenarios. Moreover, F-score increased at decreasing differential level, which indicates the ability of BayesIso in detecting less differentially expressed genes. The recall of EBSeq was higher than BayesIso; however, the precision of EBSeq was much lower, as too many genes were detected as differentially expressed. The precision of DSS and Ballgown were higher than BayesIso, as they detected much fewer differentially expressed genes, which can be inferred by their poor performance in recall. Therefore, BayesIso has been demonstrated to outperform the other competing methods on genes with different differential levels.



(a)

(b)

Figure 2.13 Performance comparison on differential analysis on MAQC data with TaqMan qRT-PCR measurements as benchmark: (a) Overall performance evaluated by F-score; (b) Recall and precision.

2.4.3 Identification of differentially expressed isoforms associated with breast cancer recurrence

We applied BayesIso to breast cancer data acquired by The Cancer Genome Atlas (TCGA) project [59]. The study was designed to identify the differentially expressed isoforms associated with breast cancer recurrence. 93 estrogen receptor positive (ER+) tumors from patients were collected for this study, where 61 patients were still alive with follow-up longer than 5 years, labeled as ‘Alive’; 32 patients were dead within 5 years, labeled as ‘Dead’. The ‘Dead’ and ‘Alive’ groups represent the ‘early recurrence’ group and the ‘late/non-recurrence’ group, respectively.

We downloaded the sequencing data (Level 1) profiled by Illumina HiSeq 2000 RNA Sequencing Version 2 from the TCGA data portal, and then performed alignment using ‘TopHat 2 (TopHat v2.0.12)’ with UCSC hg19 as the reference sequence. With the annotation file of isoform structure (RefSeq genes) downloaded from the UCSC genome browser database [73],

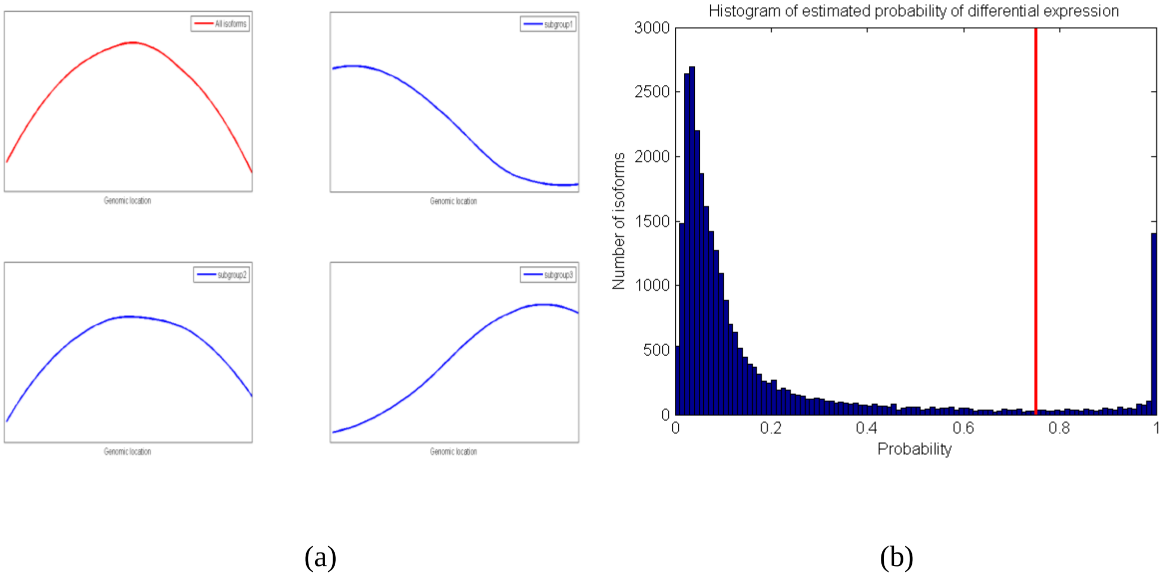


Figure 2.14 Results of BayesIso on TCGA BRCA tumor samples. (a) Estimated bias patterns of the sequencing reads. The mean bias pattern of all of the isoforms is shown by the red curve in the up-left figure. However, different sets of isoforms exhibit varying bias patterns. The three blue curves show the mean bias patterns of different groups of isoforms. The isoforms are grouped according to their bias patterns. (b) Histogram of estimated probability that the isoforms are differentially expressed. Red line denotes Prob(d=1)=0.75.

we applied our method to identify differentially expressed isoforms by analyzing samples from the two groups: the ‘Dead’ group vs. the ‘Alive’ group. Consistent with the observation, various bias patterns along the genomic location were captured by our proposed model, as shown in Figure 2.14(a). While the overall bias pattern of all isoforms was high in the middle, different subgroups of isoforms had varying bias patterns. The histogram of the estimated probability that the isoforms were differentially expressed was shown in Figure 2.14(b). With threshold ‘Probability > 0.75’, 2,299 isoforms of 1,905 genes were identified as differentially expressed. We also calculated the SNR of the identified differentially expressed isoforms. The SNRs had a mode value around -5dB, indicating that most of the identified isoforms were moderately differentially expressed. The low mean SNR value was consistent with the high variability of expression level observed across the samples. Thus, the detection power on moderately differential isoforms was critical for differential analysis of breast cancer RNA-seq data.

We compared BayesIso with isoform-level differential analysis methods Cuffdiff 2 and Ballgown in terms of identified differential genes. Differential genes were defined as genes with at least one differentially expressed isoform. Since Cuffdiff 2 is too conservative, we used a loose threshold to include more differential genes for further investigation. For a fair comparison, p-value<0.05 was used as the criterion for differentially expressed isoforms detected by Cuffdiff 2 and Ballgown, and thus, 1,719 and 5,399 genes, respectively, were identified as differential. Figure 2.15(a) shows the overlap and difference of the gene sets identified by the three methods. As we can see from the figure, Cuffdiff 2 detected a much less number of differential genes than Ballgown did; among the differential genes identified by BayesIso, 30% were uniquely identified by our method as compared with that from Cuffdiff 2 and Ballgown. The unique set of differential genes helped reveal several signaling pathways such as the PI3K/AKT/mTOR signaling and PTEN signaling pathways as shown in Figure 2.15(b). Figure 2.15(b1) shows the PI3K/AKT/mTOR signaling pathway, the hyperactivation of which has been demonstrated being associated with the tumorigenesis of ER-positive breast cancer [74, 75]. PIK3R2, a member of the PI3K protein family participating in the regulatory subunit, was detected by BayesIso as down-regulated in the ‘Dead’ group. The loss of expression of PIK3R2 is crucial to the hyperactivation of the PI3K/AKT/mTOR signaling pathway by regulating AKT2. The

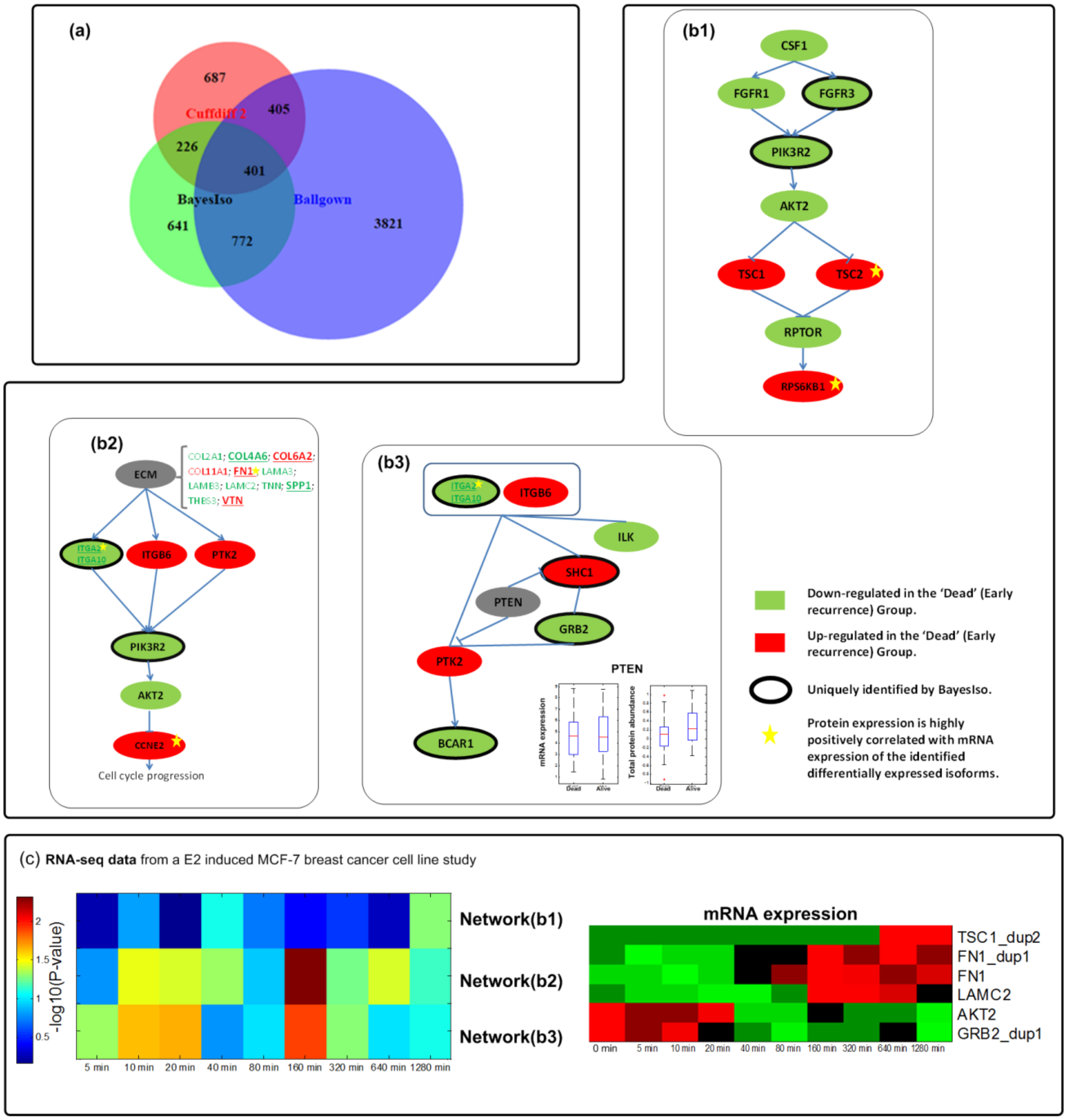


Figure 2.15 Comparison of BayesIso with Cuffdiff 2 and Ballgown on breast cancer study. (a) Venn diagram of identified differential genes (genes with differentially expressed isoforms) by the three methods: BayesIso, Cuffdiff 2, and Ballgown. (b) Three networks of differential genes detected by BayesIso: b1 - a network related to PI3K/AKT/mTOR signaling pathway; b2 - a network related to cell cycle progression of PI3K/AKT signaling pathway; b3 - a part of PTEN signaling pathway. The color of nodes represents the expression change between the two phenotypes: green means down-regulated in the 'Dead' group; red mean up-regulated in the 'Dead' group. Genes marked by bold circle or underlined are uniquely detected by BayesIso. Genes marked by yellow star have consistent protein/phosphoprotein expression. (c) Enrichment analysis of three networks using a time-course E2 induced MCF-7 breast cancer cell line data (collected at 10 time points: 0, 5, 10, 20, 40, 80, 160, 320, 640, 1280 mins, with one sample at each time point): left - enrichment analysis of the three networks; right - expression of transcripts with significant pattern change.

dysfunction of AKT2 in inhibiting the expression of TSC1 and TSC2 activated the mTOR signaling, as indicated by the over-expression of RPS6KB1, a downstream target of mTOR. The overexpression of TSC2 and RPS6KB1 was further validated by their protein/phosphoprotein expression measured by reverse phase protein array (RPPA) on a subset of the TCGA breast cancer samples, which consists of 45 samples in the 'Alive' group and 27 samples in the 'Dead' group. Specifically, the expression of NM_001114382, a differentially expressed isoform of TSC2, was positively correlated with its phosphoprotein expression at pT1462 (p-value = 0.02); the expression of NM_001272044, a differentially expressed isoform of RPS6KB1, was positively correlated with its phosphoprotein expression at pT389 (p-value = 0.0081). Note that the FPKM (expression) of NM_001272044 estimated by Cuffdiff 2 was not correlated with its phosphoprotein expression, indicating that the isoform expression estimated by BayesIso was more consistent with protein expression than Cuffdiff 2. The network shown in Figure 2.15(b2) reveals part of the PI3K/AKT signaling pathway leading to cell cycle progression. FN1 and ITGA2 were uniquely detected by BayesIso, which led to the overexpression of CCNE2 in the "Dead" group. The total protein expressions of FN1, ITGA2 and CCNE2 were highly correlated with the estimated expressions of their isoforms, respectively. Figure 2.15(b3) shows a part of the PTEN signaling, the underexpression of which results in hyperactivation of PI3K/AKT signaling pathway in breast cancer [76, 77]. Though the mRNA expression of PTEN was not differential, the total protein did have lower expression level in the 'Dead' group as shown in the boxplots. BayesIso also detected SHC1, GRB2, and BCAR1, three critical components in PTEN signaling.

We further analyzed the expression of identified transcripts in the three networks (Figure 2.15b) on a time-course of estrogen (E2) induced transcription in MCF-7 breast cancer cells (RNA-seq data; GSE62789). We performed an enriched analysis of the three networks. Specifically, for each time point, we obtained the fold change of transcript expression in log2 scale comparing with the sample at time 0, and then used the mean of fold change as the test statistic for each network. We calculated the p-value of each network from a significance test, where the null distributions were generated by calculating test statistics from randomly sampled gene sets of the same size of the network (100,000 iterations). The enrichment scores, defined as the negative of the logarithm of p-value to base 10, are shown on the left panel of Figure 2.15(c).

As we can see from the figure, two networks (Figure 2.15(b2) and Figure 2.15(b3)) were enriched at early time points (<160 minutes). Moreover, the differential isoforms of TSC1, FN1, LAMC2, AKT2, and GRB2 had significant expression pattern changes along the time, as shown on the right panel of Figure 2.15(c).

To study the interaction of the identified differential genes, we mapped the differentially expressed genes to the Protein-Protein interaction (PPI) network from the Human Protein Reference Database (HPRD) [78] and then filtered out extremely low abundant isoforms according to the abundance relative to all of the isoforms of the same gene. With the criterion of median relative abundance > 10%, 359 isoforms from 308 genes were identified as differentially expressed, among which 195 genes had multiple isoforms according to the annotation file with isoform structure. Furthermore, comparing with a gene-level analysis with the same criterion, 'Prob(d)>0.75', for identifying differentially expressed genes, 133 multiple-isoform genes were differential at the isoform level but non-differential at the gene level, as shown by the pie chart in Figure 2.16.

We conducted a functional enrichment analysis of the identified genes using Ingenuity Pathway Analysis (IPA; http://www.qiagen.com/ingenuity); it turned out that many of the genes were known to be associated with the following cellular functions: proliferation, cell death, and cell migration. We further performed an enrichment analysis of the associated sets of isoforms on a time-course of estrogen (E2) induced transcription in MCF-7 breast cancer cells (RNA-seq data; GSE62789). The two sets of isoforms associated with proliferation and migration of cells were significantly enriched with p-value = 0.043 and p-value = 0.021, respectively; the p-value of the isoforms associated with cell death is 0.07.

In the PPI network of 308 genes, several hub genes (ESR1, BRCA1, CREBBP, ERBB2, LCK) are known to play critical roles in breast cancer development (Additional file 1: Fig. S12). Also important are TNFRSF17, TNFRSF18, TNFRSF4, members of the Tumor Necrosis Factor Receptor superfamily that bind to various TRAF family members and can regulate tumor cell proliferation and death [78]. Moreover, from the functional enrichment analysis using DAVID (the Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/home.jsp), the identified genes participate in several signaling pathways such as Jak-STAT, mTOR, MAPK, and Wnt signaling. Studies on Jak-STAT signaling pathway and mTOR signaling pathway have elucidated their roles in various cellular processes

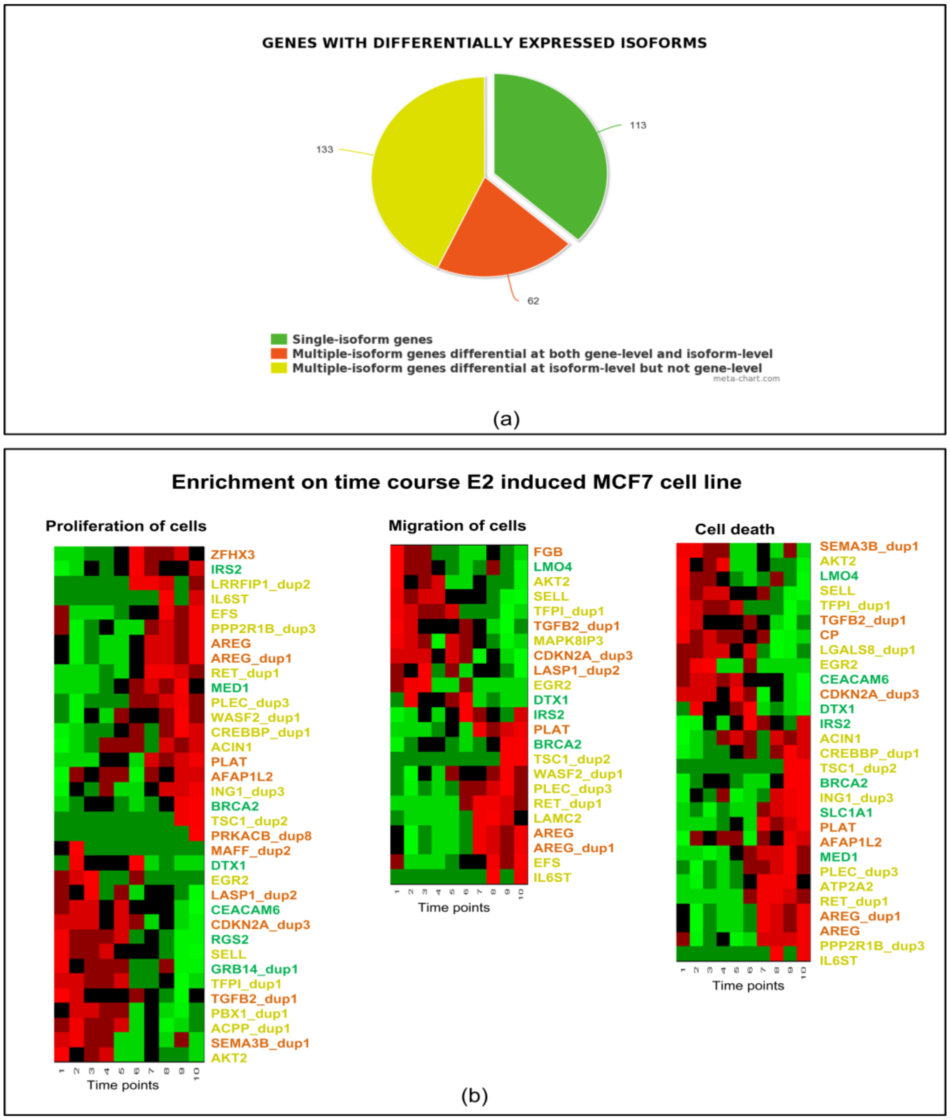


Figure 2.16 Enrichment analysis of the identified differentially expressed isoforms overlapped with PPI network. (a) The identified genes are categorized as single-isoform genes (genes with only one isoform) and multiple-isoform genes (genes with multiple isoforms). The multiple-isoform genes are further divided into two groups: differential at both gene-level and isoform-level, differential at the isoform level only. (b) Heatmaps of genes associated with proliferation of cells, migration of cells, and cell death, showing expression pattern change in a time-course E2 induced MCF-7 cell line data. The gene symbols of the heatmaps are color-coded according to the grouping in (a).

such as proliferation, apoptosis, and migration, that can contribute to malignancy [79, 80]. Many genes associated with the signaling pathways are only differential at the isoform level yet non-differential from gene-level analysis, such as PDPK1, TSC1, TSC2, PIK3R2, AKT2 in the mTOR signaling pathway and HSP90AA1, HSP90AB1 in the PI3K/AKT signaling pathway. Thus, isoform-level differential analysis is much needed to provide critical information for revealing biological mechanisms associated with cancer recurrence. While HSP90AA1 has two isoforms from alternative splicing, only NM_005348 (RefSeq_id) was overexpressed in the 'Dead' group. HSP90AB1 has five isoforms, among which NM_007355 was detected as overexpressed in the 'Dead' group whereas NM_001271971 was overexpressed in the 'Alive' group (Figure 2.17). HSP90AA1 and HSP90AB1 are Heat Shock Proteins (HSPs) that play an important role in tumorigenesis [81, 82]. The overexpression of HSP90AA1 and HSP90AB1 leads to activation of cell viability of tumor cell lines and provides an escape mechanism for cancer cell from apoptosis. Functional analysis using IPA has shown that the down-regulation of HSP90AB1 leads to activation of cell death of immune cells [83]. Collectively, these findings suggest that the change in differential expression pattern of the isoforms might contribute to different functions of cancer development.

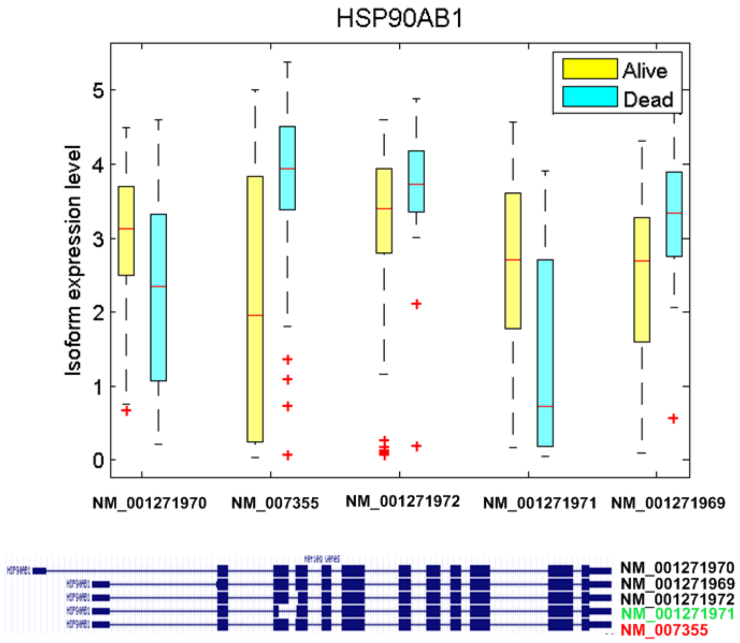


Figure 2.17 Estimated abundance of isoforms of HSP90AB1.

2.4 Discussion and Conclusion

We have developed a Bayesian approach, BayesIso, for the identification of differentially expressed isoforms. A hierarchical model, with differential states as hidden variables, is devised to account for both between-sample variability and within-sample variability. Specifically, a Poisson-Lognormal model is used to model the within-sample variability specific to each transcript. The expression level of transcripts is modeled to follow a Gamma distribution so as to capture the between-sample variability, including both over-dispersion and under-dispersion, by the model parameters. The shape parameter of the Gamma distribution is further assumed to follow a second Gamma distribution. Differential states of the transcripts are embedded into the Gamma-Gamma model as hidden variables, affecting the distribution of transcript expressions in each group or condition.

The main advantages of our proposed method, BayesIso, can be summarized as follows. First, it is a fully probabilistic approach that estimates the differential states and other model parameters iteratively. Assuredly accurate estimation of transcript expression is critical for differential state detection by comparing two groups of samples. At the same time, accurate detection of differential states benefits the estimation of transcript expressions, especially when the variance of transcript expressions among the samples (i.e., the between-sample variability) is high. Therefore transcript expressions and differential states are tightly coupled when performing differential analysis of groups of samples. In the existing two-step approaches on isoform-level differential analysis (e.g., Cuffdiff 2 and Ballgown), the detection of differentially expressed isoforms is carried out after expression estimation, not in a cooperative manner. BayesIso uses a joint model that combines the variability modeling of transcript expression with the differential state of transcripts for differential analysis of isoforms; the posterior probability of differential state is estimated jointly with other model parameters capturing expression variability through an MCMC sampling procedure.

Second, the differential state is embedded into the hierarchical model in a probabilistic way, in contrast to that conventional statistical tests calculate the significance of the difference between two groups of samples. Specifically, in BayesIso, differential state is measured by the probability that the transcript expressions in two groups of samples come from two distributions rather than from the same distribution. The probabilistic formulation makes it possible to embed the differential state into the joint modeling of expressions of two groups of samples; the probability of differential state estimated jointly with other model parameters, in turn, provides an improved performance in detecting isoforms of less differentially expressed.

Third, the Poisson-Lognormal model in BayesIso provides a flexible way to model the various biases within each transcript. Count-based methods for gene-level differential analysis (DESeq, edgeR, DSS, EBSeq) do not consider within-sample variability as the read count of each locus is used as the input. Cuffdiff 2 has tried to account for the bias along genomic location with a few known sources. Specific models are built for positional and sequence-specific biases, respectively. The parameters of the bias models are estimated in a global perspective, i.e., using all of the genes or a set of genes. However, the variability of real RNA-seq data may be much more complicated than that from known sources. The globally estimated parameters of the bias models are insufficient to account for the difference in biases among transcripts. In BayesIso, the variability of each exon/segment of each transcript is captured by one model parameter in the Poisson-Lognormal distribution, rather than estimated globally. Therefore, BayesIso models the within-sample variability specific to each transcript without relying on known sources.

We have compared the performance of BayesIso with both count-based methods (DESeq, edgeR, DSS, EBSeq) and isoform-level differential analysis methods (Cuffdiff 2 and Ballgown) using extensive simulation studies. BayesIso consistently outperforms the other methods in all of the scenarios studied, with AUC of ROC curve and F-score as the metrics to evaluate the overall performance in differential isoform identification. Cuffdiff 2 is very conservative as it can detect a few differentially expressed isoforms at high precision but missing a lot of others. Ballgown and the count-based methods can detect more differential isoforms at the cost of precision. BayesIso has significantly increased the performance in detecting true differential isoforms. From further experiments on isoforms with various differential levels, BayesIso has been demonstrated to be more effective in detecting isoforms of less differentially expressed. The improved performance is gained from iteratively estimating the differential state and the parameters of the joint modeling of both types of expression variability. At the same time, the various biases of the transcripts are well captured, leading to more accurate estimation of transcript expression. Overall, the hierarchical Bayesian framework contributes to the improved performance of BayesIso. The superiority of BayesIso comparing to the other methods has further been demonstrated using two real RNA-seq datasets with benchmarks.

We have applied BayesIso to breast cancer RNA-seq data to identify differentially expressed isoforms associated with breast cancer recurrence. The diverse bias patterns along transcripts and the generally low differential level have been observed from the real breast cancer data, indicating their importance in differential analysis of RNA-seq data. The differentially expressed isoforms detected by BayesIso are enriched in cell proliferation, apoptosis, and migration, uncovering the mechanism related to breast cancer recurrence. Moreover, the unique set of differential genes identified by BayesIso has helped reveal several signaling pathways such as the PI3K/AKT/mTOR signaling and PTEN signaling pathways. The identified down-regulated genes in the early recurrence group, e.g., NFATC1, participate in the immune system, which may indicate the role of immune system in breast cancer recurrence.

As a final note, it is a non-trivial task to model the sequencing bias for RNA-seq data analysis. The bias patterns are complicated and cannot be well explained by known sources. In the BayesIso method, we have used a flexible model to account for the bias independent of any particular pattern. However, we have also observed that certain bias patterns (such as bias to the 3’ end, or high in the middle) occur more frequently than others. Moreover, we have further observed that the bias patterns may be affected by the expression level. In the future work, we will incorporate certain bias patterns as prior knowledge into the model, which can help estimate the bias pattern of some isoforms more accurately hence to improve the performance on differential analysis of isoforms.

3 Differential methylation detection using a hierarchical Bayesian model exploiting local dependency

3.1 Introduction

DNA methylation [15] is a molecular modification of DNA, which typically occurs at the cytosine nucleotides in CpG sites. The CpG sites, shorted for '-C-phosphate-G-', are regions of DNA where a cytosine nucleotide is linked to a guanine nucleotide by one phosphate. As the most well-studied epigenetic mark, DNA methylation has been demonstrated to play a crucial role in regulating gene expression without alterations in the DNA sequence [15]. Although the underlying mechanism is still not completely known, DNA methylation is essential for cell differentiation and it is associated with various key biological processes such as embryonic development and genomic imprinting [17]. Besides its important role in normal cell development, recent studies show that DNA methylation abnormalities are associated with various diseases including cancer [26, 27]. There is strong evidence that tumor-suppressor genes may be silenced because of hypermethylation; growth-promoter genes may be activated due to hypomethylation, consequently inducing cancer development [28]. Therefore, the identification of abnormalities in DNA methylation is of increasing interests in the field of cancer research. Moreover, DNA methylation is heritable and reversible [84], which makes it a promising target for new therapeutic approaches in cancer treatment [20].

In the past decade, the development of high-throughput technologies provides the opportunity to obtain whole genome-wide DNA methylation mapping with high resolution. Let us take the Illumina Infinium HumanMethylation450 BeadChip Kit (Illumina 450k) as an example, which is one of the most popular, high-quality, cost-effective techniques for DNA methylation study. Illumina 450k measures >485,000 CpG sites per sample at single-nucleotide resolution, which covers 99% of RefSeq genes with multiple sites in the functional regions, such as promoter, 5'UTR, 1st exon, gene body, and 3'UTR. The high coverage and low cost of the

Illumina 450k array [31, 32] make it a very powerful platform for exploring genome-wide DNA methylation landscape. By virtue of the high-throughput techniques, the methylation level of each gene is measured at multiple CpG sites across the genomic location, providing more comprehensive measurements for a methylation event.

Despite the advantage of high-throughput profiling, the high resolution poses challenges to computational analysis for detecting differentially methylated genes from the huge number of measured CpG sites. Early approaches attempted to identify differentially methylated sites by statistic tests. However, the statistical power is limited due to the problem of multiple hypothesis testing; moreover, it is biologically difficult to interpret individual CpG sites without considering the neighbors. Thus, the detection of differentially methylated regions (DMRs) is of prime interest, and several methods have been proposed, falling into two categories: annotation-based methods and *de novo* methods. In the annotation-based methods, the regions are predefined according to the annotation of CpG site location. IMA [85] is a well-known annotation based pipeline, which first generates an index of the methylation value of predefined regions, and then uses statistical tests, such as limma [86], to identify differentially methylated regions. The regions can be predefined as genes, genes' promoters, CpG islands, etc. The index of the methylation value of a region is derived from the methylation value of the involved CpG sites with metrics such as mean, median, and so on. As an alternative, *de novo* methods do not rely on predefined regions for DMR detection. Bumphunter [87] first estimates the association between the methylation level and the phenotypes for each site and then identifies DMRs after a smoothing operation. DMRcate [88] is another approach agnostic to predefined regions. It first calculates a statistic from differential test for each site and then uses a Gaussian kernel to incorporate the neighboring information for DMR detection. Comb-P [89] combines spatially assigned p-values to find regions of enrichment. Probe Lasso [90] is a window-based approach that detects DMRs using neighboring significant-signals. The region-based methods have demonstrated their capability in detecting biologically meaningful differential methylation events. However, most of the existing DMR detection methods are based on statistic tests, and the neighboring information is not jointly considered when estimating the methylation change of CpG sites.

In this work, we develop a novel method, **D**ifferential **M**ethylation detection using a hierarchical **B**ayesian model exploiting **L**ocal **D**ependency (DM-BLD), to identify differentially methylated genes based on a Bayesian framework. In DM-BLD, CpG sites are first mapped or linked to genes according to their location information. For each gene, we then use a Gaussian Markov random field (GMRF) model, Leroux conditional autoregressive (CAR) structure [91], to capture the varying degree of dependency among nearby CpG sites. Based on the local dependency, it is reasoned that genes involving a sequence of CpG sites with methylation change are more likely to exhibit abnormal methylation activity. We use a discrete Markov random field (DMRF) [92] to model the dependency of methylation changes (differential states) of neighboring CpG sites. A hierarchical Bayesian model is developed to fully take into account the local dependency for differential analysis, in which differential states are embedded as hidden variables. A Gibbs sample procedure is then developed to estimate the methylation change of CpG sites jointly with other model parameters based on their conditional distributions, respectively. As a next step, the differential methylation scores of the genes are calculated from the estimated methylation changes of the involved CpG sites. Permutation-based statistical tests are designed to assess the significance of the detected differentially methylated genes. The proposed DM-BLD approach is a fully probabilistic approach with a hierarchical Bayesian model to account for the local dependency of CpG sites in both methylation level and differential state, capable of detecting methylated genes of less differential accurately and effectively.

3.2 Method

3.2.1 Spatial correlation among CpG sites in a neighborhood

Both Beta-value and M-value statistics are conventionally used as metrics to measure the methylation levels of CpG sites profiled by methylation microarray platforms. Beta-value is the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities) [93]; while M-value is the log2 ratio of the methylated probe intensity versus unmethylated probe intensity [94]. Thus, the relationship between Beta-value and M-value is a Logit transformation defined by Eq. (3.1) [95],

$$Beta_i = \frac{2^{M_i}}{2^{M_i}+1};\ M_i = \log_2\left(\frac{Beta_i}{1-Beta_i}\right), \tag{3.1}$$

where $Beta_i$ and M_i are the Beta-value and M-value of CpG site *i*, respectively. As shown in [95], the standard deviation of M-value is approximately consistent, which makes it much more appropriate for the homoscedastic assumptions of most statistical models used for microarray analysis. Therefore, we model the distribution of M-value for differential methylation identification.

Rather than being independent, the methylation of CpG sites located in a neighboring region along the genome induces a spatially dependent structure. It has been shown in the literature that the methylation levels of CpG sites within 1,000 bases are significantly correlated [96]. We studied the dependency among neighboring CpG sites using three data sets: cell line data, our in-house human data, and TCGA breast cancer data. The cell line data set consists of two cell lines (LCC1 and LCC9), with three biological replicates in each cell line; the in-house human data set consists of samples from two phenotypes, with 6 samples and 5 samples from each group, respectively; the TCGA breast cancer data set consists of 61 estrogen receptor (ER)-positive breast cancer samples divided into two groups, with 41 samples and 20 samples from each group. All of the three data sets were profiled by Illumina 450K, which measured 485,512 CpG sites covering 21,227 genes. The median of the number of CpG sites in each gene is 15, and the median distance between two consecutive CpG sites is about 300 bps. We used correlation coefficient as the metric, and compared the correlation of CpG sites in the following three scenarios: 1) randomly selected CpG sites within 1,000 bases; 2) randomly selected CpG site and its 50 closest neighboring sites; 3) randomly selected CpG sites across the genome. Figure 3.1 shows the correlation of CpG sites in the three scenarios using the three data sets. We can see that the correlation between CpG sites within 1,000 bases was significantly higher than that between randomly selected CpG sites; in all of the three data sets, a statistical significance (p-value) less than 10^{-89} was obtained as calculated from one-sided two-sample Kolmogorov-

Smirnov (K-S) test, which indicates the substantial spatial correlation among CpG sites in a neighboring region. Note that the correlation level observed from the TCGA data set (of human tumor samples) was not as high as that of the cell line data set, which might be caused by much more complex factors such as high level of variability in replicates. Therefore, it is important to incorporate the spatial correlation information for estimating the methylation levels of CpG sites hence to improve the accuracy in identifying differentially methylated genes. In the next subsection, we will describe the framework of our DM-BLD approach taking into account the intrinsic local dependency among CpG sites.

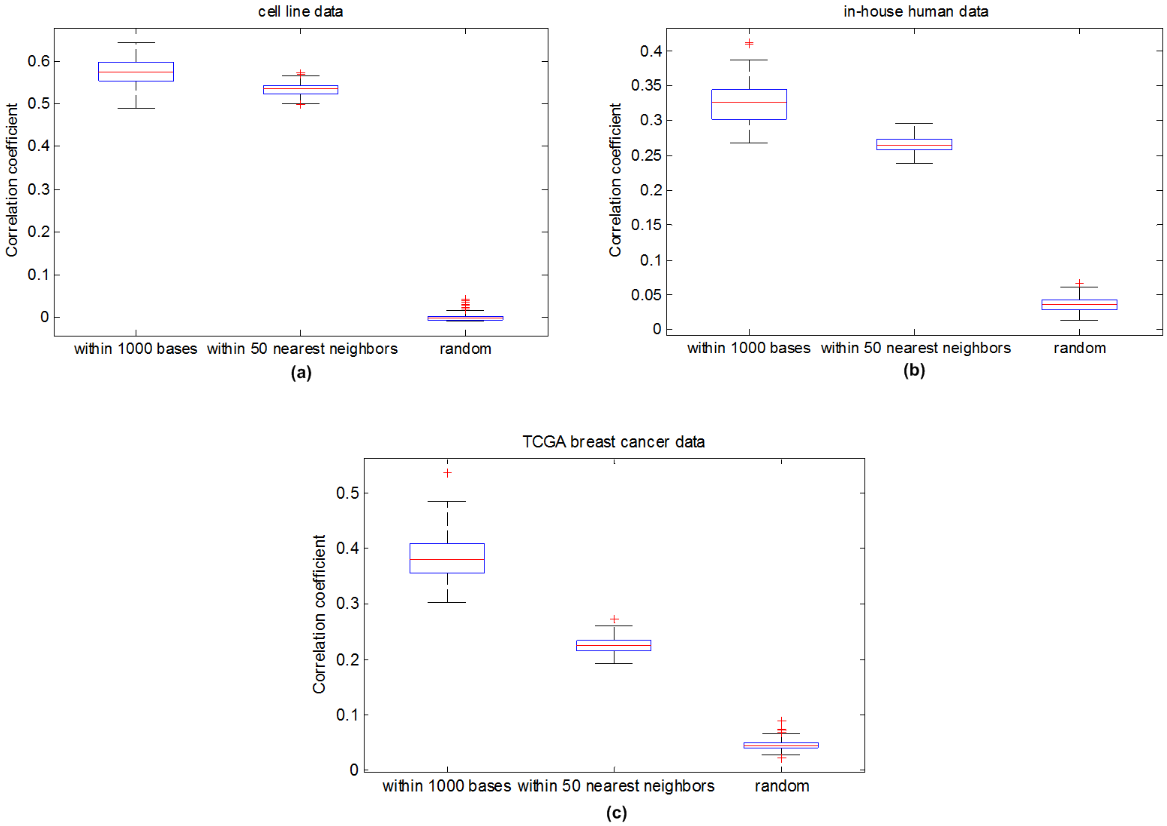


Figure 3.1 Correlation of CpG sites calculated from: (a) cell line data; (b) in-house human data; (c) TCGA breast cancer data.

3.2.2 Framework of MRF-based Bayesian model

Based on the observation that nearby CpG sites are significantly correlated, we propose to develop a probabilistic model incorporating the local dependency of CpG sites to identify differentially methylated genes. In our proposed method, CpG sites are first mapped to the genes according to their genomic location. As is provided by the annotation file of the Illumina 450k microarray platform, the probes/CpG sites are associated with RefSeq genes of the reference genome hg19. For each gene, all of the CpG sites located within 1500bp from TSS to 3'UTR are used for differential analysis. Three major steps of the proposed method, DM-BLD, are summarized as follows (see Figure 3.2): (1) within each gene, estimating the methylation level of CpG sites by modeling the local spatial correlation of methylation level and the dependency of methylation change among neighboring CpG sites; (2) calculating the differential methylation score of genes from the estimated methylation change of CpG sites; (3) performing permutation-based significance tests on potential differentially methylated genes. Specifically, in the first step, we use the Leroux model [91], an advanced conditional autoregressive (CAR) structure, to capture the dependency of methylation among neighboring CpG sites. Comparing with the intrinsic conditional autoregressive (ICAR) model [97], the Leroux model is capable of accounting for different levels of correlation [98, 99], which helps improve the accuracy in estimating the methylation level of CpG sites. We then use a discrete Markov random field [92] to model the dependency of methylation changes (via differential states) of neighboring CpG sites. With differential states embedded as hidden variables, we use a hierarchical Bayesian model to take into account the local dependency fully for differential analysis. A Gibbs sampling procedure, based on conditional distributions, is designed to estimate the methylation level and other model parameters. In the second step, the differential methylation score of a gene is calculated from the estimated methylation change of involved CpG sites. Genes can be prioritized according to the differential methylation score for further biological validation. Finally, in the third step, permutation-based hypothesis tests are implemented and performed to assess the significance of the identified differentially methylated genes for real data analysis. More details of the three steps will be given in the following subsections.

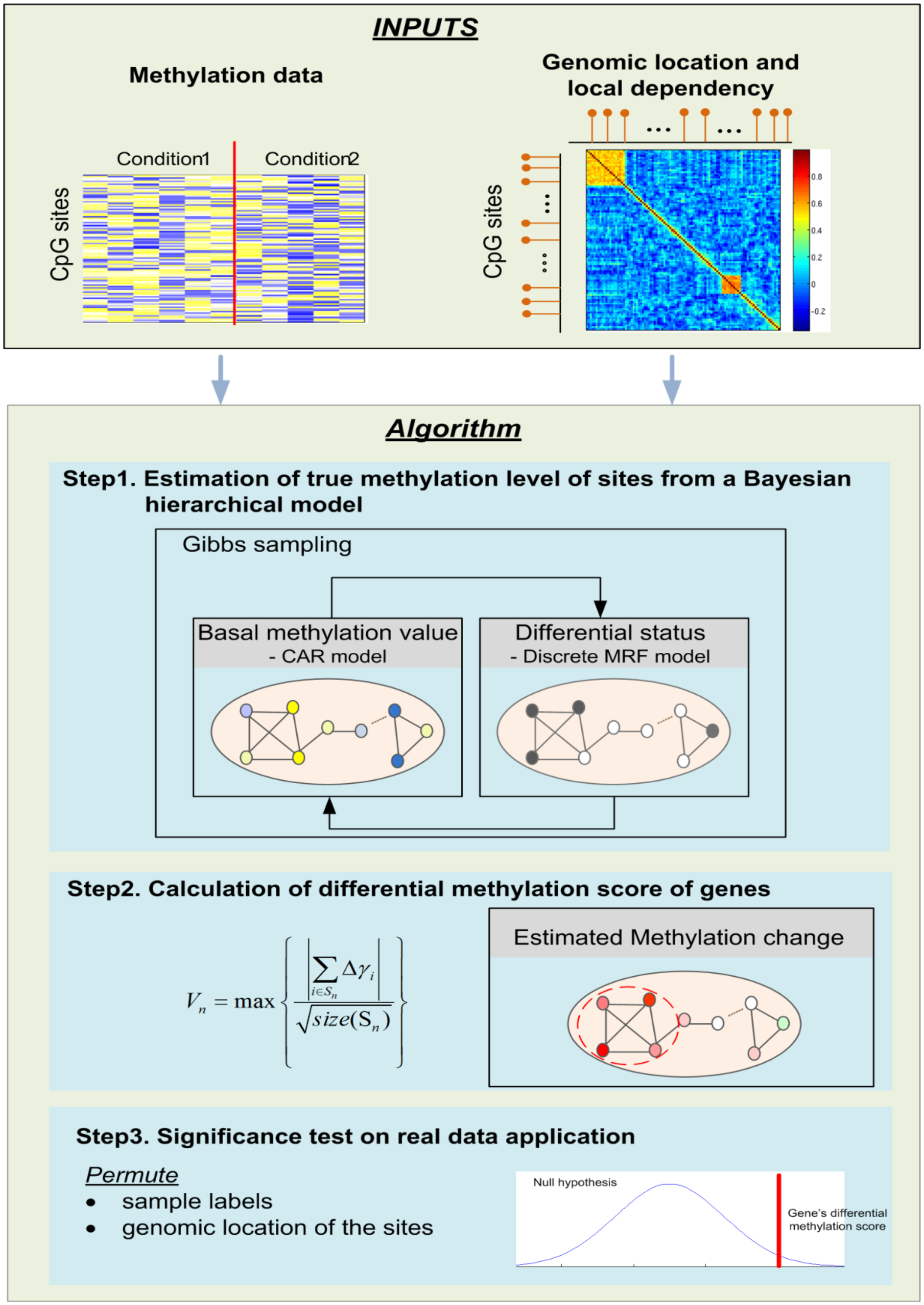

INPUTS
Methylation data
Condition1
Condition2
CpG sites
Genomic location and local dependency
CpG sites
0.8
0.6
0.4
0.2
0
-0.2
Algorithm
Step1. Estimation of true methylation level of sites from a Bayesian hierarchical model
Gibbs sampling
Basal methylation value
- CAR model
Differential status
- Discrete MRF model
Step2. Calculation of differential methylation score of genes
$V_n = \max \left\{ \frac{\left| \sum_{i \in S_n} \Delta \gamma_i \right|}{\sqrt{size(S_n)}} \right\}$
Estimated Methylation change
Step3. Significance test on real data application
Permute
• sample labels
• genomic location of the sites
Null hypothesis
Gene's differential methylation score

Figure 3.2 Flowchart of the proposed DM-BLD approach. DM-BLD consists of the following three major steps: (1) methylation level estimation; (2) differential methylation score calculation; (3) permutation-based significance tests.

3.2.3 Hierarchical Bayesian model exploiting the local dependency of sites

We first estimate the methylation level of CpG sites by taking into account the neighboring CpG sites. Beta-value is the ratio of the methylated probe intensity and the overall intensity [31], and thus, it is represented as a proportion value bounded by zero and one, which can be modeled by a logit-normal distribution [100]. Thus, as the logit transform of Beta-value, M-value can be modeled as normal-distributed.

Assume that there are N genes and gene n has M_n CpG sites. Let us denote $Beta_{i,j}$ as the beta-value of the i^{th} ($i = 1, 2, ..., M_n$) CpG site of gene n ($n = 1, 2, ..., N$) in sample j ($1 \le j \le J$), which follows a logit-normal distribution. $J = J_1 + J_2$ is the total number of samples associated with two biological phenotypes (or conditions), where J_1 and J_2 are the numbers of samples for phenotype 1 and phenotype 2, respectively. For gene n, denote $y_{i.j}$ as the logit transform of $Beta_{i,j}$, as shown in Eq. (3-2). $y_{i.j}$ (also called M-value) follows a normal distribution with mean γ_i and precision τ_e (Eq. (3-3)). γ_i (as defined by Eq. (3-4)) represents the true methylation level of CpG site i in gene n, where θ_i represents the basal methylation level of CpG site i, while μ_0 represents the methylation level change of gene n between two conditions. d_i is a binary value representing the differential state of site i. If site i is differentially methylated, $d_i = 1$; otherwise, $d_i = 0$. The above-mentioned equations are listed as follows:

$$y_{i,j} = \log\left(\frac{Beta_{i,j}}{1 - Beta_{i,j}}\right); \qquad (3\text{-}2)$$

$$y_{i,j} \sim N\left(\gamma_i, {}^{1}\!/_{\tau_e}\right); \qquad (3\text{-}3)$$

$$\gamma_i^{(1)} = \mu^{(1)} + \theta_i, \text{ and } \gamma_i^{(2)} = \mu^{(2)} + \theta_i, \qquad (3\text{-}4)$$

where

$$\mu^{(1)} = \mu^{(2)} = 0 \text{, if } d_i = 0 \text{;}$$

$$\mu^{(1)} = 0; \mu^{(2)} = \mu_0 \text{, if } d_i = 1 \text{.}$$

Thus, for non-differentially methylated CpG sites, the methylation levels under two phenotypes are the same, $\gamma^{(1)} = \gamma^{(2)} = \theta$; for differentially methylated CpG sites, $\gamma^{(2)} = \gamma^{(1)} + \mu_0 = \theta + \mu_0$.

As observed from real DNA methylation data sets, the methylation level of CpG sites display spatial dependency. For each gene, we adopt the conditional autoregressive model used in Leroux's approach [91], termed as Leroux model, to account for the spatial dependency of methylation level of neighboring CpG sites. Specifically, Leroux model is used to specify the between-site correlation of $\boldsymbol{\theta} = [\theta_i, i = 1, 2, ..., M_n]$, where the methylation level of a CpG site depends on that of its neighbors but is independent of that of all other CpG sites. Denote ∂i as the set of the neighboring sites of CpG site i . Under the Leroux model, the conditional distribution of θ_i given $\theta_{\partial i}$ is defined by

$$\theta_i \mid \theta_{\partial i}, \rho, \tau \sim \mathrm{N}\left(\frac{\rho \sum_{k=1}^{M_n} w_{k,i} \theta_k}{\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho}, \ \frac{1}{\tau(\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho)} \right), \tag{3-5}$$

where ρ controls the dependency level among the nearby CpG sites and τ controls the variance. $\mathbf{w}$ is a predefined design matrix for the neighborhood structure. $w_{i,j} = 1$, if CpG site i and CpG site j locate within a neighborhood; $w_{i,j} = 0$, otherwise. Comparing to other conditional autoregressive models regarding the dependency information, the Leroux model has some advantages. Comparing to intrinsic model [101] that captures strong dependency of neighbors, the Leroux model is capable of accounting for different levels of dependency with parameter ρ [99]. ρ close to 1 indicates strong dependency; the CpG sites are independent when $\rho = 0$. Convolution model [101] and Cressie model [102] also consider different levels of dependency.

However, the conditional variance in the Leroux model is more theoretically appealing since the conditional variance is influenced by dependency level. When there is a strong dependency, that is, ρ close to 1, the conditional variance is close to $1\Big/\tau\sum_{k=1}^{M_n} w_{k,i}$; when there is no dependency, that is, $\rho = 0$, the conditional variance is $1/\tau$, which is no longer affected by the neighboring information.

Differential state $\mathbf{d} = [d_i, i = 1, 2, ..., M_n]$ is modeled by a discrete Markov random field (DMRF), which can be defined by the following equation [92, 103]:

$$p\left(d_i = 1 \mid d_{\partial i}, a, b\right) = \frac{\exp\left(a + b\left(n_i(1) - n_i(0)\right)\Big/\sum_{k=1}^{M_n} w_{k,i}\right)}{1 + \exp\left(a + b\left(n_i(1) - n_i(0)\right)\Big/\sum_{k=1}^{M_n} w_{k,i}\right)}, \tag{3-6}$$

where

$$n_i(1) = \sum_{k=1}^{M_n}\left(w_{k,i} d_k\right), \text{ and } n_i(0) = \sum_{k=1}^{M_n}\left(w_{k,i}\left(1 - d_k\right)\right).$$

In Eq. (3-6), a and b are model parameters. Parameter b controls the consistency of differential state in DMRF. The larger b is, the more consistent the differential state is in a neighborhood. Note that in our implementation, we set $a = 0$ and $b = 3$ as default values for DMRF to control the neighborhood consistency of differential state. $n_i\left(1\right)$ ($n_i\left(0\right)$) is the number of neighboring CpG sites of site i with differential state 1 (0). More details on the equivalence between the discrete Markov random field model and Gibbs random field are introduced in Appendix B.

For differential methylation analysis, we devise a Bayesian approach to estimate the methylation levels ($\gamma^{(1)}$ and $\gamma^{(2)}$) in two phenotypes, respectively. A hierarchical Bayesian model is used to model the dependency of random variables $\boldsymbol{\theta}$ (basal methylation level), μ_0

(methylation level change if differential) and **d** (differential state), which fully determine the methylation levels ($\gamma^{(1)}$ and $\gamma^{(2)}$). The dependency graph of the model is shown in Figure 3.3, where the differential states are embedded as hidden variables in the model. According to Bayes' rule, the joint posterior distribution is given by

$$\begin{aligned}&P\left(\boldsymbol{\theta},\mathbf{d},\mu_0,\tau_e,\rho,\tau\,|\,\mathbf{y}\right)\\&\sim P\left(\mathbf{y}\,|\,\boldsymbol{\theta},\mathbf{d},\mu_0,\tau_e,\rho,\tau\right)\times P\left(\boldsymbol{\theta}\,|\,\rho,\tau\right)\times P\left(\rho\right)\times P\left(\tau\right)\times P\left(\mathbf{d}\right)\times P\left(\tau_e\right)\times P\left(\mu_0\right)\end{aligned}, \tag{3-7}$$

where τ_e, τ and ρ are termed model parameters. For mathematical convenience, we further assume conjugate prior distributions [104] for model parameters τ_e, τ and variable μ_0, and discrete uniform distribution as the prior distribution of parameter ρ. The prior distributions (set as non-informative with hyper-parameters α, β, τ_0, and $\mathbf{q}=[q_i, i=1,2,...,r]$) are defined by the following equations:

$$\tau_e,\tau \sim Gamma(\alpha,\beta)\,;\;\; \rho \sim \text{discrete uniform}(\mathrm{q}_1,...,\mathrm{q}_r)\,; \tag{3-8}$$

$$\mu_0 \sim \mathrm{N}(0,\tau_0)\,. \tag{3-9}$$

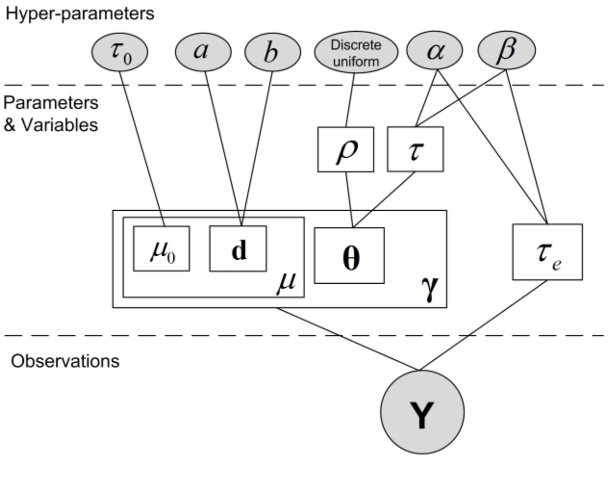


Figure 3.3 Dependency graph of the hierarchical Bayesian model in DM-BLD.

3.2.4 Estimation of methylation change of CpG sites via Gibbs sampling

Due to the complexity of the probabilistic model, we have developed a Markov Chain Monte Carlo (MCMC) method to jointly estimate the variables ($\boldsymbol{\theta}$, μ_0 and $\mathbf{d}$) and model parameters (τ_e, τ and ρ). In particular, we use Gibbs sampling to iteratively draw samples from the conditional distributions of the model variables/parameters. By virtue of the sampling process, the marginal posterior distribution can be approximated by the samples drawn. The conditional posterior distributions of the model variables/parameters can be derived as follows.

The conditional posterior distribution of the basal methylation level θ_i of CpG site *i* is

$$
\begin{aligned}
&P\left(\theta_i \mid \mathbf{y}_i, \theta_{\partial i}, \mu_0, d_i, \rho, \tau, \tau_e\right) \sim P\left(\mathbf{y}_i \mid \theta_i, \mu_0, d_i, \tau_e\right) \times P\left(\theta_i \mid \theta_{\partial i}, \rho, \tau\right) \\
&\sim \mathrm{N}\left(\frac{\tau\rho \sum_{k=1}^{M_n} w_{k,i}\theta_k + \tau_e \left(\sum_{j=1}^{J_1} y_{i,j} + \sum_{j=1}^{J_2} \left(y_{i,j} - d_i \mu_0 \right) \right)}{\tau \left(\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho \right) + J\tau_e}, \frac{1}{\tau(\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho) + J\tau_e} \right)
\end{aligned}
\tag{3-10}
$$

The conditional posterior distribution of parameter τ in the Leroux model is

$$
\begin{aligned}
&P\left(\tau \mid \boldsymbol{\theta}, \rho\right) \sim P\left(\boldsymbol{\theta} \mid \tau, \rho\right) P\left(\tau\right) \\
&\sim \mathrm{Gamma}\left(\alpha + \frac{M_n}{2}, \beta + \frac{\sum_{i=1}^{M_n} \left(\left(\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho \right) \left(\theta_i - \frac{\rho \sum_{k=1}^{M_n} w_{k,i}\theta_k}{\rho \sum_{k=1}^{M_n} w_{k,i} + 1 - \rho} \right)^2 \right)}{2} \right)
\end{aligned}
\tag{3-11}
$$

The conditional posterior distribution of parameter ρ in the Leroux model is

$$
P\left(\rho \mid \boldsymbol{\theta}, \tau\right) \sim P\left(\boldsymbol{\theta} \mid \tau, \rho\right) P\left(\rho\right) \tag{3-12}
$$

The conditional posterior distribution of parameter τ_e, which controls the noise level of replicates, is

$$\begin{aligned} P\left(\tau_e \mid \mathbf{y}, \mu_0, \mathbf{d}, \boldsymbol{\theta}\right) &\sim P\left(\mathbf{y} \mid \mu_0, \mathbf{d}, \boldsymbol{\theta}, \tau_e\right) P\left(\tau_e\right) \\ &\sim \text{Gamma}\left(\alpha + \frac{J \times M_n}{2}, \beta + \frac{\sum_{i=1}^{M_n}\left(\sum_{j=1}^{J_1}\left(y_{i,j} - \theta_i\right)^2 + \sum_{j=1}^{J_2}\left(y_{i,j} - \theta_i - d_i\mu_0\right)^2 \right)}{2} \right) \end{aligned} \tag{3-13}$$

The conditional posterior distribution of differential state d_i of site *i* is

$$\begin{aligned} P\left(d_i \mid \mathbf{y}_i, \mu_0, \theta_i, \tau_e, d_{\partial i}\right) &\sim P\left(\mathbf{y}_i \mid \theta_i, d_i, \mu_0, \tau_e\right) \times P\left(d_i \mid d_{\partial i}\right) \\ &\sim \prod_{j=1}^{J_2} \mathrm{N}\left(y_{i,j} \middle| d_i\mu_0 + \theta_i, \frac{1}{\tau_e} \right) \times P\left(d_i \mid d_{\partial i}\right) \end{aligned} \tag{3-14}$$

The conditional posterior distribution of parameter μ_0, which controls the methylation change between two phenotypes, is

$$\begin{aligned} P\left(\mu_0 \mid \mathbf{y}, \mathbf{d}, \boldsymbol{\theta}, \tau_e\right) &\sim P\left(\mathbf{y} \mid \boldsymbol{\theta}, \mu_0, \tau_e, \mathbf{d}\right) P(\mu_0) \\ &\sim \mathrm{N}\left(\frac{\tau_e \sum_{i=1}^{M_n}\sum_{j=1}^{J_2}\left(d_i \times \left(y_{i,j} - \theta_i\right)\right)}{\tau_0 + \tau_e \times J_2 \sum_{i=1}^{M_n} d_i}, \frac{1}{\tau_0 + \tau_e \times J_2 \sum_{i=1}^{M_n} d_i} \right) \end{aligned} \tag{3-15}$$

With the derived conditional distributions, we develop a Gibbs sampling method for parameter/variable estimation. Samples for the variables ($\boldsymbol{\theta}$, μ_0 and $\mathbf{d}$) and model parameters (τ_e, τ and ρ) are drawn iteratively from their conditional distributions. Specifically, samples for variables $\boldsymbol{\theta}$, μ_0 and parameters τ_e, τ are randomly drawn from the corresponding Gaussian or Gamma distribution. For variable $\mathbf{d}$ and parameter ρ, the conditional posterior distributions do not have closed form. Since $\mathbf{d}$ and ρ are finite discrete values, the corresponding conditional probabilities are calculated first, and then new samples of $\mathbf{d}$ and ρ are randomly selected according to the probabilities. Eventually, the Gibbs sampler produces Markov chains of samples of the parameters/variables, from which the estimates of the parameters/variables can be obtained from their marginal distributions. The Gibbs sampling procedure can be summarized as follows:

INPUT: methylation data $\mathbf{y}$, neighborhood structure $\mathbf{w}$, number of iterations N

OUTPUT: Estimates of true methylation level in each group and other parameters in the probabilistic model

Algorithm:

Step 1. Initialization: each parameter is set an arbitrary value and non-informative prior knowledge is used for the parameters

Step 2. Draw samples iteratively from conditional distributions of the parameters using Gibbs sampling:

Sample $\boldsymbol{\theta}$ from a Gaussian distribution;

Sample τ and τ_e from the corresponding Gamma distribution;

Sample discrete variable $\mathbf{d}$ and ρ by first calculating the conditional probabilities and then randomly generating new samples according to the probabilities;

Sample μ_0 from a Gaussian distribution.

Step 3. Estimate true methylation level $\boldsymbol{\gamma}$ as well as all other model parameters from the samples (after the burn-in period) generated from the sampling procedure. Then, for each CpG site, the estimated methylation change is calculated by $\Delta\hat{\gamma} = \hat{\gamma}_i^{(2)} - \hat{\gamma}_i^{(1)}$, which will be used in the next step to calculate the differential methylation score of genes.

Gibbs sampling, as a stochastic sampling technique, can provide an improved ability for the DM-BLD algorithm to escape local optima in which deterministic approaches (such as the expectation-maximization (EM) algorithm [105]) may get trapped. Gibbs sampling is an iterative procedure with one model parameter sampled at a time according to its conditional distribution given the other parameters. Gibbs sampling can be understood as a stochastic analogy of EM [106], a well-known method for finding maximum likelihood (ML) or maximum a posteriori

(MAP) estimates of the parameters. EM is a deterministic algorithm (that can be trapped in local optima), i.e., starting with the same initial parameters will always converge to the same solution. One approach to alleviate this limitation is to start the EM algorithm multiple times from different initial parameters. Different from EM, Gibbs sampling is a stochastic algorithm, i.e., it may arrive at different solutions from the same initial parameters. Theoretically, Gibbs sampling has the ability to escape local optima via exploiting randomized searching to a much greater degree [106].

It is worth noting that Gibbs sampling does not guarantee to find the optimal solution to the parameter estimation problem (especially when running for a finite number of iterations). In order to alleviate the potential problem of being trapped in local optima, we have implemented the Gibbs sampling algorithm with an option of multiple runs in addition to one long run of sampling. Specifically, multiple independent runs of Gibbs sampling, with different initializations of the parameters and different random seeds, are implemented in the algorithm. In our DM-BLD software package, we provide an option for multiple independent runs to be used for Gibbs sampling. When using multiple runs, the distributions generated from the runs are checked in the algorithm. In particular, we conduct a fixed number of runs (e.g., five times) and check whether a specific number of different runs (e.g., three times) generate samples from the same distribution. If so, all of the samples from all runs are used for parameter estimation. If not, another set of fixed number of runs will be conducted continually. With the solutions from multiple runs, comparisons of the solutions can indicate whether a global, optimum solution is likely to have been achieved.

3.2.5 Calculation of the differential methylation score for each gene

We further assume that gene is more likely to have abnormal methylation activity if it involves a sequence of CpG sites with methylation change. Thus, a searching method is used to determine the region of the sequence of CpG sites with methylation change, and the differential methylation score of the detected region represents the differential level of the corresponding gene.

With the estimate $\Delta\hat{\boldsymbol{\gamma}} = \left[\Delta\hat{\gamma}_1, \Delta\hat{\gamma}_2, \cdots, \Delta\hat{\gamma}_{M_n}\right]$, the differential methylation score V_n of gene n is defined by Eq. (3-16), which contributes to highlighting the genes with more neighboring CpG sites with methylation change:

$$V_n = \max\left\{ \frac{\left|\sum_{i \in S_n} \Delta\hat{\gamma}_i\right|}{\sqrt{size(\mathrm{S}_n)}} \right\}, \tag{3-16}$$

where S_n denotes a subset of sequential CpG sites in a neighboring area within the genomic location of gene n. Finally, genes are ranked according to their differential methylation scores.

3.2.6 Significance test on differentially methylated genes

When applying our proposed DM-BLD approach to real data, the confidence of the identified genes is a critical problem. In order to assess the significance of the identified differentially methylated genes, we perform permutation-based significance tests. Specifically, we first rearrange the sample labels as well as the location of the CpG sites in 100 random trials and then perform DM-BLD on the perturbed methylation data. The permutation of sample label disrupts the association between samples and phenotypes; the permutation of the CpG site locations disrupts the dependency among neighboring CpG sites. We perform two significance tests over the 100 random trials as follows:

- In the first test, the observed (or estimated) differential methylation score of each gene was tested against the ‘global’ null distribution; the ‘global’ null distribution was estimated from the differential methylation scores of all the genes in consideration, as obtained with the 100 random trials. Note that the ‘global’ null distribution was the aggregated distribution calculated from all the genes, which can be used to assess whether the gene is randomly selected from all genes.

- In the second test, the observed (or estimated) differential methylation score of each gene was tested against its corresponding ‘local’ null distribution; the ‘local’ null distribution was estimated from the differential methylation score of the gene obtained in the 100 random trials. Note that the ‘local’ null distribution was gene-specific, i.e., each gene had its own null distribution, which is conventionally used to assess the significance of the differential level of gene comparing samples from two phenotypes.

In the significance test, the null hypothesis was that the observed methylation score was drawn from the null distribution, and the p-value for each gene was calculated by assuming the null distributions were Gaussian-distributed. Benjamini-Hochberg correction [107] was used to estimate the FDR-adjusted p-value. Using both adjusted p-values, we can assess the significance of gene in terms of differential methylation comparing two phenotypes as well as comparing to the whole gene set, which is consistent with Efron’s effort on testing the significance of a gene set using restandardization [108].

3.3 Simulation studies

To systematically evaluate the performance of DM-BLD, we simulated multiple DNA methylation data sets with different scenarios. In each experiment, the methylation values of all 450K probes were generated for 20 samples in two conditions, each with 10 samples. 30% out of the 20758 genes with CpG sites in the promoter region were randomly selected as true differentially methylated genes, half hypermethylated and half hypomethylated. For each differentially methylated gene, a promoter-associated region was randomly selected as differentially methylated. The neighborhood of each CpG site was defined as the CpG sites of both sides located within 1000bp from its location. The methylated value of the CpG sites was simulated in two different ways: the first one is based on the simulation scheme used in DMRcate; the second one is based on our proposed Leroux model. In each scenario, ten random experiments were performed to assess the variance of the performance measure.

We compared the performance of DM-BLD in detecting differentially methylated genes with six existing region-based approaches. We implemented the existing methods, and calculated p-values of the genes from the detected DMRs as follows:

- ***Student's t-test on mean value:*** the mean methylation value of all involved CpG sites was calculated as the methylation value of the gene; Student's t-test was used to for differential analysis.
- ***Student's t-test on median value:*** the median methylation value of all involved CpG sites was calculated as the methylation value of the gene; Student's t-test was used to for differential analysis.
- ***Bumphunter:*** default setting with 100 permutations, where the cutoff was determined from the 100 permutations at the default setting. We used the reported p-value of the area as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum p-value of the DMRs associated with (or mapped to) the gene.
- ***DMRcate:*** default settings with pcutoff = 1, lambda = 1000, and C = 2. We used the reported meanpval as the p-value of the detected DMRs and assigned a gene's p-value as the minimum p-value of the DMRs associated with the gene.
- ***Probe Lasso:*** default setting with adjPval = 1 and DMRpval = 1. We used the reported dmr.Pval as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum P value of the DMRs associated with the gene.
- ***comb-P:*** default setting with p-value from Limma as the input, with seed = 0.5. We used the reported z_sidak_p as the p-value of the detected DMRs, and assigned a gene's p-value as the minimum p-value of the overlapped DMRs mapped/linked to the gene.

For the genes with no DMRs detected, their p-values were set as 1.0. To evaluate the performance of the competing methods, genes were ranked by their differential methylation scores (DM-BLD) and their p-values (all six competing methods). Area-under-the precision-recall curve (AUCpr) was used to assess the overall performance on differentially methylated gene identification. Moreover, we calculated signal-to-noise ratio (SNR) to show the differential level of the generated simulation data in different scenarios. SNR measures the differential level of the CpG sites taking into account both the methylation difference and the variance of the data, calculated as follows:

$$SNR = 20\log_{10}\left(\frac{\left|\text{mean}(\mathbf{y}^{(1)}) - \text{mean}(\mathbf{y}^{(2)})\right|}{\sqrt{\text{var}(\mathbf{y}^{(1)}) + \text{var}(\mathbf{y}^{(2)})}}\right),$$

where $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are methylation value in the two conditions, respectively.

3.3.1. Selection of differentially methylated genes and DMRs

Among the RefSeq genes covered by the Illumina 450K platform, 20758 genes contained probes/CpG sites in the promoter region. In each simulated data set, 30% of the 20758 genes were randomly selected as differentially methylated, with 15% hypermethylated and 15% hypomethylated in the case group, respectively. The other 70% of the 20758 genes were assigned as non-differentially methylated. For each differentially methylated gene, a promoter-associated neighborhood was randomly assigned as differentially methylated region, while the CpG sites outside the selected regions were assigned as non-differentially methylated. For each CpG site, its neighbors were defined as the CpG sites located within 1000 bps (of both sides) from it. The randomly selected promoter-associated DMRs contained a varying number of CpG sites and can be just part of the promoter region of the genes. With the randomly selected DMRs, the methylated values of the CpG sites (within DMRs or outside DMRs) were simulated in the following two different ways: (1). the simulation scheme used in DMRcate; (2). the proposed Leroux model, as described next.

3.3.2 Simulation studies following *DMRcate*

We first generated the methylation values of CpG sites following the simulation scheme used in DMRcate. For the CpG sites in each DMR, two base beta levels were randomly chosen with a predefined difference $\Delta\beta$, and the beta values within the DMR were randomly generated from a beta distribution with its mode equal to the base beta level and with predefined variability controlled by parameter *K*. The methylation levels of the CpG sites outside DMRs were randomly selected from two predefined beta distributions defined by a parameter set $\{a_0, b_0, a_1, b_1\}$, to mimic unmethylated and methylated sites, respectively. Following this simulation scheme, we generated multiple simulation data sets to evaluate the performances of the competing methods at different noise levels. Table 3-1 shows the parameters used to generate the simulation data sets.

Table 3-1. Parameter settings for simulation data generated by the simulation scheme used in DMRcate.

	Proportion of DM	$\Delta\beta$	**K**	a_0	b_0	a_1	b_1
Scenario 1	**30%**	**0.2**	**100**	**2.4**	**20**	**14**	**3**
Scenario 2	**30%**	**0.2**	**20**	**1.4**	**5**	**5.5**	**2**

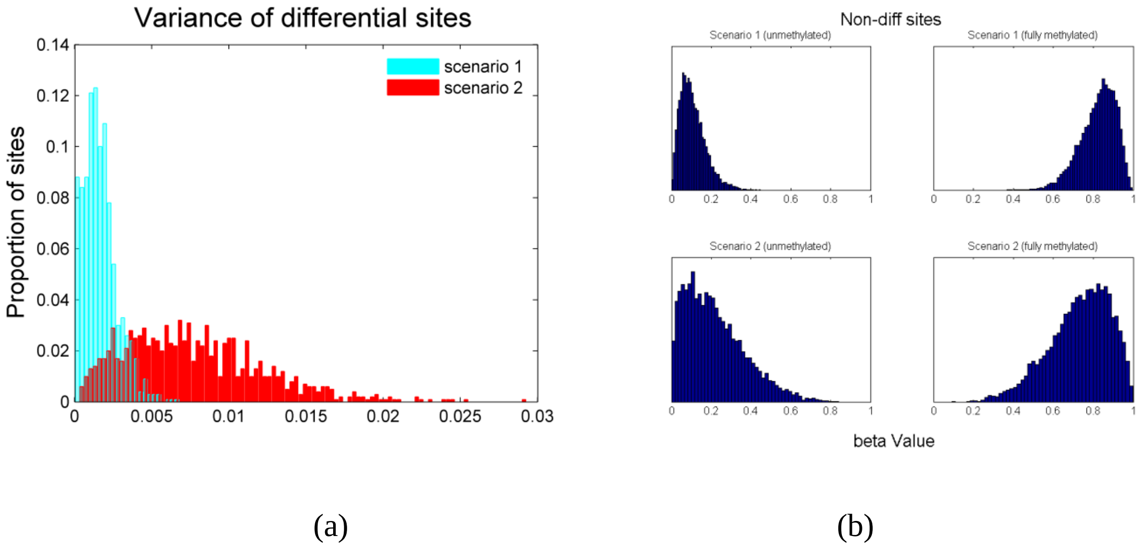


Figure 3.4 Simulation data at different parameter settings: (a) variance of differentially methylated sites; (b) beta distribution for non-differentially methylated sites.

Comparing to scenario 1, the variances of the beta distributions for the generation of simulation data were higher in scenario 2, as shown in Figure 3.4. Figure 3.4(a) shows the variance of true differentially methylated sites among samples in the same condition. By decreasing K, the variances of true differentially methylated sites were higher in scenario 2. Given predefined true difference level $\Delta\beta$, the differential level of the true differentially methylated sites decreased in scenario 2. Figure 3.4(b) shows the beta distributions for the non-differentially methylated sites in the two scenarios, which indicates that the variance of non-differentially methylated sites in scenario 2 was higher than scenario 1. Thus, scenario 2 was more challenging for identifying differentially methylated genes.

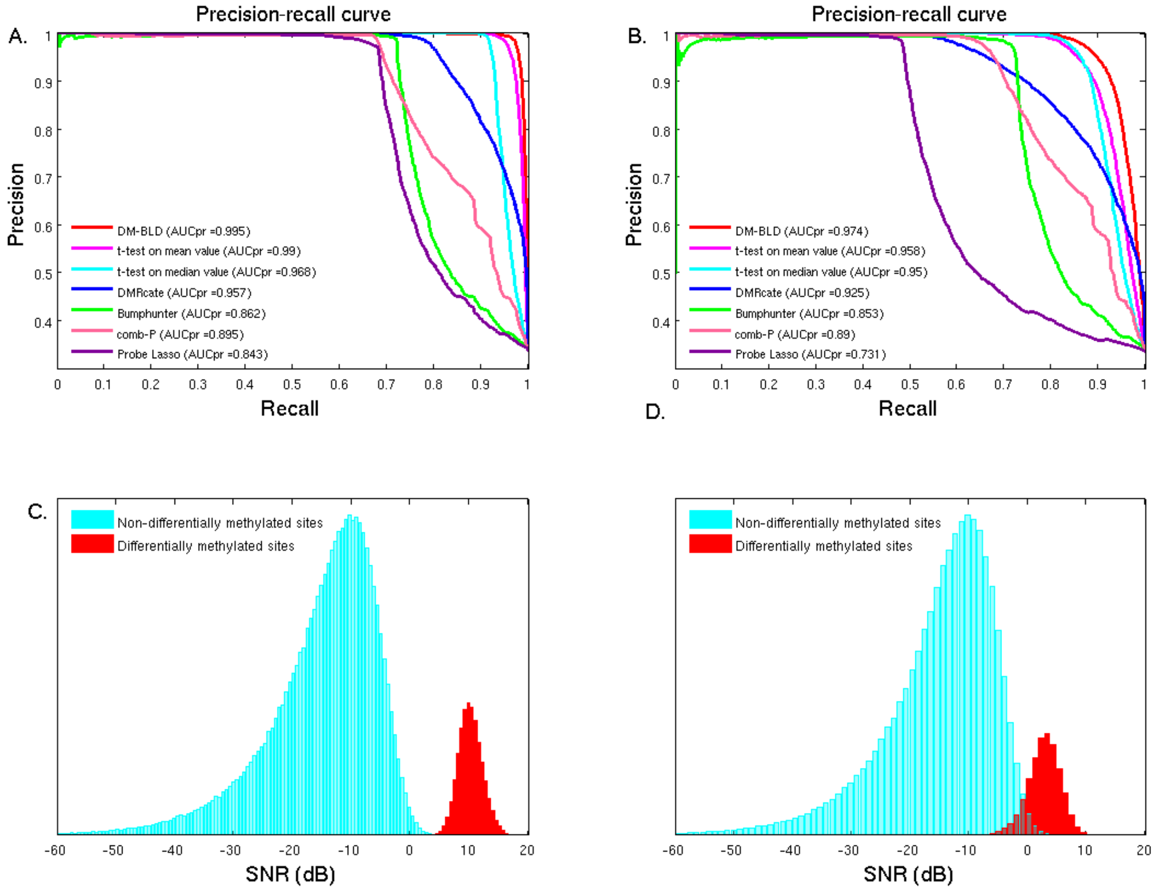


Figure 3.5 Performance on the detection of differentially methylated genes on simulation data generated by the DMRcate scheme: (A) and (C): Precision-recall curve on highly and moderately differential data; (B) and (D): SNRs of non-differentially methylated and differentially methylated sites in (A) and (C), respectively.

Figure 3.5(A) and (B) show the performance of the competing methods in the two scenarios with different noise levels. For each scenario, the SNRs of the differentially and non-differentially methylated sites were compared to show the differential level of the simulation data, as shown in Figure 3.5(C) and (D). The differential levels of the true differentially methylated sites, indicated by SNR, decreased in the second scenario. Figure 3.6 is the boxplot of the AUCpr of the 7 competing methods in 10 random trials in the two scenarios. We can see from the figures that DM-BLD outperformed the other six methods in both scenarios, especially when the differential level was lower (as shown in Figure 3.5(B) and Figure 3.6). The competing methods, such as Bumphunter and comb-P, were quite effective in detecting a subset of DMRs with multiple sites of high differential level, yet missed the others (including many of less differential). DM-BLD was specifically designed based on an MRF framework, where the differential level was estimated considering the differential status of neighboring CpG sites. Hence, DM-BLD was more effective than other competing methods on data with high level of noise.

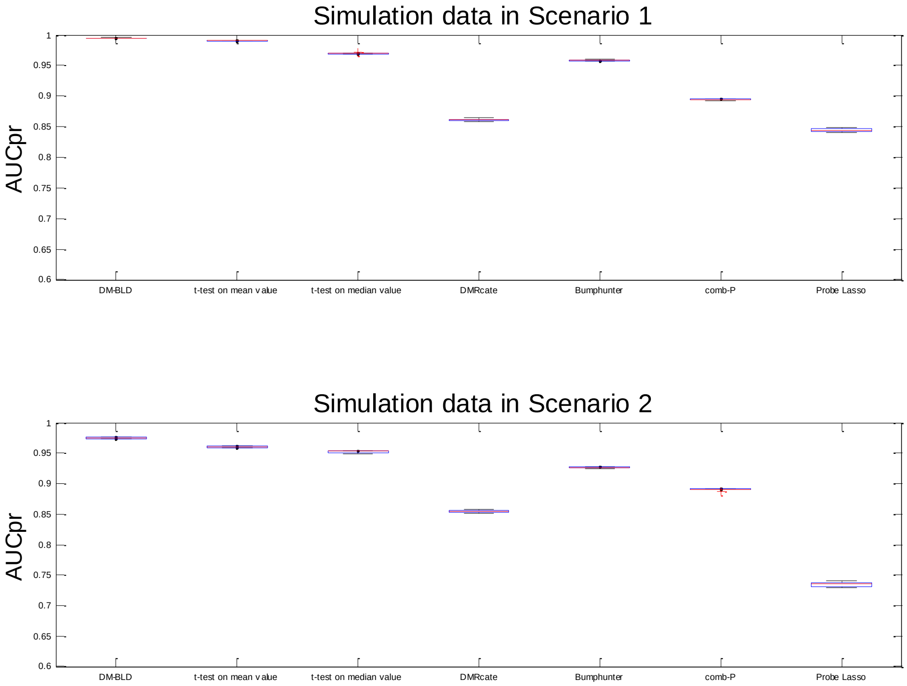


Figure 3.6 Performance of 10 random trials on simulation data generated by the DMRcate scheme.

3.3.3. Simulation study following the proposed model

In the simulation data generated by the DMRcate scheme, the methylation levels of the CpG sites outside the DMRs were generated randomly from two predefined beta-distributions. To better mimic real methylation data where the methylation values of neighboring CpG sites were dependent, we generated simulation data sets by the following steps.

1) Sample base methylation beta value of all CpG sites.

A base methylation beta value of all CpG sites was randomly drawn from the distribution of methylation beta value obtained from a real data set (TCGA breast cancer data set), as shown in Figure 3.7.

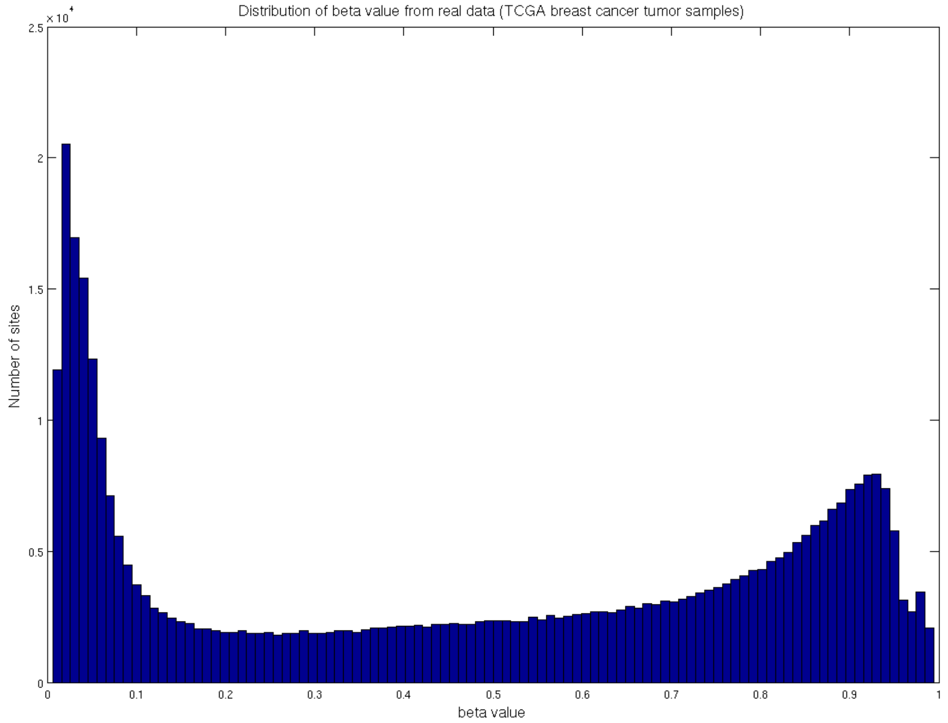


Figure 3.7 Distribution of methylation beta value from TCGA breast cancer tumor samples.

2) Sample true methylation M-value of all CpG sites for each condition.

The true methylation M-value (logit transform of beta value) of all CpG sites in the control group was generated using the Leroux model. Specifically, the true methylated M-values were randomly sampled from a multivariate Gaussian distribution with mean value defined as the logit transform of the base methylation beta value, and variance determined by the neighboring sites. For the non-differentially methylated CpG sites, the true methylation M-value in the case group was the same as in the control group. For CpG sites in the hypermethylated and hypomethylated DMRs, the true methylation M-value in the case group were larger or smaller than the control group with difference μ_0 , respectively.

3) Generate methylation M-value for all samples in each condition

In each condition, the methylation M-value for all samples were drawn from normal distribution with the mean set as the true methylation M-value, and variance $1/\tau_e$.

Thus, τ_e and μ_0 control the differential level of the simulation data set. Higher τ_e indicates lower variance among the samples in the same phenotype, resulting in higher differential level; higher μ_0 indicates the larger difference between two phenotypes, contributing to higher differential level. In the simulation study, we varied τ_e and μ_0 to generate simulation datasets with different levels of noise and methylation change. Specifically, $\tau_e = 5,\ 2,\ 1$ with $\mu_0 = 0.7$, $\tau = 1$, and $\rho = 0.3$; $\mu = 2,\ 1,\ 0.8$ with $\tau_e = 1$, $\tau = 1$, and $\rho = 0.3$. Figure 3.8 shows the SNRs of non-differentially methylated sites and differentially methylated sites in the six different scenarios.

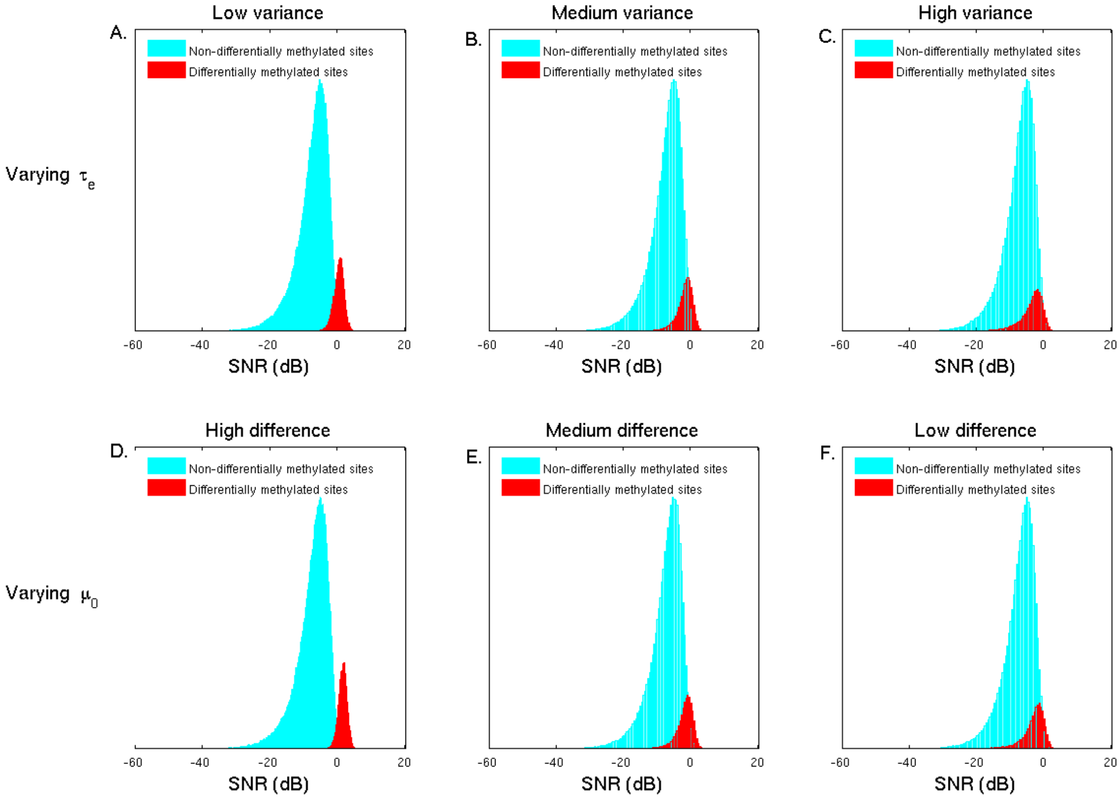


Figure 3.8 SNRs of non-differentially methylated and differentially methylated sites in six scenarios at different noise levels and methylation changes.

We first evaluated the performance of the competing methods on simulation data with different variance (noise) leveled generated by varying τ_e. Figure 3.9 (A), (B), and (C) are the precision-recall curves at low, medium, and high variance levels, respectively, and Figure 3.9(D) presents the AUCpr of the seven competing methods with error bar calculated from 10 random experiments. In the medium and high variance scenarios, Probe Lasso could not detect any differentially methylated regions, and thus, it was not included in those two scenarios. We can see that the performance of all methods dropped with a decrease of τ_e. However, DM-BLD consistently outperformed the other methods. With decreasing τ_e, the variance (noise) among the replicates in the same phenotype increased, which makes it more difficult to estimate the true methylation levels as well as to detect differentially methylated genes. The improved performance of DM-BLD can be attributed to its dependency modeling that borrows information from neighboring sites at estimated dependency level, thus becoming more effective in dealing with noise in replicates.

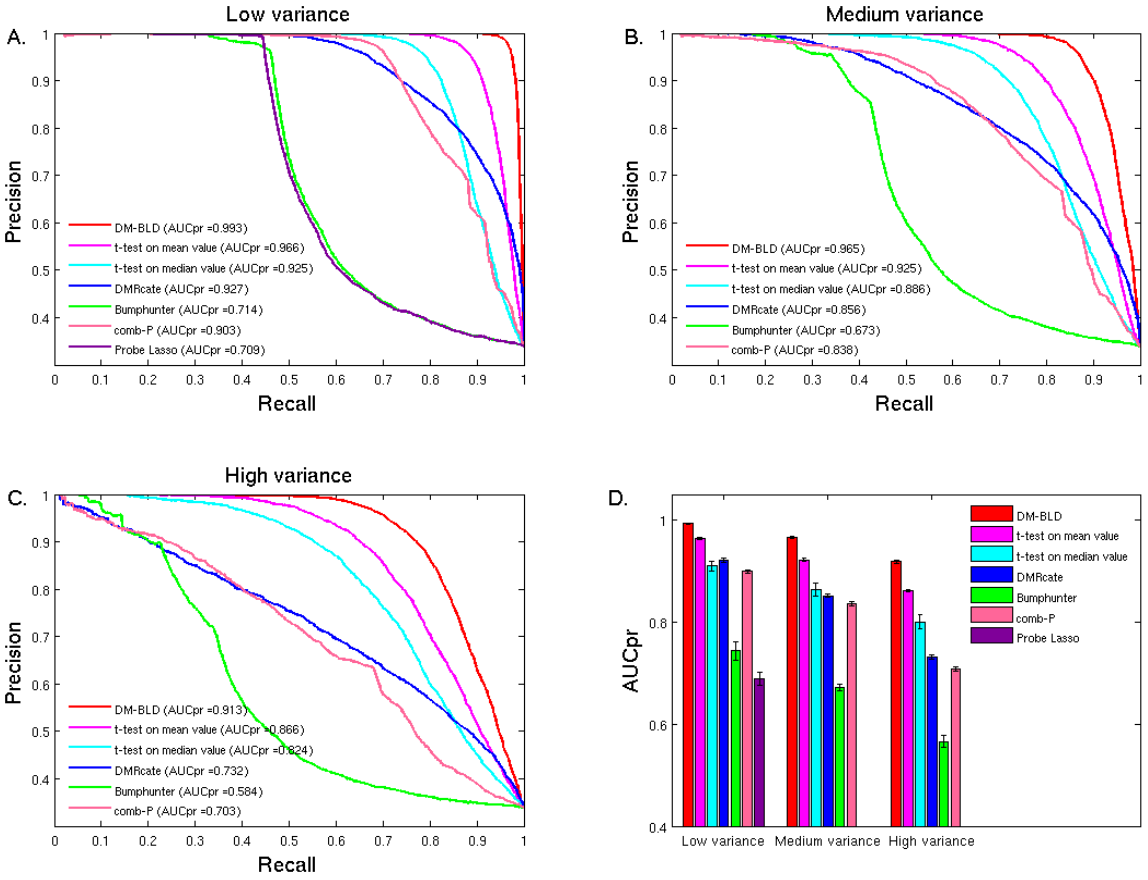


Figure 3.9 Performance on the detection of differentially methylated genes at varying noise levels. Precision-recall curves: (A).low variance; (B) medium variance; (C) high variance. (D) AUCpr in each scenario with ten experiments performed.

We also varied parameter μ_0 to evaluate the performance on varying differential levels between two phenotypes. Decreasing μ_0 lowered SNR of true differentially methylated sites, as shown in Figure 3.8, since the difference of differentially methylated CpG sites between two phenotypes was reduced. The performance of all competing methods degraded when μ_0 decreases, as shown in Figure 3.10. However, DM-BLD achieved a much better performance than that of all other methods, even when the methylation change of genes was moderate. DM-BLD, with the full probabilistic model in a Bayesian framework, was evidently more effective in detecting moderate changes.

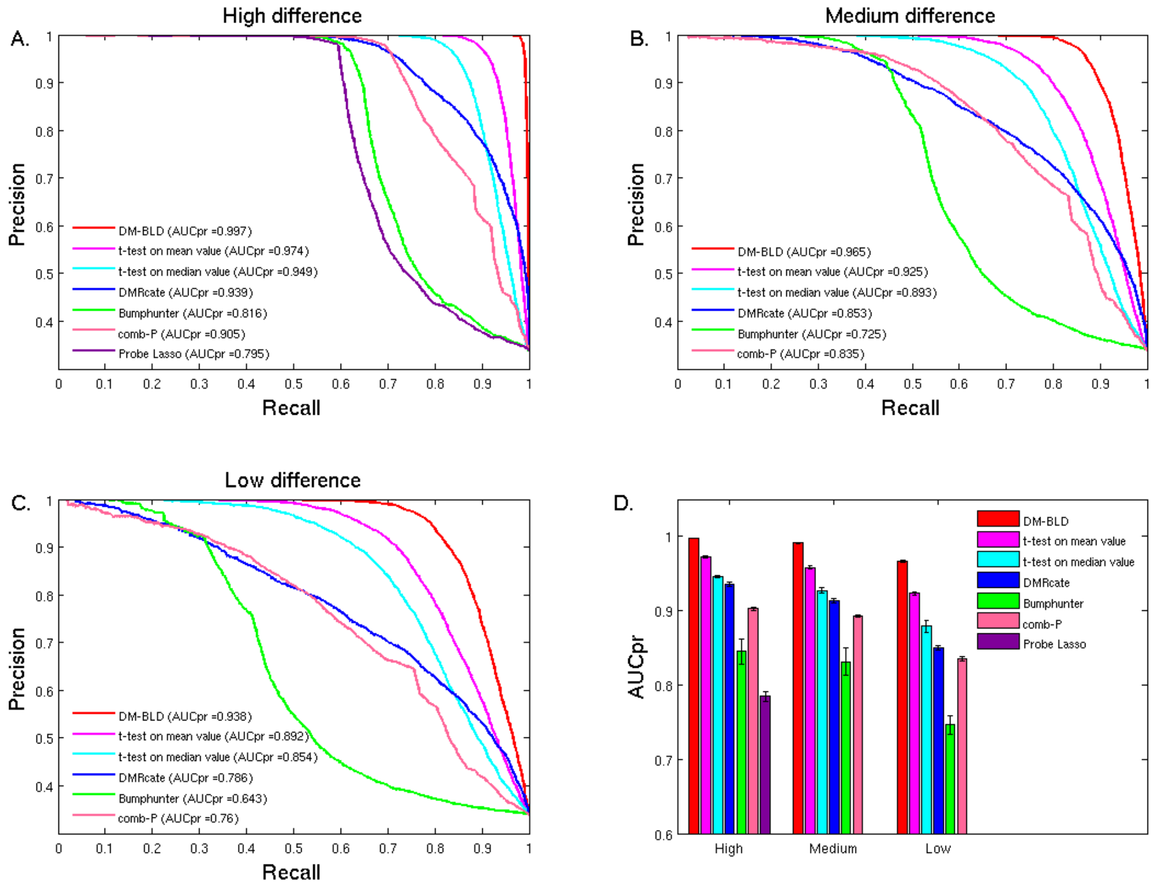
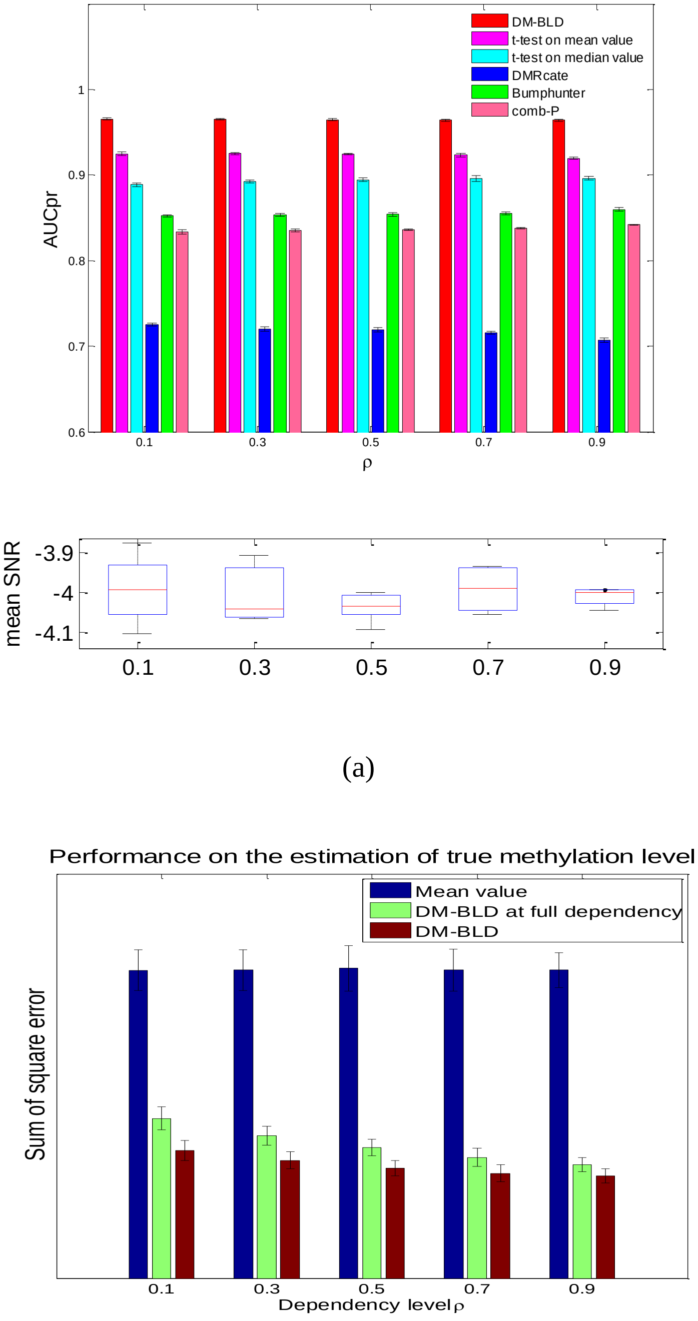


Figure 3.10 Performance on the detection of differentially methylated genes at different levels of methylation change between two phenotypes. Precision-recall curves: (A) high difference; (B) medium difference; (C) low difference. (D) AUCpr in each scenario with ten experiments performed.

To assess the effectiveness of DM-BLD on various levels of local dependency, we further varied parameter ρ to generate simulation data sets. Specifically, ρ varied from 0.01 to 0.9 with interval 0.2, while $\mu_0 = 1$, $\tau = 1$, $\tau_e = 1$. The higher ρ is, the higher the local dependency. As shown in Figure 3.11(a), varying local dependency levels did not directly affect the differential level of CpG sites. However, it impacted on the estimation of the methylation level of CpG sites, as shown in Figure 3.11(b). Since the mean value of the samples did not take dependency into account, the performance was similar among all different dependency levels. The performance of DM-BLD increased with increasing dependency levels since more information can be incorporated from the neighbors. When the dependency level was low, the performance of DM-BLD was much better than that of DM-BLD at full dependency (i.e., where ρ was simply set as 0.999), indicating that the dependency level needed to be correctly estimated. The performance of DM-BLD on the identification of differentially methylated genes consistently outperformed the other methods across different scenarios, as shown in Figure 3.11(a).



(a)

(b)

Figure 3.11 Performance comparison on varying dependency level ρ . (a) AUCpr for the performance on differentially methylated gene detection; (b) performance on the estimation of true methylation level of the CpG sites

3.4 Identification of differentially methylated genes associated breast cancer recurrence

We applied the proposed method, DM-BLD, to breast cancer data acquired from The Cancer Genome Atlas (TCGA) project [59]. The study was designed for the identification of differentially methylated genes associated with breast cancer recurrence. 61 estrogen receptor positive (ER+) tumors were collected from patients for this study, where 41 patients were still alive with the follow-up time longer than 5 years, labeled as ‘Alive’; 20 patients were dead within 5 years, labeled as ‘Dead’. The ‘Dead’ and ‘Alive’ groups represent the ‘early recurrence’ group and the ‘late recurrence’ group, respectively. We applied our method to identify differentially methylated genes by analyzing samples from the two groups. The significance of the differential level was calculated from two permutation tests. With adjusted p-value<0.05 in both permutation tests, DM-BLD detected 1543 differentially methylated genes.

3.4.1 Comparison with existing methods

We compared our DM-BLD method with Bumphunter (v1.6.0), DMRcate (v1.2.0), comb-P and Probe Lasso (part of the ChAMP (v1.4.1) package) onto the breast cancer data. We used the same parameter settings as in the simulation studies for the competing methods. Probe Lasso did not report any differentially methylated regions; thus, it was not included in the comparison. Bumphunter reported methylation regions associated with 1246 genes, where 236 genes were differential with p-value < 0.05. Comb-P reported methylation regions associated with 748 genes, where 721 genes were differential with p-value < 0.05. DMRcate reported 3347 differentially methylated genes with p-value < 0.05. The Venn diagram of the genes detected by the four methods was shown in Figure 3.12. Consistent with the simulation studies and what reported in [88], Bumphunter and comb-P are more conservative than DMRcate and DM-BLD.

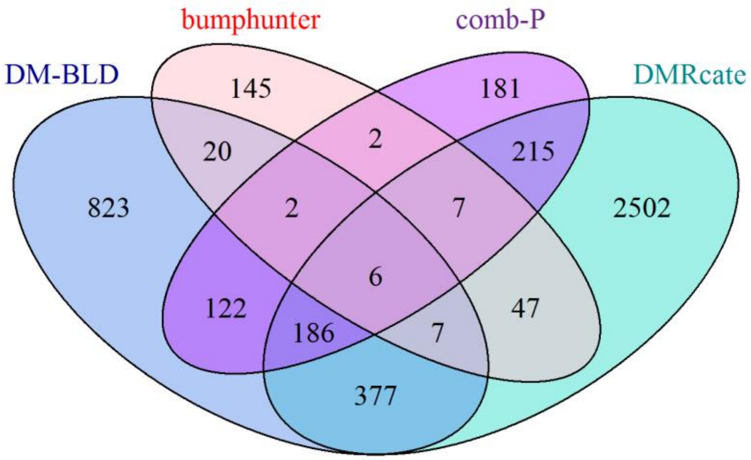


Figure 3.12 Venn diagram of the differentially methylated genes detected by the competing methods.

3.4.2 Characterization of the common and unique gene sets

Among 1543 differentially methylated genes detected by DM-BLD, 720 (common) genes were also detected by other methods yet 823 (unique) genes were detected by DM-BLD only. We compared CpG sites in the detected DMRs associated with the common genes with those associated with the unique genes in terms of noise level and number of CpG sites. First, we compared the absolute difference of beta value and the SNR of the CpG sites in the detected DMRs of the two sets of genes, as shown in Figure 3.13(A) and (B). From one-tail two-sample Kolmogorov–Smirnov (K-S) test, the absolute difference of beta value and SNR were significantly lower in the unique gene set. We also tested on the number of CpG sites across the whole gene region and the number of CpG sites in DMRs, as shown in Figure 3.13(C) and (D). K-S test showed that the number of sites across the gene was significantly higher in the common gene set and the number of sites in the DMR was significantly higher in the unique gene set, as shown in Table 3-2. This observation supported that DM-BLD was more effective in detecting genes of less differentially methylated by virtue of its capability of detecting regions consisting of a sequence of sites with moderate methylation change (resulted in part from relatively large variance observed among tumor samples). Moreover, DM-BLD was effective in detecting genes with fewer measured CpG sites that might be missed by other methods biased to dense CpG regions.

Table 3-2. p-values from K-S test.

	p-value from K-S test
Absolute difference of beta value (“>”: larger in common genes than in unique genes)	**5.79e-89**
SNR (“>”: larger in common genes than in unique genes)	**6.44e-200**
Number of sites in each gene (“>”: larger in common genes than in unique genes)	**1.77e-7**
Number of sites in DMRs (“<”: less in common genes than in unique genes)	**2.36e-15**

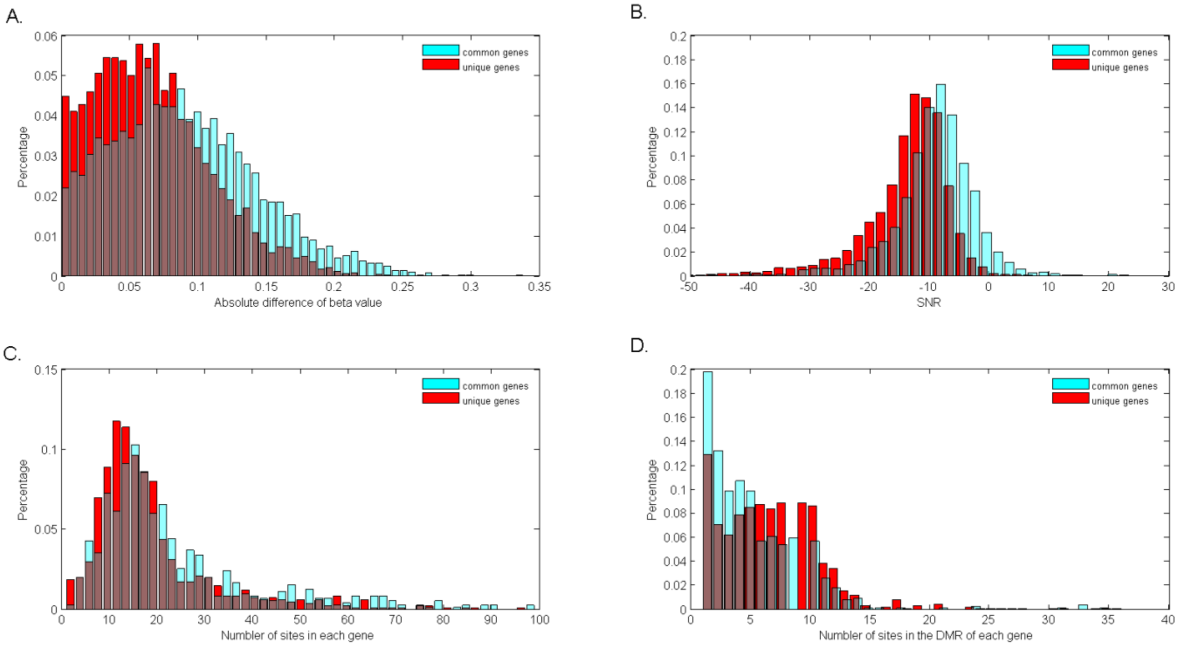


Figure 3.13 Common differentially methylated genes versus Unique differentially methylated genes detected by DM-BLD. (A) the absolute difference of beta value; (B) SNR; (C) number of CpG sites associated with gene; (D) number of CpG sites in DMR.

3.4.3 Performance evaluation via differentially expressed genes

As there was no ground truth of differentially methylated genes for real data, we used the corresponding mRNA expression change as the benchmark to assess the performance. We detected differentially expressed genes from RNA-seq data of the same set of samples. In detail, we downloaded the RNA-seq data (Level 1) of all of the 61 samples profiled by Illumina HiSeq 2000 RNA Sequencing Version 2 analysis from the TCGA data portal, and then performed alignment using 'TopHat 2 (TopHat v2.0.12)' (http://ccb.jhu.edu/software/tophat/index.shtml) with UCSC hg19 as the reference sequence. With the isoform structure annotation file (RefSeq genes) downloaded from the UCSC genome browser database (http://genome.ucsc.edu/), we applied the Cuffdiff 2 method (Cuffdiff 2.2.1; http://cole-trapnell-lab.github.io/cufflinks/) to identify differentially expressed isoforms by analyzing samples from the two groups: the 'Dead' group vs. the 'Alive' group. Differentially expressed genes were defined as genes with differentially expressed isoforms with p-value less than 0.05. As a result, 1101 differentially expressed genes were identified.

Figure 3.14 shows the proportion of differentially expressed genes among the top differentially methylated genes among the top ranked differentially methylated genes detected by the four competing methods, where genes were ranked by the p-value obtained from each method. We can see from the figure that DM-BLD detected more genes with mRNA expression change, which were benchmarked as functional differentially methylated genes.

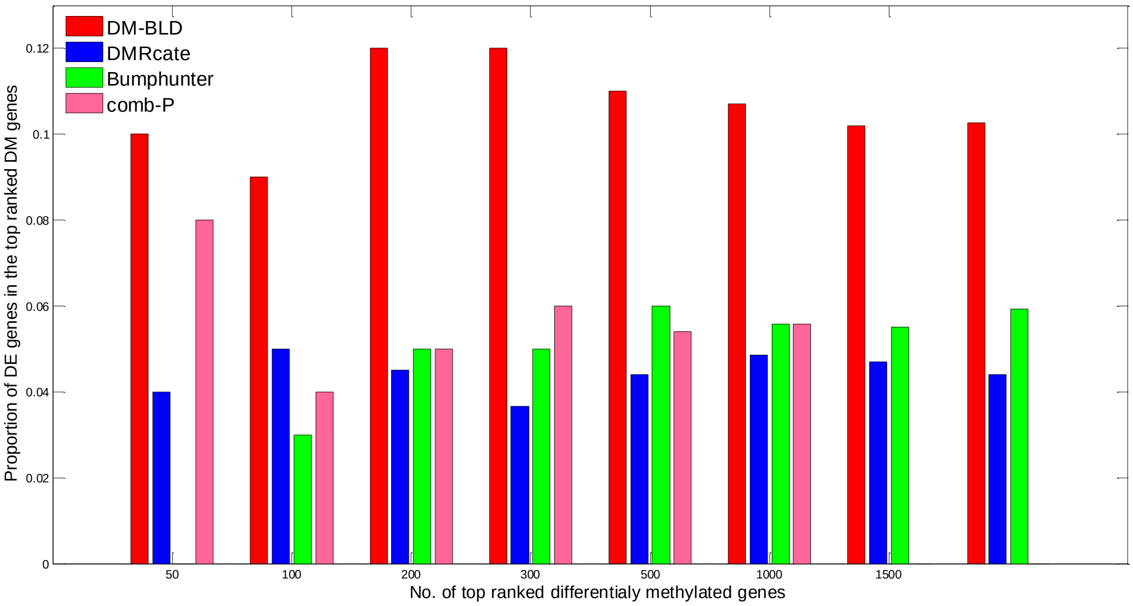


Figure 3.14 Proportion of differentially expressed (DE) genes among the top ranked differentially methylated (DM) genes detected by DM-BLD, DMRcate, Bumphunter, or comb-P.

3.4.4 Functions of the identified genes

Among the differentially methylated genes detected by DM-BLD, 523 genes were hypermethylated in the promoter region. From functional annotation clustering using DAVID (the Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/home.jsp), the set of hypermethylated genes was enriched in transcription factor activity (82 genes, p-value = 4.5E-20), and homeobox (39 genes, p-value = 4.5E-19). The enrichment in transcription factor activity may indicate the interplay between transcription factor and DNA methylation. Methylation of homeobox genes has been reported as a frequent and early epigenetic event in breast cancer [109]. Moreover, 159 genes out of the 523 genes were Polycomb target genes (hypergeometric p-value = 3.19E-29). The Polycomb target genes were the common genes detected from the ChIP-seq data of EZH2, SUZ12, H3K4me3 and H3K27me3 in embryonic stem cells (which were acquired in the ENCODE project (http://www.encodeproject.org/)). Specifically, we first downloaded the ChIP-seq data of EZH2, SUZ12, H3K4me3 and H3K27me3 in embryonic stem cells from ENCODE (https://www.encodeproject.org/). Then, we used MACS [110] with default setting for peak calling. Finally, we matched the peaks to genes using GREAT [111] with the regulatory region defined as $\pm 2K$ from the transcriptions start sites (TSS). The gene sets identified from the four ChIP-seq studies were shown in Figure 3.15. As a result, 2,589 common genes from the four studies were detected as the Polycomb target genes. Polycomb group proteins are well known epigenetic regulators that silenced the target genes. The significantly large overlap between the identified hypermethylated genes in the promoter region and the Polycomb target genes indicates that the two key epigenetic repression systems jointly regulate gene expression [112].

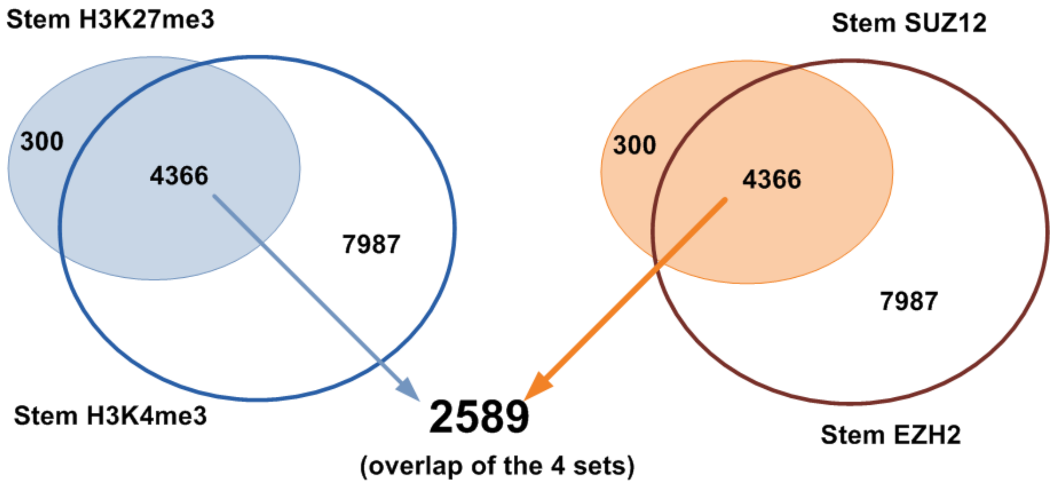


Figure 3.15 Number of genes identified from four ChIP-seq studies on stem cell.

To study the interaction of the identified genes, we first mapped the differentially expressed genes to the Protein-Protein interaction (PPI) network from the Human Protein Reference Database (HPRD) [78]. Figure 3.16(A) shows that the major connected network is largely downregulated in the 'Dead' group as compared to that in the 'Alive' group. In the PPI network of differentially expressed genes, there are two modules of interacting genes with differential methylation activity between two groups. The two modules, potentially regulated by DNA methylation, are shown in Figure 3.16(B) and highlighted by yellow and blue in Figure 3.16(A).

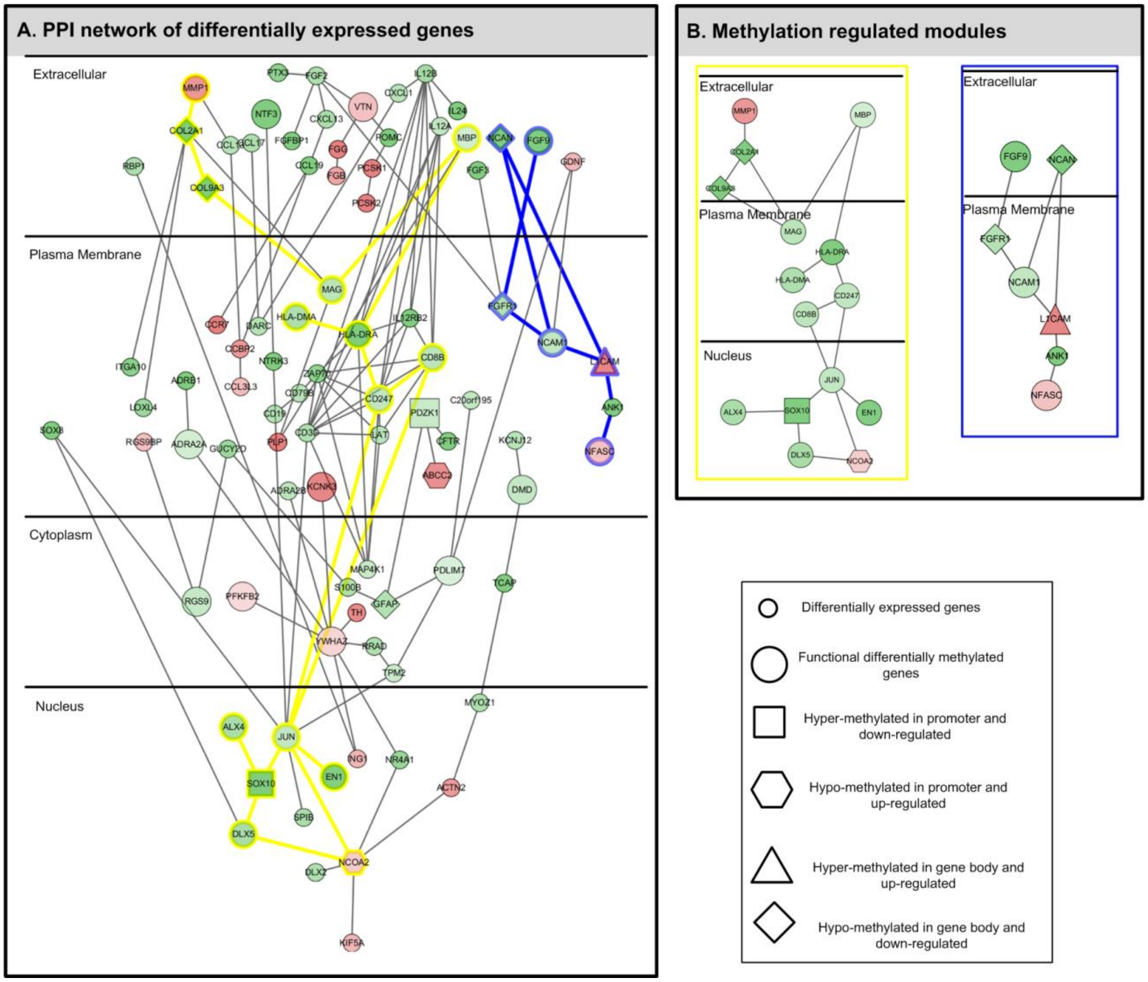


Figure 3.16 Network of differentially expressed genes and methylation regulated modules. (a) A PPI network of differentially expressed genes; (b) methylation regulated modules with interacting genes that are differentially expressed and also differentially methylated.

We further looked into functional differentially methylated genes, which are differentially methylated genes that are also differentially expressed. By incorporating the mRNA expression change estimated from the corresponding RNA-seq data, we detected 158 functional differentially methylated genes, which are enriched in cell adhesion, cell morphogenesis, cell to cell signaling, transcription factor activity, and so on. Literature has shown that hypermethylation in the promoter region repressed gene expression, contributing to cancer development. Thus, we focused on genes that are hypermethylated in the promoter region and down-regulated in the ‘Dead’ group (N=52). 18 genes are also Polycomb target genes, which may be regulated by Polycomb group protein and DNA methylation jointly. Among the 18 genes, DBC1 and SLC5A8 are tumor suppressor genes. DBC1 has been demonstrated participating in cell cycle control [113], and it was reported that the hypermethylation of DBC1 was an effective biomarker in predicting breast cancer [114, 115]. SLC5A8, a putative tumor suppressor, was found that it inhibited tumor progression [116]; the inactivation of SLC5A8 might result in tumor development [117]. HTRA3 was reported as a candidate tumor suppressor and TGF-beta signaling inhibitor, which might be regulated by transcription factor CREB3L1 to affect the development of breast cancer [118]. CMTM3, as a CMTM family protein linking chemokines and the transmenbran-4 superfamily, exerted tumor-suppressive function in tumor cells [119]. The silencing of CMTM3 due to hypermethylation would result in loss of function in inhibiting tumor cell growth and inducing apoptosis with caspase-3 activation. Note that DBC1, HTRA3, and CMTM3 were detected by our DM-BLD method only, yet missed by the other methods.

3.5 Discussion and Conclusion

It is important to accurately detect differentially methylated genes, yet with remarkable challenges, particularly in the field of cancer research where the variability of methylation among replicates/samples is high. We have developed a Bayesian approach, DM-BLD, for the identification of differentially methylated genes. A hierarchical and probabilistic model, with differential states as hidden variables, is devised to account for the local dependency of CpG sites and the variability among the samples/replicates of the same phenotype. Specifically, the Leroux model, which is capable of capturing varying degrees of local dependency, is used to model the unknown correlation among the true methylation levels of CpG sites in the neighboring region. A discrete Markov random field is then used to model the dependency of methylation changes (via differential states) of neighboring CpG sites. A hierarchical Bayesian model, with differential states embedded as hidden variables, is then developed to take into account the local dependency for differential analysis.

The main advantages of our proposed method, DM-BLD, can be summarized as follows. First, it is a fully probabilistic approach that jointly models the methylation level of CpG sites and the differential state of methylation between two phenotypes. Rather than calculating the significance of the difference between two groups of samples using statistical tests (e.g., DMRcate, comb-P), DM-BLD uses a Bayesian framework to estimate the true methylation level and the differential state in a probabilistic way. Second, the varying local dependency among neighboring CpG sites is modeled by the Leroux model, an advanced CAR structure that can account for different levels of correlation. By virtue of using information from neighborhood with local dependency, the accuracy of the estimated methylation level is greatly improved, particularly when the variability among the replicates is high. Third, the Leroux model is embedded into the Bayesian framework. Thus, the posterior distributions of the true methylation levels in each group, as well as other parameters, are estimated jointly with the local correlation levels through a Gibbs sampling procedure, which provides an improved performance in detecting CpG sites of less differentially methylated. Finally, with the estimated methylation change of CpG sites between two groups, we detect differentially methylated genes as genes with a sequence of CpG sites exhibiting methylation change and calculate the significance of differentially methylated genes by permutation tests.

We have compared the performance of DM-BLD with the existing methods using extensive simulation studies. DM-BLD consistently outperforms the other methods, particularly when the difference between two groups is less and the noise among the replicates is high. We have also compared DM-BLD with another version of DM-BLD at full dependency and demonstrated that the estimation of local dependency has improved the performance, particularly in cases of high variance among replicates and low dependency level. Moreover, we have applied DM-BLD to breast cancer data to identify differentially methylated genes associated with breast cancer recurrence and demonstrated the advantage of DM-BLD as evaluated by the consistency with the differential expression of mRNA. The differentially methylated genes identified by DM-BLD are enriched in transcription factor activity and consisted of a significant portion of Polycomb target genes. Moreover, several differentially methylated genes such as DBC1, HTRA3, and CMTM3, revealing the underlying biological mechanism related to breast cancer recurrence, have been uniquely identified by our DM-BLD method.

4 Contribution, Future work and Conclusion

4.1 Summary of original contribution

We have developed new computational methods to model RNA-seq data and methylation data for differential analysis and have demonstrated their advantages using multiple simulation data sets and real data sets. In this chapter, we briefly summarize the original contribution of this dissertation research.

4.1.1 A novel Bayesian model for differential analysis of RNA-seq data

A novel Bayesian approach, BayesIso, has been developed to identify differentially expressed isoforms from RNA-seq data. In this approach, a hierarchical model is devised for comprehensive modeling of the variability of RNA-seq data, aiming for accurate detection of differentially expressed isoforms. The variability of RNA-seq data can be introduced by the biases across genomic loci on one hand, termed within-sample variability; on the other hand, it can be caused by the large variance of replicates in a phenotype group, particularly in cancer research, termed between-sample variability. Both within-sample variability and between-sample variability are captured by the hierarchical model, with differential states embedded as hidden variables in a Bayesian framework. The main advantages of BayesIso can be summarized as follows.

- BayesIso is a fully probabilistic approach in a Bayesian framework that carries out the estimation of isoform expression level and the detection of differentially expressed isoforms in a cooperative manner. The expression levels and differential states of isoforms, as well as other model parameters, are estimated iteratively in an MCMC sampling procedure. In each iteration of the sampling procedure, accurate estimation of isoform expression is conducive to accurate estimation of differential state, which in turn benefits isoform expression estimation. The coupling of transcript expression and differential state contributes to more accurate estimation of the model parameters and the hidden differential state, providing an improved performance in detecting isoforms of less differentially expressed.

- The differential state is embedded into the hierarchical model in a probabilistic way, in contrast to that conventional statistical tests calculate the significance of the difference between two groups of samples. The probabilistic formulation makes it possible to estimate the posterior distribution of differential state. An MCMC sampling procedure is designed to estimate the parameters and hidden variables iteratively from samples drawn from the conditional distributions. Thus, by virtue of the probabilistic model and the MCMC sampling procedure, more accurate estimation of differential state is obtained from its posterior distribution, by joint estimation with isoform expression level and other model parameters.

- The Poisson-Lognormal model in BayesIso provides a flexible way to model various biases within each transcript without relying on known sources. In the Poisson-Lognormal model, one model parameter is assigned to each exon/segment of each isoform to capture the diverse variability at different loci. Thus, it is capable of accounting for different bias patterns of transcripts. The comprehensive modeling of biases at different loci is beneficial to more accurate estimation of isoform expression level and eventually, contributes to more accurate estimation of differential state.

We have applied BayesIso to multiple simulation data sets in different scenarios, and compared the performance of BayesIso with both count-based methods (DESeq, edgeR, DSS, EBSeq) and isoform-level differential analysis methods (Cuffdiff 2 and Ballgown). We have generated simulation data using two simulators: one is our simulator and the other is an existing simulator. Synthetic datasets with different levels of transcript expression and expression change between two phenotypes have been simulated to evaluate the performance of BayesIso and other competing methods. BayesIso consistently outperforms the other methods in all of the scenarios, with AUC of ROC curve and F-score as the metrics to evaluate the overall performance on the identification of differentially expressed isoforms. We have further investigated the performance of the competing methods on transcript sets of different differential levels measured by SNR and have demonstrated that BayesIso is more effective in detecting isoforms of less differentially expressed. The improved performance of BayesIso as compared to the other methods has further been demonstrated using two real RNA-seq datasets with benchmarks.

We have finally applied BayesIso to breast cancer RNA-seq data to identify differentially expressed isoforms associated with breast cancer recurrence. The diverse bias patterns along transcripts and the generally low differential level have been observed from the real breast cancer data, indicating the importance of their accurate modeling in differential analysis of RNA-seq data. The differentially expressed isoforms detected by BayesIso are enriched in cell proliferation, apoptosis, migration, and signaling pathways, shedding light on the roles of these differentially expressed isoforms in driving breast cancer recurrence. Moreover, the unique set of differential genes identified by BayesIso has helped reveal several signaling pathways such as the PI3K/AKT/mTOR signaling and PTEN signaling pathways. The enrichment of identified modules has been validated on a time course E2 induced MCF-7 breast cancer cell line data. Moreover, the immune response genes that are down-regulated in the early recurrence group may indicate the role of immune system in breast cancer recurrence.

4.1.2 Differential methylation detection from methylation data exploiting local dependency

We have developed a novel Bayesian approach, DM-BLD, for the identification of differentially methylated genes. In this approach, we model the methylation level of CpG sites of replicates in different phenotypes using MRF models: the Leroux model is used to account for the local dependency of the methylation levels of nearby CpG sites; a discrete MRF model is used to capture the dependency of methylation change (via differential states) of nearby CpG sites. The main advantages of our proposed method, DM-BLD, can be summarized as follows.

- DM-BLD is a fully probabilistic approach that jointly models the methylation level and the differential state of CpG sites between two phenotypes. The probabilistic formulation makes it possible to estimate the methylation level of CpG sites jointly with differential state and other model parameters in a Bayesian framework.

- MRF models are used to account for the dependency of CpG sites from different perspectives. Leroux model is used to capture the dependency of methylation level of nearby CpG sites, aiming for more accurate estimation of the methylation level of CpG sites. As an advanced CAR structure, Leroux model is capable of accounting for different levels of dependency. A discrete MRF model is incorporated to capture the dependency of methylation state of nearby CpG sites, aiming to detect nearby CpG sites with consistent methylation change. Accurate estimation of differential state contributes to improved performance on methylation level estimation. Thus, by virtue of using information from neighborhood with local dependency, the accuracy of estimated methylation level is greatly improved, particularly when the variability among replicates is high, which eventually provides an improved performance in detecting CpG sites of less differentially methylated.

- A Gibbs sampling procedure is designed to jointly estimate the posterior distributions of model parameters and variables. Gibbs sampling, as a stochastic sampling technique, has the ability to escape local optima via exploiting randomized searching to a much greater degree comparing to deterministic approaches. Due to a finite number of iterations in the implementation, Gibbs sampling may not achieve optimal. In order to alleviate the potential problem of being trapped in local optima, we have implemented the Gibbs sampling algorithm with an option of multiple runs in addition to one run of sampling. Samples from multiple runs are checked to see whether a global, optimum solution is likely to have been achieved.

- Finally, with the estimated methylation change of CpG sites between two groups, we detect differentially methylated genes as genes with a sequence of CpG sites exhibiting methylation change and calculate the significance of differentially methylated genes by permutation tests. Both ‘global’ and ‘local’ null distributions are generated to assess the significance of differentially methylated genes.

We have applied DM-BLD to extensive simulation data sets and compared the performance to the existing methods. Synthetic datasets with different levels of variance across replicates and methylation change between two conditions have been generated for comprehensive performance evaluation. DM-BLD consistently outperforms the other methods, particularly when the difference between two groups is less and the noise among the replicates is high. We have also applied DM-BLD and the competing methods to a breast cancer data set and demonstrated the advantage of DM-BLD benchmarked by the differentially expressed genes identified from differential analysis on the RNA-seq data generated from the same set of tumor samples. By further characteristic of the differentially methylated genes detected by DM-BLD only and the genes that are also detected by another method, we demonstrate that DM-BLD is more effective in detecting genes with a sequence of CpG sites with moderate but consistent methylation change.

The differentially methylated genes detected by DM-BLD in the breast cancer study help reveal the mechanism underlying breast cancer recurrence. The hypermethylated genes in the promoter region are significantly enriched in transcription factor activity and consist of a significant portion of Polycomb target genes, which indicates the joint function of the two key epigenetic repression systems in regulating gene expression. Moreover, several identified genes that are hypermethylated in the promoter region in the ‘early recurrent’ samples, such as DBC1, HTRA3, and CMTM3, are down-regulated in the ‘early recurrent’ samples as shown by the RNA-seq data, which may reveal the function of the aberrant methylation change related to breast cancer recurrence.

4.2 Future work

4.2.1 Biological validation of novel discoveries

We have applied the proposed methods to breast cancer tumor samples, in order to reveal the mechanism underlying breast cancer recurrence. We have initially demonstrated the potential role of the identified differential genes by functional enrichment analysis. As future work, biological experiments are needed to validate the importance of identified differential genes in breast cancer recurrence. Particularly, the identified functional genes, which are both differentially methylated and differentially expressed, are good candidates to be biologically validated as crucial biomarkers for breast cancer recurrence.

The biological experiments for the validation of identified differentially expressed transcripts can be performed in the following steps. First, the expression level of identified transcripts can be measured by other techniques, such as qRT-PCR, to demonstrate the change of transcript expression in different phenotypes. Second, since transcripts are functional in biological systems by translating into proteins, the abundance of proteins corresponding to the identified transcripts can be measured to further demonstrate the potential functional role of identified transcripts. Third, specific experiments can be designed to verify the functions of identified transcripts in cancer development. For example, the overexpressed transcript can be knocked down to see whether it affects tumor development.

4.2.2 Network identification from RNA-seq data

Identification of dysfunctional isoform networks from RNA-seq data and domain-domain interaction network

The proposed BayesIso method is developed to identify differentially expressed isoforms to understand the mechanism of diseases, particularly cancer recurrence. In the Gamma-Gamma model, i.e., a part of the joint model, the expression of each individual isoform of samples in two phenotypes is investigated to see whether it is differentially expressed. In the conventional gene expression study using microarray data, it has been demonstrated that gene networks [120-122], i.e., the interaction of genes, is more informative in understanding the molecular mechanisms associated with complex human diseases. Prior knowledge, such as protein-protein interaction network, can be incorporated for the identification of active subnetworks. The subnetwork based identification is superior to individual gene-based identification in revealing biological systems, as the subnetwork based approach is effective in detecting interacted gene sets with moderate but consistent change, as well as hub genes that show little changes in expression comparing with their downstream genes but are biologically important as a connecting point in the network. By virtue of the advantage of RNA-seq data in isoform-level analysis, the identification of isoform networks is of great interest.

Analogous to PPI network that provides information on the interactions among genes, the protein Domain-Domain interactions (DDI) between the domains in each pair of transcripts can be used as the prior knowledge on isoform interactions. The DDI network can be generated from several databases, such as Pfam database [123] and domain-domain interaction databases. The interactions between the domains can also be predicted by existing tools, such as DOMINE[124]. Moreover, it has been demonstrated that DDI network can be incorporated to improve the performance on isoform quantification [125]. Thus, it would be beneficial to incorporate DDI network in the identification of isoform networks that consist of differentially expressed isoforms associated with disease.

The current design of BayesIso is under a Bayesian framework with differential state embedded as hidden variable. The Bayesian design brings convenience to incorporate more information that can be formulated as prior knowledge. To be specific, the DDI network can be used to determine the prior distribution of differential state. Instead of non-informative prior for differential state used in the current model, the prior distribution of differential state can be modeled by a discrete Markov random field generated from the DDI network. Thus, the differential state of an isoform is impacted by the differential state of its interacting isoforms. The incorporation of DDI network may improve the performance on differential analysis and detect more biologically meaningful aberrant expression events.

4.2.3 Integration of methylation data and RNA-seq data

Identification of methylation regulated differentially expressed isoforms

In this dissertation research, we have developed DM-BLD to detect differentially methylated genes from methylation data. In the real data study on breast cancer recurrence, with the detected differentially methylated genes, we have further analyzed the expression change of the genes using RNA-seq data from the same samples of the two phenotypes, and have detected a set of functional genes (both differentially methylated and differentially expressed) which have been studied in literature for their potential role in breast cancer development. The detected functional genes may indicate the essential role of methylation change in disease, especially cancer, via regulating gene expression. However, it is not an ideal approach to detect functional genes by overlapping the differentially methylated genes and the differentially expressed genes identified from methylation data and RNA-seq data separately. To further understand the function of DNA methylation in regulating transcript expression, a new method that incorporates methylation data and RNA-seq data is more promising.

In high-throughput methylation profiling techniques, such as Illumina 450K, most of the genes are measured at multiple CpG sites along the genomic loci including the promoter region, first exon, gene body, etc. It has been shown that methylation at different regions of the genes may have different influence on expression change. In specific, hypermethylation in the promoter regions and hypo-methylation in gene body may suppress gene expression. Such understanding on the relation between methylation and expression can be incorporated into the model to detect aberrant events that are more biologically meaningful. Moreover, RNA-seq permits to study in isoform-level. Taking advantage of high-throughput methylation data and RNA-seq data, the relation between the methylation level of CpG sites and the expression of isoforms can be investigated taking advantage of CpG site locations annotated specific to the genomic location of isoforms.

To identify methylation regulated differentially expressed genes, given isoform expression and methylation data, a linear regression model can be built, in which isoform expression can be modeled as a linear combination of methylation level of the associated CpG sites. The following factors should be taken into account in the linear model: 1) methylation change of the CpG sites in the promoter regions and that in the gene body has opposite effects on gene expression; 2) neighboring CpG sites along the genome have local dependency; 3) the number of samples may be less than the number of CpG sites. These factors can be used to determine the prior distribution of the coefficients in the linear model. Finally, the coefficients, as well as other model parameters, can be estimated from a Gibbs sampling procedure. Thus, the methylation regulated differentially expressed genes can be identified by integrating methylation data and RNA-seq data.

4.3 Conclusion

Transcriptomics and epigenomics are two important and highly related studies in the field of computational molecular biology. In this dissertation research, we work on detecting aberrant events in each of the two disciplines through differential analysis of RNA-seq data and methylation data. We have developed a novel Bayesian model to identify differentially expressed isoforms from RNA-seq data, which comprehensively models the variability of RNA-seq data with embedded differential state. For differential analysis on methylation data, we have developed a hierarchical model exploiting the local dependency of CpG sites in terms of methylation level as well as methylation change. Comprehensive simulation studies and experiments on real data sets with benchmarks have demonstrated the advantages of proposed methods. The application of the proposed methods to breast cancer studies has also demonstrated the feasibility of using the proposed methods to reveal the molecular mechanisms underlying breast cancer recurrence.

Appendix A. Journal manuscript in preparation and conference publication

Manuscripts under Review

X. Wang, J. Gu, L. Hilakivi-Clarke, R. Clarke, J. Xuan, “DM-BLD: Differential Methylation detection via a hierarchical Bayesian model exploiting Local Dependency,” Bioinformatics, 2016 (*Revision submitted*).

X. Wang, X. Shi, J. Gu, A. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, J. Xuan, “BayesIso: Bayesian identification of differentially expressed isoforms using a novel joint model of RNA-seq data,” BMC Genomics, 2016. (*under review*)

Leena Hilakivi-Clarke, Anni Wärri, Kerrie B Bouker, Xiyuan Zhang, Katherine L Cook, Lu Jin, Alan Zwart, Nguyen Nguyen, Rong Hu, M Idalia Cruz, Sonia de Assis, **Xiao Wang**, Jason Xuan, Yue Wang, Bryan Wehrenberg, and Robert, "Effects of in utero exposure to ethinyl estradiol on tamoxifen resistance and breast cancer recurrence in a preclinical model," (*Revision submitted*)

Conference Publication

X. Wang, J. Gu, R. Clarke, L. Hilakivi-Clarke, and J. Xuan, "A Markov random field-based Bayesian model to identify genes with differential methylation," in Proc. IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, Hawaii, USA, May 2014.

X. Wang, J. Gu, L. Chen, A. N. Shajahan, R. Clarke, and J. Xuan, "Sampling-based Subnetwork Identification from Microarray Data and Protein-protein Interaction Network," in Proc. ICMLA 2012: Machine Learning in Health Informatics Workshop, Vol. 2, pp. 158 - 163, Boca Raton, Florida, Dec. 2012.

J. Gu, J. Xuan, **X. Wang**, A. N. Shajahan, L. Hilakivi-Clarke, and R. Clarke, "Reconstructing transcriptional regulatory networks by probabilistic network component analysis," in Proc. The fourth ACM International Conference on Bioinformatics, Computational Biology, and Biomedical Informatics (ACM BCB 2013), pp. 96-105, Washington, DC, Sept. 2013.

X. Chen, **X. Wang**, and J. Xuan, "Tracking Multiple Moving Objects Using Unscented Kalman Filtering Techniques," in Proc. Int'l Conf. on Engineering and Applied Science, Beijing, China, July 2012.

Peer-reviewed Journal Publication

J. Gu, **X. Wang**, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, "BADGE: A novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data," BMC Bioinformatics, 15(Suppl 9):S6, 2014.

X. Shi, **X. Wang**, A. Shajahan, L. Hilakivi-Clarke, R. Clarke and J. Xuan* , "BMRF-MI: integrative identification of protein interaction network by modeling the gene dependency," BMC Genomics , 16(Suppl 7):S10, 2015.

Y. Liu, L. Hilakivi-Clarke, Y. Zhang, **X. Wang**, Y. Pan, J. Xuan, S. Fleck, D. Doerge, W. Helferich, "Isoflavones in soy flour diet have different effects on whole-genome expression patterns than purified isoflavone mix in human MCF-7 breast tumors in ovariectomized athymic nude mice," Molecular Nutrition and Food Research, 59(8):1419-1430, 2015.

Manuscripts in Preparation

X. Wang, X. Shi, A. Shajahan-Haq, L. Hilakivi-Clarke, R. Clarke, J. Xuan, “BayesDenovo: A more accurate *de novo* transcriptome assembler exploiting a Bayesian model,” (*to be submitted*).

Appendix B. The equivalence between Gibbs random field and the discrete Markov random field used in DM-BLD

As observed from real DNA methylation data sets, the methylation change pattern of the neighboring CpG sites is not independent. For example, if one CpG site has methylation change comparing two phenotypes, i.e., the CpG site is differentially methylated, it is more likely that its neighboring CpG sites are also differentially methylated. In order to explicitly account for the dependency of methylation change patterns over the CpG sites in a neighborhood, we first assume the differential states of CpG sites follow a Gibbs random field:

$$p\left(\mathbf{d}\right)=\frac{1}{Z}\exp\left(-\frac{1}{T}U(\mathbf{d})\right), \tag{B-1}$$

where $Z=\sum_{\mathbf{d}\in\mathbf{D}}\exp\left(-\frac{1}{T}U(\mathbf{d})\right)$. T is a constant value called temperature, which is usually assumed to be 1. $U(\mathbf{d})$ is the energy function, which can be calculated as follows:

$$U(\mathbf{d})=-(a_0n_0+a_1n_1)+bn_{01}, \tag{B-2}$$

where $\{a_0,a_1,b\}$ is the set of parameters for the MRF model; n_0 is the number of sites at state 0; n_1 is the number of sites at state 1; n_{01} is the number of connections linking two CpG sites with different states. a_0 and a_1 are arbitrary parameters and b is required to be larger than 0 to discourage inconsistency of states between neighboring CpG sites. Thus, the energy function is determined by both the potential of each CpG site and the consistency between each pair of neighboring CpG sites. The Gibbs random field is equivalent to a discrete Markov random field (DMRF). Thus, the differential state of CpG sites can be modeled by a discrete MRF model as follows.

$$p\left(\mathbf{d}\,|\,a_0,a_1,b\right)\propto\exp\left(a_0n_0+a_1n_1-bn_{01}\right). \tag{B-3}$$

By considering any two realizations of differential states which only differ at CpG site *i*, the conditional probability of differential state of CpG site *i* given all its neighbors can be derived as

$$p\left(d_i = 1 \mid d_{\partial i}, a_0, a_1, b\right) \propto \exp\left((a_1 - a_0) + b\left(n_i(1) - n_i(0)\right)\right), \tag{B-4}$$

where $n_i(1)$ is the number of neighbors of CpG site *i* with state 1, and $n_i(0)$ is the number of neighbors of CpG site *i* with state 0. In order to eliminate the effect caused by different numbers of neighboring sites for different CpG sites, we modify the conditional probability as follows:

$$p\left(d_i = 1 \mid d_{\partial i}, a_0, a_1, b\right) \propto \exp\left(a_1 - a_0 + b\left(n_i(1) - n_i(0)\right) \Big/ \sum_{k=1}^{M_n} w_{k,i}\right). \tag{B-5}$$

Specifically, we defined the conditional distribution of the differential state of CpG site *i* as

$$p\left(d_i = 1 \mid d_{\partial i}, a, b\right) = \frac{\exp\left(a + b\left(n_i(1) - n_i(0)\right) \Big/ \sum_{k=1}^{M_n} w_{k,i}\right)}{1 + \exp\left(a + b\left(n_i(1) - n_i(0)\right) \Big/ \sum_{k=1}^{M_n} w_{k,i}\right)}, \tag{B-6}$$

where

$$n_i(1) = \sum_{k=1}^{M_n}\left(w_{k,i} d_k\right), \text{ and } n_i(0) = \sum_{k=1}^{M_n}\left(w_{k,i}\left(1 - d_k\right)\right).$$

a and b are model parameters for the DMRF model. Parameter b controls the consistency of differential state in DMRF. The larger b is, the more consistent the differential state is in a neighborhood.

Bibliography

[1] J. L. Snoep, F. Bruggeman, B. G. Olivier, and H. V. Westerhoff, "Towards building the silicon cell: a modular approach," *Biosystems*, vol. 83, pp. 207-16, Feb-Mar 2006.

[2] U. Sauer, M. Heinemann, and N. Zamboni, "Genetics. Getting closer to the whole picture," *Science,* vol. 316, pp. 550-1, Apr 27 2007.

[3] A. Ma'ayan, "Introduction to network analysis in systems biology," *Sci Signal,* vol. 4, p. tr5, Sep 13 2011.

[4] H. Kitano, "Computational systems biology," *Nature,* vol. 420, pp. 206-10, Nov 14 2002.

[5] National Human Genome Research Institute. (Feb. 14, 2014, Apr. 7). *A Brief Guide to Genomics*. Available: http://www.genome.gov/19016904

[6] W. S. Klug, *Concepts of genetics*, 10th ed. San Francisco: Pearson Education, 2012.

[7] J. Pevsner, *Bioinformatics and functional genomics*, 2nd ed. Hoboken, N.J.: Wiley-Blackwell, 2009.

[8] R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, and D. Altshuler, "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms," *Nature,* vol. 409, pp. 928-33, Feb 15 2001.

[9] B. S. Shastry, "SNP alleles in human disease and evolution," *J Hum Genet,* vol. 47, pp. 561-6, 2002.

[10] J. Staaf, G. Jonsson, M. Ringner, J. Vallon-Christersson, D. Grabau, A. Arason, H. Gunnarsson, B. A. Agnarsson, P. O. Malmstrom, O. T. Johannsson, N. Loman, R. B. Barkardottir, and A. Borg, "High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer," *Breast Cancer Res,* vol. 12, p. R25, 2010.

[11] S. Temam, H. Kawaguchi, A. K. El-Naggar, J. Jelinek, H. Tang, D. D. Liu, W. Lang, J. P. Issa, J. J. Lee, and L. Mao, "Epidermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer," *J Clin Oncol,* vol. 25, pp. 2164-70, Jun 1 2007.

[12] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nat Rev Genet,* vol. 7, pp. 85-97, Feb 2006.

[13] D. Lakich, H. H. Kazazian, Jr., S. E. Antonarakis, and J. Gitschier, "Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A," *Nat Genet,* vol. 5, pp. 236-41, Nov 1993.

[14] A. Bird, "Perceptions of epigenetics," *Nature,* vol. 447, pp. 396-8, May 24 2007.

[15] A. Bird, "DNA methylation patterns and epigenetic memory," *Genes Dev,* vol. 16, pp. 6-21, Jan 1 2002.

[16] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," *Nature,* vol. 403, pp. 41-5, Jan 6 2000.

[17] A. Meissner, "Epigenetic modifications in pluripotent and differentiated cells," *Nat Biotechnol,* vol. 28, pp. 1079-88, Oct 2010.

[18] N. Song, J. Liu, S. An, T. Nishino, Y. Hishikawa, and T. Koji, "Immunohistochemical Analysis of Histone H3 Modifications in Germ Cells during Mouse Spermatogenesis," *Acta Histochem Cytochem,* vol. 44, pp. 183-90, Aug 27 2011.

[19] S. B. Baylin and J. G. Herman, "DNA hypermethylation in tumorigenesis: epigenetics joins genetics," *Trends Genet,* vol. 16, pp. 168-74, Apr 2000.

[20] M. Kulis and M. Esteller, "DNA methylation and cancer," *Adv Genet,* vol. 70, pp. 27-56, 2010.

[21] J. Dopazo, E. Zanders, I. Dragoni, G. Amphlett, and F. Falciani, "Methods and approaches in the analysis of gene expression data," *J Immunol Methods,* vol. 250, pp. 93-112, Apr 2001.

[22] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet,* vol. 10, pp. 57-63, Jan 2009.

[23] N. L. Anderson and N. G. Anderson, "Proteome and proteomics: new technologies, new concepts, and new words," *Electrophoresis,* vol. 19, pp. 1853-61, Aug 1998.

[24] S. Mathivanan, B. Periaswamy, T. K. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. L. Ramachandra, and A. Pandey, "An evaluation of human protein-protein interaction data in the public domain," *BMC Bioinformatics,* vol. 7 Suppl 5, p. S19, 2006.

[25] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Res,* vol. 39, pp. D561-8, Jan 2011.

[26] A. P. Feinberg, "Phenotypic plasticity and the epigenetics of human disease," *Nature,* vol. 447, pp. 433-40, May 24 2007.

[27] S. B. Baylin and P. A. Jones, "A decade of exploring the cancer epigenome - biological and translational implications," *Nat Rev Cancer,* vol. 11, pp. 726-34, Oct 2011.

[28] A. P. Feinberg and B. Tycko, "The history of cancer epigenetics," *Nat Rev Cancer,* vol. 4, pp. 143-53, Feb 2004.

[29] M. Esteller, "CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future," *Oncogene,* vol. 21, pp. 5427-40, Aug 12 2002.

[30] J. G. Herman and S. B. Baylin, "Gene silencing in cancer in association with promoter hypermethylation," *N Engl J Med,* vol. 349, pp. 2042-54, Nov 20 2003.

[31] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J. B. Fan, and R. Shen, "High density DNA methylation array with single CpG site resolution," *Genomics,* vol. 98, pp. 288-95, Oct 2011.

[32] J. Sandoval, H. Heyn, S. Moran, J. Serra-Musach, M. A. Pujana, M. Bibikova, and M. Esteller, "Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome," *Epigenetics,* vol. 6, pp. 692-702, Jun 2011.

[33] A. P. Bauer, D. Leikam, S. Krinner, F. Notka, C. Ludwig, G. Langst, and R. Wagner, "The impact of intragenic CpG content on gene expression," *Nucleic Acids Res,* vol. 38, pp. 3891-908, Jul 2010.

[34] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, pp. 139-40, Jan 1 2010.

[35] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol,* vol. 11, p. R106, 2010.

[36] H. Wu, C. Wang, and Z. Wu, "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data," *Biostatistics,* vol. 14, pp. 232-43, Apr 2013.

[37] N. Leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, B. M. Smits, J. D. Haag, M. N. Gould, R. M. Stewart, and C. Kendziorski, "EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments," *Bioinformatics,* vol. 29, pp. 1035-43, Apr 15 2013.

[38] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nat Biotechnol,* vol. 31, pp. 46-53, Jan 2013.

[39] A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, and J. T. Leek, "Ballgown bridges the gap between transcriptome assembly and expression analysis," *Nat Biotechnol,* vol. 33, pp. 243-6, Mar 2015.

[40] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat Biotechnol,* vol. 28, pp. 511-5, May 2010.

[41] Z. H. Zhang, D. J. Jhaveri, V. M. Marshall, D. C. Bauer, J. Edson, R. K. Narayanan, G. J. Robinson, A. E. Lundberg, P. F. Bartlett, N. R. Wray, and Q. Y. Zhao, "A comparative study of techniques for differential expression analysis on RNA-Seq data," *PLoS One,* vol. 9, p. e103207, 2014.

[42] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter, "Improving RNA-Seq expression estimates by correcting for fragment bias," *Genome Biol,* vol. 12, p. R22, 2011.

[43] M. Hu, Y. Zhu, J. M. Taylor, J. S. Liu, and Z. S. Qin, "Using Poisson mixed-effects model to quantify transcript-level gene expression in RNA-Seq," *Bioinformatics,* vol. 28, pp. 63-8, Jan 1 2012.

[44] Z. Wei and H. Li, "A Markov random field model for network-based analysis of genomic data," *Bioinformatics,* vol. 23, pp. 1537-44, Jun 15 2007.

[45] B. P. Carlin and S. Chib, "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 57, pp. 473-484, 1995.

[46] W. R. Gilks, "Markov Chain Monte Carlo," in *Encyclopedia of Biostatistics*, ed: John Wiley & Sons, Ltd, 2005.

[47] S. Chib and E. Greenberg, "Understanding the Metropolis-Hastings Algorithm," *The American Statistician,* vol. 49, pp. 327-335, 1995.

[48] G. Casella and E. I. George, "Explaining the Gibbs Sampler," *The American Statistician,* vol. 46, pp. 167-174, 1992.

[49] K. D. Robertson, "DNA methylation and human disease," *Nat Rev Genet,* vol. 6, pp. 597-610, Aug 2005.

[50] L. Bullinger, M. Ehrich, K. Dohner, R. F. Schlenk, H. Dohner, M. R. Nelson, and D. van den Boom, "Quantitative DNA methylation predicts survival in adult acute myeloid leukemia," *Blood,* vol. 115, pp. 636-42, Jan 21 2010.

[51] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nat Rev Genet,* vol. 12, pp. 87-98, Feb 2011.

[52] B. T. Wilhelm and J. R. Landry, "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing," *Methods,* vol. 48, pp. 249-57, Jul 2009.

[53] A. Oshlack, M. D. Robinson, and M. D. Young, "From RNA-seq reads to differential expression results," *Genome Biol,* vol. 11, p. 220, 2010.

[54] J. Eswaran, A. Horvath, S. Godbole, S. D. Reddy, P. Mudvari, K. Ohshiro, D. Cyanam, S. Nair, S. A. Fuqua, K. Polyak, L. D. Florea, and R. Kumar, "RNA sequencing of cancer reveals novel splicing alterations," *Sci Rep,* vol. 3, p. 1689, 2013.

[55] Z. Wu, X. Wang, and X. Zhang, "Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq," *Bioinformatics,* vol. 27, pp. 502-8, Feb 15 2011.

[56] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics,* vol. 13, pp. 204-16, Apr 2012.

[57] K. D. Hansen, S. E. Brenner, and S. Dudoit, "Biases in Illumina transcriptome sequencing caused by random hexamer priming," *Nucleic Acids Res,* vol. 38, p. e131, Jul 2010.

[58] J. Gu, X. Wang, L. Halakivi-Clarke, R. Clarke, and J. Xuan, "BADGE: a novel Bayesian model for accurate abundance quantification and differential analysis of RNA-Seq data," *BMC Bioinformatics,* vol. 15 Suppl 9, p. S6, 2014.

[59] "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, pp. 61-70, Oct 4 2012.

[60] V. G. Cheung, R. R. Nayak, I. X. Wang, S. Elwyn, S. M. Cousins, M. Morley, and R. S. Spielman, "Polymorphic cis- and trans-regulation of human gene expression," *PLoS Biol,* vol. 8, 2010.

[61] D. Bottomly, N. A. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann, "Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays," *PLoS One,* vol. 6, p. e17820, 2011.

[62] H. Edgren, A. Murumagi, S. Kangaspeska, D. Nicorici, V. Hongisto, K. Kleivi, I. H. Rye, S. Nyberg, M. Wolf, A. L. Borresen-Dale, and O. Kallioniemi, "Identification of fusion genes in breast cancer by paired-end RNA-sequencing," *Genome Biol,* vol. 12, p. R6, 2011.

[63] M. A. Newton, C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui, "On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data," *J Comput Biol,* vol. 8, pp. 37-52, 2001.

[64] W. Li and T. Jiang, "Transcriptome assembly and isoform expression level estimation from biased RNA-Seq reads," *Bioinformatics,* vol. 28, pp. 2914-21, Nov 15 2012.

[65] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nat Methods,* vol. 5, pp. 621-8, Jul 2008.

[66] S. Tarazona, P. Furio-Tari, A. Ferrer, and A. Conesa, "NOISeq: Exploratory analysis and differential expression for RNA-seq data," *R package version 2.0.0.,* 2012.

[67] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, and F. Jaffrezic, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Brief Bioinform,* vol. 14, pp. 671-83, Nov 2013.

[68] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, Z. Su, T. M. Chu, F. M. Goodsaid, L. Pusztai, J. D. Shaughnessy, Jr., A. Oberthuer, R. S. Thomas, R. S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B. D. Gallas, X. Ge, D. B. Megherbi, W. F. Symmans, M. D. Wang, J. Zhang, H. Bitter, B. Brors, P. R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Chou, T. S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D. J. Dix, J. Dopazo, K. C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K. R. Hess, H. Hong, J. Huan, R. A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C. G. Lambert, L. Li, Y. Li, Z. Li, S. M. Lin, G. Liu, E. K. Lobenhofer, J. Luo, W. Luo, M. N. McCall, Y. Nikolsky, G. A. Pennello, R. G. Perkins, R. Philip, V. Popovici, N. D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S. J. Wang, J. Wu, Y. Wu, Q. Xie, W. A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A. B. Lucas, F. Berthold, R. J. Brennan, A. Buness, J. G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J. C. Fuscoe, L. Gatto, W. Ge, D. R. Goldstein, L. Guo, D. N. Halbert, J. Han, S. C. Harris, C. Hatzis, D. Herman, J. Huang, R. V. Jensen, R. Jiang, C. D. Johnson, G. Jurman, Y. Kahlert, S. A. Khuder, M. Kohl, J. Li, M. Li, Q. Z. Li, S. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R. A. Moffitt, D. Montaner, P. Mukherjee, G. J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G. P. Page, J. Parker, R. M. Parry, X. Peng, R. L. Peterson, J. H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A. H. Roter, F. W. Samuelson, M. M. Schumacher, J. D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T. H. Stokes, Q. Sun, P. Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S. C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J. C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, M. Zhang, C. Zhao, R. K. Puri, U. Scherf, W. Tong and R. D. Wolfinger, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nat Biotechnol,* vol. 28, pp. 827-38, Aug 2010.

[69] L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. de Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T. M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X. H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K.

Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. LeClerc, S. Levy, Q. Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. McDaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, S. Zhong, Y. Zong and W. Slikker, Jr., "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nat Biotechnol,* vol. 24, pp. 1151-61, Sep 2006.

[70] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, C. E. Mason, N. D. Socci, and D. Betel, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biol,* vol. 14, p. R95, Sep 10 2013.

[71] R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid, "Evaluation of DNA microarray results with quantitative gene expression platforms," *Nat Biotechnol,* vol. 24, pp. 1115-22, Sep 2006.

[72] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics,* vol. 11, p. 94, 2010.

[73] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, T. R. Dreszer, P. A. Fujita, L. Guruvadoo, M. Haeussler, R. A. Harte, S. Heitner, A. S. Hinrichs, K. Learned, B. T. Lee, C. H. Li, B. J. Raney, B. Rhead, K. R. Rosenbloom, C. A. Sloan, M. L. Speir, A. S. Zweig, D. Haussler, R. M. Kuhn, and W. J. Kent, "The UCSC Genome Browser database: 2014 update," *Nucleic Acids Res,* vol. 42, pp. D764-70, Jan 2014.

[74] E. M. Ciruelos Gil, "Targeting the PI3K/AKT/mTOR pathway in estrogen receptor-positive breast cancer," *Cancer Treat Rev,* vol. 40, pp. 862-71, Aug 2014.

[75] E. Paplomata and R. O'Regan, "The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers," *Ther Adv Med Oncol,* vol. 6, pp. 154-66, Jul 2014.

[76] L. A. DeGraffenried, L. Fulcher, W. E. Friedrichs, V. Grunwald, R. B. Ray, and M. Hidalgo, "Reduced PTEN expression in breast cancer cells confers susceptibility to inhibitors of the PI3 kinase/Akt pathway," *Ann Oncol,* vol. 15, pp. 1510-6, Oct 2004.

[77] A. R. Panigrahi, S. E. Pinder, S. Y. Chan, E. C. Paish, J. F. Robertson, and I. O. Ellis, "The role of PTEN and its signalling pathways, including AKT, in breast cancer; an assessment of relationships with other prognostic factors and with outcome," *J Pathol,* vol. 204, pp. 93-100, Sep 2004.

[78] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V.

Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human Protein Reference Database--2009 update," *Nucleic Acids Res,* vol. 37, pp. D767-72, Jan 2009.

[79] S. S. Azab, "Targeting the mTOR Signaling Pathways in Breast Cancer: More Than the Rapalogs," *Journal of Biochemical and Pharmacological Research,* vol. 1 pp. 75-83, June 2013.

[80] S. J. Thomas, J. A. Snowden, M. P. Zeidler, and S. J. Danson, "The role of JAK/STAT signalling in the pathogenesis, prognosis and treatment of solid tumours," *Br J Cancer,* vol. 113, pp. 365-71, Jul 28 2015.

[81] A. Ozgur, L. Tutar, and Y. Tutar, "Regulation of Heat Shock Proteins by miRNAs in human breast cancer," *Microrna,* vol. 3, pp. 118-35, 2014.

[82] L. C. Cooper, E. Prinsloo, A. L. Edkins, and G. L. Blatch, "Hsp90alpha/beta associates with the GSK3beta/axin1/phospho-beta-catenin complex in the human MCF-7 epithelial breast cancer model," *Biochem Biophys Res Commun,* vol. 413, pp. 550-4, Oct 7 2011.

[83] C. C. Kuo, C. M. Liang, C. Y. Lai, and S. M. Liang, "Involvement of heat shock protein (Hsp)90 beta but not Hsp90 alpha in antiapoptotic effect of CpG-B oligodeoxynucleotide," *J Immunol,* vol. 178, pp. 6100-8, May 15 2007.

[84] S. Ramchandani, S. K. Bhattacharya, N. Cervoni, and M. Szyf, "DNA methylation is a reversible biological signal," *Proc Natl Acad Sci U S A,* vol. 96, pp. 6107-12, May 25 1999.

[85] D. Wang, L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia, and S. Liu, "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data," *Bioinformatics,* vol. 28, pp. 729-30, Mar 1 2012.

[86] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, ed: Springer, 2005, pp. 397-420.

[87] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *Int J Epidemiol,* vol. 41, pp. 200-9, Feb 2012.

[88] T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, V. L. R, S. J. Clark, and P. L. Molloy, "De novo identification of differentially methylated regions in the human genome," *Epigenetics Chromatin,* vol. 8, p. 6, 2015.

[89] B. S. Pedersen, D. A. Schwartz, I. V. Yang, and K. J. Kechris, "Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values," *Bioinformatics,* vol. 28, pp. 2986-8, Nov 15 2012.

[90] L. M. Butcher and S. Beck, "Probe Lasso: a novel method to rope in differentially methylated regions with 450K DNA methylation data," *Methods,* vol. 72, pp. 21-8, Jan 15 2015.

[91] B. G. Leroux, X. Lei, and N. Breslow, "Estimation of disease rates in small areas: a new mixed model for spatial dependence," in *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, ed: Springer, 2000, pp. 179-191.

[92] P. Wei and W. Pan, "Network-based genomic discovery: application and comparison of Markov random field models," *J R Stat Soc Ser C Appl Stat,* vol. 59, pp. 105-125, Jan 1 2010.

[93] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. W. Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, and J. B. Fan, "High-throughput DNA methylation profiling using universal bead arrays," *Genome Res,* vol. 16, pp. 383-93, Mar 2006.

[94] R. A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddeloh, B. Wen, and A. P. Feinberg, "Comprehensive high-throughput arrays for relative methylation (CHARM)," *Genome Res,* vol. 18, pp. 780-90, May 2008.

[95] P. Du, X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics,* vol. 11, p. 587, 2010.

[96] F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, "DNA methylation profiling of human chromosomes 6, 20 and 22," *Nat Genet,* vol. 38, pp. 1378-85, Dec 2006.

[97] D. Clayton and J. Kaldor, "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics,* vol. 43, pp. 671-81, Sep 1987.

[98] D. M. Lee, Richard, "Locally adaptive spatial smoothing using conditional autoregressive models," *eprint arXiv:1205.3641*, 05/2012.

[99] D. Lee, "A comparison of conditional autoregressive models used in Bayesian disease mapping," *Spat Spatiotemporal Epidemiol,* vol. 2, pp. 79-89, Jun 2011.

[100] J. Atchison and S. M. Shen, "Logistic-normal distributions: Some properties and uses," *Biometrika,* vol. 67, pp. 261-272, 1980.

[101] J. Besag, J. York, and A. Molli é, "Bayesian image restoration, with two applications in spatial statistics," *Annals of the institute of statistical mathematics,* vol. 43, pp. 1-20, 1991.

[102] N. Cressie, "Statistics for spatial data: Wiley series in probability and statistics," *Wiley-Interscience, New York,* vol. 15, pp. 105-209, 1993.

[103] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 259-302, 1986.

[104] A. Gelman, *Bayesian data analysis*, 2nd ed. Boca Raton, Fla.: Chapman & Hall/CRC, 2004.

[105] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological),* pp. 1-38, 1977.

[106] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science,* vol. 262, pp. 208-14, Oct 8 1993.

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

[108] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The annals of applied statistics,* pp. 107-129, 2007.

[109] S. Tommasi, D. Karm, X. Wu, Y. Yen, and G. Pfeifer, "Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer," *Breast Cancer Research,* vol. 11, p. R14, 2009.

[110] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biol,* vol. 9, p. R137, 2008.

[111] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "GREAT improves functional interpretation of cis-regulatory regions," *Nat Biotechnol,* vol. 28, pp. 495-501, May 2010.

[112] E. Vire, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, and F. Fuks, "The Polycomb group protein EZH2 directly controls DNA methylation," *Nature,* vol. 439, pp. 871 - 874, 2006.

[113] H. Nishiyama, J. H. Gill, E. Pitt, W. Kennedy, and M. A. Knowles, "Negative regulation of G(1)/S transition by the candidate bladder tumour suppressor gene DBCCR1," *Oncogene,* vol. 20, pp. 2956-64, May 24 2001.

[114] Z. Li, X. Guo, Y. Wu, S. Li, J. Yan, L. Peng, Z. Xiao, S. Wang, Z. Deng, L. Dai, W. Yi, K. Xia, L. Tang, and J. Wang, "Methylation profiling of 48 candidate genes in tumor and matched normal tissues from breast cancer patients," *Breast Cancer Res Treat,* vol. 149, pp. 767-79, Feb 2015.

[115] V. K. Hill, L. B. Hesson, T. Dansranjavin, A. Dallol, I. Bieche, S. Vacher, S. Tommasi, T. Dobbins, D. Gentle, D. Euhus, C. Lewis, R. Dammann, R. L. Ward, J. Minna, E. R. Maher, G. P. Pfeifer, and F. Latif, "Identification of 5 novel genes methylated in breast and other epithelial cancers," *Mol Cancer,* vol. 9, p. 51, 2010.

[116] V. Coothankandaswamy, S. Elangovan, N. Singh, P. D. Prasad, M. Thangaraju, and V. Ganapathy, "The plasma membrane transporter SLC5A8 suppresses tumour progression through depletion of survivin without involving its transport function," *Biochem J,* vol. 450, pp. 169-78, Feb 15 2013.

[117] S. Elangovan, R. Pathania, S. Ramachandran, S. Ananth, R. N. Padia, S. R. Srinivas, E. Babu, L. Hawthorn, P. V. Schoenlein, T. Boettger, S. B. Smith, P. D. Prasad, V. Ganapathy, and M. Thangaraju, "Molecular mechanism of SLC5A8 inactivation in breast cancer," *Mol Cell Biol,* vol. 33, pp. 3920-35, Oct 2013.

[118] M. Rose, C. Schubert, L. Dierichs, N. T. Gaisa, M. Heer, A. Heidenreich, R. Knuchel, and E. Dahl, "OASIS/CREB3L1 is epigenetically silenced in human bladder cancer facilitating tumor cell spreading and migration in vitro," *Epigenetics,* vol. 9, pp. 1626-40, Dec 2014.

[119] Y. Wang, J. Li, Y. Cui, T. Li, K. M. Ng, H. Geng, H. Li, X. S. Shu, W. Liu, B. Luo, Q. Zhang, T. S. Mok, W. Zheng, X. Qiu, G. Srivastava, J. Yu, J. J. Sung, A. T. Chan, D. Ma, Q. Tao, and W. Han, "CMTM3, located at the critical tumor suppressor locus 16q22.1, is silenced by CpG methylation in carcinomas and inhibits tumor cell growth through inducing apoptosis," *Cancer Res,* vol. 69, pp. 5194-201, Jun 15 2009.

[120] L. Chen, J. Xuan, R. B. Riggins, Y. Wang, and R. Clarke, "Identifying protein interaction subnetworks by a bagging Markov random field-based method," *Nucleic Acids Res,* vol. 41, p. e42, Jan 2013.

[121] X. Wang, J. Gu, J. Xuan, A. N. Shajahan, R. Clarke, and L. Chen, "Sampling-Based Subnetwork Identification from Microarray Data and Protein-Protein Interaction Network," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, 2012, pp. 158-163.

[122] T. Ideker and N. J. Krogan, "Differential network biology," *Mol Syst Biol,* vol. 8, p. 565, 2012.

[123] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, "The Pfam protein families database," *Nucleic Acids Res,* vol. 40, pp. D290-301, Jan 2012.

[124] S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, and R. Jothi, "DOMINE: a comprehensive collection of known and predicted domain-domain interactions," *Nucleic Acids Res,* vol. 39, pp. D730-5, Jan 2011.

[125] W. Zhang, J. W. Chang, L. Lin, K. Minn, B. Wu, J. Chien, J. Yong, H. Zheng, and R. Kuang, "Network-Based Isoform Quantification with RNA-Seq Data for Cancer Transcriptome Analysis," *PLoS Comput Biol,* vol. 11, p. e1004465, Dec 2015.