STATISTICAL METHODS TO DETECT ALLELE SPECIFIC EXPRESSION, ALTERATIONS OF ALLELE SPECIFIC EXPRESSION AND DIFFERENTIAL EXPRESSION

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

JING XIE

JIANGUO (TONY) SUN, Dissertation Supervisor

MAY 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

STATISTICAL METHODS TO DETECT ALLELE SPECIFIC EXPRESSION, ALTERATIONS OF ALLELE SPECIFIC EXPRESSION AND DIFFERENTIAL EXPRESSION

presented by Jing Xie,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Jianguo (Tony) Sun

Dr. Tieming Ji

Dr. Chong He

Dr. Rocio M. Rivera

Dr. Yushu Shi

ACKNOWLEDGMENTS

Foremost, I would like to give my sincerest gratitude to my advisors, Dr. Jianguo (Tony) Sun and Dr. Tieming Ji. It was Dr. Ji who led me into the world of statistical genetics and taught me how to conduct research. The mentorship I have received from Dr. Sun is one of the most valuable things during my doctoral study. I cannot image how I would survive the long journey of the doctoral study without the tremendous advice and support, both personally and professionally, from Dr. Ji and Dr. Sun.

Besides, I would like to express my heartfelt thanks to my dissertation committee members, Dr. Chong He, Dr. Rocio M. Rivera and Dr. Yushu Shi, for their insightful comments and questions.

Meanwhile, financial support from National Science Foundation, Curator's Grant-in-Aid Scholarship, and the Data Science and Analytics master program is greatly thanked.

Additionally, I truly appreciate the help from other professors, staff, and graduate students. Their selfless support provided me with sufficient courage to overcome all the difficulties I had encountered these years. Those include but not limited to Dr. Christopher K. Wikle, Dr. Sakis Micheas, Judy Dooley, Abbie R. Van Nice, Kaiyi Chen, Shirley Rojas Salazar, Yuanyuan Guo, Ruiwen Zhou, Dr. Dayu Sun, Dian Yang, Dr. Wenyang Wang, Peng Shao, Yahan Li, Dr. Jiaxun Chen and Dr. Dongyan Yan. I sincerely apologize for those whom I may have left out.

Last but not least, I would like to thank my caring, loving, and supportive husband Dr. Zhengyang Wang, and other family members. Thank you all for always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Statistical Methods to Detect Allele Specific Expression, Alterations of Allele Specific Expression and Differential Expression

Jing Xie

Dr. Jianguo (Tony) Sun, Dissertation Supervisor

ABSTRACT

The advent of next-generation sequencing (NGS) technology has facilitated the recent development of RNA sequencing (RNA-seq), which is a novel mapping and quantifying method for transcriptomes. By RNA-seq, one can measure the expression of different features such as gene expression, allelic expression, and intragenic expression in the forms of read counts. These features have provided new opportunities to study and interpret the molecular intricacy and variations that are potentially associated with the occurrence of specific diseases. Therefore, there has been an emerging interest in statistical method to analyze the RNA-seq data from different perspectives. In this dissertation, we focus on three important challenges: identifying allele specific expression (ASE) on the gene level and single nucleotide polymorphism (SNP) level simultaneously, the detection of ASE regions in the control group and regions of ASE alterations in case group simultaneously, and detecting genes whose expression levels are significantly different across treatment groups (DE genes).

In Chapter 2, we propose a method to test ASE of a gene as a whole and variation in ASE within a gene across exons separately and simultaneously. A generalized linear mixed model is employed to incorporate variations due to genes, SNPs, and

biological replicates. To improve reliability of statistical inferences, we assign priors on each effect in the model so that information is shared across genes in the entire genome. We utilize the Bayes factor to test the hypothesis of ASE for each gene and variations across SNPs within a gene. We compare the proposed method to competing approaches through simulation studies that mimicked the real datasets. The proposed method exhibits improved control of the false discovery rate and improved power over existing methods when SNP variation and biological variation are present. Besides, the proposed method also maintains low computational requirements that allows for whole genome analysis. As an example of real data analysis, we apply the proposed method to four tissue types in a bovine study to *de novo* detect ASE genes in the bovine genome, and uncover intriguing predictions of regulatory ASEs across gene exons and across tissue types.

In Chapter 3, we propose a new and powerful algorithm for detecting ASE regions in a healthy control group and regions of ASE alterations in a disease/case group compared to the control. Specifically, we develop a bivariate Bayesian hidden Markov model (HMM) and an expectation-maximization inferential procedure. The proposed algorithm gains advantages over existing methods by addressing their limitations and by recognizing the complexity of biology. First, the bivariate Bayesian HMM detects ASEs for different mRNA isoforms due to alternative splicing and RNA variants. Second, it models spatial correlations among genomic observations, unlike existing methods that often assume independence. At last, the bivariate HMM draws inferences simultaneously for control and case samples, which maximizes the utilization of available information in data. Real data analysis and simulation studies that mimic real data sets are shown to illustrate the improved performance and practical utility of the proposed method.

In Chapter 4, we present a new method to detect DE genes in any sequencing

experiment. The number of read counts for different treatment groups are modelled by two Negative Binomial distributions which may have different means but share the same dispersion parameter. We propose a mixture prior model for the dispersion parameters with a point mass at zero and a lognormal distribution. The mixture model allows shrinkage across genes within each of the two mixture components, thus prevents the overcorrection resulting from shrinkage across all genes. The simulation studies demonstrate that the proposed method yields a better dispersion estimation and FDR control, and a higher accuracy in gene ranking. In addition, the proposed method exhibits robustness to the misspecification of the bimodal distribution for the dispersion parameters, thus is flexible and can be easily generalized.

Chapter 1

General Introduction

1.1 RNA-seq data

The advent of next-generation sequencing (NGS) technology has facilitated the recent development of RNA sequencing (RNA-seq), which is a novel mapping and quantifying method for transcriptomes. In a typical RNA-seq experiment, a population of RNA is at first converted to a library of cDNA fragments. With or without amplification, in the second step, each molecule is sequenced in a high-throughput manner to obtain short sequences (reads). Depending on the DNA-sequencing technology used, length of the reads are usually between 30 and 400 base pairs (bps). Example DNA-sequencing platforms include the Illumina IG [73, 62, 123], Applied Biosystems SOLiD [16, 123], and Roche 454 Life Science [23, 123], etc. The last step following sequencing is to align the resulting reads to either a reference genome or reference transcripts. The end-product of a RNA-seq experiment is a genome-scale transcription map that consists of both the transcriptional structure and/or level of expression for each gene [123].

Compared with other transcriptomics methods, such as hybridization-based approach (microarrays), sanger sequencing of cDNA or EST sequencing, and tag-based sequencing approach (SAGE, CAGE, MPSS) [46, 120, 123], RNA-seq is based on high-throughput sequencing and has several advantages. First, RNA-seq is not limited to detecting transcripts that correspond to existing genomic sequence. Second, since DNA sequences can been unambiguously mapped to unique regions of the genome, RNA-seq is believed to be more accurate with a lower background noise. Thrid, the required amount of RNA and the cost for mapping transcriptomes of large genomes using RNA-seq technique are relatively low [120]. These advantages makes RNA-seq gradually replace microarrays and become the new standard for transcriptomics studies [123].

By RNA-seq, one can measure the expression of different features such as gene expression, allelic expression, and intragenic expression in the forms of read counts. Table 1.1 shows a table of read counts from a hypothetical RNA-seq experiment. In this example, there are two biological replicates (samples) for each of two treatment groups. As the hypothetical table indicates, the typical data from RNA-seq experiment has three important features: (1) there are a large number of genes available, usually tens of thousands; (2) the number of replicates is limited, often three or four due to the high expense associated with acquiring biological replicates and high-throughput sequencing experiments; (3) the data is discrete rather than continuous. These features has led to an emerging interest in statistical method to analyze the data from different perspective, such as differential expression analysis, and allelic expression analysis.

Table 1.1: Table of read counts from a hypothetical RNA-seq experiment.

Gene	SNP	Control group						Case group					
		sample 1			sample 2			sample 1			sample 2		
		M	P	Total	M	P	Total	M	P	Total	M	P	Total
1	1	50	89	139	45	79	124	100	110	210	107	123	230
	2	48	90	138	40	80	120	110	100	210	120	112	232
	3	55	80	135	48	75	123	106	90	196	110	105	215
2	1	12	40	52	20	38	58	10	45	55	16	39	55
	2	14	42	56	22	40	62	12	47	59	18	41	59
	3	11	39	50	19	37	56	9	44	53	15	38	53
	4	22	50	72	30	48	78	20	55	75	26	49	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10,000	1	10	12	22	13	11	24	11	40	51	10	38	48
	2	20	18	38	19	18	37	10	30	40	12	29	41

1.2 Allele specific expression

Allele specific expression (ASE) in diploid genomes refers to a phenomenon that the two alleles of a gene express substantially differently. One such example involves imprinted genes whose allele expression is based on the parent of origin [11, 7]. Imprinted genes are mainly or completely expressed from either the maternally or paternally inherited allele but not both, so the total expression from genomic copies is the appropriate amount for healthy and viable organisms [100]. Another prominent example is X-chromosome inactivation in mammals [128, 54], where one copy of the X chromosome is inactivated in female cells to maintain the same dosage of X-linked genes compared to male cells. The choice of which X chromosome is silenced is random initially, but once chosen, the same X chromosome remains inactive in subsequent cell divisions. In a third and rather random case, allelic imbalance occurs when there are mutations in *cis*-regulatory regions of one allele, leading to differential expression of two alleles [27, 78].

ASE presents in both healthy and tumor tissues [29, 103] across species. Studies that analyzed genome-wide ASE in humans [63, 117], mouse [31], bovine [129], and drosophila [31] have discovered hundreds to a few thousands of genes that exhibit significant ASE [36]. ASE is essential for normal development and many cellular processes, and may lead to phenotypic variation depending on the function of the genes or result in disease if impaired [95, 8]. However, it is not biologically clear what series of mechanisms a cell employs to precisely initiate ASE during fetal development and consistently maintain it through a lifetime. Therefore, a tremendous recent interest has been focused on understanding the underlying mechanisms of ASE, and on utilizing ASE as a direct approach to connect genotype and disease susceptibility. For example, [80] used ASE analysis as a marker to identify candidate genes associated with alcohol use disorders. [20] found that colorectal cancer risk increased as the imbalance of ASE in gene adenomatous polyposis coli increased.

To understand diseases associated with ASE or aberrations in ASE, a fundamental approach is to compare ASE status between a healthy control group and a case/disease group. That is, to seek for the answers to two questions. First, which genes exhibit ASE in the normal samples. Second, which regions of a specific chromosome exhibit ASE and how the ASE status varies across normal and case samples. To that end, researchers need to estimate genome-wide ASE status in the control group and whole genome ASE status changes between the case and control groups.

High-throughput sequencing experiments, which can determine allele origins, have been used to assess genome-wide ASE. Despite the amount of data generated from high-throughput experiments such as RNA-seq, statistical methods are either too simplistic to understand the complexity of gene expression, or too computational expensive to apply to the entire genome. [70], [35] and [97] are exanples of the existing methods which can answer the first aforementioned question. However, [70]

and [35] assume that ASE of a gene or a region is constant across single nucleotide polymorphisms (SNPs), and they do not test ASE of a gene as a whole and variation in ASE within a gene across exons separately and simultaneously. [97] adopts a Markov chain Monte Carlo (MCMC) method to compute posterior probabilities for inferences of genes and SNPs and thus requires an extensive computational power. Alternatively, we propose a generalized linear mixed model to incorporate variations due to genes, SNPs, and biological replicates. The proposed approach can detect ASE genes and ASE variations within genes simultaneously while maintaining a low computational requirement. Coupled with exon and RNA transcript information, the statistical predictions provided by the proposed method can produce detailed, biologically relevant, intriguing results that enable researchers to examine the molecular mechanisms of ASE regulation in detail.

For the exploration of the answer to the second question, existing methods usually fail to identify boundaries of ASE regions precisely [89, 77, 66, 70, 35, 24], and ignore the potential spatial correlation between SNPs and the correlation between ASE in the control sample and ASE alterations in the case samples [97, 35, 70, 24]. To close these research gaps, we further take into account the spatical correlation between adjcent SNPs, and develop a bivariate Bayesian hidden Markov model (HMM) and an expectation-maximization (EM) method for detecting ASE in control group and alternations of ASE in case group simultaneously. We refer to the proposed method as hmmASE algorithm. The hmmASE algorithm gains advantages over existing methods by addressing the limitations of current practice and recognizing the complexity of biology. Specifically, the bivariate Bayesian HMM detects ASE at the exon level (SNPs) rather than testing ASE for a whole gene. In addition, it models spatial correlations among SNPs, unlike existing methods that often assume independence across observations. The bivariate HMM draws inferences simultaneously for control and

case samples that exploits information among observations at the same SNP locus.

1.3 Differential expression

As the most common form of transcriptome analysis, differential expression analysis plays an important role in the characterization and understanding of the molecular basis of phenotypic variation in biology, including diseases [101]. Differential expression analysis involves searching for a set of genes in the whole genome whose mean expression levels are substantially different across treatment conditions, such as control versus disease. With the development of novel high-throughput DNA sequencing methods (RNA-seq), developing statistical methods for detecting differentially expressed (DE) genes has been an extensively studied research area.

Negative Binomial has been widely used to model the gene expression data in the form of read counts by virtue of its flexibility to embrace overdispersion. To overcome the small sample size problem in dispersion estimation, Bayesian method by shrinking dispersion estimates towards the mean dispersion estimates of all genes has been universally adopted by many statistical methods to detect DE genes [88, 87, 84, 69, 2, 64, 34]. However, biologically and empirically speaking, genes have intrinsic dispersions that are related with their functions. Forcing dispersions of all genes to be "similar" to each other could introduce biases for genes intrinsically with high or low dispersions. We propose a mixture prior model for dispersion parameters with a point mass at zero and a lognormal distribution, where the former models genes whose dispersion parameter is low and close to zero; and the latter models genes exhibiting uncontrolled high dispersion due to technical and biological replications. The mixture model allows shrinkage across genes within each of the two mixture compo-

nents, preventing overcorrection resulting from shrinking across all genes. Through real data analysis and simulation studies, we demonstrate the flexibility of the proposed model, and previous statistical methods with lognormal prior (or normal prior) are special cases of the proposed mixture model. Simulation studies also suggest improved estimation of dispersion parameters and power in differentially expressed gene detection.

1.4 Organization of dissertation

In the rest of this dissertation, each main chapter corresponds to a proposed method targeting at each of the aforementioned challenges. Chapter 2 presents the proposed Bayesian linear mixed regression model for the detection of ASE genes and ASE variations across SNPs within a gene. Chapter 3 extends the work in Chapter 2 by taking into account the spatial correlation among SNPs and by incorporating the comparison between control samples and case samples. The new proposed method hmmASE is based on a bivariate Bayesian Hidden Markov Model, and can simultaneously detect the ASE region in the control group and region of ASE alterations in the case group. In Chapter 4, we switch the focus from allelic expression to differential expression. We propose a mixture prior model on dispersion parameters with a point mass at zero and a lognormal probability distribution, The former mixture component represents genes whose dispersion parameter is small and close to zero, and the latter corresponds to genes with high uncontrolled dispersion across biological replicates. For each of the proposed method depicted in Chapters 2-4, simulation studies and real data analysis are conducted to evaluate the performance and empirical utility. Chapter 5 discusses the potential directions for future research.

Chapter 2

Modeling Allele Specific Expression at the Gene and SNP Levels Simultaneously by a Bayesian Logistic Mixed Regression Model

2.1 Introduction

In a diploid cell, the two alleles of a gene inherited from maternal and paternal parents express roughly equally for most genes. However, research has uncovered a group of genes in the genome where two copies of a gene express substantially differently, a phenomenon known as allelic imbalance. One such example involves imprinted genes whose allele expression is based on the parent of origin [11, 7]. Imprinted genes are mainly or completely expressed from either the maternally or paternally inherited allele but not both, so the total expression from genomic copies is the appropriate amount for healthy and viable organisms [100]. Another prominent example is X-

chromosome inactivation in mammals [128, 54], where one copy of the X chromosome is inactivated in female cells to maintain the same dosage of X-linked genes compared to male cells. The choice of which X chromosome is silenced is random initially, but once chosen, the same X chromosome remains inactive in subsequent cell divisions. In a third and rather random case, allelic imbalance occurs when there are mutations in *cis*-regulatory regions of one allele, leading to differential expression of two alleles [27, 78].

Allelic imbalance affects approximately 5-10% of genes in the mammalian genome [54], but it is not biologically clear what series of mechanisms a cell employs to precisely initiate allele-specific expression (ASE) during fetal development and consistently maintain it through a lifetime. Several common congenital human disorders are caused by mutations or deletions within these ASE regions, such as Beckwith-Wiedemann syndrome (BWS) [17, 124], which characterizes an array of congenital overgrowth phenotypes; Angelman syndrome [3], which characterizes nervous system disorders; and Prader-Willi syndrome, in which infants suffer from hyperphagia and obesity.

To understand the molecular mechanisms underlying ASEs and human developmental defects due to misregulated ASE regions, a powerful and accurate computational algorithm to detect genome-wide ASEs is urgently needed. The binomial exact test, employed in AlleleSeq [89], is one of the most widely used methods to test ASEs due to its simplicity. [77] uses analysis of variance (ANOVA) in their proposed pipeline Allim. [66] fits a mixture of folded Skellam distributions to the absolute values of read differences between two alleles. However, these abovementioned statistical methods draw conclusions based on observations produced from one gene; due to the expensive cost of acquiring tissue samples and sequencing experiments, most laboratories can only afford three or four biological replicates. Depending on sequencing

depth, genes may also have low read counts, limiting the power of the aforementioned methods.

In searching for more powerful and reliable ASE detection methods, several groups have proposed Bayesian approaches to share information across genes and thus improve gene-related inferences on average. For instance, the MBASED method [70] and the QuASAR method [35] all assume the read counts follow binomial distributions with a beta prior on the probability parameter. In their statistical models, they assume that ASE of a gene or a region is constant across SNPs. However, ASE is known to vary within a gene due to alternative splicing [75, 119], which is essentially universal in human multi-exon genes that comprise 94% of genes overall [30, 119]. Therefore, a highly desirable feature of ASE detection methods is identification of ASE genes and ASE variations within genes across multiple exons. [97] developed a flexible statistical framework that satisfied this requirement. It assumes a binomial distribution with a beta prior. Additionally, it places a two-component mixture prior on the parameters of the beta-binomial model. A Markov chain Monte Carlo (MCMC) method was adopted to compute posterior probabilities for inferences of genes and SNPs. However, due to the extensive computational power required in the MCMC calculation for one gene and the large number of genes in the entire genome, this method is not empirically appealing. Other relevant methods include the EAGLE method [50] that detects associations between environmental variables and ASEs, the WASP method [116] that addresses incorrect genotype calls, and the RASQUAL method [51] that detects gene regulatory effects.

In this chapter, we propose a new statistical method that addresses the above-mentioned challenges. Specifically, the proposed approach can detect ASE genes and ASE variations within genes simultaneously while maintaining a low computational requirement. Coupled with exon and RNA transcript information, the statistical

predictions provided by the proposed method can produce detailed, biologically relevant, intriguing results that enable researchers to examine the molecular mechanisms of ASE regulation in detail.

Particularly, we model the logistic transformation of the probability parameter in the binomial model as a linear combination of the gene effect, single nucleotide polymorphism (SNP) effect, and biological replicate effect. The random SNP effect permits ASE to vary within a gene; the random replicate effect accounts for extra dispersion among biological replicates beyond binomial variation. To overcome the low number of biological replicates and/or low number of read counts of a gene, we propose a hierarchical model with a Gaussian prior on the fixed gene effect and inverse gamma priors, respectively, on the variance components of the random SNP and replicate effects. We test hypotheses via Bayesian model selection method based on model posterior probabilities. To compute posterior probabilities, we propose combining the empirical Bayes method and Laplace approach to approximate integrations, leading to substantially reduced computational power requirements compared to MCMC. We illustrate the utility of the proposed method by applying it to the bovine genome in [13], which motivated this chapter; findings reveal for the first time highly detailed information regarding the testing results for whole-genome ASEs, unveiling inspiring ASE variations across exons and across tissue types. To compare the proposed method with existing approaches, we simulate data that mimic real datasets to ensure that the comparison results can be reproduced in practice. The proposed method outperforms existing methods in false discovery rate (FDR) control of detecting ASEs and variations therein across SNPs. We call the proposed method the Bayesian Logistic Mixed Regression Model (BLMRM) method. The R package, BLMRM, for the proposed method is publicly available for download at https://github.com/JingXieMIZZOU/BLMRM.

2.2 Bayesian generalized linear mixed model

Let n_{gjk} denote the total number of read counts for the kth biological replicate of gene g at its jth SNP, where $g = 1, 2, \ldots, G, j = 1, 2, \ldots, J_g,$ and $k = 1, 2, \ldots, K$. Let y_{gjk} denote the number of read counts from the maternal allele of replicate k. We model $y_{gjk} \sim \text{Binomial}(n_{gjk}, p_{gjk})$, where p_{gjk} denotes the proportion of gene expression from the maternal allele for gene g at SNP j of replicate k. It is known that using the RNA-seq approach to detect ASEs can produce bias during mapping because reads from the reference allele are more likely to be mapped due to fewer number of mismatches compared to reads from alternative alleles [21]. Potential solutions have been proposed in [21, 92, 13] to correct mapping bias. Here and throughout this dissertation, n_{gjk}'s and y_{gjk}'s denote the read counts after bias correction.

The objective of this chapter is to detect genes and regions within a gene whose expression is significantly different between the maternal and paternal alleles. Most existing methods assumed equal gene expression across all SNPs of a given gene; however, research discoveries have disproven this assumption for several reasons [10, 18], including alternative splicing and RNA variants. Thus, we model y_{gjk} as

$$\begin{aligned} y_{gjk} &\sim \text{Binomial}(n_{gjk}, p_{gjk}), \text{ and} \\ \log\frac{p_{gjk}}{1-p_{gjk}} &= \beta_g + S_{gj} + R_{gk}, \end{aligned} \tag{2.1}$$

where β_g is the fixed gene effect; S_{gj} is the random SNP effect and $S_{gj} \overset{iid}{\sim} \text{N}(0, \sigma^2_{sg})$; R_{gk} is the random replicate effect and $R_{gk} \overset{iid}{\sim} \text{N}(0, \sigma^2_{rg})$. We also assume S_{gj}'s and R_{gk}'s are mutually independent. Therefore, the null hypothesis $H_0 : \beta_g = 0$ is to test whether gene g exhibits imbalanced allelic expression. Furthermore, $H_0 : \sigma^2_{sg} = 0$ is to examine whether maternal (and/or paternal) gene expression percentage is the

same across all SNPs of a gene.

Due to the expense of sample collection and sequencing experiments, most laboratories can only afford a few biological replicates, such as $K = 3$ or 4. In addition, the number of available SNPs in a gene also depends on the diversity between parental alleles. Often, only a small number of genes contain a large number of SNPs. Thus, for most genes, the estimates of β_g, σ^2_{sg}, and σ^2_{rg} are not robust, leading to unreliable statistical inferences. To improve estimation accuracy, we assume hierarchical priors on β_g, σ^2_{sg}, and σ^2_{rg} to share information across all genes in the genome. Specifically, we assume $\sigma^2_{sg} \overset{iid}{\sim} \text{IG}(a_s, b_s)$, $\sigma^2_{rg} \overset{iid}{\sim} \text{IG}(a_r, b_r)$, and a Gaussian prior on the gene effect $\beta_g \overset{iid}{\sim} \text{N}(\mu, \sigma^2)$. The hyperparameters a_s, b_s, a_r, b_r, μ, and σ^2 no longer have the subscript g because they are estimated by pooling observations from all genes. Given that there are tens of thousands of genes in the genome, the estimates of these prior hyperparameters are accurate.

2.3 Detection of imbalanced allelic gene expression through Bayesian model selection

Next, we describe the proposed Bayesian model selection method to detect ASE at the gene level and corresponding variations across SNPs. Based on model (2.1), there are four models, indexed by $m \in \{1, 2, 3, 4\}$, in model space $\mathcal{M}$, where $\beta_g = 0$ and $\sigma^2_{sg} = 0$ in Model 1; $\beta_g \neq 0$ and $\sigma^2_{sg} = 0$ in Model 2; $\beta_g = 0$ and $\sigma^2_{sg} \neq 0$ in Model 3; and $\beta_g \neq 0$ and $\sigma^2_{sg} \neq 0$ in Model 4. For each gene g, we select model m in $\mathcal{M}$, which has the largest posterior probability defined as

$$P(m|\mathbf{y}^g, \mathbf{n}^g) = \frac{P(m)P(\mathbf{y}^g|m, \mathbf{n}^g)}{\sum_{m=1}^{4} P(m)P(\mathbf{y}^g|m, \mathbf{n}^g)}$$

$$\propto P(m)P(\mathbf{y}^g|m,\mathbf{n}^g), \tag{2.2}$$

where $\mathbf{y}^g = (y_{g11},\ldots,y_{gJ_gK})'$ and $\mathbf{n}^g = (n_{g11},\ldots,y_{gJ_gK})'$. $P(m)$ denotes the prior probability of model m. Without prior information, we assume a uniform prior on space $\mathcal{M}$. Thus, our objective is to select a model m in $\mathcal{M}$ that maximizes the marginal likelihood $P(\mathbf{y}^g|m,\mathbf{n}^g)$, which, when comparing two models, is equivalent to choosing the model m using the Bayes factor. Let $\mathbf{b}_g$ denote all random effects; that is, $\mathbf{b}_g = (S_{g1},\ldots,S_{gJ_g},R_{g1},\ldots,R_{gK})'$. Hence,

$$\begin{aligned} P(\mathbf{y}^g|m,\mathbf{n}^g) = \iiiint & P(\mathbf{y}^g|\beta_g,\mathbf{b}_g,\mathbf{n}^g,m)P(\beta_g)\times \\ & P(\mathbf{b}_g|\sigma^2_{sg},\sigma^2_{rg})P(\sigma^2_{sg},\sigma^2_{rg})\times \\ & d\beta_g\, d\mathbf{b}_g\, d\sigma^2_{sg}\, d\sigma^2_{rg}. \end{aligned} \tag{2.3}$$

A direct integration of (2.3) is difficult because an analytical result of the density is not a closed form. An alternative approach is to use Laplace approximation to iteratively approximate each integral; however, in our experience, this leads to error accumulated through each layer of integration and thus affects the accuracy of results. To overcome this problem, we propose a combination of empirical Bayes estimation and Laplace approximation. Inspired by the approach in [82], we obtain the following empirical Bayes estimators.

$$\widetilde{\beta}_g = E(\beta_g|\widehat{\beta}_g) \approx \frac{\widehat{\mathrm{Var}(\beta_g)}\widehat{\mu} + \widehat{\sigma}^2\widehat{\beta}_g}{\widehat{\mathrm{Var}(\beta_g)} + \widehat{\sigma}^2}, \tag{2.4}$$

$$\widetilde{\sigma}^2_{sg} = E(\sigma^2_{sg}|\widehat{\sigma}^2_{sg}) \approx \frac{d_{sg}\widehat{\sigma}^2_{sg} + 2\widehat{b}_s}{d_{sg} + 2\widehat{a}_s}, \text{ and} \tag{2.5}$$

$$\widetilde{\sigma}^2_{rg} = E(\sigma^2_{rg}|\widehat{\sigma}^2_{rg}) \approx \frac{d_{rg}\widehat{\sigma}^2_{rg} + 2\widehat{b}_r}{d_{rg} + 2\widehat{a}_r}, \tag{2.6}$$

where $\widetilde{\beta}_g$, $\widetilde{\sigma}^2_{sg}$, and $\widetilde{\sigma}^2_{rg}$ denote the empirical Bayes estimates of β_g, σ^2_{sg}, and σ^2_{rg}, respectively. $\widehat{\beta}_g$, $\widehat{\mathrm{Var}(\beta_g)}$, $\widehat{\sigma}^2_{sg}$, and $\widehat{\sigma}^2_{rg}$ are maximum likelihood estimates from model (2.1). $\widehat{\mu}$, $\widehat{\sigma}^2$, $\widehat{a}_r$, $\widehat{b}_r$, $\widehat{a}_s$, and $\widehat{b}_s$ are estimated hyperparameters whose estimation method will be introduced in detail later in this section. d_{rg} and d_{sg} are degrees of freedom of the random SNP and random replicate effect, respectively, with $d_{sg} = J_g - 1$ and $d_{rg} = K - 1$. We enter these empirical Bayes estimates directly into (2.3), obtaining the approximation:

$$P(\mathbf{y}^g|m,\mathbf{n}^g) \approx \int P(\mathbf{y}^g|\widetilde{\beta}_g,\mathbf{b}_g,m,\mathbf{n}^g)\times P(\mathbf{b}_g|\widetilde{\sigma}^2_{sg},\widetilde{\sigma}^2_{rg})\,d\mathbf{b}_g. \tag{2.7}$$

Accordingly, (2.3) is reduced to (2.7), which requires only one step of Laplace approximation. The objective in combining empirical Bayes estimates and Laplace approximation is to develop a method with improved power and accuracy while maintaining affordable computational power that allows for empirical application. In the simulation study, we compared the proposed approach with the method using pure Laplace approximation. We found that the proposed method is superior than purely using Laplace approximation with respect to FDR control and true positive rate (see simulation results section). This approach also greatly decreases computational requirements compared to MCMC, considering there are tens of thousands of genes in an entire genome [42]. For instance, the method in [97] employs an MCMC algorithm for identifying ASE. With the default setting, their approach took approximately 1.5 hour to analyze 50 genes, whereas the proposed method took approximately 3 minutes.

We still need to estimate hyperparameters μ, σ^2, a_s, b_s, a_r, and b_r. To avoid extreme values that produce unstable estimates, we first let $y^*_{gjk} = y_{gjk} + 1$ and

$n^*_{gjk} = n_{gjk} + 2$. Then, based on y^*_{gjk}'s and n^*_{gjk}'s, μ and σ^2 are estimated by the method of moments using significant $\widehat{\beta}_g$ via likelihood ratio tests when controlling FDR at 0.05. a_s, b_s, a_r, and b_r are estimated based on y^*_{gjk}'s and n^*_{gjk}'s by the maximum likelihood method, where a_s and b_s are based on significant estimates of $\widehat{\sigma}^2_{sg}$'s via likelihood ratio tests and controlling FDR at 0.05, and a_s and b_s are based on $\widehat{\sigma}^2_{rg}$'s from all genes.

Finally, we test $H_0 : \beta_g = 0$ and $H_0 : \sigma^2_{sg} = 0$ for gene g by choosing Model m, where $m = \underset{\gamma \in \{1,2,3,4\}}{\arg\max}\, P(\gamma|\mathbf{y}^g, \mathbf{n}^g)$ for $g = 1, \ldots, G$. Let $P(g \in \{m\}|\mathbf{y}^g, \mathbf{n}^g)$ denote the posterior probability of gene g being sampled from Model m. The posterior probability of a gene exhibiting an ASE gene effect is $P(g \in \{2, 4\}|\mathbf{y}^g, \mathbf{n}^g)$. Similarly, the posterior probability of a gene exhibiting ASE variations across SNPs is $P(g \in \{3, 4\}|\mathbf{y}^g, \mathbf{n}^g)$. Finally, the posterior probability of a gene exhibiting an ASE gene effect and ASE variations across SNPs simultaneously is $P(g \in \{4\}|\mathbf{y}^g, \mathbf{n}^g)$. We adopt the following method to control FDR that have been used in [42, 19]. To control the FDR when testing the ASE gene effect, we order $P(g \in \{2, 4\}|\mathbf{y}^g, \mathbf{n}^g)$, $g = 1, \ldots, G$, from largest to smallest. Let $g_{(1)}, \ldots, g_{(G)}$ be the ordered genes; then, we find the largest l such that $\sum_{i=1}^{l}(1 - P(g_{(i)} \in \{2, 4\}|\mathbf{y}^{g_{(i)}}, \mathbf{n}^{g_{(i)}}))/l \leq \alpha$, where α is a pre-defined FDR threshold. We declare the first l genes are significant for testing $H_0 : \beta_g = 0$ when FDR is controlled at α level. The same strategy is used to control FDR for testing ASE variations among SNPs and gene and SNP variation effects simultaneously.

2.4 Simulation study

Simulation studies based on real datasets can best evaluate empirical usage and performance. In this section, we describe the simulation design which is based on the real dataset in [13], and compare the proposed BLMRM method with the binomial test, ANOVA, MBASED, generalized linear mixed model (GLMM), and the BLMRM method with pure Laplace approximation.

In each simulation, we simulated 4,000 genes in total with 1,000 genes for each of the four models in $\mathcal{M}$. To base the simulation study upon real datasets, we randomly selected 4,000 genes from liver tissue in the real dataset and used the numbers of SNPs of these genes as the numbers of SNPs for the 4,000 simulated genes. To ensure consistency with the real dataset, we set the number of biological replicates to be four.

Real data from liver tissue in [13] indicates a linear relationship between the logarithm of average total read counts and that of the sample standard deviation of total read counts within a gene across SNPs. Real data also indicates a roughly linear relationship between the logarithm of average total read counts and that of the sample standard deviation of total read counts within a SNP across four replicates. To simulate n_{gjk}, we utilized these two linear relationships. Specifically, let $\bar{n}_g$ denote the sample average of the total read count of gene g across SNPs; that is, $\bar{n}_g = \sum_{j=1}^{J_g}(\bar{n}_{gj})/J_g$ where $\bar{n}_{gj} = \sum_{k=1}^{K} n_{gjk}/\mathrm{K}$. For the liver tissue in real data, by regressing $\log S(\bar{n}_g)$ on $\log(\bar{n}_g)$ with a simple linear model where $S(\cdot)$ denotes the sample standard deviation, we obtained fitted intercept $\widehat{\alpha}_1 = -0.36$ and slope $\widehat{\alpha}_2 = 0.97$. Hence, for each simulated gene, we independently sampled $\log \bar{n}_{g1}, \ldots, \log \bar{n}_{gJ_g} \sim \mathrm{N}$ ($\mu = \log \bar{n}_g$, and $\sigma = \widehat{\alpha}_1 + \widehat{\alpha}_2 \log \bar{n}_g$), where $\bar{n}_g$'s were computed from the 4,000 genes randomly selected from the real dataset. Next, we fit a linear regression model between $\log S(\bar{n}_{gj})$ and

log $(\bar{n}_{gj})$, which yielded an estimated intercept $\widehat{\alpha}_3 = -0.53$ and slope $\widehat{\alpha}_4 = 0.77$. Similarly, we simulated $n_{gj1}, \ldots, n_{gj4} \sim \text{N}\ (\mu = \log\ \bar{n}_{gj}, \sigma = \widehat{\alpha}_3 + \widehat{\alpha}_4 \log\ \bar{n}_{gj})$. We rounded the simulated values to ensure n_{gjk}'s were integers.

Given the simulated n_{gjk}'s, to simulate y_{gjk}'s, we needed to simulate p_{gjk}'s. We simulated gene effect β_g uniformly from $\{-4.39, -1.20, -0.41, 0.41, 1.20, 4.39\}$ for genes where $\beta_g \neq 0$. 0.41, 1.20, and 4.39 are the 10th, 50th, and 90th percentiles of absolute values of $\widehat{\beta}_g$'s, respectively, when significant gene ASEs are reported by the GLMM in (2.1). We simulated $\sigma^2_{sg} \overset{iid}{\sim} \text{IG}\ (\widehat{a}_s, \widehat{b}_s)$, $S_{gj} \overset{iid}{\sim} \text{N}\ (0, \sigma^2_{sg})$, and simulated $\sigma^2_{rg} \overset{iid}{\sim} \text{IG}\ (\widehat{a}_r, \widehat{b}_r)$, $R_{gk} \overset{iid}{\sim} \text{N}\ (0, \sigma^2_{rg})$, where $\widehat{a}_s$, $\widehat{b}_s$, $\widehat{a}_r$, and $\widehat{b}_r$ are hyperparameter estimates from the liver tissue whose values are given in real data analysis section. p_{gjk} was computed as $\exp(\beta_g + S_{gj} + R_{gk})/(1 + \exp(\beta_g + S_{gj} + R_{gk}))$. At last, we simulated $y_{gjk} \sim \text{Binomial}(n_{gjk}, p_{gjk})$. We repeated such simulation 10 times to assess variations in performance.

We compared the BLMRM method with the binomial test, ANOVA test in [77], MBASED method in [70], and GLMM in (2.1) without Bayesian priors. The binomial test and ANOVA test only detect the gene effect; the MBASED method can detect gene ASE and SNP variation separately but not simultaneously; and the GLMM and BLMRM methods can detect the gene effect, SNP variation, and gene ASE and SNP variation simultaneously. For the binomial, ANOVA, MBASED, and GLMM methods, we applied Storey's method [108] to estimate and control FDR. The FDR control for the BLMRM method was described in the Method section.

For the proposed BLMRM method, the hyperparameter estimation is accurate and stable across 10 simulations. The mean of absolute biases across 10 simulations are 0.61, 0.12, 0.08, and 0.06, respectively, for $\widehat{a}_s$, $\widehat{b}_s$, $\widehat{a}_r$, and $\widehat{b}_r$; and the standard deviations of these 10 absolute biases are 0.17, 0.08, 0.04, and 0.00.

Table 2.1 summarizes the average true FDR and average true positive rate (TPr)

Table 2.1: Assess of FDR control and TPr when controlling estimated FDR at 0.05.

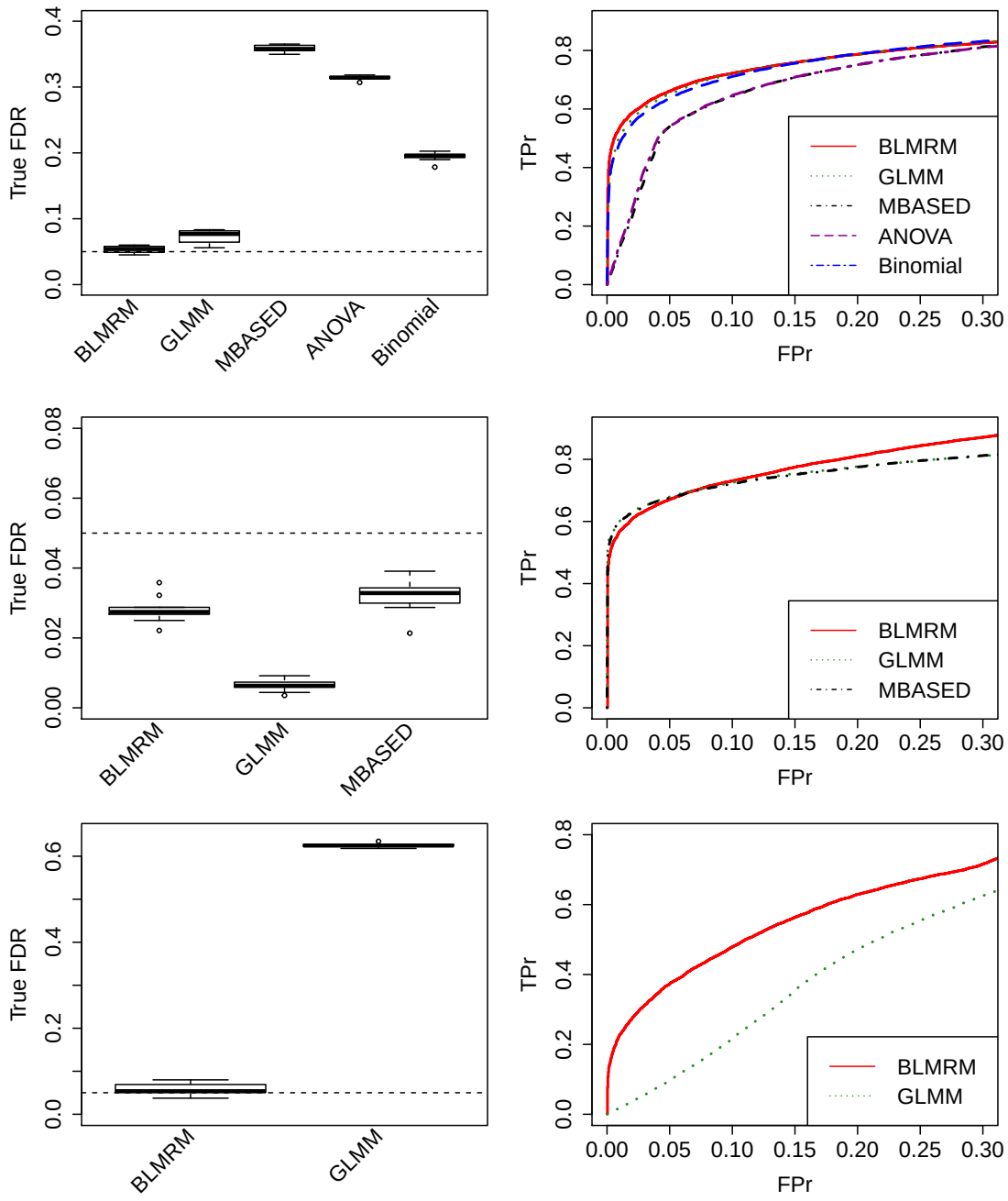
Method	True FDR			TPr(%)		
	gene	SNP	gene-SNP	gene	SNP	gene-SNP
BLMRM	0.053	0.028	0.059	66.37	60.82	17.51
	(0.006)	(0.004)	(0.014)	(0.87)	(1.80)	(1.65)
BLMRM	0.060	0.030	0.094	68.87	56.82	17.50
(pure Laplace)	(0.006)	(0.002)	(0.008)	(0.29)	(1.19)	(0.91)
GLMM	0.073	0.006	0.625	68.66	57.20	86.72
	(0.010)	(0.002)	(0.004)	(1.52)	(1.49)	(0.86)
MBASED	0.358	0.032	-	91.34	64.32	-
	(0.006)	(0.005)	-	(0.54)	(1.51)	-
ANOVA	0.194	-	-	82.02	-	-
	(0.007)	-	-	(1.04)	-	-
Binomial	0.314	-	-	88.26	-	-
	(0.003)	-	-	(0.80)	-	-

across 10 simulations when we control the estimated FDR at 0.05. Numbers in parentheses are sample standard deviations. Results suggested that among all methods under investigation, only the proposed BLMRM method controlled FDR at the nominal level. The BLMRM method with pure Laplace approximation does not control FDR for simultaneous test on both gene effect and SNP variation. In addition, the proposed BLMRM method also has slightly higher TPr than the pure Laplace approximation approach in testing SNP variation. This suggests that the combined method of empirical Bayes and one time of Laplace approximation provides more accurate results than three layers of Laplace approximation. The GLMM method was slightly liberal in testing gene ASE, overly conservative in testing the random SNP effect, and overly liberal in testing simultaneous gene ASE and SNP variation. The MBASED and binomial test methods did not control FDR when testing the gene effect. The MBASED method can not test gene ASE and ASE variation across SNPs simultaneously. Thus, under our simulation scenario, the MBASED method does not

correctly separate observed variations among multiple sources of variations; i.e., gene ASE, SNP variation, biological variation, and error variation.

We plotted the box plots of true FDRs across 10 simulations in the left panel of Figure 2.1, respectively, on testing the gene effect, SNP effect, and gene and SNP effects simultaneously when controlling the estimated FDR at 0.05, which represents same conclusions on FDR control in Table 2.1. The right panel in Figure 2.1 displays the ROC curves when the false positive rate (FPr) was between 0 and 0.3. Compared to the other competing methods, the BLMRM method showed greater partial area under the ROC curves (AUCs) in testing gene ASE, SNP variation in ASE, and gene and SNP variation simultaneously. The GLMM and BLMRM methods were competitive for gene ranking when testing gene and SNP variation; however, the BLMRM method substantially outperformed the GLMM method in gene ranking when detecting simultaneous ASE gene effect and ASE variation within a gene.

Figure 2.1: FDR and ROC comparison. Top row shows results for testing the gene effect; middle row shows results for testing SNP variation within a gene; bottom row shows results for simultaneously testing gene ASE and SNP variation. Left panel shows box plots of true FDR across 10 simulations when controlling estimated FDR = 0.05; right panel presents ROC curves.



2.5 Real data analysis

Most of the imprinted genes identified to date have been in the mouse [125]. Original work, identified the non-equivalency of the parental alleles by generating embryos which only had maternal chromosomes (gynogenotes and parthenogenotes) or paternal chromosomes (androgenotes) [71, 112]. By doing this, investigators identified which genes are expressed exclusively from each chromosome. In mouse genome, there are various types of genetic rearrangements including translocations, duplications and deletions . Other studies that used mice have noticed that the direction in which the allele was inherited (either through the mother or the father) mattered for the successful development and wellbeing of the offspring [6]. Subsequent work turned to genetic manipulations to identify the function of imprinted genes in mice. More recently, with the advent of genome wide approaches, investigators have generated large datasets from F1 individuals generated from the breeding of two inbred (homozygous) strains of mice [37]. An advantage of using mice to do this type of work is that most strains have been sequenced and all animals within a strain will have the same maternal and paternal DNA sequence. While useful, the mouse model does not always faithfully represent other mammals [76]. In addition, most laboratory mice are inbred (homozygous) while other mammals are heterozygous which incorporates complexity to the analysis of identifying parental alleles. As imprinted gene expression is species-specific, tissue-specific, and developmental stage specific [125], investigators would have to do monetary and animal expensive studies to identify novel imprinted genes and their potential function in health and disease.

A current limitation for investigators working in the area of genomic imprinting in heterozygote animals such as bovine, is the difficulty to assess whether a gene or a region in a gene has ASE for the entire genome. For example, in the case in which

4 fetuses are obtained from the breeding of one cow and one bull, each of the fetuses may have a specific combination of alleles (penitentially 4 combinations), making the identification of imprinted gene expression a daunting task, not to mention extremely expensive. Therefore, new computational tools and analyses must be devised in order to provide investigators knowledge of allelic imbalances in the transcriptome which may then be used to do locus-specific wet bench work to determine the accuracy of the predictions.

Specifically, [13] measured gene expressions of four normal female F1 conceptuses (fetus and placenta) generated from the mating of Bos taurus taurus (mother) and Bos taurus indicus (father). Tissues were retrieved from the brain, kidney, liver, skeletal muscle, and placenta of these four conceptuses. RNA-seq experiments were conducted on each tissue type for each replicate.

Aligning RNA-seq reads to a non-identical reference genome has been shown to introduce alignment bias [21, 106]. To address the mapping bias problem, [13] combined the reference genome (i.e., the *B. t. taurus* reference genome UMD3.1 build) and the pseudo *B. t. indicus* genome to create a custom diploid genome. Specifically, the sire's DNA was subjected to next generation sequencing (DNA-seq) to identify all SNPs between his genome and the *B. t. taurus* reference genome. Then Genome Analysis Toolkit (GATK) [72] and SAMtools [59] pipelines were applied for SNP calling and only SNPs identified by both pipelines were used to generate a pseudo *B. t. indicus* genome. At last, RNA-seq reads from the *B. t. indicus* $\times$ *B. t. taurus* F1 conceptuses were mapped to the diploid genome using both the HISAT2 [49] and BWA [60] pipelines and only variants identified by both methods were retained to minimize the potential effects of false positives. The resulting datasets are publicly available at the Gene Expression Omnibus database under accession number GSE63509.

We used the BLMRM method to separately analyze liver, kidney, muscle, and

brain tissue data from [13]. Missing values are not uncommon in real datasets, especially when dealing with heterozygous species (for example, cattle and humans), as not all replicates share the same set of SNPs among parental alleles. We first filtered out genes containing only one SNP or for which all SNPs were not represented by at least two individuals. We also removed genes for which the observed maternal and paternal expression percentages were constant across all replicates and all SNPs as statistical inferences are straightforward in such a scenario. In total, 9,748 genes remained for analysis, among which many had low numbers of total RNA-seq read counts.

We then applied the proposed BLMRM method to these 9,748 genes. Hyperparameters were estimated using the method described in the Method section. For example, for liver tissue, we have $\widehat{\mu} = 0.43$, $\widehat{\sigma}^2 = 4.62$, $\widehat{a}_s = 2.35$, $\widehat{b}_s = 1.37$, $\widehat{a}_r = 2.03$, and $\widehat{b}_r = 0.09$.

We identified several examples containing varied and informative patterns of tissue-specific and/or exon-specific ASEs. Here, we present four genes: *AOX1*, *HACL1*, *TMEM50B*, and *IGF2R*. Aldehyde oxidase 1 (*AOX1*; XLOC_003018) is a cytosolic enzyme expressed at high levels in the liver, lung, and spleen but at a much lower level in many other organs since this gene plays a key role in metabolizing drugs containing aromatic azaheterocyclic substituents [61, 26]. By controlling FDR at 0.05, the BLMRM method identified gene *AOX1* as exhibiting ASE at the gene level in the brain, kidney, and muscle, and biallelically expressed in the liver (top panel in Figure 1). The vertical axis in Figure 1 indicates the observed sample average percentage of gene expression from the maternal allele. The bar around each sample average denotes the 95% confidence interval at each SNP. SNPs are drawn with ascending genomic locations in a chromosome. The bottom of each panel in Figure 1 shows the distribution of SNPs in exons from annotated RefSeq transcripts of this gene. Conclusions

from the proposed BLMRM method coincide with *AOX1* gene functional analysis. Using the binomial exact test, [13] only found that *AOX1* had preferential paternal expression in bovine muscle and failed to detect ASE in the brain and kidney. The proposed method also suggests significant ASE variations across SNPs in the liver, kidney, and muscle with FDR at the 0.05 level. Interestingly, regions in the liver showing ASE variations corresponded to the 16th, 17th, and 18th exons housing the 5-7th and 14-16th SNPs. Given this exon- and tissue-specific information, biologists can examine the ASE regulatory mechanism in detail.

2-hydroxyacyl-CoA lyase (*HACL1*; XLOC_001524) is involved in perixosomal branched fatty acids oxidation and primarily expressed in the liver [25]. The proposed BLMRM method identified *HACL1* as exhibiting significant ASE at the gene level and its variations across SNPs. Figure 1 Panel 2 visualizes our observations and shows a clear maternal preference of expression for the first 15 SNPs, whereas the remaining six suggest biallelic expression of this gene. This surprising finding spurred further investigation, upon which we identified that the first 15 SNPs belong to exon 17 of alternative splice variant XM_010801748.2 while the last SNPs are shared between two or three splice isoforms (i.e. NM_001098949.1, XM_015474169.1, and XM_010801748.2). No further information is available regarding the ASE mechanism of this gene, as this is the first time we have retrieved such detailed statistical results for each gene in an entire genome within a short computational window. Future work will identify whether this ASE gene is a novel imprinted gene and if, in fact, this gene shows variant-specific imprinted expression as has been documented for other genes [104].

Transmembrane protein 50B (*TMEM50B*; XLOC_000329) is a ubiquitously expressed housekeeping gene. The proposed method identified this gene to be biallelically expressed in all analyzed tissues (Figure 1, Panel 3) as expected for a housekeeping gene. Interestingly, The proposed method also predicted significant variations

across SNPs in each of these four tissue types. Upon investigating detailed activity of this gene, Figure 1 indicates that a portion of the 3' UTR of this transcript appears to have maternal preference. The consistent pattern across tissues motivated us to understand the importance of this SNP variation. We hypothesize that this corresponds to a specific RNA variant required for maintaining cellular function.

Finally, insulin-like growth factor 2 receptor (*IGF2r*; XLOC_018398) is a well-known maternally expressed mannose receptor that targets IGF2 for degradation [22]. This gene is imprinted in the liver, kidney, and muscle (Figure 1, Panel 4) but has biallelic expression in the brain of mice and cattle [67, 12]. In addition, *IGF2r* is lowly expressed in the cattle brain [12]. Prediction results from the proposed method coincide with the literature.

By controlling FDR at 0.05, Figure 2 summarizes the numbers of detected ASE genes, numbers of genes with ASE variations across SNPs, and numbers of genes exhibiting ASE at the gene level and ASE variations across SNPs simultaneously, respectively, among the four tissues. We conducted some further analysis on these detected genes. For instance, in the top Venn diagram, among the 37 detected ASE genes shared by all four tissue types, 11 of them cannot be mapped to the set of annotated genes using the UMD 3.1 build. Among the rest of 26 annotated and detected ASE genes, we found that three of them had been documented as imprinted genes across all or most of these four tissue types. These three imprinted genes are (1) *GSTK1* that is maternally expressed in human placenta but unknown in other human tissues [1], paternally expressed in mouse kidney, liver, muscle, and maternally expressed in mouse brain [79], maternally expressed in bovine oocyte and unknown in other bovine tissues [90]; (2) *PLAGL1* that is paternally expressed in human kidney, muscle, and unknown in other human tissues [45], paternally expressed in mouse muscle, kidney, and brain [79], and paternally expressed in bovine brain, kidney, mus-

cle, and liver [83]; (3) *BEGAIN*, which is unknown in human genome, preferentially expressed from the paternal allele in mouse neonatal brain [114], paternally expressed in bovine kidney and muscle with strong statistical evidence though no biological verification yet [12], and found to be paternally expressed in sheep kidney, liver, muscle, and brain (all four) tissue types [98]. Excluding these three documented imprinted genes, the other 23 annotated ASE genes detected by the BLMRM method are *de novo* detected ASE genes and their biological relevance await experimental verification.

Collecting all ASE genes from the first Venn diagram in Figure 2.3, we summarized the number of detected ASE genes on each chromosome (see Supplementary Table 1). We found several interesting patterns. For instance, chromosomes 11 and 21 tend to have more ASE genes than other chromosomes for all tissue types. Besides, the X chromosome has more ASE genes in brain tissue than other tissue types. Supplementary Figure 2.2 plots distributions of these ASE genes in each chromosome, revealing several ASE clusters. Among all detected ASE genes, most ASE genes show preference of the maternal allele than the paternal allele. Specifically, 79%, 74%, 68%, and 71% ASE genes show maternal preference in the brain, liver, kidney, and muscle tissues, respectively.

At this stage, we are not able to statistically distinguish imprinted genes from other type of ASE genes as further experiment data are required to separate imprinting from other ASE molecular mechanisms. However, collecting all the detected ASE genes from all three Venn diagrams in Figure 2.3, we found that seven *de novo* detected ASE genes are highly likely to be imprinted in the bovine genome but they have not been documented in any bovine study. They are: (1) *GATM*, *SNX14*, and *NT5E*, which are imprinted in mouse [91, 38]; (2) *IGF1R* and *RCL1*, which are imprinted in human [111, 105]; and (3) *KLHDC10* and *SLC22A18*, which are imprinted in both

human and mouse [33, 5]. These genes are involved in varied physiological functions. For example, *GATM* encodes an arginine glycine amidinotransferase (AGAT) which is involved in creatine synthesis [44, 107]. *NT5E* encodes the protein CD73 (cluster of differentiation 73), a cell surface anchored molecule with ectoenzymatic activity that catalyzes the hydrolysis of AMP into adenosine and phosphate and has been shown to mediate the invasive and metastatic properties of cancers [131, 28]. *SNX14* is a protein coding gene involved in maintaining normal neuronal excitability and synaptic transmission [38] and may be involved in intracellular trafficking [113]. *IGF1R* is a receptor tyrosine kinase that mediates the actions of insulin-like growth factor 1 (IGF1). *IGF1R* is involved in cell growth and survival and has a crucial role in tumor transformation and survival of malignant cells [122, 48]. *RCL1* is a protein-coding gene with roles in 18 S rRNA biogenesis and in the assembly of the 40 S ribosomal subunit [9, 47]. The Kelch repeat protein *KLHDC10* activates the apoptosis signal-regulating kinase 1 (ASK1) through the suppression of protein phophatase 5 [94] and activation of the ASK1 contributes in oxidative stress-mediated cell death through the activation of the JNK and p38 MAPK pathways [99]. *SLC22A18* plays a role in lipid metabolism [40] and also acts as a tumor suppressor [93]. Visualization of significant expression pattern of these seven genes are plotted in Supplementary Figure 2 along with its significance level assessed by FDR.

So far, no existing statistical methods can provide simultaneous inferences at both gene and exon (SNPs) levels for the entire genome in a short computational window, like the *de novo* detection for the bovine genome shown here. We are able to achieve this goal because we model multiple sources of variations (i.e., genes, SNPs, biological replicates, error variation) in one statistical model and adopt an efficient estimation method (i.e., a combination of empirical Bayes and Laplace approximation) for model selection, that is designed for whole genome analysis.

Figure 2.2: Percentage of gene expression from maternal allele in brain, liver, kidney, and muscle, respectively. The top panel shows gene *AOX1*. The second panel shows gene *HACL1*. The third panel shows gene *TMEM50B*, and the bottom panel shows gene *IGF2r*. SNPs are drawn with ascending genomic locations. The bottom of each panel shows distribution of SNPs in exons from all RefSeq annotated transcripts of this gene. Rectangles represent exons (only those with SNPs are shown) with exon numbers indicated under each rectangle. Lengths of exons are not drawn to scale.

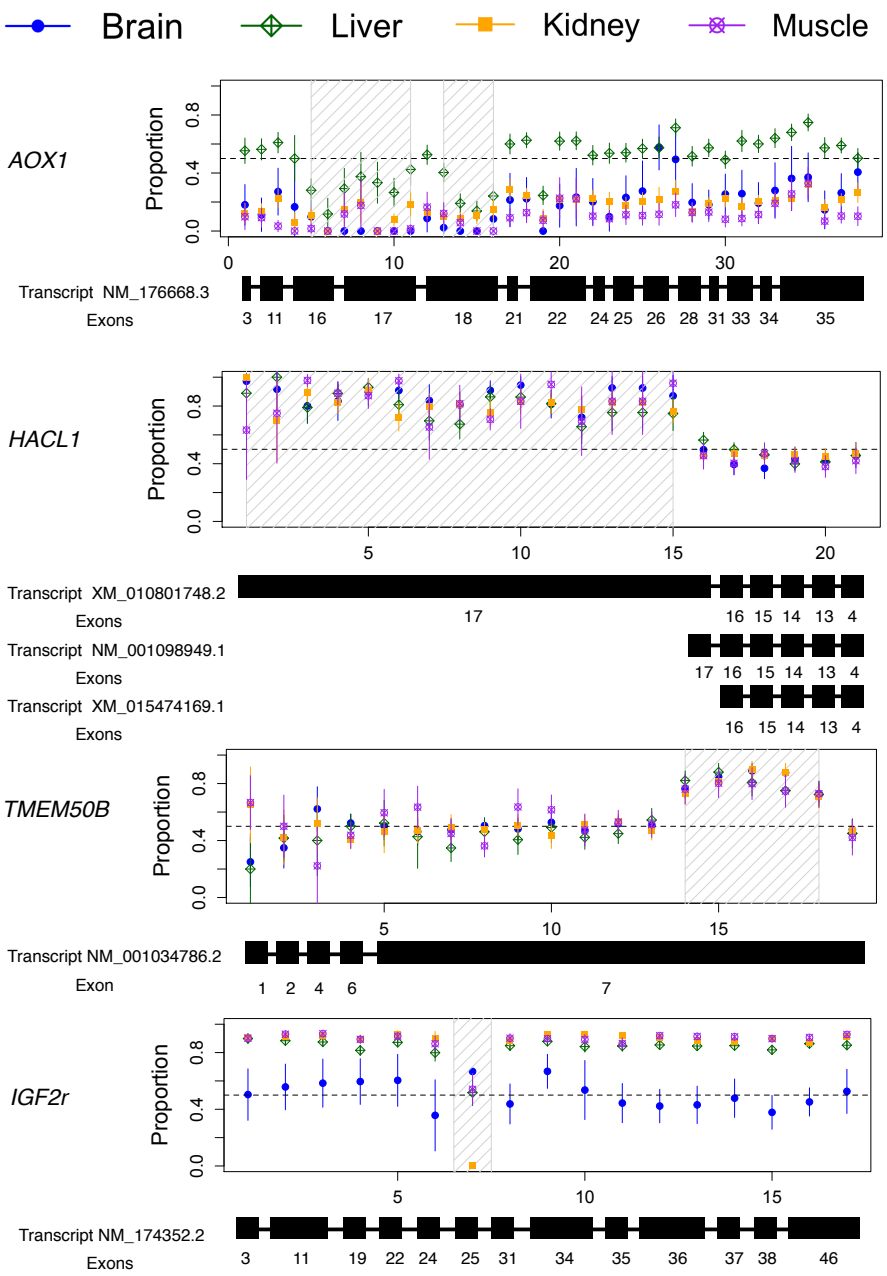
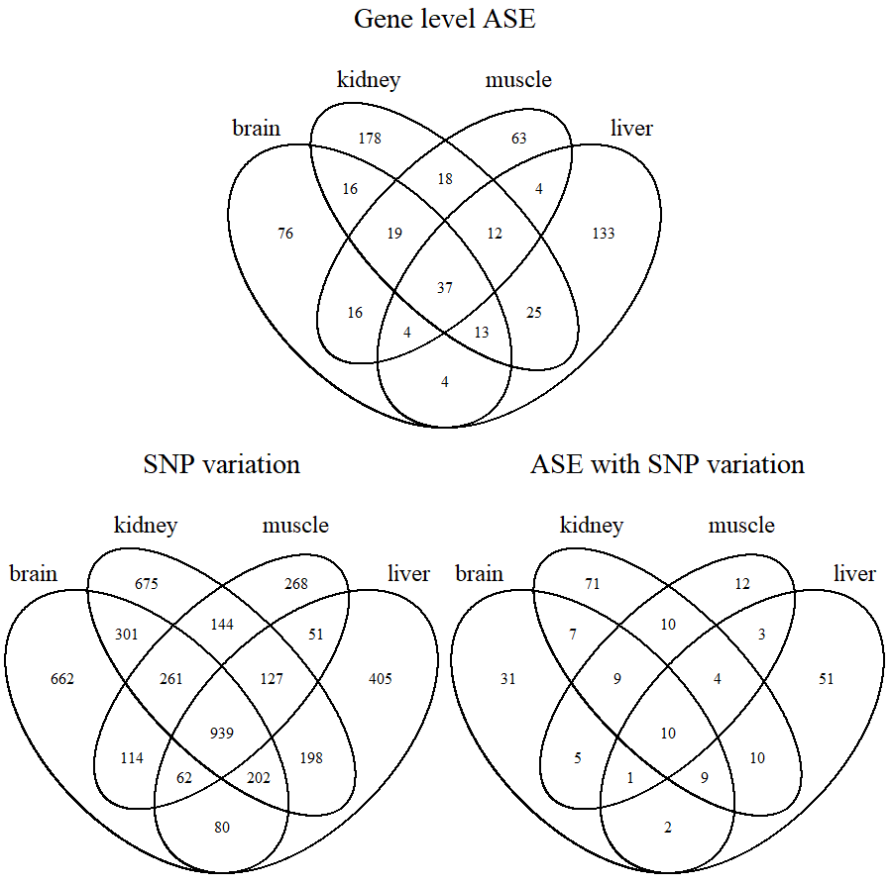


Figure 2.3: Venn Diagram of detected ASEs across tissue types. Number of significant genes (estimated FDR=0.05) across four tissue types when testing ASE at the gene level, testing ASE variations across SNPs, and testing ASE gene and ASE variations within a gene simultaneously.



2.6 Conclusions

We have proposed a new method, BLMRM, to detect ASE for any RNA-seq experiment. Specifically, we propose a Bayesian logistic mixed regression model that accounts for variations from genes, SNPs, and biological replicates. To improve the reliability of inferences on ASE, we assign hyperpriors on genes, SNPs, and replicates, respectively. The hyperprior parameters are empirically estimated using observations from all genes in an entire genome. We then develop a Bayesian model selection method to test the ASE hypothesis on genes and variations of SNPs within a gene. To select a fitting model based on Bayes factors, we adopt a combination of the empirical Bayesian method and Laplace approximation method to substantially accelerate computation. To illustrate the utility of the proposed method, we have applied it to the bovine study that motivated this chapter; findings reveal the potential of the proposed method for application to real data analysis. We also conduct simulation studies that mimic the real data structure. The real data application and simulation study demonstrate the improved power, accuracy, and empirical utility of the proposed method compared to existing approaches. The R package, BLMRM, based on the proposed method is available to download via Github at https://github.com/JingXieMIZZOU/BLMRM.

The model specified in (2.1) is based on three fundamental assumptions. First, the numbers of read counts from maternal allele for different genes are independent, i.e., $y_{gjk} \perp y_{g'jk}$ when $g \neq g'$. This assumption allows the gene-wise detection of ASE and ASE variations across SNPs. When a group of genes share biological functions under common regulatory control, their expression patterns may be similar or correlated. However, the correlation of the expression profiles between genes is out of the scope of this chapter. Other research areas such gene expression correlation analysis [74] and

gene co-expression network [109] focus more on this issue. Second, the observations at different SNPs are assumed to be independent, i.e., S_{gj}'s are independent and identically distributed, in consideration of simplicity. The distribution of SNPs within a gene is not necessarily homogenous. Thus, it is reasonable to assume neighboring SNPs that are close in their genomic location tend to have correlation. Negligence of the potential spatial correlation among SNPs is the limitation of model (2.1). This limitation will be addressed in Chapter 3 of this dissertation by a bivariate Bayesian hidden Markov model. Third, we assume that there is no interaction between the random SNP effect and random replicate effect, i.e., $R_{gk} \perp S_{gj}$. This assumption is more trivial than the aforementioned two assumptions since the number of biological replicates and the number of SNPs are usually small, and it is not biologically meaningful to study the interaction.

Chapter 3

Detecting Allele Specific Expression and Alterations of Allele Specific Expression by a Bivariate Bayesian Hidden Markov Model

3.1 Introduction

Allele specific expression (ASE) in diploid genomes refers to a phenomenon that the two alleles of a gene express substantially differently. An extreme case of ASE is monoallelic expression, where only one allele of a gene expresses whereas the other is completely silent. In a less extreme case, both alleles of a gene are transcribed with one allele contributing significantly more mRNAs than the other. ASE presents in both healthy and tumor tissues [29, 103] across species. Studies that analyzed genome-wide ASE in humans [63, 117], mouse [31], bovine [129], and drosophila [31] have discovered hundreds to a few thousands of genes that exhibit significant ASE

[36].

Recently, to utilize ASE as a direct approach to understand genotype and disease susceptibility has attracted an increasing attention. For example, [80] used ASE analysis as a marker to identify candidate genes associated with alcohol use disorders. [20] found that colorectal cancer risk increased as the imbalance of ASE in gene adenomatous polyposis coli increased. To understand diseases associated with aberrations in ASE, a fundamental approach is to compare ASE status between a healthy control group and a case/disease group. To that end, researchers need to estimate genome-wide ASE status in the control group and whole genome ASE status changes between the case and control groups.

Current statistical algorithms provide several opportunities for analytical improvement for detecting ASE regions in a control group and regions of ASE alterations between case and control groups. First, most methods search for ASE *genes* rather than actual ASE *regions*. For example, the binomial exact test in [89], the ANOVA method in [77], the mixture model in [66], the MBASED method [70], the QuASAR method [35], and the ASEP method [24]. However, genes with multiple isoforms due to alternative splicing and RNA variants result in ASE regions that contain one or more but not all exons of a gene [75, 119]. In fact, alternative splicing is universal in human multi-exon genes that comprise approximately 94% of genes overall [30, 119]. Therefore, a highly desirable feature of detecting ASE regions and ASE alterations is to incorporate such biological complexity into model construction and to identify boundaries of ASE regions precisely.

Second, existing methods universally treat observations as independent samples across single nucleotide polymorphisms (SNPs). As mentioned, an ASE region may contain one or more consecutive SNPs in exons that exhibit significant imbalanced gene expression; it is thus reasonable to model correlations among adjacent SNPs.

Existing methods use Bayesian priors to introduce correlation structures, such as the methods in [97], [70], [35], [129], and [24]. However, these hierarchical models lead to equal correlations among SNPs. A more reasonable strategy is to model correlations based on genomic distance such that neighboring SNPs that are closer are modeled with higher positive correlations compared to distant ones.

Third, existing methods tend to separately detect ASE regions in a control group and ASE regions with alterations between control and case groups. This strategy is time-consuming, practically inconvenient, and wasteful of resources. We will show that ASE observations under the control condition and the amount of ASE change between case and control conditions are likely correlated variables that become independent in special cases. Therefore, it is advantageous to construct a bivariate model and draw conclusions simultaneously. When only control group samples are available, or when we are only interested in differences between case and control groups, the bivariate model is naturally reduced to a univariate model.

To address aforementioned limitations and to properly model the complexity of biology, we develop a new statistical approach, called the hmmASE algorithm, to jointly detect ASE regions in control samples and ASE regions of alteration in case samples. The hmmASE algorithm is based on a bivariate Bayesian Hidden Markov model that accounts for spatial correlations among SNPs and jointly models data from control and case groups. hmmASE detects regions precisely among different gene isoforms and RNA variants and simultaneously discovers ASE regions in the control group and ASE regions of alterations between case and control groups. We illustrate the proposed method by a real data analysis in [13]. We compare the proposed hmmASE method with existing approaches through simulation studies that mimic real data sets. Simulation study suggests the hmmASE algorithm substantially outperforms other methods in power, accuracy, and precision.

3.2 Bivariate Bayesian hidden Markov model

Let X_{ijk} denote the number of sequencing reads from maternal allele for the SNP s in the j^{th} biological replicate of the i^{th} treatment group, where $i = 1, 2$, $j = 1, \ldots, J_i$, $s = 1, \ldots, S$. Here $i = 1$ denotes the control group and $i = 2$ denotes the case group. The proposed model does not require equal sample size between control and case groups, nor does it require multiple samples in either of the groups; as such J_1 and J_2 can both be 1 or more, and they are not necessarily of the same value. Let N_{ijs} denote the total number of sequencing read counts for the SNP s in the j^{th} biological replicate of the i^{th} treatment group, which is the summation of the number of read counts from the maternal and paternal alleles. We assume that X_{ijk} and N_{ijk} are read counts after bias correction. We further define P_{is} as the observed gene expression percentage in the i^{th} treatment group of the s^{th} SNP by averaging across biological replicates, where $P_{is} = [\sum_{j=1}^{j=J_i}(X_{ijs}+0.5)]/[\sum_{j=1}^{j=J_i}(N_{ijs}+1)]$. The numbers 0.5 and 1 constitute a Haldane-Anscombe correction to avoid extreme values. To eliminate the range limit, we model the observations in the logistic scale, and define $Y_{1s} = \log(P_{1s}/(1-P_{1s}))$, $Y_{2s} = \log(P_{2s}/(1-P_{2s}))$, and $D_s = Y_{2s} - Y_{1s}$. Intuitively, D_s denotes the sample mean difference of maternal expression percentage between the case and control groups for SNP s after logistic transformation.

The objective of the proposed method is to simultaneously detect ASE in control group and ASE alterations in case group compared with control group. Therefore, we model on a sequence of bivariate observations $\boldsymbol{O}_s = (Y_{1s}, D_s)'$, where the first dimension Y_{1s} assesses ASE in the control group, and the second dimension D_s evaluates ASE changes between the case group and control group. The two dimensions Y_{1s} and D_s are correlated with an exception when $\mathrm{Cov}(Y_{1s}, Y_{2s}) = \mathrm{Var}(Y_{1s})$, thus a suitable joint analysis promotes inferences of both. We assume that the bivari-

ate observations $\boldsymbol{O}_s$ are controlled by hidden bivariate states $\boldsymbol{L}_s$ that are discrete and unobservable. To explore the bivariate state space, we first consider the two dimensions separately. At each SNP, the control samples have three possible states. Specifically, here M represents the state where maternal allele expresses significantly more amount of RNA than the paternal allele, P denotes the state where paternal allele expresses significantly more amount of RNA than the maternal allele, and N represents the state where maternal and paternal alleles have the same expression level. Similarly, for ASE alterations in case samples, we let state 1 denote increased expression from the maternal allele when comparing between case and control groups, state 2 denote reduced expression from the maternal allele, and state 0 denote no change. Collectively, there are nine states in the bivariate state space $\mathcal{M} = \{(M,1),(M,2),(M,0),(P,1),(P,2),(P,0),(N,1),(N,2),(N,0)\}$.

A Bayesian HMM in this application is a bivariate random process of $\{\boldsymbol{O}_s, \boldsymbol{L}_s\}$, where $\boldsymbol{O}_s$ is the observation at SNP s and $\boldsymbol{L}_s \in \mathcal{M}$ is the true hidden state of SNP s. The goal is to infer the best sequence of the hidden states $\boldsymbol{L}_s$ for a chromosome based on the information provided by observations $\boldsymbol{O}_s$. Given the hidden state at SNP s, the probability of seeing the observation $\boldsymbol{O}_s$ is $P(\boldsymbol{O}_s|\boldsymbol{L}_s)$, which is the emission probability and is specified in (3.1).

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s = (M,1)', o_s^m, \sigma_m, d_s^1, \tau_1, \rho_1) \sim \mathrm{N}\left(\begin{pmatrix} o_s^m \\ d_s^1 \end{pmatrix}, \begin{pmatrix} \sigma_m^2 & \rho_1\sigma_m\tau_1 \\ \rho_1\sigma_m\tau_1 & \tau_1^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s = (M,2)', o_s^m, \sigma_m, d_s^2, \tau_2, \rho_2) \sim \mathrm{N}\left(\begin{pmatrix} o_s^m \\ d_s^2 \end{pmatrix}, \begin{pmatrix} \sigma_m^2 & \rho_2\sigma_m\tau_2 \\ \rho_2\sigma_m\tau_2 & \tau_2^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s = (M,0)', o_s^m, \sigma_m, d_s^0, \tau_0, \rho_3) \sim \mathrm{N}\left(\begin{pmatrix} o_s^m \\ d_s^0 \end{pmatrix}, \begin{pmatrix} \sigma_m^2 & \rho_3\sigma_m\tau_0 \\ \rho_3\sigma_m\tau_0 & \tau_0^2 \end{pmatrix}\right)$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,1)', o_s^p, \sigma_p, d_s^1, \tau_1, \rho_4) \sim \mathrm{N}\left(\begin{pmatrix} o_s^p \\ d_s^1 \end{pmatrix}, \begin{pmatrix} \sigma_p^2 & \rho_4\sigma_p\tau_1 \\ \rho_4\sigma_p\tau_1 & \tau_1^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,2)', o_s^p, \sigma_p, d_s^2, \tau_2, \rho_5) \sim \mathrm{N}\left(\begin{pmatrix} o_s^p \\ d_s^2 \end{pmatrix}, \begin{pmatrix} \sigma_p^2 & \rho_5\sigma_p\tau_2 \\ \rho_5\sigma_p\tau_2 & \tau_2^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,0)', o_s^p, \sigma_p, d_s^0, \tau_0, \rho_6) \sim \mathrm{N}\left(\begin{pmatrix} o_s^p \\ d_s^0 \end{pmatrix}, \begin{pmatrix} \sigma_p^2 & \rho_6\sigma_p\tau_0 \\ \rho_6\sigma_p\tau_0 & \tau_0^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,1)', o_s^n, \sigma_n, d_s^1, \tau_1, \rho_7) \sim \mathrm{N}\left(\begin{pmatrix} o_s^n \\ d_s^1 \end{pmatrix}, \begin{pmatrix} \sigma_n^2 & \rho_7\sigma_n\tau_1 \\ \rho_7\sigma_n\tau_1 & \tau_1^2 \end{pmatrix}\right),$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,2)', o_s^n, \sigma_n, d_s^2, \tau_2, \rho_8) \sim \mathrm{N}\left(\begin{pmatrix} o_s^n \\ d_s^2 \end{pmatrix}, \begin{pmatrix} \sigma_n^2 & \rho_8\sigma_n\tau_2 \\ \rho_8\sigma_n\tau_2 & \tau_2^2 \end{pmatrix}\right), \text{ and}$$

$$P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,0)', o_s^n, \sigma_n, d_s^0, \tau_0, \rho_9) \sim \mathrm{N}\left(\begin{pmatrix} o_s^n \\ d_s^0 \end{pmatrix}, \begin{pmatrix} \sigma_n^2 & \rho_9\sigma_n\tau_0 \\ \rho_9\sigma_n\tau_0 & \tau_0^2 \end{pmatrix}\right), \quad (3.1)$$

where o_s^m, o_s^p, and o_s^n represent the mean maternal allele expression in control group at SNP s. More specifically, they denote the mean maternal allele expression in control group at SNP s after logistic transformation for SNP s (i.e., Y_{1s}) given hidden states M, P, and N, respectively. Similarly, d_s^1, d_s^2, and d_s^0 denote the mean maternal expression percentage change after logistic transformation in case group compared with control group for SNP s (i.e., D_s) given hidden states 1, 2, and 0, respectively. The subscript s allows these variables to differ among SNPs. Based on biological interpretation of each state in $\mathcal{M}$, we further assume truncated multivariate normal priors on these model parameters as follows. The truncated normal priors restrict increased maternal expression percentage corresponds to a positive change and reduced maternal expression percentage corresponds to a negative change.

$$
\begin{aligned}
&\begin{pmatrix} o_s^m \\ d_s^1 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^+ \\ d^+ \end{pmatrix}, \begin{pmatrix} \gamma_+^2 & \alpha_1\gamma_+\beta_+ \\ \alpha_1\gamma_+\beta_+ & \beta_+^2 \end{pmatrix}\right) I(o^+ > 0, d^+ > 0), \\
&\begin{pmatrix} o_s^m \\ d_s^2 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^+ \\ d^- \end{pmatrix}, \begin{pmatrix} \gamma_+^2 & \alpha_2\gamma_+\beta_- \\ \alpha_2\gamma_+\beta_- & \beta_-^2 \end{pmatrix}\right) I(o^+ > 0, d^- < 0), \\
&\begin{pmatrix} o_s^m \\ d_s^0 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^+ \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma_+^2 & \alpha_3\gamma_+\beta_0 \\ \alpha_3\gamma_+\beta_0 & \beta_0^2 \end{pmatrix}\right) I(o^+ > 0), \\
&\begin{pmatrix} o_s^p \\ d_s^1 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^- \\ d^+ \end{pmatrix}, \begin{pmatrix} \gamma_-^2 & \alpha_4\gamma_-\beta_+ \\ \alpha_4\gamma_-\beta_+ & \beta_+^2 \end{pmatrix}\right) I(o^- < 0, d^+ > 0), \\
&\begin{pmatrix} o_s^p \\ d_s^2 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^- \\ d^- \end{pmatrix}, \begin{pmatrix} \gamma_-^2 & \alpha_5\gamma_-\beta_- \\ \alpha_5\gamma_-\beta_- & \beta_-^2 \end{pmatrix}\right) I(o^- < 0, d^- < 0), \\
&\begin{pmatrix} o_s^p \\ d_s^0 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} o^- \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma_-^2 & \alpha_6\gamma_-\beta_0 \\ \alpha_6\gamma_-\beta_0 & \beta_0^2 \end{pmatrix}\right) I(o^- < 0), \\
&\begin{pmatrix} o_s^n \\ d_s^1 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} 0 \\ d^+ \end{pmatrix}, \begin{pmatrix} \gamma_0^2 & \alpha_7\gamma_0\beta_+ \\ \alpha_7\gamma_0\beta_+ & \beta_+^2 \end{pmatrix}\right) I(d^+ > 0), \\
&\begin{pmatrix} o_s^n \\ d_s^2 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} 0 \\ d^- \end{pmatrix}, \begin{pmatrix} \gamma_0^2 & \alpha_8\gamma_0\beta_- \\ \alpha_8\gamma_0\beta_- & \beta_-^2 \end{pmatrix}\right) I(d^- < 0), \text{ and} \\
&\begin{pmatrix} o_s^n \\ d_s^0 \end{pmatrix} \overset{i.i.d.}{\sim} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \gamma_0^2 & \alpha_9\gamma_0\beta_0 \\ \alpha_9\gamma_0\beta_0 & \beta_0^2 \end{pmatrix}\right). \qquad (3.2)
\end{aligned}
$$

Notice that parameters o^+, o^-, d^+, d^-, γ_+, γ_-, γ_0, β_+, β_-, and β_0 no longer have subscripts associated with a particular SNP s because these hyperparameters are estimated using observations from all SNPs in a chromosome. The hierarchical model in (3.2) ensures information is shared across SNPs and regions, which is helpful for ASE analysis as most genes contain a few number of SNPs (e.g., 5 to 10) and a small number of genes contain a large number of SNPs (e.g., 100). The genes (and regions)

with fewer SNPs will benefit more from information sharing and show more improvement in power. The normality assumption in model (3.1) will be validated by real data in the Supplementary Material Appendix A.3. By integrating out parameters o_s^m, o_s^p, o_s^n, d_s^1, d_s^2, and d_s^0, the marginal distributions of $\boldsymbol{O}_s$ given hidden states follow bivariate normal distributions as follows.

$$
\begin{aligned}
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(M,1)',o^+,d^+,\sigma_+,\tau_+,\omega_1) &\sim \mathrm{N}\left(\begin{pmatrix} o^+ \\ d^+ \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & \omega_1\sigma_+\tau_+ \\ \omega_1\sigma_+\tau_+ & \tau_+^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(M,2)',o^+,d^-,\sigma_+,\tau_-,\omega_2) &\sim \mathrm{N}\left(\begin{pmatrix} o^+ \\ d^- \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & \omega_2\sigma_+\tau_- \\ \omega_2\sigma_+\tau_- & \tau_-^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(M,0)',o^+,\sigma_+,\tau,\omega_3) &\sim \mathrm{N}\left(\begin{pmatrix} o^+ \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_+^2 & \omega_3\sigma_+\tau \\ \omega_3\sigma_+\tau & \tau^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,1)',o^-,d^+,\sigma_-,\tau_+,\omega_4) &\sim \mathrm{N}\left(\begin{pmatrix} o^- \\ d^+ \end{pmatrix}, \begin{pmatrix} \sigma_-^2 & \omega_4\sigma_-\tau_+ \\ \omega_4\sigma_-\tau_+ & \tau_+^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,2)',o^-,d^-,\sigma_-,\tau_-,\omega_5) &\sim \mathrm{N}\left(\begin{pmatrix} o^- \\ d^- \end{pmatrix}, \begin{pmatrix} \sigma_-^2 & \omega_5\sigma_-\tau_- \\ \omega_5\sigma_-\tau_- & \tau_-^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(P,0)',o^-,\sigma_-,\tau,\omega_6) &\sim \mathrm{N}\left(\begin{pmatrix} o^- \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_-^2 & \omega_6\sigma_-\tau \\ \omega_6\sigma_-\tau & \tau^2 \end{pmatrix}\right) \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,1)',d^+,\sigma,\tau_+,\omega_7) &\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ d^+ \end{pmatrix}, \begin{pmatrix} \sigma^2 & \omega_7\sigma\tau_+ \\ \omega_7\sigma\tau_+ & \tau_+^2 \end{pmatrix}\right), \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,2)',d^-,\sigma,\tau_-,\omega_8) &\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ d^- \end{pmatrix}, \begin{pmatrix} \sigma^2 & \omega_8\sigma\tau_- \\ \omega_8\sigma\tau_- & \tau_-^2 \end{pmatrix}\right), \text{ and} \\
P(\boldsymbol{O}_s|\boldsymbol{L}_s=(N,0)',\sigma,\tau,\omega_9) &\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \omega_9\sigma\tau \\ \omega_9\sigma\tau & \tau^2 \end{pmatrix}\right), \qquad (3.3)
\end{aligned}
$$

where $\sigma_+^2 = \sigma_m^2 + \gamma_+^2$, $\sigma_-^2 = \sigma_p^2 + \gamma_-^2$, and $\sigma^2 = \sigma_n^2 + \gamma_0^2$; similarly, $\tau_+^2 = \tau_1^2 + \beta_+^2$,

$\tau_-^2 = \tau_2^2 + \beta_-^2$, and $\tau^2 = \tau_0^2 + \beta_0^2$.

We restrict parameters $\sigma_+^2 = \sigma_-^2 = \tau_+^2 = \tau_-^2$, $\sigma^2 = \tau^2$, and $o^+ = |o^-| = d^+ = |d^-|$. We restrict $o^+ = |o^-| = d^+ = |d^-|$ because, biologically speaking, if the mean magnitude of imbalance in gene expression is $o^+ = |o^-|$ when detecting ASE in control, then the mean magnitude of alterations of ASE in the case group should be on average the same to be detected as a significant change; i.e., $o^+ = d^+ = |o^-| = |d^-|$. Driven by similar logic, we restrict $\sigma_+^2 = \sigma_-^2 = \tau_+^2 = \tau_-^2$ and $\sigma^2 = \tau^2$. These restrictions together also effectively eliminate false positives and speed up the convergence of the hidden Markov chain [32, 43]. Besides, we also want to restrict $o^+ = |o^-| = d^+ = |d^-|$ to be greater than a positive constant. Generally speaking, a 75% of gene expression coming from one allele captures attention from researchers. Given a SNP s, when 75% of gene expression is from the maternal allele, $Y_{1s} = 1.1$; when 75% of gene expression is from the paternal allele, $Y_{1s} = -1.1$. Therefore, we assume the mean parameters $o^+ = d^+ \geq 1.1$ and $o^- = d^- \leq -1.1$ in the posterior distributions.

The proposed Bayesian HMM assumes the hidden states $\boldsymbol{L}_s$ behave as a Markov process, i.e., $P(\boldsymbol{L}_{s+1}|\boldsymbol{L}_1, \ldots, \boldsymbol{L}_s) = P(\boldsymbol{L}_{s+1}|\boldsymbol{L}_s)$, $s = 1, \ldots, S-1$. The Markov chain is governed by the probability $t_{\boldsymbol{kl}}(s) = P(\boldsymbol{L}_{s+1} = \boldsymbol{l}|\boldsymbol{L}_s = \boldsymbol{k})$ which is defined as the transition probability from state $\boldsymbol{k}$ at SNP s to state $\boldsymbol{l}$ at SNP $(s+1)$, $\boldsymbol{k}, \boldsymbol{l} \in \mathcal{M}$. By construction, we have $\sum_{\boldsymbol{l} \in \mathcal{M}} t_{\boldsymbol{kl}}(s) = 1$ for any $\boldsymbol{k} \in \mathcal{M}$ at SNP s. To incorporate correlation between adjacent SNPs, we model the transition probability as a function of genomic distance. Specifically, let c_s be the genomic distance between SNP s and SNP $(s+1)$. $t_{\boldsymbol{kl}}(s)$ is a 9-by-9 transition matrix from SNP s to SNP $(s+1)$ defined as

$$t_{\boldsymbol{kl}}(s) = P(\boldsymbol{L}_{s+1} = \boldsymbol{l}|\boldsymbol{L}_s = \boldsymbol{k}) = \begin{cases} p_{\boldsymbol{kl}} \left(1 - \exp(-\alpha c_s)\right), & \boldsymbol{l} \neq \boldsymbol{k}, \\ 1 - \left(\sum_{\boldsymbol{l} \neq \boldsymbol{k}} p_{\boldsymbol{kl}}\right) \left(1 - \exp(-\alpha c_s)\right), & \boldsymbol{l} = \boldsymbol{k}. \end{cases} \tag{3.4}$$

Here, α is a positive-valued parameter that determines the effect of genomic distance in the transition matrix. The parameter $p_{\boldsymbol{kl}}$ affects transition probabilities from state $\boldsymbol{k}$ to state $\boldsymbol{l}$ where $\boldsymbol{k} \neq \boldsymbol{l}$ with constraints $p_{\boldsymbol{kl}} \in (0,1)$ and $\sum_{\boldsymbol{l} \neq \boldsymbol{k}} p_{\boldsymbol{kl}} < 1$. When two adjacent SNPs are extremely far away from each other, i.e., $c_s \geq 40,000$ base pairs, we let $t_{\boldsymbol{kl}}(s) = \frac{1}{9} \, \forall \, (k,l)$.

3.3 Detecting ASE regions in a control group and regions of ASE alteration in a case group

Given the Bayesian HMM, this section depicts the inferential procedure for parameter estimation and simultaneous detection of ASE regions in a control group and ASE alteration regions in a case group. Specifically, we let $\boldsymbol{\theta} = (\boldsymbol{\theta}_{emiss}, \boldsymbol{\theta}_{trans})^T$ be the vector containing all parameters to be estimated including emission probability parameters $\boldsymbol{\theta}_{emiss} = (o^+, o^-, d^+, d^-, \sigma_+^2, \sigma_-^2, \sigma^2, \tau_+^2, \, \tau_-^2, \tau^2, \omega_1, \ldots, \omega_9)^T$ and transition probability parameters $\boldsymbol{\theta}_{trans} = (p_{\boldsymbol{kl}} \text{ where } \boldsymbol{k}, \boldsymbol{l} \in \mathcal{M}), \alpha)^T$. The incomplete-data and complete-data likelihoods are denoted by $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{O})$ and $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{O}, \boldsymbol{L})$, respectively, where $\boldsymbol{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_S)^T$, $\boldsymbol{L} = (\boldsymbol{L}_1, \ldots, \boldsymbol{L}_S)^T$. Additionally, $\boldsymbol{p} = p_{\boldsymbol{kl}}$'s where $\boldsymbol{k}, \boldsymbol{l} \in \mathcal{M}$, and $\pi_{\boldsymbol{L}_1} = P(\boldsymbol{L}_1 = \boldsymbol{k})$ $(\boldsymbol{k} \in \mathcal{M})$ is the probability distribution for the first SNP.

It is difficult to estimate $\boldsymbol{\theta}$ by maximizing the incomplete-data likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{O})$ directly due to high dimensionality of $\boldsymbol{\theta}$; we choose to use the expectation-maximization (EM) algorithm to iteratively estimate $\boldsymbol{\theta}$. At the m-th iteration of the EM algorithm, the E step assumes the parameter vector $\boldsymbol{\theta}$ is equal to $\boldsymbol{\theta}^{(m-1)}$, and then evaluates the expectation of the complete-data likelihood (in (3.5)) with respect to the conditional distribution of hidden states based on observations. In the M step, we update $\boldsymbol{\theta}^{(m-1)}$ to $\boldsymbol{\theta}^{(m)}$ by finding the parameter vector that maximizes the expec-

tation of the complete-data likelihood. Particularly, we update all parameters in the following approaches; we update $\boldsymbol{p}$ and $\boldsymbol{\theta}_{emiss}$ by solving partial derivatives of the goal function; we update α by a grid search; at last we let $\pi^{(m)}_{\boldsymbol{L}_1=\boldsymbol{k}} = P(\boldsymbol{L}_1 = \boldsymbol{k} \mid \boldsymbol{O}, \boldsymbol{\theta}^{(m-1)})$. The details of the EM algorithm are elaborated in Supplementary Material Appendix A.1. The EM algorithm converges when the difference between $\boldsymbol{\theta}^{(m)}$ and $\boldsymbol{\theta}^{(m-1)}$ is smaller than a pre-defined threshold. We use the estimates at the last iteration as the proposed parameter estimates and denote it by $\hat{\boldsymbol{\theta}}$.

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{L}|\boldsymbol{O},\boldsymbol{\theta}^{(m-1)}}[\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{O}, \boldsymbol{L}) \mid \boldsymbol{O}, \boldsymbol{\theta}^{(m-1)}] = \sum_{\boldsymbol{k}} P(\boldsymbol{L} = \boldsymbol{k} \mid \boldsymbol{O}, \boldsymbol{\theta}^{(m-1)}) \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{O}, \boldsymbol{L}) \\
&= \sum_{\boldsymbol{k}} P(\boldsymbol{L} = \boldsymbol{k} \mid \boldsymbol{O}, \boldsymbol{\theta}^{(m-1)}) \pi_{\boldsymbol{L}_1} \prod_{s=1}^{S-1} t_{\boldsymbol{L}_s \boldsymbol{L}_{s+1}}(s \mid \boldsymbol{\theta}^{(m-1)}_{trans}) \prod_{s=1}^{S} P(\boldsymbol{O}_s \mid \boldsymbol{L}, \boldsymbol{\theta}^{(m-1)}_{emiss}). \quad (3.5)
\end{aligned}$$

To initiate the aforementioned iterations, we need to assign initial values $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$. We first calculate the sample standard deviations of the two dimensions of $\boldsymbol{O}$, respectively, from all SNPs, and assign them as initial values for variance components. Namely, $\sigma^{(0)}_{+} = \sigma^{(0)}_{-} = \sigma^{(0)} = \sqrt{\sum_{s=1}^{S}(Y_{1s} - \bar{Y}_1)^2/(N-1)}$, and $\tau^{(0)}_{+} = \tau^{(0)}_{-} = \tau^{(0)} = \sqrt{\sum_{s=1}^{S}(D_s - \bar{D})^2/(N-1)}$ where $\bar{Y}_1$ and $\bar{D}$ are respectively the sample averages of Y_{1s}'s and D_s's. Then SNPs with $Y_{1s} > \sigma^{(0)}$, $Y_{1s} < -\sigma^{(0)}$, and $-\sigma^{(0)} \leq Y_{1s} \leq \sigma^{(0)}$ are assigned to states M, P, and N respectively; SNPs with $D_s > \tau^{(0)}$, $D_s < -\tau^{(0)}$, and $-\tau^{(0)} \leq D_s \leq \tau^{(0)}$ are assigned to states 1, 2, and 0 respectively. We assign $o^{+(0)} = -|o^{-(0)}|$ and equals to the average of Y_{1s} for SNPs in state M and absolute values of Y_{1s} for SNPs in state P. Similarly, we assign $d^{+(0)} = -|d^{-(0)}|$ and equals to the average of D_s for SNPs in state 1 and absolute values of D_s for SNPs in state 2. Initial values of $\omega_1, \ldots, \omega_9$ are sample correlation coefficients based on the aforementioned initial state assignment. We let $p^{(0)}_{\boldsymbol{kl}} = 0.01$ where $\boldsymbol{k} \neq \boldsymbol{l}$, $\alpha^{(0)} = 0.05$, and $\pi^{(0)}_{\boldsymbol{L}_1=\boldsymbol{k}} = 1/9$ for $\boldsymbol{k} \in \mathcal{M}$.

Before declaring an ASE region or a region of ASE alteration, we need to infer the best possible sequence of hidden states given estimated parameters by decoding the HMM. We utilize the Viterbi algorithm [118] to find the sequence of hidden states $\boldsymbol{L}$ that maximizes $P(\boldsymbol{L}, \boldsymbol{O} \mid \hat{\boldsymbol{\theta}})$; that is, $\hat{\boldsymbol{L}} = \arg\max_{\boldsymbol{L}} P(\boldsymbol{L}, \boldsymbol{O} \mid \hat{\boldsymbol{\theta}})$. The last step is to claim an ASE region with adjacent SNPs which have the same hidden states in one or both dimensions. For example, the three consecutive SNPs with $\boldsymbol{L}_s \in \{(M,0),(M,1),(M,2)\}$ form an ASE region with maternal preference in the control group. Analogously, the three consecutive SNPs with $\boldsymbol{L}_s \in \{(M,1),(N,1),(P,1)\}$ constitute a region of ASE alterations that exhibits increased maternal allele expression in the case group compared with the control group. We require that each ASE region contains no two adjacent SNPs that are further than 5,000 base pairs apart. An ASE region with a gap longer than 5,000 base pairs is segmented into two ASE regions.

3.4 Simulation study

We utilize the 261 genes (total 2,035 SNPs) in chromosome 9 in the liver tissue in [13] to facilitate the simulation study. Guided by the real data analysis results, we randomly assign 170 of 261 genes to $(N,0)$, and other 91 genes are randomly assigned to the other eight hidden states in $\mathcal{M}$ with equal probability. We let SNPs that belong to the same gene have the same state. However, as shown in the method and real data analysis sections, the proposed hmmASE method applies to both regions that are within a gene or across multiple genes.

We set up two simulation scenarios. In the first scenario, we let $o^+ = -o^- = 1.16$, $d^+ = -d^- = 1.16$, $\sigma^2_+ = \sigma^2_- = \tau^2_+ = \tau^2_- = 0.89$, $\sigma^2 = \tau^2 = 0.19$, and fix correlation

to 0.1; i.e., $\omega_1 = \cdots = \omega_9 = 0.1$. These parameters are chosen as the average of parameter estimates from the 30 chromosomes in real data analysis. Under these parameters, we expect ASE SNPs in the control group to have on average 76.1% percent expression from one allele. In the second scenario, we set $o^+ = -o^- = d^+ = -d^- = 1.73$ and keep other parameters the same with scenario one. Under this setting, we expect ASE SNPs in the control group to have on average 85% percent expression from one allele.

Based on the parameter settings, we first simulate the sequence of bivariate observations $\boldsymbol{O}$, which is used to compute P_{1s}'s and P_{2s}'s. In the simulation study, we assume four biological replicates, same with real data in [13]. We utilize the real data N_{ijs}'s in the simulation. Then, we compute X_{ijs}'s based on N_{ijs}'s, P_{1s}'s and P_{2s}'s. By applying the Bayesian HMM model, we categorize each SNP into nine categories. Figure 3.1 illustrates the Bayesian HMM results on one set of simulated data under each of the two scenarios. The left column shows fitting results of the first dimension (Y_{1s}'s). The right column shows fitting results of the second dimension (D_s's). In both of the scenarios, the proposed Bayesian HMM fits the simulated data well.

We then compare the proposed hmmASE method with the MBASED method [70], the T2ER method [56], and the exact test which consists of a binomial test on detecting ASEs in the control group and a Fisher exact test on the difference between case and control groups. We choose these methods to compare as they are able to detect ASE in control as well as detect expression status change between case and control groups. To conduct the comparison, we simulated 20 sets of data under each of two scenarios. We calculate true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR) according to formula (3.6). Here, 'significant SNPs' are either significant SNPs showing ASE in control or significant SNPs showing substantial change in expression between case and control groups. An

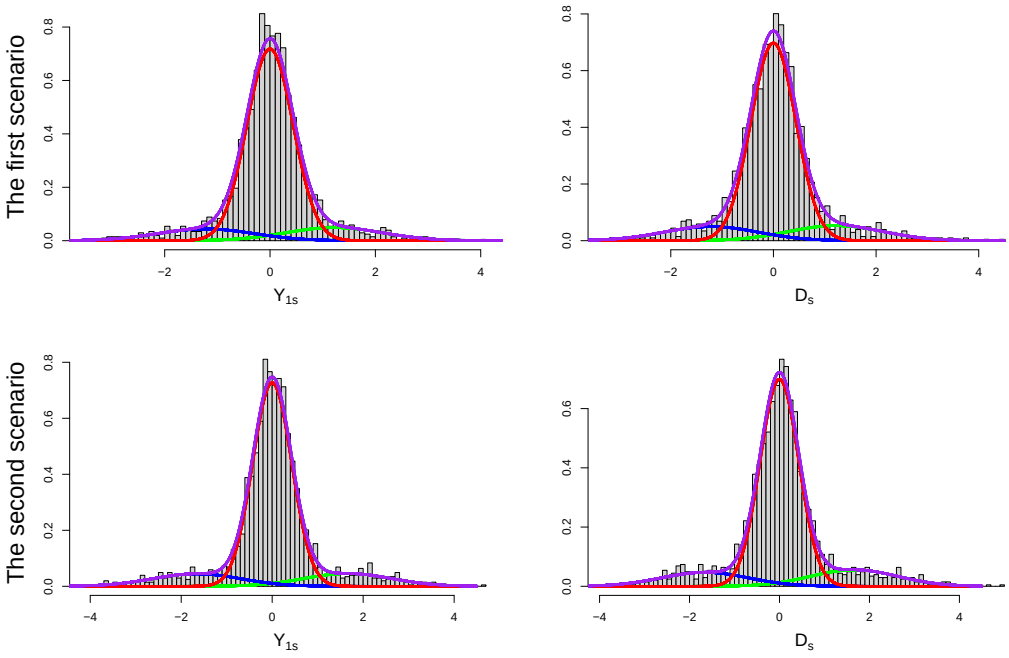


Figure 3.1: An illustration of fitting results for one set of simulated data under two scenarios. The histograms are simulated data; the blue, green, and red curves represent the three normal components. Purple curve is the fitted mixture distribution.

ASE detection is considered correct if both the detection and direction of change are correct.

$$\begin{aligned}
\text{TPR} &= \frac{\text{\# of correctly identified significant SNPs by type}}{\text{\# of significant SNPs by type}}, \\
\text{FPR} &= \frac{\text{\# of incorrectly identified significant SNPs}}{\text{\# of true null SNPs}}, \\
\text{TNR} &= \frac{\text{\# of identified true null SNPs}}{\text{\# of true null SNPs}}, \\
\text{FNR} &= \frac{\text{\# of incorrectly identified true null SNPs}}{\text{\# of significant SNPs}}.
\end{aligned} \tag{3.6}$$

Table 3.1 and Table 3.2 summarize average TPR, FPR, TNR, and FNR on detecting ASE in the control group and ASE changes between case and control groups across

twenty times of simulations, respectively. The numbers in parenthesis are variations across simulations. The proposed hmmASE method shows substantially improved TPR and TNR compared to other approaches. The hmmASE method also shows smaller variations across repeated simulations for each of the performance statistics. In contrast, the other three methods suffer from lack of power in detecting true signals as well as large variations across repeated simulations.

Table 3.1: TPRs, FPRs, TNRs, and FNRs comparisons in the first scenario.

State		hmmASE		Exact		MBASED		T2ER	
M	TPr	0.859	(0.067)	0.130	(0.038)	0.150	(0.036)	0.047	(0.026)
	FPr	0.007	(0.003)	0.026	(0.005)	0.029	(0.005)	0.007	(0.002)
P	TPr	0.861	(0.069)	0.132	(0.059)	0.155	(0.058)	0.043	(0.031)
	FPr	0.007	(0.003)	0.025	(0.004)	0.027	(0.003)	0.007	(0.002)
N	TNr	0.967	(0.006)	0.879	(0.020)	0.873	(0.020)	0.918	(0.022)
	FNr	0.081	(0.028)	0.729	(0.045)	0.685	(0.043)	0.916	(0.029)
1	TPr	0.860	(0.096)	0.088	(0.045)	0.113	(0.046)	0.030	(0.023)
	FPr	0.008	(0.002)	0.009	(0.002)	0.011	(0.002)	0.002	(0.001)
2	TPr	0.836	(0.062)	0.090	(0.036)	0.115	(0.046)	0.030	(0.023)
	FPr	0.008	(0.002)	0.009	(0.002)	0.012	(0.003)	0.002	(0.001)
0	TNr	0.967	(0.005)	0.881	(0.017)	0.875	(0.014)	0.912	(0.019)
	FNr	0.093	(0.045)	0.861	(0.041)	0.817	(0.042)	0.950	(0.019)
(M, 1)	TPr	0.846	(0.127)	0.049	(0.048)	0.056	(0.046)	0.000	(0.000)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.002)	0.000	(0.000)
(M, 2)	TPr	0.802	(0.110)	0.086	(0.057)	0.103	(0.058)	0.037	(0.046)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.001)	0.000	(0.000)
(M, 0)	TPr	0.909	(0.052)	0.277	(0.117)	0.316	(0.104)	0.112	(0.065)
	FPr	0.010	(0.005)	0.032	(0.006)	0.036	(0.006)	0.010	(0.003)
(P, 1)	TPr	0.848	(0.126)	0.094	(0.057)	0.118	(0.060)	0.032	(0.026)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.001)	0.001	(0.001)
(P, 2)	TPr	0.824	(0.108)	0.051	(0.047)	0.058	(0.051)	0.000	(0.000)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.001)	0.000	(0.000)
(P, 0)	TPr	0.867	(0.134)	0.231	(0.111)	0.263	(0.105)	0.086	(0.060)
	FPr	0.008	(0.004)	0.031	(0.005)	0.035	(0.005)	0.009	(0.003)
(N, 1)	TNr	0.861	(0.162)	0.125	(0.110)	0.166	(0.107)	0.063	(0.066)
	FNr	0.008	(0.003)	0.007	(0.002)	0.011	(0.003)	0.003	(0.001)
(N, 2)	TNr	0.849	(0.131)	0.140	(0.072)	0.195	(0.087)	0.055	(0.045)
	FNr	0.011	(0.004)	0.007	(0.003)	0.011	(0.003)	0.003	(0.001)
(N, 0)	TNr	0.972	(0.005)	0.935	(0.006)	0.924	(0.006)	0.981	(0.004)
	FNr	0.051	(0.020)	0.723	(0.040)	0.662	(0.040)	0.906	(0.024)

Table 3.2: TPRs, FPRs, TNRs, and FNRs comparisons in the second scenario.

State		hmmASE		Exact		MBASED		T2ER	
M	TPr	0.955	(0.033)	0.192	(0.051)	0.243	(0.042)	0.061	(0.032)
	FPr	0.004	(0.002)	0.027	(0.005)	0.030	(0.005)	0.008	(0.003)
P	TPr	0.949	(0.030)	0.195	(0.078)	0.243	(0.075)	0.056	(0.038)
	FPr	0.003	(0.002)	0.025	(0.003)	0.028	(0.003)	0.008	(0.002)
N	TNr	0.983	(0.004)	0.881	(0.019)	0.876	(0.019)	0.918	(0.021)
	FNr	0.030	(0.015)	0.627	(0.047)	0.541	(0.040)	0.907	(0.037)
1	TPr	0.952	(0.038)	0.136	(0.054)	0.187	(0.053)	0.047	(0.025)
	FPr	0.003	(0.002)	0.008	(0.002)	0.012	(0.002)	0.002	(0.001)
2	TPr	0.946	(0.033)	0.159	(0.058)	0.204	(0.064)	0.053	(0.029)
	FPr	0.004	(0.001)	0.009	(0.002)	0.012	(0.003)	0.002	(0.001)
0	TNr	0.983	(0.004)	0.886	(0.015)	0.881	(0.012)	0.912	(0.018)
	FNr	0.028	(0.020)	0.812	(0.051)	0.733	(0.050)	0.933	(0.024)
(M, 1)	TPr	0.959	(0.035)	0.073	(0.063)	0.089	(0.061)	0.001	(0.002)
	FPr	0.000	(0.001)	0.003	(0.002)	0.003	(0.001)	0.000	(0.000)
(M, 2)	TPr	0.942	(0.063)	0.172	(0.085)	0.211	(0.092)	0.057	(0.052)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.001)	0.000	(0.000)
(M, 0)	TPr	0.957	(0.037)	0.356	(0.118)	0.462	(0.088)	0.132	(0.069)
	FPr	0.004	(0.002)	0.036	(0.007)	0.042	(0.007)	0.010	(0.003)
(P, 1)	TPr	0.930	(0.071)	0.159	(0.078)	0.216	(0.078)	0.063	(0.044)
	FPr	0.000	(0.001)	0.003	(0.001)	0.003	(0.001)	0.001	(0.001)
(P, 2)	TPr	0.940	(0.051)	0.072	(0.060)	0.092	(0.066)	0.002	(0.007)
	FPr	0.001	(0.001)	0.003	(0.001)	0.003	(0.001)	0.000	(0.000)
(P, 0)	TPr	0.953	(0.043)	0.329	(0.149)	0.394	(0.127)	0.095	(0.077)
	FPr	0.004	(0.002)	0.035	(0.006)	0.041	(0.007)	0.009	(0.003)
(N, 1)	TNr	0.953	(0.066)	0.182	(0.122)	0.260	(0.119)	0.084	(0.070)
	FNr	0.004	(0.002)	0.006	(0.002)	0.011	(0.003)	0.002	(0.001)
(N, 2)	TNr	0.935	(0.093)	0.235	(0.115)	0.317	(0.116)	0.100	(0.073)
	FNr	0.004	(0.002)	0.007	(0.002)	0.012	(0.004)	0.003	(0.001)
(N, 0)	TNr	0.985	(0.005)	0.932	(0.006)	0.920	(0.008)	0.980	(0.005)
	FNr	0.016	(0.008)	0.632	(0.045)	0.523	(0.032)	0.893	(0.029)

3.5 Real data analysis

In this section, we apply the proposed hmmASE method in a Large offspring syndrome (LOS) study. LOS is a congenital overgrowth syndrome in ruminants, which is phenotypically similar to Beckwith-Wiedemann syndrome (BWS) in humans. BWS is the most common congenital overgrowth disorder with an estimated worldwide frequency of 1 in 13,700 natural births [17]. Shared phenotypes between LOS and BWS include excessive birth weight, large tongue, breathing difficulties, umbilical hernia, hypoglycemia, and visceromegaly.

To investigate the molecular mechanism of LOS, which infers etiology of BWS, [13] used cattle as the study animal, and generated allele-specific gene expression data by RNA sequencing experiments. For each treatment group (control and LOS), four biological replicates (four cattle offsprings) were collected. [13] addressed the mapping bias problem by combining the reference genome (i.e., the *B. t. taurus* reference genome UMD3.1 build) and the pseudo *B. t. indicus* genome to create a custom diploid genome [13, 129]. The RNAseq reads of brain, kidney, liver, and skeletal muscle after bias correction are publicly available at the Gene Expression Omnibus database under accession number GSE63509. Here, we use liver tissue as an example to illustrate the application of the proposed approach. SNPs with extremely low read counts, i.e., the summation of total number of reads across four biological replicates is less than 6, are filtered out from the study. The goal is to detect shared ASE in control samples as well as shared ASE alterations in LOS group compared to the control samples.

Note that the EM algorithm greedily maximizes the expected log likelihood at each iteration. As such, outliers in Y_{1s}'s drive estimates of states M and P to extreme values. In practice, to avoid converging to boundary values in the parameter space, we

eliminate extreme values before applying the proposed method. When more than 98% of gene expression at a SNP comes from one (maternal or paternal) allele, $Y_{1s} \geq 3.89$ or $Y_{1s} \leq -3.89$. We remove SNPs with $|Y_{1s}| > 3.9$, and apply the hmmASE method to the remaining SNPs. We then infer the hidden states of these removed SNPs that are of extreme values of Y_{1s}. Specifically, for SNP s with extreme values, its hidden state is inferred as $\boldsymbol{k} = \arg\max_{\boldsymbol{k}} P(\boldsymbol{O}_s, \boldsymbol{L}_s \mid \hat{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{k}} P(\boldsymbol{O}_s \mid \boldsymbol{L} = \boldsymbol{k}, \hat{\boldsymbol{\theta}}) P(\boldsymbol{L} = \boldsymbol{k} \mid \boldsymbol{O}, \hat{\boldsymbol{\theta}})$, where $P(\boldsymbol{L} = \boldsymbol{k} \mid \boldsymbol{O}, \hat{\boldsymbol{\theta}})$'s ($\boldsymbol{k} \in \mathcal{M}$) are estimated empirically as proportions of SNPs identified in each of the nine states in the state space.

After the EM algorithm converges, each chromosome has a distinct set of parameters. For example, for chromosome 23, estimated parameters are $\hat{o}^+ = 1.163$, $\hat{o}^- = -1.163$, $\hat{d}^+ = 1.163$, $\hat{d}^- = -1.163$, $\hat{\sigma}^2_+ = \hat{\sigma}^2_- = \hat{\delta}^2_+ = \hat{\delta}^2_- = 0.971$, $\hat{\sigma}^2 = \hat{\delta}^2 = 0.209$, $\hat{\omega}_1 = 0.045$, $\hat{\omega}_2 = 0.766$, $\hat{\omega}_3 = 0.027$, $\hat{\omega}_4 = -0.017$, $\hat{\omega}_5 = -0.108$, $\hat{\omega}_6 = -0.092$, $\hat{\omega}_7 = -0.086$, $\hat{\omega}_8 = -0.116$, $\hat{\omega}_9 = -0.047$, and $\hat{\alpha} = 0.02$. For chromosome 21, estimated parameters are $\hat{o}^+ = 1.154$, $\hat{o}^- = -1.154$, $\hat{d}^+ = 1.154$, $\hat{d}^- = -1.154$, $\hat{\sigma}^2_+ = \hat{\sigma}^2_- = \hat{\delta}^2_+ = \hat{\delta}^2_- = 1.379$, $\hat{\sigma}^2 = \hat{\delta}^2 = 0.213$, $\hat{\omega}_1 = -0.048$, $\hat{\omega}_2 = -0.125$, $\hat{\omega}_3 = -0.049$, $\hat{\omega}_4 = 0.063$, $\hat{\omega}_5 = -0.067$, $\hat{\omega}_6 = -0.144$, $\hat{\omega}_7 = -0.107$, $\hat{\omega}_8 = 0.154$, $\hat{\omega}_9 = -0.010$, and $\hat{\alpha} = 0.03$. The estimated parameter vector $\boldsymbol{p}$'s for these two chromosomes are shown in Supplementary Material Appendix A.2.

After using the proposed bivariant Bayesian HMM model, Table 3.3 shows ten ASE regions with the most numbers of SNPs in the control group (first dimension of $\boldsymbol{O}$) with either a predicted maternal expression (M state) or paternal expression (P state). The last column in the table shows mean proportion of gene expression from the maternal allele. Table 3.4 summarizes the top ten regions of ASE alterations with the most number of SNPs in the LOS group when compared to the control group.

We display three different types of declared regions in Figure 3.2, Figure 3.3, and Figure 3.4. In each figure, the horizontal axis shows genomic positions of SNPs;

Table 3.3: Ten detected ASE regions in control samples with the most number of SNPs.

Region	Chromosome	Start	End	Length	State	SNP	Proportion
1	7	22450103	22453500	3397	M	44	0.70
2	8	39924975	39929910	4935	M	36	0.74
3	14	71829133	71832560	3427	M	25	0.81
4	22	47706034	47717136	11102	P	24	0.21
5	14	47259240	47261976	2736	M	22	0.73
6	21	20303886	20308883	4997	P	22	0.09
7	8	104332849	104337031	4182	M	21	0.81
8	2	132633210	132636855	3645	M	20	0.89
9	22	58261381	58267555	6174	P	20	0.28
10	29	42075522	42083140	7618	M	20	0.73

Table 3.4: Ten detected regions of ASE alterations with the most number of SNPs.

Region	Chromosome	Start	End	Length	State	SNP	Odds Ratio
1	23	28308097	28315879	7782	2	18	0.12
2	24	57310840	57312123	1283	2	14	0.18
3	2	136251406	136253626	2220	2	11	0.23
4	3	54218782	54221715	2933	2	11	0.42
5	15	27871839	27872522	683	1	11	2.44
6	4	98836310	98840696	4386	2	10	0.32
7	23	19883217	19883773	556	2	10	0.66
8	14	10170174	10170615	441	2	10	0.33
9	30	130300790	130307410	6620	2	9	0.38
10	18	61410496	61421828	11332	2	9	0.78

vertical axis in the top panel represents observed maternal percentage of gene expression in control samples after logistic transformation (i.e.,Y_{1s}'s), and vertical axis in the bottom panel displays the difference between transformed percentage of gene expression from maternal allele in two samples (i.e., D_s's).

Figure 3.2 visualizes the fourth region in Table 3.3. It is an example of significant ASE in the control group, where the paternal allele contributes significantly more mRNAs for most of the 24 SNPs. This paternal-allele expression preference is retained in the LOS group. Figure 3.3 corresponds to the first region in Table 3.4, where the

gene expression is balanced between two alleles in the control group, but the maternal allele expresses significantly less in the LOS sample compared to the control group. A third example is the fifth region in Table 3.3 that is displayed in Figure 3.4. It exhibits both significant ASE in control and ASE alterations in LOS. In the control group, the paternal allele expresses more mRNAs than maternal allele. However, in the LOS group, the percentage of paternal allele expression is reduced.

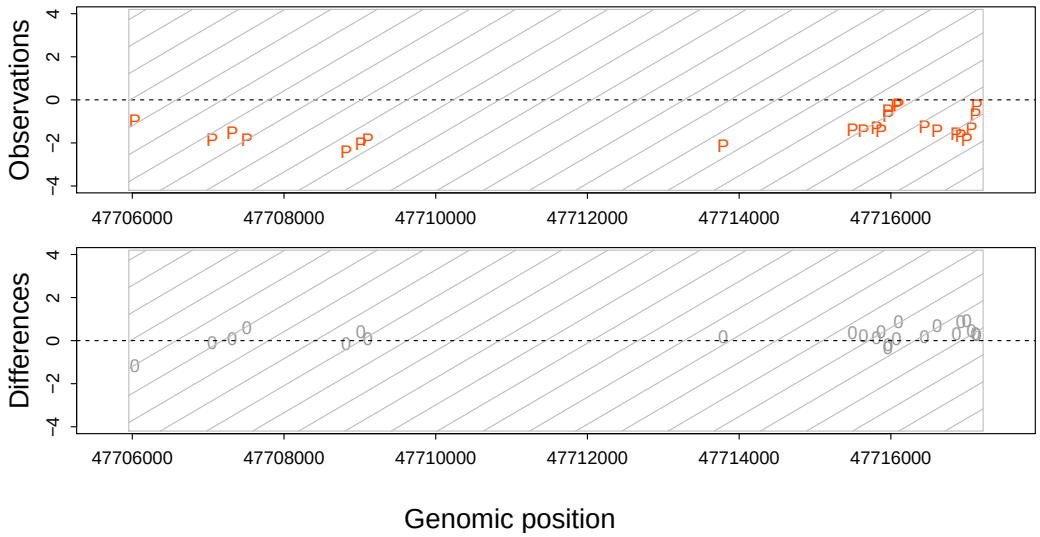


Figure 3.2: A detected ASE region on chromosome 22. The top and bottom panels respectively show the transformed observations in control samples, and the transformed difference between LOS and control samples. The shaded area represents the detected region. The leading and trailing areas represent 200 base pairs before and after the region.

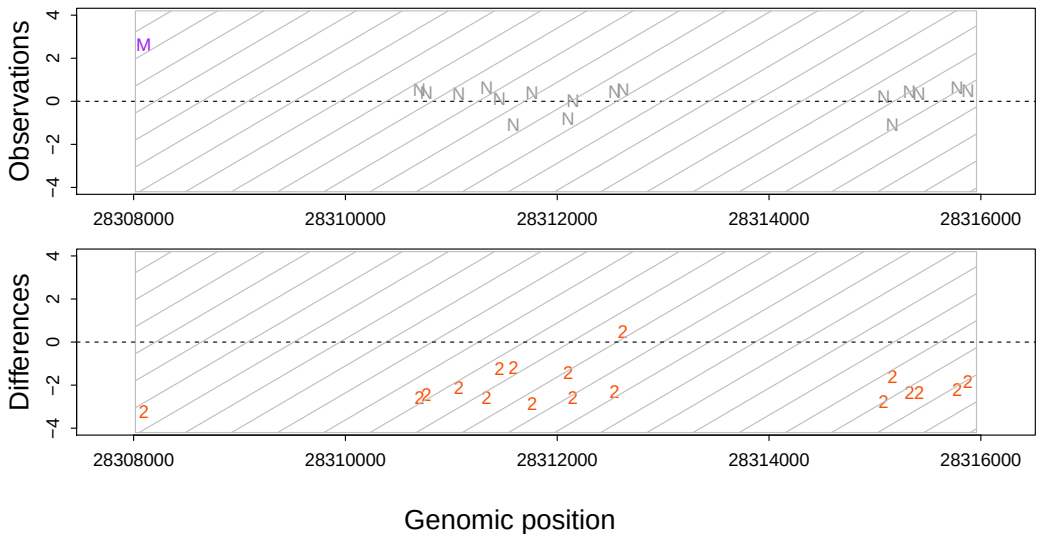


Figure 3.3: A detected region of ASE alterations on chromosome 23. The top and bottom panels respectively show the transformed observations in control sample, and the transformed difference observed between LOS and control samples. The shaded area represents the detected region. The leading and trailing areas represent 200 base pairs before and after the region.

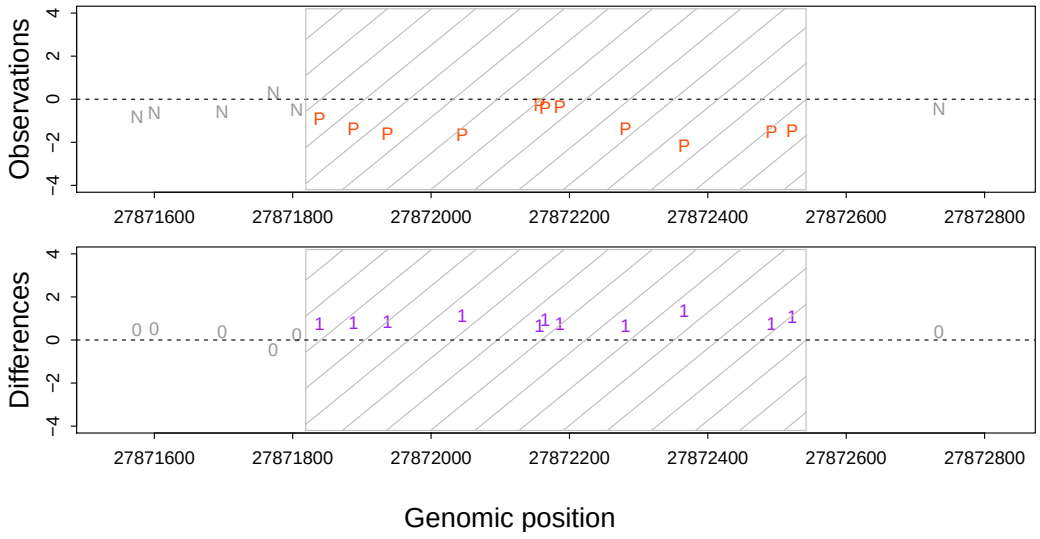


Figure 3.4: A detected region that exhibits both ASE in the control group and ASE alterations in the case group on chromosome 15. The top and bottom panels respectively show the transformed observation in control samples, and the transformed difference observed between the LOS and control samples. The shaded area represents the detected region. The leading and trailing areas represent 200 base pairs before and after the region.

We plot the inference results of the whole genome in Figure 3.5 as a bivariate contour plot, where the horizontal axis indicates transformed maternal percentage in the control group, and the vertical axis indicates difference of transformed maternal percentages between the LOS and control groups. The letters in the upper right corner of each nine small panels display the predicted hidden state. The percentage of SNPs in the whole genome in each of the nine categories displays at the panel bottom.

In sum, in the control sample, we find 83.46% of SNPs with about equal expression from two parental alleles, 9.29% of SNPs have significant more expression from maternal allele, and 7.25% of SNPs have significant more expression from paternal allele. In the LOS samples, compared with the control group, 81.28% of SNPs show no significant change in ASE, 6.74% of SNPs exhibit increased proportion of gene expression from the maternal allele, and 11.99% of SNPs have increased proportion of gene expression from the paternal allele.

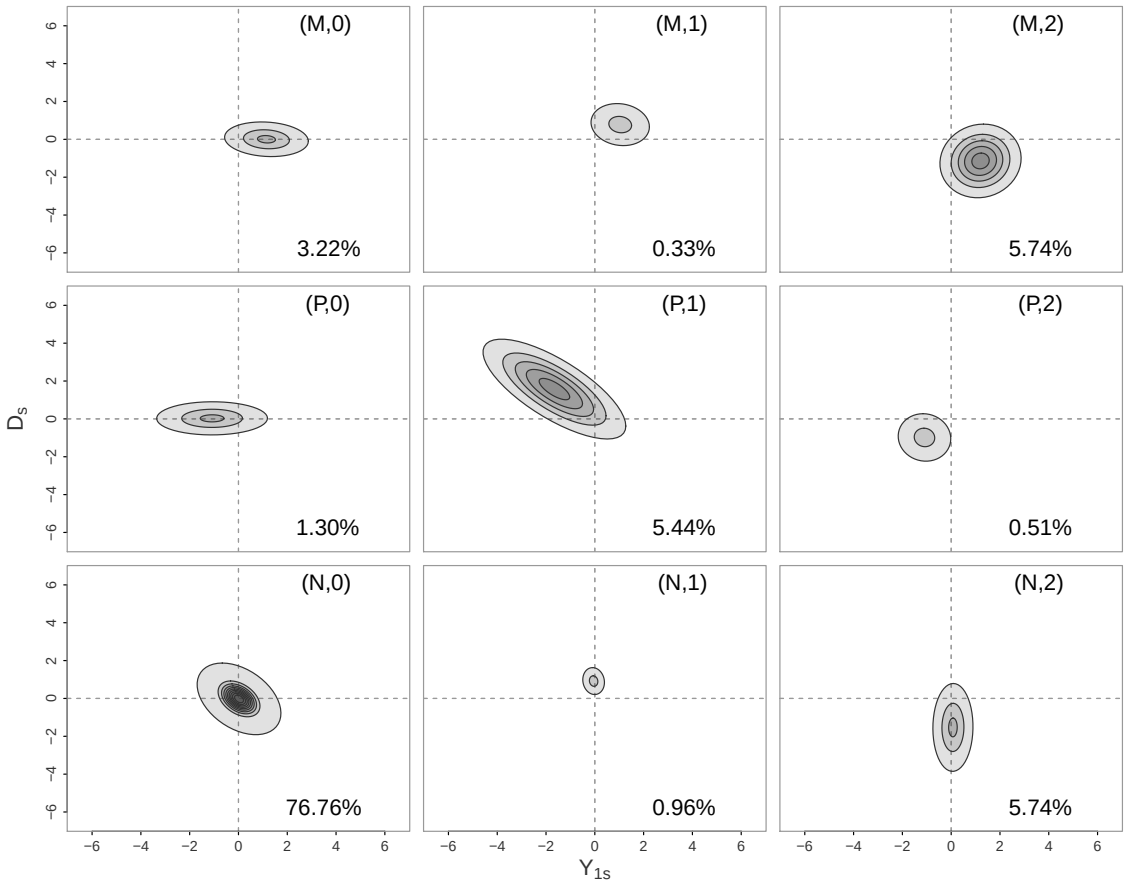


Figure 3.5: Bivariate contour plot. The horizontal axis represents transformed maternal percentage in control samples, and the vertical axis represents difference of transformed maternal percentages between LOS samples and control samples. The upper right letters in each panel display the predicted hidden state. The number at bottom right corner in each panel indicates the percent of SNPs in each category.

3.6 Conclusions

In this chapter, we have proposed a new method, hmmASE, to detect ASE in control and ASE expression changes in a case group for any sequencing experiment. We test these two variables simultaneously and incorporate information from these two dimensions. The proposed approach has substantially improved power compared to other existing methods that render the same functionality. Specifically, we utilize a Bayesian HMM with a bivariate Gaussian emission probability model. We further assume truncated bivariate Gaussian priors on the means of the emission probabilities. We exploit a transition probability model that incorporates the genomic locations of SNPs to account for correlations between adjacent observations. The EM algorithm was utilized to estimate model parameters. We decode the best sequence of hidden states by the Viterbi algorithm. Consecutive SNPs with the same state naturally form regions. In practice, we often set some restrictions, for example, we require no two adjacent SNPs in one region are further than 5,000 base pairs apart. A real data analysis on the LOS study shows the practical utility of the proposed approach. We then utilize the real data set to facilitate simulation studies, which has demonstrated a dominating power, accuracy, and precision of the proposed method compared to other approaches.

Chapter 4

A Mixture Model for Dispersion Parameters that Improves Differentially Expressed Gene Detection

4.1 Introduction

As the most common form of transcriptome analysis, differential expression analysis plays an important role in the characterization and understanding of the molecular basis of phenotypic variation in biology, including diseases [101]. Differential expression analysis involves searching for a set of genes in the whole genome whose mean expression levels are substantially different across treatment conditions, such as control versus disease. With the development of novel high-throughput DNA sequencing methods (RNA-seq), developing statistical methods for detecting differentially expressed (DE) genes has been an extensively studied research area.

Three methods - edgeR [88, 87, 84, 69], DESeq [2, 64], and baySeq [34] have made

significant earlier contributions to push the field forward. All these three methods assumed that, for each gene, the observed gene expressions were a random sample from Negative Binomial distributions for each treatment condition, respectively. To improve estimation of the dispersion parameters in Negative Binomial distributions, edgeR, DESeq, and baySeq further utilized a Bayesian method to borrow information across genes in an entire genome such that the estimation of dispersion parameter for each gene was improved. Due to the expense associated with acquiring biological replicates and high-throughput sequencing experiments, the number of biological replicates of each treatment group is small, such as 3 or 4 for most experiments. Thus, estimation of variances (or dispersion parameters in Negative Binomial distributions) that only utilizes gene-wise observations is not accurate and reliable. Meanwhile, since a large number of genes in an entire genome are available and collected together by the high-throughput sequencing technology, there is useful information about dispersion that can be shared across genes to improve dispersion estimation for each gene. Therefore, this kind of shrinkage idea has been widely accepted to empirically produce superior statistical inferences than the methods without sharing information across genes.

These pioneer methods have inspired many new statistical developments to adopt similar Bayesian approaches to devise more powerful solutions for the detection of DE genes. For example, the Cuffdiff method [115] can detect DE genes with alternative splicings. The voom [53] method is built on a similar empirical Bayes analysis pipeline with additional weights for improved sample and gene-specific dispersion estimation. The EBseq method [55] is a empirical Bayes method aims to identify differentially expressed isoforms and it is based on Negative Binomial model and has been shown to be robust in the identification of DE genes.

However, intriguing results published in [58, 57] have motivated the reconsider-

ation of the shrinkage idea, where the hierarchical models proposed in [88, 87] by borrowing information from all genes in an entire genome tend to over shrink the dispersion parameters. As pointed out by [58, 57], "regress" parameters of all genes toward the middle is a double-edged sword. One the one hand, it alleviates the small sample size problem; on the other hand, it inadvertently introduces biases for genes that have intrinsic high or intrinsic low variance. [58] examined 566 historical datasets [68] which suggested intrinsic gene-wise dispersion that was related with gene functions. For example, housekeeping genes such as translation elongation or ribosome-related genes often have low dispersion, whereas genes responsive to stimuli exhibit high dispersion. Pulling dispersion parameters of all genes towards each other leads to overcorrection. [58, 57] proposed to group genes into categories with informative priors based on historical data in the hope that genes in one category share more similar size of dispersions. However, historical data may not always be available and likely require thorough scrutiny for helpful use.

Another motivation of this chapter is from the study of the large offspring syndrome (LOS) gene expression data [12], where the empirical distribution of maximum likelihood estimates (MLE) of dispersion parameters (Figure 4.1) exhibits two groups. Specifically, one group of estimated dispersions follow a unimodal distribution, potentially modeled by a normal or lognormal distribution, and the other exhibits low level of dispersion where dispersion parameters are close to zero. This empirical distribution deviates substantially from a Gaussian prior assumed in edgeR, DESeq, baySeq, and so on, and also significantly different from a lognormal distribution assumed in the DSS method [126]. This observation reinforces the idea that not all genes have similar dispersion size, which coincides with results from [68] and [58]. Thus, grouping genes in categories and then sharing information across genes within a category can improve dispersion estimation compared with the prior methods where dispersion

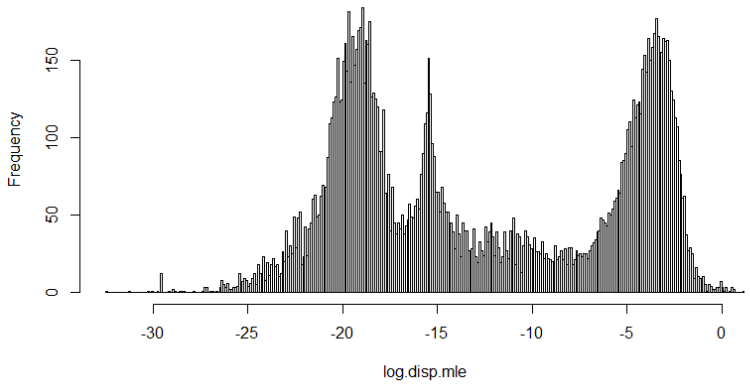


Figure 4.1: Empirical distribution of logarithm of the estimated dispersion parameters by MLE of liver tissues from the LOS study by [12].

parameters of all genes are shrunk towards the overall mean. Based on the empirical study of the LOS data [12] and prior research results [2, 68, 58, 57], we propose a mixture prior model on dispersion parameters with a lognormal probability distribution and a point mass at zero. The latter mixture component represents genes with high uncontrolled dispersion across biological replicates, and the former mixture component represents genes whose dispersion parameter is small and close to zero such that the Negative Binomial distribution is degenerated to a Poisson distribution. In fact, previous methods with a lognormal prior (or normal prior) is a special case of the proposed mixture prior when the percentage of zero component is zero in the mixture model. Thus, the proposed mixture model is flexible and can be generalized to any sequencing experiment with a prior closer to the empirical distribution than all prior unimodal approaches.

To assess the performance of the proposed method, we conduct simulation studies under different scenarios and compare the dispersion estimation and testing power with other competing methods. We use the LOS gene expression studies to illustrate the practical usage of the proposed method. The R package, mix, is developed based on the proposed method and is available for use. A copy of R code is downloadable

on GitHub at https://github.com/JingXieMIZZOU.

4.2 Mixture model for dispersion parameters

Let Y_{gjk} denote the number of read counts for gene g in the k^{th} biological replicate of the j^{th} treatment group, where $g = 1, \ldots, G$, $j = 1, 2$, $k = 1, \ldots, n_j$. Here the main focus is the comparison between two groups, where $j = 1$ represents the control group and $j = 2$ represents the treatment group, and the numbers of biological replicates in two groups can be different. A natural choice for modeling the number of read counts may be Poisson. However, the mean-variance relationship of Poisson limits the flexibility of the corresponding model, thus not appealing in practise when data exhibits overdispersion. Therefore, we choose a Negative Binomial model for the underlying probabilistic data generating mechanism:

$$Y_{gjk} \mid \mu_{gj}, \phi_g \sim \mathrm{NB}\left(\mu_{gj}, \phi_g\right). \tag{4.1}$$

where μ_{gj} denotes the mean expression for the gth gene in the jth group. ϕ_g denotes the dispersion parameter of gene g. Under this parameterization, we have $E(Y_{gjk}) = \mu_{gj}$ and $\mathrm{Var}(Y_{gjk}) = \mu_{gj} + \phi_g \mu_{gj}^2$.

Shrinkage estimators for ϕ_g have been shown to be useful in typical RNA-seq experiments when the number of replicates is small [2, 34, 84]. Since there is no conjugate prior for ϕ_g in the Negative Binomial model, the choice of prior model is usually empirical based on the real data. For example, edgeR [88] utilized a normal prior. [126] used real datasets arguing that a lognormal prior was closer to the observed empirical distributions than a normal prior, thus rendering an improved shrinkage estimator. Note that reported observations in edgeR and DSS methods can

be both case-specific. In previous empirical studies, for instance, the gene expression data from LOS studies (Figure 4.1) strongly deviates from the unimodal assumption in both the normal and lognormal models. In addition, using historical data, [58] demonstrated that edgeR and DSS methods that shared information directly across all genes likely led to overcorrection. To prevent overcorrection, we propose the mixture model (4.2). The proposed prior model is closer to the empirical distribution and is flexible to be generalized to majority of the sequencing experiments.

$$\phi_g \sim p \times \mathbf{1}_{[\phi_g=0]} + (1-p) \times \mathbf{1}_{[\phi_g>0]} \times \text{lognormal}\,(m, \tau^2). \tag{4.2}$$

where p is the mixture proportion for zero dispersion, m and τ^2 are the hyperparameters that we need to estimate. Specifically, we model ϕ_g following a mixture model with a point mass at zero and a lognormal distribution. The point mass at zero represents genes which exhibit very low dispersion that is close to zero, whereas the lognormal component represents genes with high uncontrolled dispersion across biological replicates. Over-shrinkage is avoided since dispersion information is shared only across genes within each mixture component. When $\phi_g = 0$, the distribution of the number of read counts Y_{gjk} reduces to a Poisson.

4.3 Detecting differentially expressed genes

By assuming a mixture model for dispersion parameters, this section outlines the inferential procedure for parameter estimation and the detection of DE genes. Given models in (4.1) and (4.2), for gene g, the conditional posterior distribution of ϕ_g given

all $n = n_1 + n_2$ observations and treatment mean μ_{gj} expressions is

$$
\begin{aligned}
& P(\phi_g | Y_{gjk}, \mu_{gj}, k = 1, \ldots, n) \\
\propto & P(Y_{gjk} \mid \mu_{gj}, \phi_g, k = 1, \ldots, n) P(\phi_g) \\
= & p \times \mathbf{1}_{[\phi_g = 0]} \times \prod_{k=1}^{n} P_{\text{Poisson}}(Y_{gjk} \mid \mu_{gj}) + \\
& (1 - p) \times \mathbf{1}_{[\phi_g > 0]} \times P_{\text{lognormal}}(\phi_g \mid m, \tau^2) \prod_{k=1}^{n} P_{\text{NB}}(Y_{gjk} \mid \mu_{gj}, \phi_g). \qquad (4.3)
\end{aligned}
$$

Due to the intractability of (4.3), it involves computational complexity to obtain the posterior mean of ϕ_g. Alternatively, we propose to employ the weighted average of posterior modes, which is denoted by $\tilde{\phi}_g$, to approximate the posterior mean, and utilize it as the Bayesian estimate of the dispersion parameter ϕ_g. As mentioned earlier, the mixture prior on ϕ_g makes the distribution of Y_{gjk} a mixture of Poisson and Negative Binomial. Denote the first part of the posterior associated with the Poisson by $L_1 = p \times \mathbf{1}_{[\phi_g = 0]} \times \prod_{k=1}^{n} P_{\text{Poisson}}(Y_{gjk} \mid \mu_{gj})$, which is the product of likelihood of Poisson model and the mixture proportion p. Similarly, the second piece of the posterior associated with Negative Binomial model is denoted by $L_2 = (1 - p) \times \mathbf{1}_{[\phi_g > 0]} \times P_{\text{lognormal}}(\phi_g \mid m, \tau^2) \prod_{k=1}^{n} P_{\text{NB}}(Y_{gjk} \mid \mu_{gj}, \phi_g)$. Let $\hat{\phi}_m$ denote the mode of L_2, then the proposed Bayesian estimate of ϕ_g is defined as $\tilde{\phi}_g = 0 \times \frac{L_1}{L_1 + L_2} + \hat{\phi}_m \times \frac{L_2}{L_1 + L_2} = \hat{\phi}_m \times \frac{L_2}{L_1 + L_2}$. The Newton-Raphson method was used to seek for $\hat{\phi}_m$ that maximizes L_2.

Before applying aforementioned procedure to achieve the Bayesian estimate $\tilde{\phi}_g$, we need to estimate the mean μ_{gj}, hyperparameters m, τ^2 and p. We estimate μ_{gj} by the sample mean, i.e., $\hat{\mu}_{gj} = \frac{1}{n_j} \sum_{k=1}^{n_j} Y_{gjk}$. The method of moment is utilized to estimate the hyperparameters. Recall that dispersion parameter ϕ_g follows a mixture model with a point mass at zero and a lognormal distribution, and the moments ϕ_g

are

$$E(\phi_g^n) = (1-p) \times \exp\left(nm + \frac{1}{2}n^2\tau^2\right). \tag{4.4}$$

Let $\widehat{\phi_g}$ denote the MLE of dispersion parameter ϕ_g where $g = 1, \ldots, G$. Then $Z_1 = \bar{\hat{\phi}}_g = \frac{1}{G}\sum \hat{\phi}_g$, $Z_2 = \bar{\hat{\phi}}_g^2 = \frac{1}{G}\sum \hat{\phi}_g^2$ and $Z_3 = \bar{\hat{\phi}}_g^3 = \frac{1}{G}\sum \hat{\phi}_g^3$ represent the first, second and third sample moment of $\hat{\phi}_g$ respectively. By utilizing the method of moments method, we obtain the estimate of hyperparameters p, m, and τ^2 as follows:

$$\begin{cases} \hat{p} & = 1 - exp\left\{3\,log\left(Z_1\right) - 3\,log\left(Z_2\right) + log\left(Z_3\right)\right\} \\ \hat{m} & = 4\,log\left(Z_2\right) - \frac{5}{2}\,log\left(Z_1\right) - \frac{3}{2}\,log\left(Z_3\right) \\ \hat{\tau}^2 & = log\left(Z_1\right) + log\left(Z_3\right) - 2\,log\left(Z_2\right) \end{cases} \tag{4.5}$$

Once the shrinkage estimation of the parameter ϕ_g is obtained, the last step is to perform a hypothesis test to compare the mean expression between two groups, i.e., test $H_0 : \mu_{g1} = \mu_{g2}$ VS $H_a : \mu_{g1} \neq \mu_{g2}$ for each g. There are several testing methods available such as the Wald test by [65], Exact test by [88] and Likelihood Ratio test by [4]. As studied in previous research, the empirical distribution of the Wald statistics is Gaussian-like [126]. The Wald test and the corresponding false discovery rate (FDR) estimates have been implemented in some R packages such as edgeR and DSS, thus we choose the Wald test because of its simplicity and the convenience of implementation. Specifically, the Wald test statistic is defined as

$$T_g = \frac{\hat{\mu}_{g1} - \hat{\mu}_{g2}}{\sqrt{\hat{\sigma}_{g1}^2 + \hat{\sigma}_{g2}^2}}. \tag{4.6}$$

where $\hat{\sigma}_{gj} = \frac{1}{n_j}(\hat{\mu}_{gj} + \phi_g \hat{\mu}_{gj}^2)$ with $j = 1$ and $j = 2$ represents the estimated variance for the control and treatment group, respectively.

4.4 Simulation study

In this section, we simulate data to mimic the structure of real data, and demonstrate the performance of the proposed method which is refered to as "mix". The mix method is compared with edgeR [88], DSS [126], TSPM [4] and EBseq [55] from three perspectives: parameter estimation, false discovery rate (FDR) control and gene ranking. For parameter estimation, since only DSS, edgeR and mix are based on Negative Binomial model which is parameterized by mean and dispersion parameter, and thus provide dispersion estimation, we exclude TSPM and EBseq for the comparison. Three scenarios are considered in the simulation study to show the robustness to the distributional assumption, where the dispersion parameters ϕ_g's are simulated from different mixture distributions.

In the first scenario, we simulate data from the true model. Specifically, 10,000 genes are generated with 10% of which are truly differently expressed in the treatment group compared to the control group. The dispersion parameters ϕ_g's are simulated from the mixture distribution of point mass at zero and lognormal distribution with equal weight, i.e., $\phi_g \sim p \times \mathbf{1}_{[\phi_g=0]} + (1-p) \times \mathbf{1}_{[\phi_g>0]} \times \text{lognormal}\,(m, \tau^2)$ with $p = 0.5$. Particularly, for gene g, the mean expression for control group μ_{g1} is randomly sampled from the sample means of control group in the large offspring syndrome (LOS) gene expression data [12]. For each gene, an indicator γ_g is sampled from discrete uniform [0,1], then the dispersion parameters are generated as $\phi_g|(\gamma_g = 0) = 0$ and $\phi_g|(\gamma_g = 1) \sim \text{lognormal}\,(m_1, \tau_1^2)$. $m_1 = -2.36$ and $\tau_1^2 = 0.35$ and these values are estimated from LOS data. Randomly select 1000 genes to express differently for two groups, and for which the mean expression for treatment group is specified as $\mu_{g2} = \mu_{g1} + k\delta\sqrt{\mu_{g1}(1+\mu_{g1}\phi_g)}$ where $\delta \sim \text{Beta}\,(2,\, 4)$. The constant k is to control the magnitudes of differential expression and we set it to 4 in all the simulation

scenarios. For the other 9,000 genes, set $\mu_{g2} = \mu_{g1}$. With μ_{g1}, μ_{g2} and ϕ_g, we simulate R replicates for each groups from either Poisson or Negative Binomial distribution depending on the value of ϕ_g.

The simulation settings for the second and third scenario are similar with the first scenario, except for the parametric distributions for ϕ_g. For the second scenario, we aim to explore the cases where all the genes demonstrate overdispersion by assuming a mixture distribution of two lognormal. Thus the ϕ_g's for genes with $\gamma_g = 0$ are no longer 0 but sampled from another lognormal distribution with a very small mean, i.e., $\phi_g|(\gamma_g = 0) \sim \text{lognormal}\,(m_0, \tau_0^2)$, and $\phi_g|(\gamma_g = 1) \sim \text{lognormal}\,(m_1, \tau_1^2)$. Here we set $m_0 = -10$, $\tau_0^2 = 2$. As with the third scenario, we keep the structure of true model but deviate from it by generating ϕ_g's for genes with $\gamma_g = 1$ from Gamma(α, β) instead of lognormal. The distribution of dispersion parameter is a mixture of point mass at 0 and Gamma, i.e, $\phi_g \sim p \times \mathbf{1}_{[\phi_g=0]} + (1-p) \times \mathbf{1}_{[\phi_g>0]} \times \text{Gamma}\,(\alpha, \beta)$ with $\alpha = -1.92$, $\beta = 15.22$, which are estimated from LOS data. For each of the three scenarios, we consider the situations where $R = 4$, $R = 6$, and $R = 8$, and for each situation we repeate the simulation 50 times.

Figure 4.2 shows the Boxplot for comparing the mean square errors (MSE) of the dispersion estimates across 50 simulations. To be consistent with edgeR and DSS, the MSE's in the plot are in $\phi_g/(1+\phi_g)$ scale. The upper, middle and bottom panel represents simulation scenario 1, scenario 2 and scenario 3, respectively. The first, second and third column corresponds to the cases where we have 4, 6 and 8 replicates in each group, respectively. Compared with edgeR and DSS, the proposed mix method shows improvement in the dispersion estimation in all the situations of all the three scenarios. The improvement mostly comes from the estimation for genes which are not overdispersed, i.e., the genes with $\phi_g = 0$ in the scenario 1 and scenario 2, or genes have a mild overdispersion, i.e., the genes with $\phi_g \sim \text{lognormal}\,(m_0, \tau_0^2)$

in the scenario 3. It is expected when the distribution of dispersion ϕ_g departs from the true model on which the proposed method is based, e.g., ϕ_g is simulated from a mixture of point mass at zero and Gamma as shown in the bottom panel, the MSE's shift up.

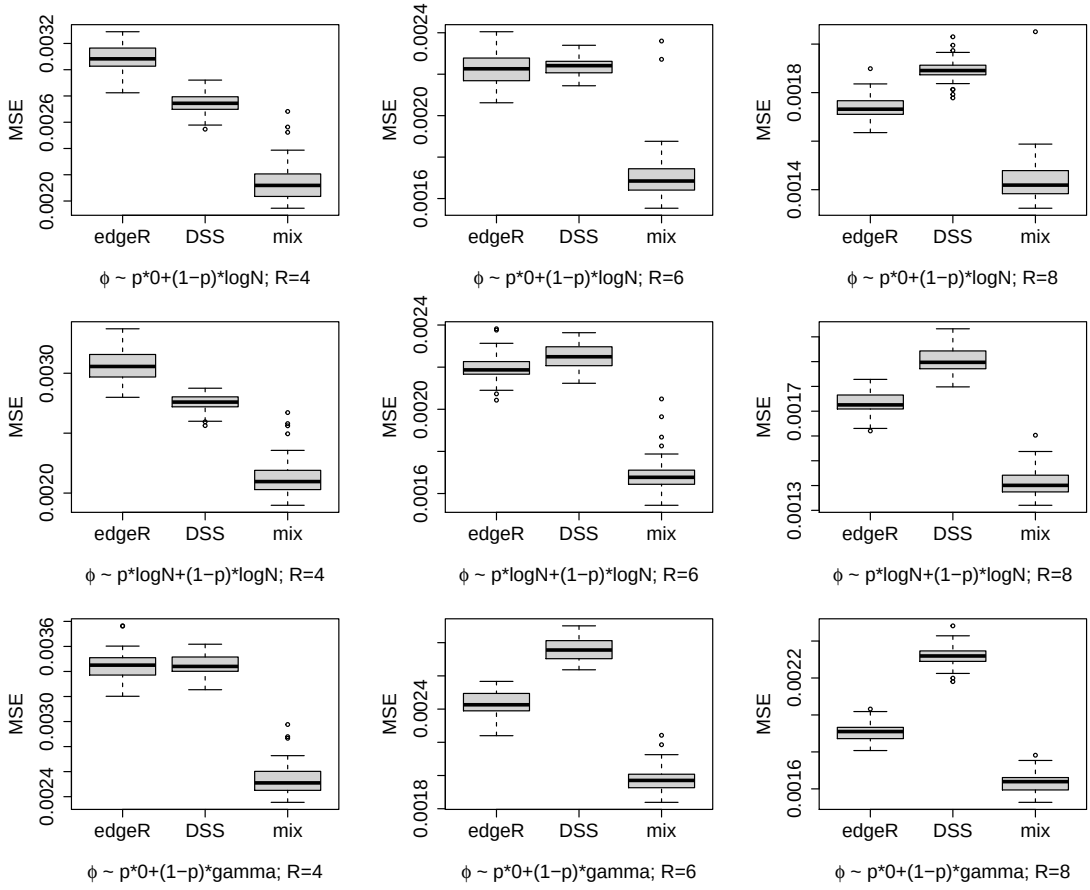


Figure 4.2: Comprison of MSE for dispersion estimations for DSS, egdgeR and proposed mix method across 50 simulations. The top, middle and bottom panel represents the first, second and third simulation scenario, respectively. The first to third columns represents cases where $R = 4$, $R = 6$, and $R = 8$ respectively. The Boxplot are on the $\phi_g/(1+\phi_g)$ scale.

Figure 4.3 presents the true discovery curves which depict the accuracy of differential expression detection for the top ranked genes. Similar with Figure 4.2, each subplot represents a distinct situation where the simulation scenario and/or the num-

ber of biological replicates are different. Genes are ranked by FDR from low to high, for mix, DSS, edgeR and EBseq; ranked by adjusted P value from low to high for TSPM. The horizental axis represents the number of top-ranked genes. The vertical axis represents the percentage of true discovery among the top-ranked genes. The ideal true discovery curve should be the horizontal line at $y = 1$, which means the method successfully rank truly significant DE genes ahead of truly non-significant genes without an error. The red curves represent the proposed mix method and they are slightly closer to the ideal curve and have grater areas under the curve than others methods in all the nine subplots, which indicates an improvement in the accuracy of of differential expression detection. Among the other competing methods, DSS has the most similar overall performance with the proposed mixed method. Particularly, when the number of replicates in each group is 4, the true discovery curve for DSS is close to that of proposed mix method, and the mixed method and DSS substantially outperform the other three methods. When the sample size becomes larger from 8 to 16, the performance of all the methods improve and the TSPM method exhibits the most significant improvement.

The last benchmark that we employ to compare the proposed mix method with other competing methods is the ability to control FDR at an imposed level. Figure 4.4 shows the Boxplots of true FDR across 50 simulations for various methods when controlling FDR at 0.05 level. In a more realistic situation where the sample size is small, the proposed mixed method control the FDR the best. Specifically, for scenario 1 and scenario 2, where the mixture components contains lognormal, the mix method control the FDR at the nominal level, whereas the edgeR is overly conservative, EBseq is slightly conservative, both DSS and TSPM are too liberal. For scenario 3, where the dispersion are from a mixture of point mass at zero and Gamma, the mix method is slightly liberal but still reasonable. Along with the increase in sample size, TSPM

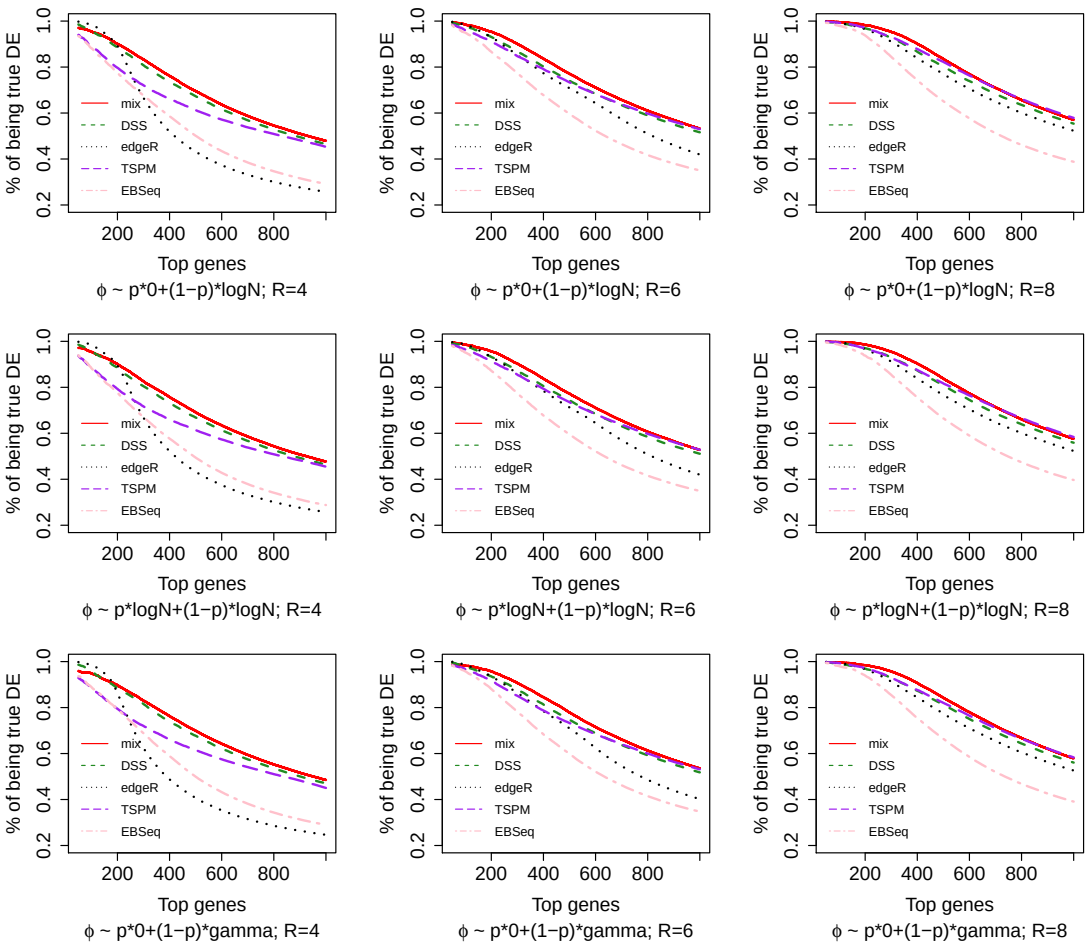


Figure 4.3: Comparison of DE detection accuracy for top 1000 ranked genes. The top, middle and bottom panel represents the first, second and third simulation scenario, respectively. The first to third columns represents cases where $R = 4, R = 6$, and $R = 8$ respectively.

performed better in terms of FDR control, which coincides with the previous study [101, 52]. It is surprising that the proposed mix method become more conservative when sample size increases. A possible explanation is that the noise introduced by more sample results in the overestimation of dispersion and also the variance, which makes the mix method underestimate the significance of these genes. By considering all these 9 situations in general, the mix and DSS method have the best overall

performance in FDR control.

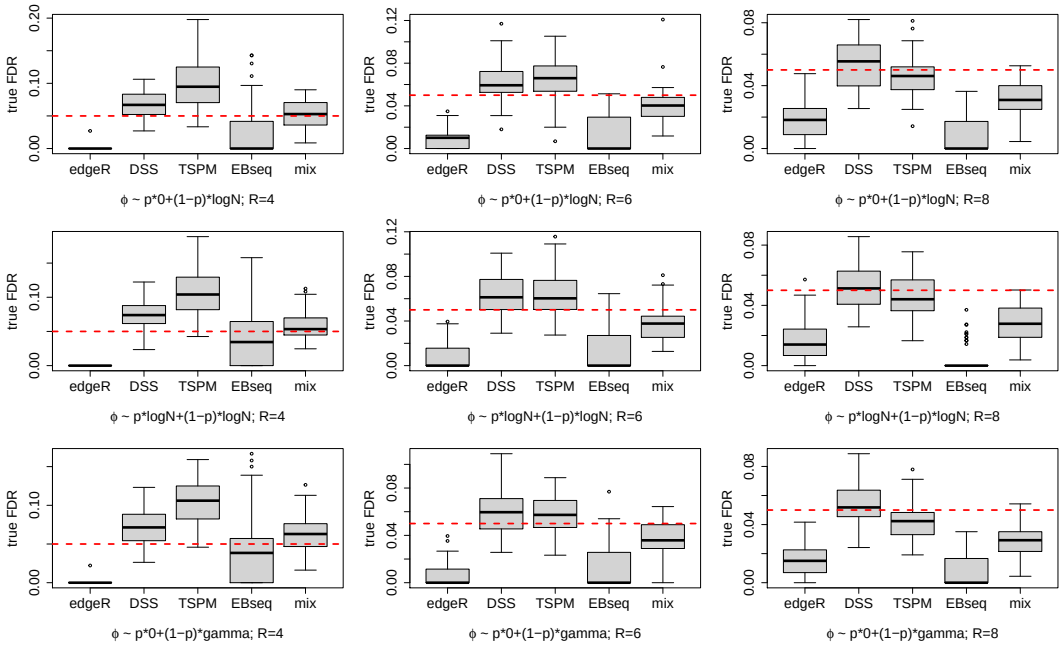


Figure 4.4: Boxplot of the true FDR when controlling estimated FDR or adjusted p value at 0.05 level for the comparison of FDR control. The top, middle and bottom panel represents the first, second and third simulation scenario, respectively. The first to third columns represents cases where $R = 4$, $R = 6$, and $R = 8$ respectively.

The three simulation scenarios indicate that the proposed mix method exhibits improved dispersion estimation, better FDR control and more accuracy in ranking the DE genes on the top of the list. Although the performance varies across simulation scenarios, the advantages of mix method over other methods have been shown in all these three scenarios. It suggests that the proposed mix method is robust to misspecification of the parametric distribution of dispersion parameters if a bimodal distribution is assumed. To study the performance of the mix method when the true underlying distribution for dispersion is unimodal, we conduct a side study where $\phi_g \sim \text{lognormal}\,(m_1, \tau_1^2)$. $m_1 = -2.36$, $\tau_1^2 = 0.35$, and other settings are the same with aforementioned three scenarios. Figure 4.5 shows the simulation results for the

side study. The upper panel shows the comparison for gene ranking for different method. It seems that when the sample size is small, the proposed mix method does not perform as well as DSS; and when the sample size increase, the gaps between the TPR curves of different method are diminished. The bottom panel shows the comparison for dispersion estimation, and it is expected that when the distribution of the dispersion is bimodal, the proposed mix method would loss its advantage.

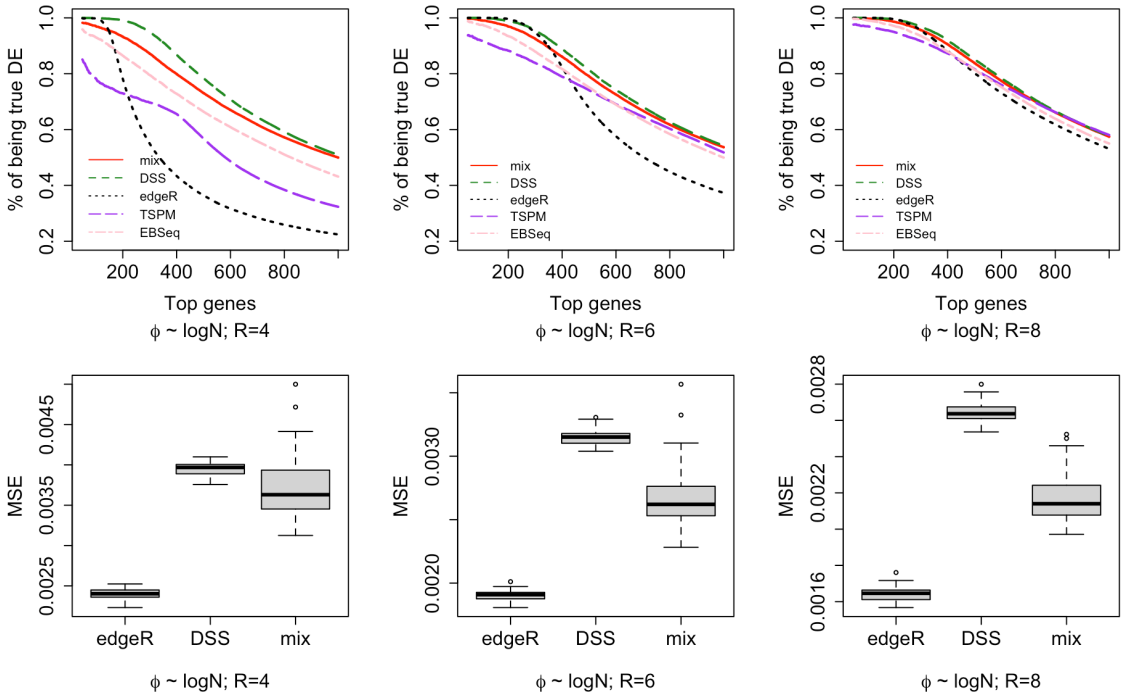


Figure 4.5: Simulation results for the side study. The upper panel shows the comparison of gene ranking curve, and the bottom panel displays the MSE for dispersion estimation. The first to third columns represents cases where $R = 4, R = 6$, and $R = 8$ respectively.

4.5 Real data analysis

To illustration the empirical utility, we apply the proposed mix method to the LOS study. LOS is a fetal overgrowth condition that exhibits variable phenotypic abnormalities including overgrowth, enlarged tonge, and abdominal wall defects. These characteristics mimic the human syndrome Beckwith Wiedemann (BWS), which is the most common pediatric overgrowth syndrome. The phenotypic and epigenetic similarities between LOS and BWS make LOS an appropriate animal model for the study of BWS and the understanding of the etiology of these overgrowth syndromes [14]. To study the underlying molecular mechanism behind LOS, [12] used cattle as the study animal, and generated RNA sequencing (RNAseq) data of skeletal muscle, liver, kidney, and brain from four control and four LOS day $\sim$ 105 (d104–106) B. t. indicus $\times$ B. t. taurus F1 fetuses.

A typical way to perform a differential expression analysis is to assume that the number of read counts for four biological replicates in each of the control and LOS group are fluctuating around a underlying mean, and significant DE genes should exhibit difference between the means in two groups. However, as LOS fetuses exhibited dramatic difference in bodyweight [15], to eliminate the confounding effect of body weight, it makes better sense to consider each LOS fetus individually. Therefore, in the real data analysis, each LOS fetus is compared to the mean of the four controls to identify differentially expressed genes, as with what has been done in [15]. That is, the sample size for control group is four whereas the sample size for LOS group is one. Separate analysis is conducted for each of the four tissue types.

Before applying the proposed method, we process the raw read counts as follows. At the first step, edgeR package is used to filter out genes that have less than 4 replicates (either from control or LOS group) with cpm greater than 2. The number

of genes kept in the study for kidney, muscle, liver and brain is 15,214, 14,258, 13,146 and 15,239, respectively. At the second step, the trimmed mean of M-values method (TMM) proposed by [86] is employed to normalize the data. Finally, the rounded normalized counts are exported as the input for the subsequent analysis.

Table 4.1 summarizes the number of significant DE genes detected by the proposed mix method based on a FDR$<$ 0.05 level in different types of tissue. The first four columns corresponds to the DE genes in each LOS fetus and the last column represents the number of genes which exhibit significantly different expression in at least one LOS fetus. If a gene is significant in at least one LOS fetus, we claim it a DE gene. For example, 1576 DE genes are detected in liver tissue, among which only 385 DE genes are found in fetus LOS 3# whereas 725 DE genes are significant in fetus LOS 1#. Similarly, the difference in the number of detected DE genes in different LOS fetus exists in muscle, kidney and brain tissues. In general, there is smallest number of detected DE genes in muscle tissue, and greatest number of detected DE genes in kidney tissue.

Table 4.1: Number of detected DE genes in different tissues

tissue	LOS#1	LOS#2	LOS#3	LOS#4	at least one LOS Fetus
liver	725	418	385	670	1576
kidney	963	974	1022	737	2474
muscle	188	257	161	309	678
brain	479	732	119	95	1033

Figure 4.6 shows the Venn diagram corresponding to the last column of Table 4.1, and it summarizes the categories of detected DE genes for each tissue, depending on if the DE genes are significant in other tissue types. For example, only 17 detected DE genes are significant in all the four types of tissue; 79 detected DE genes are significant in brain, kidney and muscle; 33 detected DE genes are significant in kidney, muscle

and brain; 56 detected DE genes are significant in kidney, liver and brain. There are relatively larger number of DE genes which only exhibit significant difference in gene expression between control and LOS in one specific type of tissue. Specifically, the number of DE genes that are only significant in liver, kidney, muscle and brain is 842, 1547, 301 and 681, respectively. We also use kidney tissue as an example to apply all the other methods that are mentioned in the simulation study. The Venn diagram of DE genes for kidney tissue detected by different methods is shown in Figure 4.7. Among the DE genes detected by the proposed mix method, 275 out of 2474 genes are also claimed as DE genes by all the other methods.

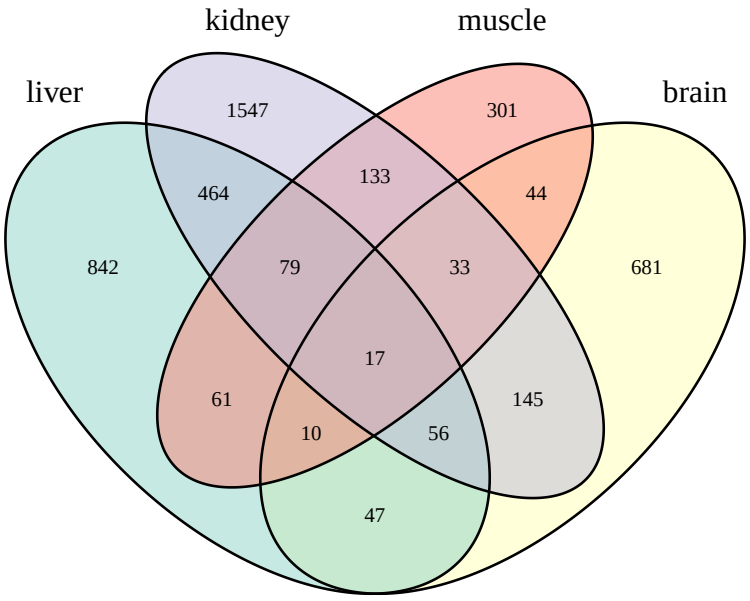


Figure 4.6: Venn diagram of detected DE genes for different tissue types

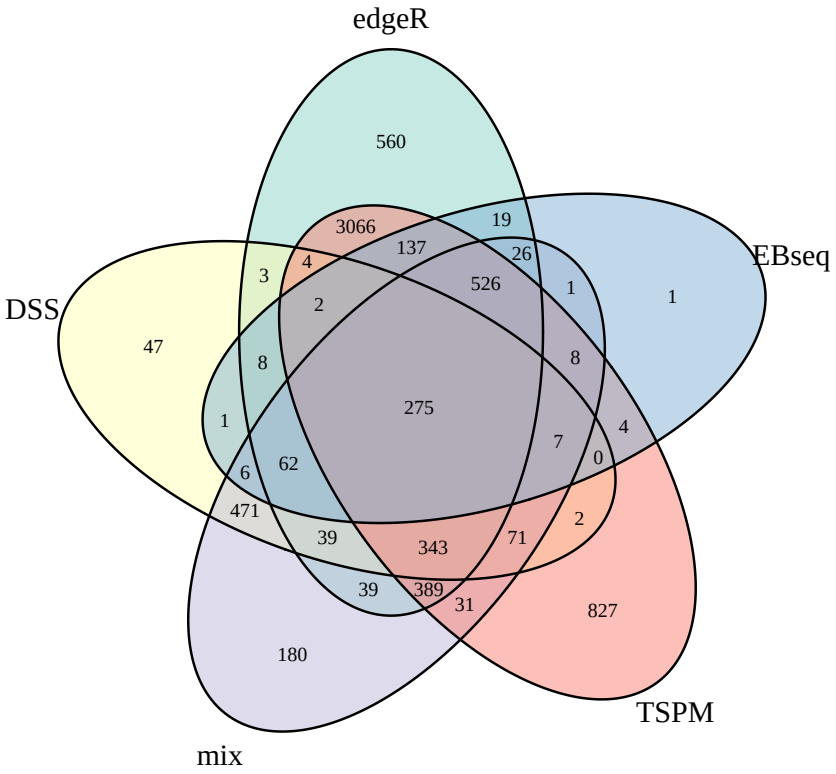


Figure 4.7: Venn diagram of DE genes for kidney tissue detected by different methods

4.6 Conclusions

In this chapter, we proposed a new method to detect the DE genes in any sequencing experiment. The number of read counts for different treatment groups are modelled by two Negative Binomial distribution which may have different means but share the same dispersion parameter. We propose a mixture prior model for the dispersion parameters with a point mass at zero and a lognormal distribution. The mixture model allows shrinkage across genes within each of the two mixture components, thus prevents overcorrection by shrinkage across all genes. In the simulation study, we demonstrate that the proposed method yields a better estimation of the dispersion, demonstrates a higher accuracy in ranking the significant genes on the top, and exhibits a better general FDR control than the other competing methods, when the distribution of dispersion is indeed bimodal. In addition, the proposed method exhibits robustness to the misspecification of the bimodal distribution for the dispersion parameters. Thus it is more flexible and can be generalized to any sequencing experiment with a prior closer the empirical distribution than the other prior unimodal appreoaches. The empirical usage of the proposed method is demonstrated in a real data analysis on a LOS study.

Chapter 5

Future Research

5.1 Combined analysis of multi-omics data

In this dissertation, we have proposed three new statistical methods, i.e., BLMRM, hmmASE and mix. BLMRM targets at the joint inferences on gene level ASE and variations of ASE across SNPs within a gene, hmmASE focuses on the simultaneous detection of ASE region in control group and regions of ASE alterations in case group, and mix aims to detect DE genes. The ultimate objective of these studies is to discover the potential connections between diseases (such as LOS), with ASE and/or differential expression. Except for ASE and differential expression, other molecular features measured by different platforms, e.g., methylation level, microRNA expression, copy number variation etc. all affect phenotype in a specific pathway [41]. Comprehensive understanding of diseases requires interpretation of molecular intricacy and variations at multiple levels [110]. In this chapter, we use LOS as an example disease to describe the potential directions for future research on the combined analysis of multi-omics data, under a hypothetical situation where methylation data and

microRNA expression data is also available.

The Chapter 2 and Chapter 3 have formed a ramified framework for ASE, which can be easily extended to the independent analysis of methylation data and microRNA expression data. However, the information provided by the study that focuses on one-dimensional omics data is limited. In addition, these omics, i.e., ASE, mehylation and microRNA expression, are interactive instead of independent [39]. Therefore, it's important to develop an integrative statistical method to analyze multi-omics data simultaneously. There are numerous progresses done in the field of multi-omics integration, for example, the matrix factorization methods by [130], [96], [81] etc., and the multiple kernel learning and multi-step analysis by [102]. [39] gives a comprehensive review on recent research.

5.2 Integrative analysis of ASE, microRNA and methylation data

As with the integrative analysis of ASE, microRNA and methylation data on LOS study, we propose two potential directions for future research. The first proposed model works for the situation where the number of sample size is limited. It is essentially a two-step method, and a natural extension of the bivariate HMM model in Chapter 3. Specifically, in the first step, we conduct a differential expression analysis on the microRNA data by the mix method proposed in Chapter 4 or DSS [127] and edgeR [85], depending on the dispersion of the data. Let G denote the set of genes which exhibit significant differential expression between the control group and LOS group. G_1 and $\bar{G}_1$ represent the subset of G which is regulated and not regulated by the microRNA, respectively. In the second step, we model the alteration of ASE and alteration of methylation simultanously by a bivariate hid-

den Markov model. Analogous to the hmmASE method, the sequence of bivariate observations $\boldsymbol{O}_b = (Y_{1b}, Y_{2b})'$ will be controlled by 9 hidden states in $\mathcal{M} = \{(M,1),(M,2),(M,0),(P,1),(P,2),(P,0),(N,1),(N,2),(N,0)\}$. The subscript b represents bin which is defined as the segments of chromosome from current SNP to next. Each bin contains one SNP and multiple CpG, and the length of bins are not homogeneous. Y_{b1} and Y_{b2} are defined as $Y_{b1} = \log\left(P_{2e,b}/\left(1-P_{2e,b}\right)\right) - \log\left(P_{1e,b}/\left(1-P_{1e,b}\right)\right)$, $Y_{b2} = \log\left(P_{2m,b}/\left(1-P_{2m,b}\right)\right) - \log\left(P_{1m,b}/\left(1-P_{1m,b}\right)\right)$ where $P_{1m,b}$, $P_{1e,b}$ denote the percentage of maternal methylation and maternal expression for normal group, $P_{2m,b}$, $P_{2e,b}$ denote the percentage of maternal methylation and maternal expression for disease group. The transition probabilities and transition matrix are defined in a similar way as in hmmASE.

The information of microRNA will be integrated as follows. Since microRNA expression affects gene expression more directly than affecting methylation, we let $\boldsymbol{O}_{b:\, b\in\bar{G}_1}$ follow the bivariate normal distribution as defined in (3.1) in Chapter 3. For $\boldsymbol{O}_{b:\, b\in G_1}$, we let the covariance matrix remain the same with that of $\boldsymbol{O}_{b:\, b\in\bar{G}_1}$ but the first dimension of all the 9 mean vector are shifted by a random variable $s \sim N(a, \sigma^2)$ where a and σ^2 are constants. Essentially, depend on if the bins belong to or overlap with a gene that is regulated by a significantly differentially expressed microRNA, we seperate the hidden Markov chain into two sub chains. These two sub chains share same transition probability but have different emission probability distributions. The inference interest now becomes (1) infer the hidden states, (2) estimate the mean effect of microRNA.

Under the circumstance where we have the data from more individuals, e.g., hundreds of samples from both control group and LOS group, we can extend the iBAG method original proposed by [121]. This extended model will consists of two components, i.e., the first component consider the direct effects of methylation and mi-

croRNA expression on ASE by regressing the measures of ASE on the methylation measurement, microRNA expression, and their potential interaction; the second component uses the information in the first component to predict if an individual will have LOS. The inference interest for this method is to select the genes significantly associated with LOS.

In summary, the two aforementioned methods, which serve different inference purposes, are the examples of two potential directions for integrated analysis of multi-omics data. The first method does not require biological replicates, and it reports the regions of ASE and methylation regions, thus is more appealing in practise. However, the information of microRNA is incorporated in the model through an indirect way and it might be difficult to link the significant differentially expressed microRNA back to genes. On the contrary, the second method models the multi-omics information directly and reports more straightforward outputs but it requires a much larger sample size.

Appendix A

Supplementary Materials for "Detecting Allele-Specific Expression and Alterations of Allele-Specific Expression by a Bivariate Bayesian Hidden Markov Model"

A.1 EM algorithm for parameter estimation

Let L_s and $\boldsymbol{O}_S$ denote the hidden states and observations, respectively. Note L_s is different with the $\boldsymbol{L}_s$ in the main paper since here $L_s \in \{1,2,3,4,5,6,7,8,9\}$, which corresponds to the nine bivariate states in the space $\mathcal{M}$ defined in the main paper, respectively. Write $\boldsymbol{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_S)^T$, $\boldsymbol{L} = (L_1, \ldots, L_S)^T$, $\boldsymbol{p} = p'_{kl}s$ where $k, l \in \{1,2,3,4,5,6,7,8,9\}$. $\pi_{L_1} = P(L_1 = k)\,(k \in \{1,2,3,4,5,6,7,8,9\})$ be the probability distribution for the first SNP, $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O})$ and $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O}, \boldsymbol{L})$ denote the incomplete-data

and complete-data likelihood, respectively. Then we have,

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L}) = \pi_{L_1} \prod_{s=1}^{S-1} P\left(L_{s+1}|L_s,\, \boldsymbol{\theta}_{\text{trans}}\right) \prod_{s=1}^{S} P\left(\boldsymbol{O}_s|L_s,\boldsymbol{\theta}_{\text{emiss}}\right) \tag{A.1}$$

Define the logarithm of the complete-data likelihood as $l(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L}) = \log(\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L}))$. We utilize the EM algorithm to iteratively update the parameter $\boldsymbol{\theta}$ and maximize the expectation of complete-data log-likelihood w.r.t the conditional distribution of the hidden status given the observations and parameters, i.e., $\mathrm{E}_{\boldsymbol{L}|\boldsymbol{O},\boldsymbol{\theta}}\left[l(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L})|\boldsymbol{O},\boldsymbol{\theta}\right]$.

The E-step

At the E-step, we evaluate the expectation at the current estimated parameters, e.g., $\mathrm{E}_{\boldsymbol{L}|\boldsymbol{O},\boldsymbol{\theta}^{m-1}}\left[l(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L})|\boldsymbol{O},\boldsymbol{\theta}^{m-1}\right]$ at the m-th iteration. For simplicity, the superscript $(m-1)$ would be omitted in Appendix A.

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{L}|\boldsymbol{O},\boldsymbol{\theta}^{m-1}}\left[l(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L})|\boldsymbol{O},\boldsymbol{\theta}^{m-1}\right] = \sum\nolimits_{k=1}^{9} P\left(L=k|\boldsymbol{O},\boldsymbol{\theta}\right) l(\boldsymbol{\theta}|\boldsymbol{O},\boldsymbol{L}) \\
&= \sum\nolimits_{k=1}^{9} P\left(L=k|\boldsymbol{O},\boldsymbol{\theta}\right) \left\{ log\pi_{L_1} + \sum_{s=1}^{S-1} \log P\left(L_{s+1}|L_s,\, \boldsymbol{\theta}_{trans}\right) + \sum_{s=1}^{S} \log P\left(\boldsymbol{O}_s|L_s,\boldsymbol{\theta}_{emiss}\right) \right\} \\
&= \sum\nolimits_{k=1}^{9} P_k(1) \log \pi_k \qquad \text{(A.2)} \\
&+ \sum_{s=1}^{S-1} \sum\nolimits_{k=1}^{9} \sum\nolimits_{l=1}^{9} P_{kl}(s) \log\left(t_{kl}\left(s|\boldsymbol{\theta}_{trans}\right)\right) \qquad \text{(A.3)} \\
&+ \sum_{s=1}^{S} \sum\nolimits_{k=1}^{9} P_k(s) \log P\left(\boldsymbol{O}_s|L_s,\boldsymbol{\theta}_{emiss}\right) \qquad \text{(A.4)}
\end{aligned}$$

where

$$P_k(s) = P\left(L_s=k|\boldsymbol{O},\boldsymbol{\theta}\right) = \frac{P\left(L_s=k,\boldsymbol{O}|\boldsymbol{\theta}\right)}{P(\boldsymbol{O}|\boldsymbol{\theta})} \tag{A.5}$$

$$P_{kl}(s) = P\left(L_s=k, L_{s+1}=l|\boldsymbol{O},\boldsymbol{\theta}\right) = \frac{P\left(L_S=k, L_{s+1}=l,\boldsymbol{O}|\boldsymbol{\theta}\right)}{P(\boldsymbol{O}|\boldsymbol{\theta})} \tag{A.6}$$

Note that denominators of $P_k(s)$ and $P_{kl}(s)$ are their numerators summed over all the possible states, i.e., $P(\boldsymbol{O}|\boldsymbol{\theta}) = \sum_k P\left(L_s = k, \boldsymbol{O}|\boldsymbol{\theta}\right)$ in (A.5) and $P(\boldsymbol{O}|\boldsymbol{\theta}) = \sum_k \sum_l P\left(L_S = k, L_{s+1} = l, \boldsymbol{O}|\boldsymbol{\theta}\right)$ in (A.6). To calculate $P\left(L_s = k, \boldsymbol{O}|\boldsymbol{\theta}\right)$ and $P\left(L_S = k, L_{s+1} = l, \boldsymbol{O}|\boldsymbol{\theta}\right)$, the forward- backward algorithm were adopted as follow.

$$
\begin{aligned}
&P\left(L_s = k, \boldsymbol{O}|\boldsymbol{\theta}\right) = P\left(O_1, \ldots, O_s, L_s = k|\boldsymbol{\theta}\right) P\left(O_{s+1}, \ldots, O_S|L_s = k, \boldsymbol{\theta}\right) \\
&= \alpha_k(s)\beta_k(s) \qquad \text{(A.7)} \\
&P\left(L_s = k, L_{s+1} = l, \boldsymbol{O}|\boldsymbol{\theta}\right) = P\left(\boldsymbol{O}, L_s = k|\boldsymbol{\theta}\right) P\left(O_{s+1}, L_{s+1} = l|L_s = k, O_1, \ldots, O_s, \boldsymbol{\theta}\right) \\
&\times P\left(O_{s+2}, \ldots, O_S|L_s = k, L_{s+1} = l, O_1, \ldots, O_{s+1}, \boldsymbol{\theta}\right) \\
&= \alpha_k(s) a_{kl}(s) P\left(O_{s+1}|L_{s+1} = l, \boldsymbol{\theta}\right) \beta_l(s+1) \qquad \text{(A.8)}
\end{aligned}
$$

where $\alpha_k(s)$ and $\beta_k(s)$ are forward probability and backward probability defined as

$$
\alpha_k(s) = P\left(O_1, \ldots, O_s, L_s = k|\boldsymbol{\theta}\right) \qquad \text{(A.9)}
$$

$$
\beta_k(s) = P\left(O_{s+1}, \ldots, O_S|S_s = k, O_1, \ldots, O_s, \boldsymbol{\theta}\right) \qquad \text{(A.10)}
$$

The forward and backward probability can be computed recursively

$$
\begin{aligned}
&\alpha_k(1) = \pi_k P\left(O_1|L_1 = k, \boldsymbol{\theta}\right) \qquad \text{(A.11)} \\
&\alpha_k(s) = \sum_l \alpha_l(s-1) t_{lk}(s-1) P\left(O_s|L_s = k, \boldsymbol{\theta}\right) \text{ for } s = 2, \ldots, S \qquad \text{(A.12)} \\
&\beta_k(S) = 1 \qquad \text{(A.13)} \\
&\beta_k(s) = \sum_l t_{kl}(s) P\left(O_{s+1}|L_{s+1} = l, \boldsymbol{\theta}\right) \beta_l(s+1) \text{ for } s = S-1, \ldots, 1 \qquad \text{(A.14)}
\end{aligned}
$$

The M-step

In the M-step, we seek for the values of parameters to maximize the expectation $\mathrm{E}_{L|O,\theta^{m-1}}\left[l(\boldsymbol{\theta}|O,L)|O,\boldsymbol{\theta}\right]$, and save the updated parameters $\boldsymbol{\theta}^{(m)}$ for the next iteration.

Define $G_1(\pi_k) = \sum\limits_k P\left(\boldsymbol{L}=\boldsymbol{k}|\boldsymbol{O},\boldsymbol{\theta}\right)\log\pi_{L_1}$, $G_2(\boldsymbol{p},\alpha) = \sum_{s=1}^{S-1}\sum\limits_k\sum\limits_l P_{kl}(s)\log\left(t_{kl}\left(s|\boldsymbol{p},\alpha\right)\right)$, $G_3(\boldsymbol{\theta_{emiss}}) = \sum_{s=1}^{S}\sum\limits_k P_k(s)\log P\left(O_s|L_s,\boldsymbol{\theta}_{emiss}\right)$, then $\mathrm{E}_{L|O,\theta}\left[l(\boldsymbol{\theta}|O,L)|O,\boldsymbol{\theta}\right] \stackrel{\triangle}{=} G_1(\pi_k)+ G_2(\boldsymbol{p},\alpha) + G_3(\boldsymbol{\theta}_{emiss})$. By taking partial derivative of the expectation against each parameter and equal them to 0, we can update the parameters in the order of $\boldsymbol{p}$, α, o^+ (o^-), d^+ (d^-), σ_+^2 $(\sigma_-^2,\ \sigma^2)$, τ_+^2 $(\tau_-^2,\ \tau^2)$, $\omega_1,...,\omega_9$, π_k. When updating each parameter, we use the most up to date values for other parameters which would be needed in the computation of the partial derivative.

Updating transition parameters

At first, we update $\boldsymbol{p}$ by solving the equation $\frac{\partial}{p_{kl}}\left(G_2(\boldsymbol{p},\alpha)\right) = 0\ l \neq k$, for $l,\ k \in \mathcal{M}$. This yields $\sum_{s=1}^{S-1} P_{kk}(s)\frac{-(1-e^{-\alpha c_s})}{1-(1-e^{-\alpha c_s})\sum_{l\neq k}p_{kl}} + \sum_{s=1}^{S-1}\frac{P_{kl}(s)(1-e^{-\alpha c_s})}{p_{kl}(1-e^{-\alpha c_s})} = 0$ for $k,\ l \in \mathcal{M}$ and $l \neq k$. Specifically, for each k, we have $\sum_{s=1}^{S-1}\frac{P_{kl}(s)}{p_{kl}} = \sum_{s=1}^{S-1}\frac{P_{kk}(s)(1-e^{-\alpha c_s})}{1-(1-e^{-\alpha c_s})\sum_{l\neq k}p_{kl}}$ for every $l \in \mathcal{M}$. Let $h_k = \sum_{s=1}^{S-1}\frac{P_{kl}(s)}{p_{kl}}$, then find the h_k by maximizing

$$\sum_{s=1}^{S-1} P_{kk}(s) log\left(1-\left(\frac{\sum_{l\neq k}\sum_{s=1}^{S-1}P_{kl}(s)}{h_k}\right)(1-e^{-\alpha c_s})\right) + \sum_{s=1}^{S-1}\sum_{l\neq k} P_{kl}(s) log\left(\frac{\sum_{s=1}^{S-1}P_{kl}(s)}{h_k}(1-e^{-\alpha c_s})\right)$$

for each k. Then $p_{kl}^m = \frac{\sum_{s=1}^{S-1}P_{kl}(s)}{h_k}\ l \neq k,\ l,\ k \in \mathcal{M}$. To get the complete transition matrix, we still need to update parameter α, which can be achieved by a direct grid search over the range [0.01, 10].

Updating emission parameters

Let $k = 1, ..., 9$ correspond to the nine states in the bivariate state space $\mathcal{M} = (M,1),(M,2),(M,0),(P,1),(P,2),(P,0),(N,1),(N,2),(N,0)$. $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial o^+} = 0$ yields

$$o^+ \left\{ \frac{\sum_{s=1}^S P_1(s)}{(1-\omega_1^2)\sigma_+^2} + \frac{\sum_{s=1}^S P_2(s)}{(1-\omega_2^2)\sigma_+^2} + \frac{\sum_{s=1}^S P_3(s)}{(1-\omega_3^2)\sigma_+^2} \right\} = \frac{\sum_{s=1}^S P_1(s)Y_{1s}}{(1-\omega_1^2)\sigma_+^2} - \frac{\omega_1 \sum_{s=1}^S P_1(s)D_s}{(1-\omega_1^2)\sigma_+\tau} + \frac{\sum_{s=1}^S P_2(s)Y_{1s}}{(1-\omega_2^2)\sigma_+^2} - \frac{\omega_2 \sum_{s=1}^S P_2(s)(D_s - d^+)}{(1-\omega_2^2)\sigma_+\tau_+} + \frac{\sum_{s=1}^S P_3(s)Y_{1s}}{(1-\omega_3^2)\sigma_+^2} - \frac{\omega_3 \sum_{s=1}^S P_3(s)(D_s - d^-)}{(1-\omega_3^2)\sigma_+\tau_-} \tag{A.15}$$

Similarly, $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial o^-} = 0$ yields

$$o^- \left\{ \frac{\sum_{s=1}^S P_4(s)}{(1-\omega_4^2)\sigma_-^2} + \frac{\sum_{s=1}^S P_5(s)}{(1-\omega_5^2)\sigma_-^2} + \frac{\sum_{s=1}^S P_6(s)}{(1-\omega_6^2)\sigma_-^2} \right\} = \frac{\sum_{s=1}^S P_4(s)Y_{1s}}{(1-\omega_1^2)\sigma_-^2} - \frac{\omega_4 \sum_{s=1}^S P_4(s)D_s}{(1-\omega_1^2)\sigma_-\tau} + \frac{\sum_{s=1}^S P_5(s)Y_{1s}}{(1-\omega_5^2)\sigma_-^2} - \frac{\omega_5 \sum_{s=1}^S P_5(s)(D_s - d^+)}{(1-\omega_5^2)\sigma_-\tau_+} + \frac{\sum_{s=1}^S P_6(s)Y_{1s}}{(1-\omega_6^2)\sigma_-^2} - \frac{\omega_6 \sum_{s=1}^S P_6(s)(D_s - d^-)}{(1-\omega_6^2)\sigma_-\tau_-} \tag{A.16}$$

Denote the roots of equation (A.15) and (A.16) by $\hat{o}^+$ and $\hat{o}^-$ respectively, then $o^{+(m)} = -o^{-(m)} = \frac{1}{2}(\hat{o}^+ + \hat{o}^-)$.

$\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial d^+} = 0$ and $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial d^-} = 0$ yield

$$d^+ \left\{ \frac{\sum_{s=1}^S P_2(s)}{(1-\omega_2^2)\tau_+^2} + \frac{\sum_{s=1}^S P_5(s)}{(1-\omega_5^2)\tau_+^2} + \frac{\sum_{s=1}^S P_8(s)}{(1-\omega_8^2)\tau_+^2} \right\} = \frac{\sum_{s=1}^S P_2(s)D_s}{(1-\omega_2^2)\tau_+^2} - \frac{\omega_2 \sum_{s=1}^S P_2(s)(Y_{1s} - o^+)}{(1-\omega_2^2)\sigma_+\tau_+} + \frac{\sum_{s=1}^S P_5(s)D_s}{(1-\omega_5^2)\tau_+^2} - \frac{\omega_5 \sum_{s=1}^S P_5(s)(Y_{1s} - o^-)}{(1-\omega_5^2)\sigma_-\tau_+} + \frac{\sum_{s=1}^S P_8(s)D_s}{(1-\omega_8^2)\tau_+^2} - \frac{\omega_8 \sum_{s=1}^S P_8(s)Y_{1s}}{(1-\omega_8^2)\sigma\tau_+} \tag{A.17}$$

and

$$d^{-}\left\{\frac{\sum_{s=1}^{S}P_3(s)}{(1-\omega_3^2)\tau_-^2}+\frac{\sum_{s=1}^{S}P_6(s)}{(1-\omega_6^2)\tau_-^2}+\frac{\sum_{s=1}^{S}P_9(s)}{(1-\omega_9^2)\tau_-^2}\right\}=\frac{\sum_{s=1}^{S}P_3(s)D_s}{(1-\omega_3^2)\tau_-^2}-$$
$$\frac{\omega_3\sum_{s=1}^{S}P_3(s)(Y_{1s}-o^{+})}{(1-\omega_3^2)\sigma_+\tau_-}+\frac{\sum_{s=1}^{S}P_6(s)D_s}{(1-\omega_6^2)\tau_-^2}-\frac{\omega_6\sum_{s=1}^{S}P_6(s)(Y_{1s}-o^{-})}{(1-\omega_6^2)\sigma_-\tau_-}+$$
$$\frac{\sum_{s=1}^{S}P_9(s)D_s}{(1-\omega_9^2)\tau_-^2}-\frac{\omega_9\sum_{s=1}^{S}P_9(s)Y_{1s}}{(1-\omega_9^2)\sigma\tau_-} \tag{A.18}$$

Let $\hat{d}^{+}$ and $\hat{d}^{-}$ denote the root of equation equation (A.17) and (A.18) respectively, then $d^{+(m)}=-d^{-(m)}=\frac{1}{2}(\hat{d}^{+}+\hat{d}^{-})$.

$\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\sigma_+^2}=0$ yields

$$\frac{1}{\sigma_+^2}\left\{\frac{\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^{+})^2}{(1-\omega_1^2)}+\frac{\sum_{s=1}^{S}P_2(s)(Y_{1s}-o^{+})^2}{(1-\omega_2^2)}+\frac{\sum_{s=1}^{S}P_3(s)(Y_{1s}-o^{+})^2}{(1-\omega_3^2)}\right\}-$$
$$\frac{1}{\sigma_+}\left\{\frac{\omega_1\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^{+})D_s}{(1-\omega_1^2)\tau}+\frac{\omega_2\sum_{s=1}^{S}P_2(s)(Y_{1s}-o^{+})(D_s-d^{+})}{(1-\omega_2^2)\tau_+}+\right.$$
$$\left.\frac{\omega_3\sum_{s=1}^{S}P_3(s)(Y_{1s}-o^{+})(D_s-d^{-})}{(1-\omega_3^2)\tau_-}\right\}=\sum_{s=1}^{S}\left(P_1(s)+P_2(s)+P_3(s)\right) \tag{A.19}$$

Similarly, $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\sigma_-^2}=0$ and $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\sigma^2}=0$ yields

$$\frac{1}{\sigma_-^2}\left\{\frac{\sum_{s=1}^{S}P_4(s)(Y_{1s}-o^{-})^2}{(1-\omega_4^2)}+\frac{\sum_{s=1}^{S}P_5(s)(Y_{1s}-o^{-})^2}{(1-\omega_5^2)}+\frac{\sum_{s=1}^{S}P_6(s)(Y_{1s}-o^{-})^2}{(1-\omega_6^2)}\right\}-$$
$$\frac{1}{\sigma_-}\left\{\frac{\omega_4\sum_{s=1}^{S}P_4(s)(Y_{1s}-o^{-})D_s}{(1-\omega_4^2)\tau}+\frac{\omega_5\sum_{s=1}^{S}P_5(s)(Y_{1s}-o^{-})(D_s-d^{+})}{(1-\omega_5^2)\tau_+}+\right.$$
$$\left.\frac{\omega_6\sum_{s=1}^{S}P_6(s)(Y_{1s}-o^{-})(D_s-d^{-})}{(1-\omega_6^2)\tau_-}\right\}=\sum_{s=1}^{S}\left(P_4(s)+P_5(s)+P_6(s)\right) \tag{A.20}$$

and

$$\frac{1}{\sigma^2}\left\{\frac{\sum_{s=1}^{S}P_7(s)Y_{1s}^2}{(1-\omega_7^2)}+\frac{\sum_{s=1}^{S}P_8(s)Y_{1s}^2}{(1-\omega_8^2)}+\frac{\sum_{s=1}^{S}P_9(s)Y_{1s}^2}{(1-\omega_9^2)}\right\}-$$

$$\frac{1}{\sigma}\left\{\frac{\omega_7\sum_{s=1}^{S}P_7(s)Y_{1s}D_s}{(1-\omega_7^2)\tau}+\frac{\omega_8\sum_{s=1}^{S}P_8(s)Y_{1s}(D_s-d^+)}{(1-\omega_8^2)\tau_+}+\frac{\omega_9\sum_{s=1}^{S}P_9(s)Y_{1s}(D_s-d^-)}{(1-\omega_9^2)\tau_-}\right\}$$
$$=\sum_{s=1}^{S}(P_7(s)+P_8(s)+P_9(s)) \tag{A.21}$$

Let $\hat{\sigma}_+^2$, $\hat{\sigma}_-^2$ and $\hat{\sigma}^2$ denote the root of equation (A.19), (A.20) and (A.21) respectively, then $\sigma_+^{2(m)}=\sigma_-^{2(m)}=\frac{1}{2}(\hat{\sigma}_+^2+\hat{\sigma}_-^2)$, and $\sigma^{2(m)}=\hat{\sigma}^2$.

τ_+^2, τ_-^2 and τ^2 can be updated in a analogous way as $\delta_+^{2(m)}=\delta_-^{2(m)}=\frac{1}{2}(\hat{\delta}_+^2+\hat{\delta}_-^2)$, and $\delta^{2(m)}=\hat{\delta}^2$ where $\hat{\tau}_+^2$, $\hat{\tau}_-^2$ and $\hat{\tau}^2$ are the root of equations (A.22), (A.23) and (A.24), which are yielded by $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\tau_+^2}=0$, $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\tau_-^2}=0$ and $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\tau^-}=0$ respectively.

$$\frac{1}{\tau_+^2}\left\{\frac{\sum_{s=1}^{S}P_2(s)(D_s-o^+)^2}{(1-\omega_2^2)}+\frac{\sum_{s=1}^{S}P_5(s)(D_s-o^+)^2}{(1-\omega_5^2)}+\frac{\sum_{s=1}^{S}P_8(s)(D_s-o^+)^2}{(1-\omega_8^2)}\right\}-$$
$$\frac{1}{\tau_+}\left\{\frac{\omega_2\sum_{s=1}^{S}P_2(s)(Y_{1s}-d^+)(D_s-o^+)}{(1-\omega_2^2)\sigma_+}+\frac{\omega_5\sum_{s=1}^{S}P_5(s)(Y_{1s}-d^-)(D_s-o^+)}{(1-\omega_5^2)\sigma_-}+\right.$$
$$\left.\frac{\omega_8\sum_{s=1}^{S}P_8(s)Y_{1s}(D_s-o^+)}{(1-\omega_8^2)\sigma}\right\}=\sum_{s=1}^{S}(P_2(s)+P_5(s)+P_8(s)) \tag{A.22}$$

$$\frac{1}{\tau_-^2}\left\{\frac{\sum_{s=1}^{S}P_3(s)(D_s-o^-)^2}{(1-\omega_3^2)}+\frac{\sum_{s=1}^{S}P_6(s)(D_s-o^-)^2}{(1-\omega_6^2)}+\frac{\sum_{s=1}^{S}P_9(s)(D_s-o^-)^2}{(1-\omega_9^2)}\right\}-$$
$$\frac{1}{\tau_-}\left\{\frac{\omega_3\sum_{s=1}^{S}P_3(s)(Y_{1s}-d^+)(D_s-o^-)}{(1-\omega_3^2)\sigma_+}+\frac{\omega_6\sum_{s=1}^{S}P_6(s)(Y_{1s}-d^-)(D_s-o^-)}{(1-\omega_6^2)\sigma_-}+\right.$$
$$\left.\frac{\omega_9\sum_{s=1}^{S}P_9(s)Y_{1s}(D_s-o^-)}{(1-\omega_9^2)\sigma}\right\}=\sum_{s=1}^{S}(P_3(s)+P_6(s)+P_9(s)) \tag{A.23}$$

$$\frac{1}{\tau^-}\left\{\frac{\sum_{s=1}^{S}P_1(s)D_s^2}{(1-\omega_1^2)}+\frac{\sum_{s=1}^{S}P_4(s)D_s^2}{(1-\omega_4^2)}+\frac{\sum_{s=1}^{S}P_7(s)D_s^2}{(1-\omega_7^2)}\right\}-$$

$$\frac{1}{\tau}\left\{\frac{\omega_1\sum_{s=1}^{S}P_1(s)(Y_{1s}-d^+)D_s}{(1-\omega_1^2)\sigma_+}+\frac{\omega_4\sum_{s=1}^{S}P_4(s)(Y_{1s}-d^-)D_s}{(1-\omega_4^2)\sigma_-}+\frac{\omega_7\sum_{s=1}^{S}P_7(s)Y_{1s}D_s}{(1-\omega_7^2)\sigma}\right\}$$
$$=\sum_{s=1}^{S}(P_1(s)+P_4(s)+P_7(s)) \tag{A.24}$$

To update ω_1, ... ω_9, we take partial derivative of $G_3(\boldsymbol{\theta}_{emiss})$ against these parameters as well. In this appendix, we will only show the procedure to update ω_1, ω_2, ... ω_9 can be updated in a symmetric manner. $\frac{\partial G_3(\boldsymbol{\theta}_{emiss})}{\partial\omega_1}=0$ yields

$$\omega_1(1-\omega_1^2)\sum_{s=1}^{S}P_1(s)-2\omega_1\left\{\frac{\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^+)^2}{\sigma_+^2}+\frac{\sum_{s=1}^{S}P_1(s)D_s^2}{\tau^2}\right\}+$$
$$(1+\omega_1^2)\left\{\frac{\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^+)D_s}{\sigma_+\tau}\right\}=0 \tag{A.25}$$

Let $a=\sum_{s=1}^{S}P_1(s)$, $=\frac{\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^+)^2}{\sigma_+^2}+\frac{\sum_{s=1}^{S}P_1(s)D_s^2}{\tau^2}$ and $c=\frac{\sum_{s=1}^{S}P_1(s)(Y_{1s}-o^+)D_s}{\sigma_+\tau}$. We search the root for nonlinear equation $a\omega_1^3-c\omega_1^2-(a-2b)\omega_1-c=0$ over the range [-1,1], which is the updated $\omega_1^{(m)}$. At the last step of the Mstep, we update $\pi_k=P(S_1=k)$ by $\pi_k^{(m)}=P(S_1=k|\boldsymbol{O},\boldsymbol{\theta}^{(m-1)})$.

A.2 Estimated transition probabilities for real data analysis

k l	1	2	3	4	5	6	7	8	9
1		0.0022	0.0139	0.0000	0.1834	0.0000	0.3673	0.0000	0.0024
2	0.0919		0.4204	0.0000	0.0000	0.0000	0.0001	0.0022	0.0006
3	0.0035	0.0000		0.0000	0.0208	0.0000	0.5988	0.0000	0.0013
4	0.0000	0.0000	0.1993		0.2136	0.2307	0.1805	0.0000	0.0232
5	0.0147	0.0059	0.1101	0.0106		0.0000	0.4439	0.0000	0.2992
6	0.0000	0.0000	0.0000	0.1784	0.0110		0.5117	0.0000	0.0000
7	0.0003	0.0000	0.0607	0.0020	0.0737	0.0000		0.0000	0.0643
8	0.0006	0.0000	0.2941	0.0000	0.0000	0.0000	0.0148		0.0123
9	0.0000	0.0002	0.0367	0.0407	0.0812	0.0040	0.2672	0.0000	

Table A.1: Estimation of p_{kl} for chromosome 23 (rounded to 4 decimal places)

k l	1	2	3	4	5	6	7	8	9
1		0.0000	0.0570	0.0346	0.0000	0.0000	0.4137	0.1438	0.0000
2	0.0004		0.0709	0.2084	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0581	0.0279		0.0010	0.0580	0.0000	0.5397	0.0000	0.0791
4	0.0067	0.0000	0.0000		0.0003	0.4877	0.0000	0.0000	0.0887
5	0.0000	0.0000	0.0119	0.0000		0.0000	0.6472	0.0000	0.1620
6	0.0331	0.0000	0.0000	0.0598	0.1192		0.0000	0.0000	0.0518
7	0.0037	0.0000	0.0476	0.0000	0.0436	0.0000		0.0000	0.0344
8	0.2219	0.0000	0.4842	0.0000	0.0000	0.0000	0.0946		0.0000
9	0.0179	0.0000	0.0000	0.0000	0.0002	0.0000	0.7614	0.0000	

Table A.2: Estimation of p_{kl} for chromosome 21 (rounded to 4 decimal places)

A.3 Model validation for Normal emission probability

Our model assume the marginal distribution of observations given hidden states follow bivariate normal distributions. More specifically, the marginal distribution of the observation at a specific dimension, e.g., D_s and Y_{1s} will follow a mixture of three normal distributions. Figure A.1 displays the model fitting for an example chromosome, i.e., chromosome 23. The left panel shows the histogram of Y_{1s}, the observations at first dimension, which is the logistic transformation of the proportion of gene expression from maternal allele in control sample. The right panel shows the histogram of D_s, observations at second dimension, which is the difference between the logistic transformed proportion of gene expression from maternal allele in control and LOS group. The superimposed blue curves represents the fitted distribution, and the blue, green and red curves represents the three mixture components of the estimated distribution. It is indicated in Figure A.1 that both the distribution of Y_{1s} and D_s are well fitted by the Gaussian mixture distribution.

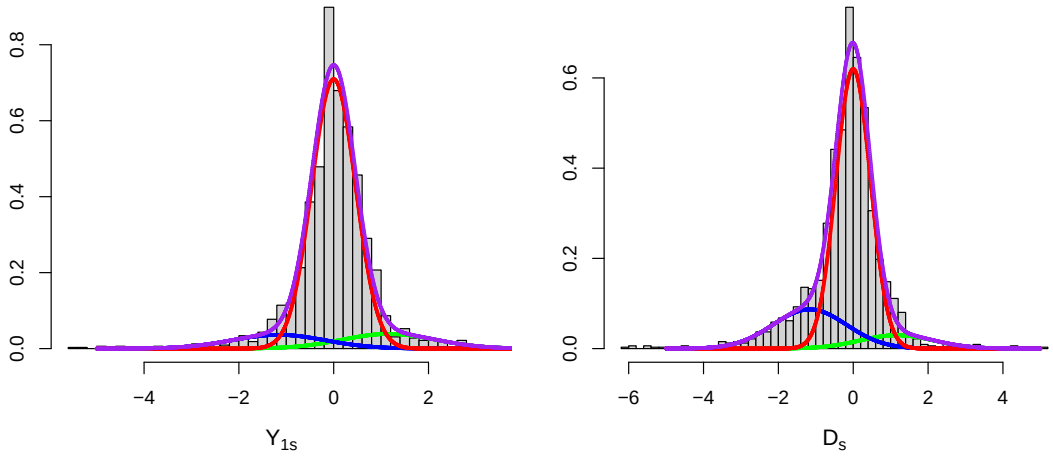


Figure A.1: Model fitting for chromosome 23. The left panel shows the histogram of the logistic transformed proportion of maternal expression in control group. The right panel shows the histogram of the difference between the transformed maternal proportion in control and LOS groups. The superimposed purple curve corresponds to the fitted mixture distribution of the data, with the blue, green and red curves representing the three normal components.

Bibliography

[1] Smallwood A, A Papageorghiou, K Nicolaides, MK Alley, A Jim, G Nargund, K Ojha, S Campbell, and S Banerjee. Temporal regulation of the expression of syncytin (herv-w), maternally imprinted peg10, and sgce in human placenta. *Biology of Reproduction*, 69:286–293, 2003.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

[3] H. Angelman. "puppet" children: A report of three cases. *Journal of Developmental Medicine and Child Neurology*, 7:681–688, 1965.

[4] P L Auer and R W Doerge. A two-stage poisson model for testing rna-seq data. *Statistical Applications in Genetics and Molecular Biology*, 10:1–28, 2011.

[5] T Babak, B DeVeale, EK Tsang, Y Zhou, X Li, KS Smith, KR Kukurba, R Zhang, JB Li, D van der Kooy, SB Montgomery, and HB Fraser. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature Genetics*, 47:544–549, 2015.

[6] DP Barlow, R Stöger, BG Herrmann, K Saito, and N Schweifer. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the tme locus. *Nature*, 349:84–87, 1991.

[7] M. S. Bartolomei. Genomic imprinting: employing and avoiding epigenetic processes. *Genes & Development*, 23:2124–2133, 2009.

[8] Christopher G Bell and Stephan Beck. Advances in the identification and analysis of allele-specific expression. *Genome medicine*, 1(5):56, 2009.

[9] E Billy, T Wegierski, F Nasr, and W Filipowicz. Rcl1p, the yeast protein similar to the rna 3'-phosphate cyclase, associates with u3 snornp and is required for 18s rrna biogenesis. *The EMBO Journal*, 19:2115–2126, 2000.

[10] N Blagitko, S Mergenthaler, U Schulz, H A Wollmann, W Craigen, T Eggermann, H-H Ropers, and V M Kalscheuer. Human grb10 is imprinted an expressed from the paternal and maternal allele in a highly tissue- and isoform-specific fashion. *Human Molecular Genetics*, 9:1587–1595, 2000.

[11] L. Carrel and H. F. Willard. X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature*, 434:400–404, 2005.

[12] Z Chen, DE Hagen, CG Elsik, T Ji, Morris CJ, LE Moon, and RM Rivera. Characterization of global loss of imprinting in fetal overgrowth syndrome induced by assisted reproduction. *Proceedings of the National Academy of Sciences*, 112(15):4618–4623, 2015.

[13] Z Chen, DE Hagen, J Wang, CG Elsik, T Ji, LG Siqueira, PJ Hansen, and RM Rivera. Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing. *Epigenetics*, 11:501–516, 2016.

[14] Z. Chen, K. M. Robbins, K. D. Wells, and R. M. Rivera. Large offspring syndrome: A bovine model for the human loss-of-imprinting overgrowth syndrome beckwith-wiedemann. *Epigenetics*, 8(6):591–601, 2013.

[15] Zhiyuan Chen, Darren E Hagen, Tieming Ji, Christine G Elsik, and Rocío M Rivera. Global misregulation of genes largely uncoupled to dna methylome epimutations characterizes a congenital overgrowth syndrome. *Scientific reports*, 7(1):1–14, 2017.

[16] Nicole Cloonan, Alistair RR Forrest, Gabriel Kolle, Brooke BA Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, et al. Stem cell transcriptome profiling via massive-scale mrna sequencing. *Nature methods*, 5(7):613–619, 2008.

[17] MM Jr. Cohen. Beckwith-wiedemann syndrome: historical, clinicopathological, and etiopathogenetic perspectives. *Pediatric and Developmental Pathology*, 8:287–304, 2005.

[18] S Croteau, M-C Charron, K E Latham, and A K Naumova. Alternative splicing and imprinting control of the meg3/gtl2-dlk1 locus in mouse embryos. *Mammalian Genome*, 14:231–241, 2003.

[19] S. Cui, S. Guha, M. A. R. Ferreira, and A. N. Tegge. hmmseq: a hidden markov model for detecting differentially expressed genes from rna-seq data. *The Annals of Applied Statistics*, 9:901–925, 2015.

[20] Maria Cristina Curia, Sabrina De Iure, Laura De Lellis, Serena Veschi, Sandra Mammarella, Marquitta J White, Jacquelaine Bartlett, Angelo Di Iorio, Cristina Amatetti, Marco Lombardo, et al. Increased variance in germline allele-specific expression of apc associates with colorectal cancer. *Gastroenterology*, 142(1):71–77, 2012.

[21] JF Degner, JC Marioni, AA Pai, JK Pickrell, E Nkadori, Y Gilad, and

JK Pritchard. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics*, 25:3207–3212, 2009.

[22] A. Denley, L. J. Cosgrove, G. W. Booker, J. C. Wallace, and B. E. Forbes. Molecular interactions of the igf system. *Cytokine & Growth Factor Reviews*, 16:421–439, 2005.

[23] Scott J Emrich, W Brad Barbazuk, Li Li, and Patrick S Schnable. Gene discovery and annotation using lcm-454 transcriptome sequencing. *Genome research*, 17(1):69–73, 2007.

[24] Jiaxin Fan, Jian Hu, Chenyi Xue, Hanrui Zhang, Katalin Susztak, Muredach P Reilly, Rui Xiao, and Mingyao Li. Asep: Gene-based detection of allele-specific expression across individuals in a population by rna sequencing. *PLoS Genetics*, 16(5):e1008786, 2020.

[25] V. Foulon, M. Sniekers, E. Huysmans, S. Asselberghs, V. Mahieu, G. P. Mannaerts, P. P. Van Veldhoven, and M. Casteels. Breakdown of 2-hydroxylated straight chain fatty acids via peroxisomal 2-hydroxyphytanoyl-coa lyase: a revised pathway for the alpha-oxidation of straight chain fatty acids. *The Journal of Biological Chemistry*, 280:9802–9812, 2005.

[26] C. Fu, L. Di, X. Han, C. Soderstrom, M. Snyder, M. D. Troutman, R. S. Obach, and H. Zhang. Aldehyde oxidase 1 (aox1) in human liver cytosols: quantitative characterization of aox1 expression level and activity relationship. *Drug Metabolism and Disposition*, 41:1797–1804, 2013.

[27] B. Ge, D. K. Pokholok, T. Kwan, E. Grundberg, L. Morcos, D. J. Verlaan, J. Le, V. Koka, K. C. Lam, V. Gagné, J. Dias, R. Hoberman, A. Montpetit, M. M.

Joly, E. J. Harvey, D. Sinnett, P. Beaulieu, R. Hamon, A. Graziani, K. Dewar, E. Harmsen, J. Majewski, H. H. Göring, A. K. Naumova, M. Blanchette, K. L. Gunderson, and T. Pastinen. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature Genetics*, 41:1216–1222, 2009.

[28] F Ghiringhelli, M Bruchard, F Chalmin, and C Rébé. Production of adenosine by ectonucleotidases: A key factor in tumor immunoescape. *Journal of Biomedicine and Biotechnology*, 2012:473712, 2012.

[29] Alexander Gimelbrant, John N Hutchinson, Benjamin R Thompson, and Andrew Chess. Widespread monoallelic expression on human autosomes. *Science*, 318(5853):1136–1140, 2007.

[30] B. R. Graveley. The haplo-spliceo-transcriptome: common variations in alternative splicing in the human population. *Trends in Genetics*, 24:5–7, 2008.

[31] RM Graze, LL Novelo, V Amin, JM Fear, G Casella, SV Nuzhdin, and LM McIntyre. Allelic imbalance in drosophila hybrid heads: exons, isoforms, and evolution. *Molecular biology and evolution*, 29(6):1521–1532, 2012.

[32] Subharup Guha, Yi Li, and Donna Neuberg. Bayesian hidden markov modeling of array cgh data. *Journal of the American Statistical Association*, 103(482):485–497, 2008.

[33] H Hamada, H Okae, H Toh, H Chiba, H Hiura, K Shirane, T Sato, M Suyama, N Yaegashi, H Sasaki, and T Arima. Allele-specific methylome and transcriptome analysis reveals widespread imprinting in the human placenta. *The American Journal of Human Genetics*, 99:1045–1058, 2016.

[34] Thomas J Hardcastle and Krystyna A Kelly. bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):1–14, 2010.

[35] CT Harvey, GA Moyerbrailean, GO Davis, X Wen, F Luca, and R Pique-Regi. Quasar: quantitative allele-specific analysis of reads. *Bioinformatics*, 31:1235–1242, 2015.

[36] Yehudit Hasin-Brumshtein, Farhad Hormozdiari, Lisa Martin, Atila Van Nas, Eleazar Eskin, Aldons J Lusis, and Thomas A Drake. Allele-specific expression and eqtl analysis in mouse adipose tissue. *BMC genomics*, 15(1):1–13, 2014.

[37] CL Hsu, CH Chou, SC Huang, CY Lin, MY Lin, CC Tung, CY Lin, IP Lai, YF Zou, NA Youngson, SP Lin, CH Yang, SK Chen, SS Gau, and HS Huang. Analysis of experience-regulated transcriptome and imprintome during critical periods of mouse visual system development reveals spatiotemporal dynamics. *Human Molecular Genetics*, 27:1039–1054, 2018.

[38] H S Huang, B J Yoon, S Brooks, R Bakal, J Berrios, R S Larsen, M L Wallace, J E Han, E H Chung, M J Zylka, and B D Philpot. Snx14 regulates neuronal excitability, promotes synaptic transmission, and is imprinted in the brain of mice. *PLoS One*, 9:e98383, 2014.

[39] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More is better: recent progress in multi-omics data integration methods. *Frontiers in genetics*, 8:84, 2017.

[40] S Ito, G Honda, Y Fujino, S Ogata, M Hirayama-Kurogi, and S Ohtsuki. Knockdown of orphan transporter slc22a18 impairs lipid metabolism and increases invasiveness of hepg2 cells. *Pharmaceutical Research*, 36:39, 2019.

[41] Elizabeth M Jennings, Jeffrey S Morris, Raymond J Carroll, Ganiraju C Manyam, and Veerabhadran Baladandayuthapani. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP Journal on Bioinformatics and Systems Biology*, 2013(1):13, 2013.

[42] T Ji, P Liu, and D Nettleton. Estimation and testing of gene expression heterosis. *Journal of Agricultural, Biological, and Environmental Statistics*, 19:319–337, 2014.

[43] Tieming Ji. A bayesian hidden markov model for detecting differentially methylated regions. *Biometrics*, 75(2):663–673.

[44] M Joncquel-Chevalier Curt, P M Voice, M Fontaine, A F Dessein, N Porchet, K Mention-Mulliez, D Dobbelaere, G Soto-Ares, D Cheillan, and J Vamecq. Creatine biosynthesis and transport in health and disease. *Biochimie*, 119:146–165, 2015.

[45] M Kamiya, H Judson, Y Okazaki, M Kusakabe, M Muramatsu, S Takada, N Takagi, T Arima, N Wake, K Kamimura, K Satomura, R Hermann, DT Bonthron, and Y Hayashizaki. The cell cycle control gene zac/plagl1 is imprinted – a strong candidate gene for transient neonatal diabetes. *Human Molecular Genetics*, 9:453–460, 2000.

[46] Sophien Kamoun, Peter Hraber, Bruno Sobral, Donald Nuss, and Francine Govers. Initial assessment of gene diversity for the oomycete pathogen phytophthora infestans based on expressed sequences. *Fungal Genetics and Biology*, 28(2):94–106, 1999.

[47] K Karbstein, S Jonas, and J A Doudna. An essential gtpase promotes assembly of preribosomal rna processing complexes. *Molecular Cell*, 20:633–643, 2005.

[48] A Kasprzak, W Kwasniewski, A Adamek, and A Gozdzicka-Jozefiak. Insulin-like growth factor (igf) axis in cancerogenesis. *Mutation Research / Reviews in Mutation Research*, 772:78–104, 2017.

[49] D Kim, B Langmead, and SL Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature Methods*, 12:357–360, 2015.

[50] DA Knowles, JR Davis, H Edgington, A Raj, MJ Favé, X Zhu, JB Potash, MM Weissman, J Shi, DF Levinson, P Awadalla, S Mostafavi, SB Montgomery, and A Battle. Allele-specific expression reveals interactions between genetic variation and environment. *Nature Methods*, 14:699–702, 2017.

[51] N Kumasaka, AJ Knights, and DJ Gaffney. Fine-mapping cellular qtls with rasqual and atac-seq. *Nature Genetics*, 48:206–213, 2016.

[52] Vanessa M Kvam, Peng Liu, and Yaqing Si. A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *American journal of botany*, 99(2):248–256, 2012.

[53] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):1–17, 2014.

[54] J. T. Lee and M. S. Bartolomei. X-inactivation, imprinting, and long noncoding rnas in health and disease. *Cell*, 152:1308–1323, 2013.

[55] Ning Leng, John A Dawson, James A Thomson, Victor Ruotti, Anna I Rissman, Bart MG Smits, Jill D Haag, Michael N Gould, Ron M Stewart, and Christina Kendziorski. Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

[56] L. León-Novelo, A. R. Gerken, R. M. Graze, L. M. McIntyre, and F. Marroni. Direct testing for allele-specific expression differences between conditions. *G3 (Bethesda, Md.)*, 8(2):447–460, 2018.

[57] Ben Li, Yunxiao Li, and Zhaohui S Qin. Improving hierarchical models using historical data with applications in high-throughput genomics data analysis. *Statistics in biosciences*, 9(1):73–90, 2017.

[58] Ben Li, Zhaonan Sun, Qing He, Yu Zhu, and Zhaohui S Qin. Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes. *Bioinformatics*, 32(5):682–689, 2016.

[59] H Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27:2987–2993, 2011.

[60] H Li and R Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25:1754–1760, 2009.

[61] M. Li-Calzi, C. Raviolo, E. Ghibaudi, L. De Gioia, M. Salmona, G. Cazzaniga, M. Kurosaki, M. Terao, and E. Garattini. Purification, cdna closing, and tissue distribution of bovine liver aldehyde oxidase. *Journal of Biological Chemistry*, 270:31037–31045, 1995.

[62] Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in arabidopsis. *Cell*, 133(3):523–536, 2008.

[63] H Shuen Lo, Zhining Wang, Ying Hu, Howard H Yang, Sheryl Gere, Kenneth H Buetow, and Maxwell P Lee. Allelic variation in gene expression is common in the human genome. *Genome research*, 13(8):1855–1862, 2003.

[64] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[65] Jun Lu, John K Tomfohr, and Thomas B Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC bioinformatics*, 6(1):1–14, 2005.

[66] R. Lu, R. M. Smith, M. Seweryn, D. Wang, K. Hartmann, A. Webb, W. Sadee, and G. A. Rempala. Analyzing allele specific rna expression using mixture models. *BMC Genomics*, 16:566, 2015.

[67] T. Ludwig, J. Eggenschwiler, P. Fisher, A. J. D'Ercole, M. L. Davenport, and A. Efstratiadis. Mouse mutants lacking the type 2 igf receptor (igf2r) are rescued from perinatal lethality in lgf2 and lgf1r null backgrounds. *Developmental Biology*, 177:517–535, 1996.

[68] Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature biotechnology*, 28(4):322–324, 2010.

[69] Aaron TL Lun, Yunshun Chen, and Gordon K Smyth. It's de-licious: a recipe for differential expression analyses of rna-seq experiments using quasi-likelihood methods in edger. In *Statistical genomics*, pages 391–416. Springer, 2016.

[70] O. Mayba, H. N. Gilbert, J. Liu, P. M. Haverty, S. Jhunjhunwala, Z. Jiang, C. Watanabe, and Z. Zhang. Mbased: allele-specific expression detection in cancer tissues and cell lines. *Genome Biology*, 15:405, 2014.

[71] J McGrath and D Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37:179–183, 1984.

[72] A McKenna, M Hanna, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, K Garimella, D Altshuler, S Gabriel, M Daly, and MA DePristo. The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research*, 20:1297–1303, 2010.

[73] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[74] Jamil Najafov and Ayaz Najafov. Geco: gene expression correlation analysis after genetic algorithm-driven deconvolution. *Bioinformatics*, 35(1):156–159, 2019.

[75] V Nembaware, KH Wolfe, F Bettoni, J Kelso, and C Seoighe. Allele-specific transcript isoforms in human. *FEBS Letters*, 577:233–238, 2004.

[76] I Okamoto, C Patrat, D Thépot, N Peynot, P Fauque, N Daniel, P Diabangouaya, JP Wolf, JP Renard, V Duranthon, and E Heard. Eutherian mammals use diverse strategies to initiate x-chromosome inactivation during development. *Nature*, 472:370–374, 2011.

[77] RV Pandey, SU Franssen, A Futschik, and C Schlötterer. Allelic imbalance

metre (allim), a new tool for measuring allele-specific gene expression with rna-seq data. *Molecular Ecology Resources*, 13:740–745, 2013.

[78] T. Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11:533–538, 2010.

[79] G Piras, A El Kharroubi, S Kozlov, D Escalante-Alcalde, L Hernandez, NG Copeland, DJ Gilbert, NA Jenkins, and CL Stewart. Zac1 (lot1), a potential tumor suppressor gene, and the gene for epsilon-sarcoglycan are maternally imprinted genes: identification by a subtractive screen of novel uniparental fibroblast lines. *Molecular and Cellular Biology*, 20:3308–3315, 2000.

[80] Xi Rao, Kriti S Thapa, Andy B Chen, Hai Lin, Hongyu Gao, Jill L Reiter, Katherine A Hargreaves, Joseph Ipe, Dongbing Lai, Xiaoling Xuei, et al. Allele-specific expression and high-throughput reporter assay reveal functional genetic variants associated with alcohol use disorders. *Molecular psychiatry*, pages 1–10, 2019.

[81] Priyadip Ray, Lingling Zheng, Joseph Lucas, and Lawrence Carin. Bayesian joint analysis of heterogeneous genomics data. *Bioinformatics*, 30(10):1370–1376, 2014.

[82] ME Ritchie, B Phipson, D Wu, Y Hu, CW Law, W Shi, and GK Smyth. *limma* powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43:e47, 2015.

[83] KM Robbins, Z Chen, KD Wells, and RM Rivera. Expression of kcnq1ot1, cdkn1c, h19, and plagl1 and the methylation patterns at the kvdmr1 and h19/igf2 imprinting control regions is conserved between human and bovine. *Journal of Biomedical Science*, 19:95, 2012.

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

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

[86] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

[87] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

[88] MD Robinson and GK Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881–2887, 2007.

[89] J Rozowsky, A Abyzov, J Wang, P Alves, D Raha, A Harmanci, J Leng, R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj, M Rubin, M Snyder, and M Gerstein. Alleleseq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7:522, 2011.

[90] NT Ruddock, KJ Wilson, MA Cooney, NA Korfiatis, RT Tecirlioglu, and AJ French. Analysis of imprinted messenger rna expression during bovine preimplantation development. *Biology of Reproduction*, 70:1131–1135, 2004.

[91] LL Sandell, XJ Guan, R Ingram, and SM Tilghman. Gatm, a creatine synthesis

enzyme, is imprinted in mouse placenta. *Proceedings of the National Academy of Sciences*, 100:4622–4627, 2003.

[92] R V Satya, N Zavaljevski, and J Reifman. A new strategy to reduce allelic bias in rna-seq readmapping. *Nucleic Acids Research*, 40:e127, 2012.

[93] C Schwienbacher, L Gramantieri, R Scelfo, A Veronese, G A Calin, L Bolondi, C M Croce, G Barbanti-Brodano, and M Negrini. Gain of imprinting at chromosome 11p15: A pathogenetic mechanism identified in human hepatocarcinomas. *Proceedings of the National Academy of Sciences*, 97:5445–5449, 2000.

[94] Y Sekine, R Hatanaka, T Watanabe, N Sono, S Immure, T Natsume, E Kuranaga, M Miura, K Takeda, and H Ichijo. The kelch repeat protein klhdc10 regulates oxidative stress-induced ask1 activation by suppressing pp5. *Molecular Cell*, 48:692–704, 2012.

[95] Lin Shao, Feng Xing, Conghao Xu, Qinghua Zhang, Jian Che, Xianmeng Wang, Jiaming Song, Xianghua Li, Jinghua Xiao, Ling-Ling Chen, et al. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proceedings of the National Academy of Sciences*, 116(12):5653–5658, 2019.

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

[97] DA Skelly, M Johansson, J Madeoy, J Wakefield, and JM Akey. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome Research*, 21:1728–1737, 2011.

[98] MA Smit, X Tordoir, G Gyapay, NE Cockett, M Georges, and C Charlier. Begain: a novel imprinted gene that generates paternally expressed transcripts in a tissue- and promotor-specific manner in sheep. *Mammalian Genome*, 16:801–814, 2005.

[99] M Soga, A Matsuzawa, and H Ichjio. Oxidative stress-induced diseases via the ask1 signaling pathway. *International Journal of Cell Biology*, 2012:439587, 2012.

[100] D Solter. Differential imprinting and expression of maternal and paternal genomes. *Annual Review of Genetics*, 22:127–146, 1988.

[101] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC bioinformatics*, 14(1):1–18, 2013.

[102] Nora K Speicher and Nico Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, 2015.

[103] Liam Spurr, Muzi Li, Nawaf Alomran, Qianqian Zhang, Paula Restrepo, Mercedeh Movassagh, Chris Trenkov, Nerissa Tunnessen, Tatiyana Apanasovich, Keith A Crandall, et al. Systematic pan-cancer analysis of somatic allele frequency. *Scientific reports*, 8(1):1–12, 2018.

[104] Y Stelzer, S Bar, O Bartok, S Afik, D Ronen, S Kadener, and N Benvenisty. Differentiation of human parthenogenetic pluripotent stem cells reveals multiple tissue- and isoform-specific imprinted transcripts. *Cell Reports*, 11:308–320, 2015.

[105] Y Stelzer, D Ronen, C Bock, P Boyle, A Meissner, and N Benvenisty. Identification of novel imprinted differentially methylated regions by global analysis of human-parthenogenetic-induced pluripotent stem cells. *Stem Cell Reports*, 1:79–89, 2013.

[106] KR Stevenson, JD Coolon, and PJ Wittkopp. Sources of bias in measures of allele-specific expression derived from rna-seq data aligned to a single reference genome. *BMC Genomics*, 14:536, 2013.

[107] A Stockler-Ipsiroglu, D Apatean, R Battini, S DeBrosse, K Dessoffy, S Edvardson, F Eichler, K Johnston, D M Keller, S Nouioua, M Tapir, A Verma, M D Dowling, K J Wierenga, A M Wierenga, V Zhang, and L J Wong. Arginine: glycine amidinotransferase (agat) deficiency: Clinical features and long term outcomes in 16 patients diagnosed worldwide. *Molecular Genetics and Metabolism*, 116:252–259, 2015.

[108] JD Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B*, 64:479–498, 2002.

[109] Joshua M Stuart, Eran Segal, Daphne Koller, and Stuart K Kim. A gene-coexpression network for global discovery of conserved genetic modules. *science*, 302(5643):249–255, 2003.

[110] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[111] J Sun, W Li, Y Sun, D Yu, X Wen, H Wang, J Cui, G Wang, AR Hoffman, and JF Hu. A novel antisense long noncoding rna within the igf1r gene locus

is imprinted in hematopoietic malignancies. *Nucleic Acids Research*, 42:9588–9601, 2014.

[112] MA Surani, SC Barton, and ML Norris. Nuclear transplantation in the mouse: heritable differences between parental genomes after activation of the embryonic genome. *Cell*, 45:127–136, 1986.

[113] R D Teasdale and B M Collins. Insights into the px (pho-homology) domain and snx (sorting nexin) protein families: structures, functions and roles in disease. *Biochemical Journal*, 441:39–59, 2012.

[114] S Tierling, G Gasparoni, N Youngson, and M Paulsen. The begain gene marks the centromeric boundary of the imprinted region on mouse chromosome 12. *Mammalian Genome*, 20:699–710, 2009.

[115] D Trapnell, BA Williams, G Pertea, A Mortazavi, G Kwan, MJ van Baren, SL Salzburg, BJ Wold, and L Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28:511–515, 2010.

[116] B van de Geijn, G McVicker, Y Gilad, and JK Pritchard. Wasp: allele-specific software for robust molecular quantitative trait locus disocvery. *Nature Methods*, 12:1061–1063, 2015.

[117] Daniel Onofre Vidal, Jorge Estefano S de Souza, Lilian Campos Pires, Cibele Masotti, Anna Christina Matos Salim, Maria Cristina Ferreira Costa, Pedro Alexandre Favoretto Galante, Sandro Jose de Souza, and Anamaria Aranha Camargo. Analysis of allelic differential expression in the human genome using allele-specific serial analysis of gene expression tags. *Genome*, 54(2):120–127, 2011.

[118] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[119] ET Wang, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–476, 2008.

[120] Lin Wang, Pinghua Li, and Thomas P Brutnell. Exploring plant transcriptomes using ultra high-throughput sequencing. *Briefings in functional genomics*, 9(2):118–128, 2010.

[121] Wenting Wang, Veerabhadran Baladandayuthapani, Jeffrey S Morris, Bradley M Broom, Ganiraju Manyam, and Kim-Anh Do. ibag: integrative bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, 29(2):149–159, 2013.

[122] Y Wang, Z Cheng, H Z Elalieh, E Nakamura, M T Nguyen, S Mackem, T L Clemens, D D Bikle, and W Chang. Igf-1r signaling in chondrocytes modulates growth plate development by interacting with the pthrp/ihh pathway. *Journal of Bone and Mineral Research*, 26:1437–1446, 2011.

[123] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[124] R Weksberg, C Shuman, and AC Smith. Beckwith-wiedemann syndrome. *American Journal of Medical Genetics Part C*, 137:12–23, 2005.

[125] JF Wilkins, F Úbeda, and J Van Cleve. The evolving landscape of imprinted

genes in humans and mice: Conflict among alleles, genes, tissues, and kin. *Bioessays*, 38:482–489, 2016.

[126] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–243, 2013.

[127] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, 14(2):232–243, 2013.

[128] A Wutz. Gene silencing in x-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews Genetics*, 12:542–553, 2011.

[129] Jing Xie, Tieming Ji, Marco AR Ferreira, Yahan Li, Bhaumik N Patel, and Rocio M Rivera. Modeling allele-specific expression at the gene and snp levels simultaneously by a bayesian logistic mixed regression model. *BMC bioinformatics*, 20(1):530, 2019.

[130] Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic acids research*, 40(19):9379–9391, 2012.

[131] X Zhi, S Cheng, P Zhou, Z Chao, L Wang, Z Ou, and L Yin. Rna interference of echo-5'-nucleotidase (cd73) inhibits human breast cancer cell growth and invasion. *Clinical & Experimental Metastasis*, 24:439–448, 2007.

VITA

Jing Xie was born in Xiantao city in the Hubei province of China. She received her Bachelor of Science in Environmental Engineering in 2012, and Master of Public Administration in 2015 from the Huazhong University of Science and Technology in Wuhan, China. She received her Master of Art in Statistics from the University of Missouri- Columbia in June 2016 and then continued to join the Ph.D. program in the same department. After graduation, she will join Biogen as a Senior Statistician.