NANYANG
TECHNOLOGICAL
UNIVERSITY

TWO CLUSTERING PROBLEMS IN ANALYZING NEXT GENERATION SEQUENCING DATA

TIAN YE

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

2016

TWO CLUSTERING PROBLEMS IN ANALYZING NEXT GENERATION SEQUENCING DATA

TIAN YE

School of Physical and Mathematical Sciences

A thesis submitted to the Nanyang Technological University

in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

2016

Acknowledgements

The four-year PhD study is a long and meaningful journey in my life. Before I reach the terminal of this journey, I would like to express my genuine thanks to all the people who help and support me.

First of all, I want to show my sincere gratitude and deepest respect to my supervisor, Prof. LIAN Heng, for his help and encouragement to my PhD studies. I am always motivated by his diligence, rigorousness, rich experience and quick-wittedness in scientific researches.

Secondly, I would like to express my thanks to Prof. CHEN Xin, for his precious advices and help in the research project to detect DMRs between paired sample genomes.

Thirdly, I wish to show my appreciation to all the examiners for examining my thesis and providing valuable suggestions and comments.

Last but not the least, I want to show my thanks to my families and friends for their support and blessings. In particular, I would like to say thanks to my dear wife, SUN Ruimin, who always stands by my side.

Contents

Abstract

As the next generation sequencing (NGS) becomes the dominating technology for studying the gene expression profiles, downstream statistical analysis tools are needed urgently. Clustering samples is an important approach to revealing smaples' relationships, such as for the discovery of new subtypes of cancer cells. To cluster high dimensional data, it is also of interest to select the variables (genes) informative for clustering. A new penalized model-based method called PMixClus is presented in this thesis to select genes and perform clustering simultaneously. The negative binomial mixture model is developed for the nonnegative and discrete count data from RNA sequencing experiments. Moreover, our method can automatically determine the number of clusters using the Bayesian information criterion. Additionally, in the PMixClus hybrid-hierarchical tree guided by the output from model-based clustering can be applied to visualize partial clustering structure in a hierarchical way. Results of both simulated and real data demonstrate that our method perform better or equally well compared to other competitive methods.

DNA methylation is a significant epigenetic modification to regulate gene transcription and plays a critical role in diseases. The whole genome bisulfite sequencing (WGBS) is a specific NGS technology for the detection of genome-wide DNA methylation at a single CpG site resolution. However, the high cost of such experiments and

the complexity of data challenges the downstream analysis. We proposed a new tool called DMReSearch to identify differentially methylated regions (DMRs) based on the WGBS data. We developed a three-dimensional rank method to pre-cluster the CpG sites, which considers CpG density, distance between centers and fluctuation of differences between two biological groups. Then we smoothed the methylation levels in each cluster with a modified local kernel smoother, carried out statistical test at each CpG by using the beta-binomial distribution and accordingly trimed and merged the identified DMRs. We compared our method to BSmooth which is the most popular method to detect DMR based on WGBS data. In simulation experiments, DMReSearch presents better receiver operating characteristic curves. Real data experiments show that DMReSearch performs better smoothing results, reports less unreasonable DMRs and presents consistency between low- and high-coverage data sets.

Chapter 1

Introduction

1.1 Introduction

DNA, RNA and proteins are the three main macromolecules that play an essential role for all living organisms. DNA (or deoxyribonucleic acids) is a double-strand molecule with anti-parallel helix structure. It carries the most heritable information that is involved in development, functioning and reproduction. In particular, only the information stored in some small pieces of DNA can be encoded for a certain functional product. Such small pieces of DNA are called genes, which are believed as the inheritable units. Genes can determine biological functions but cannot execute the corresponding functions. Most of final executors are proteins. A protein molecule is composed of amino acid sequences, which are folded into active three-dimensional structures to perform their specific functions. For eukaryotes (such as mammals, human beings), DNA molecules are usually organized into chromosomes and reside in the nucleus whereas proteins are located in cytoplasm. It implies that proteins cannot be produced directly by genes. In fact, an RNA (ribonucleic acids) is responsible for collecting genetic in-

formation from DNA and passing the information to code for proteins in cytoplasm. Such process is also known as gene expression.

The expression of a gene can be regulated according to the needs of cells by a variety of mechanisms. One of these mechanisms is epigenetic modification, which controls the activation of a gene but do not changes its DNA sequence. Among the three systems of epigenetic modifications, DNA methylation is best studied. It may turn off genes to prevent expression. Epigenetic modifications occur throughout the lifetime of an living organism in order to maintain cellular functions and adapt to changing environments.

In this thesis, I will discuss two main problems: the clustering of samples based on different gene expression profiles across distinct tissues or subtypes of cancer cells and the identification of differentially methylated regions between paired sample genomes. In the first chapter, the related biological backgrounds will be introduced, together with the corresponding analytical technologies. In Chapter 2 and 3, these two problems will be discussed, respectively. In detail, I will review some related and widely-applied methods, discuss our proposed methods step by step, and compare and evaluate the performances of our methods with other approaches by running experiments on both simulation data and real biological data. In the last chapter, summaries of these two projects will be made and future works will be discussed.

1.2 Gene Expression

Gene expression is the process through which the genetic information is used to synthesize functional gene products, including proteins and RNA. Generally, gene expression consists of two stages: transcription and translation. Below, I will introduce these two steps in the case of eukaryotes.

The transcription of a gene starts from binding the RNA polymerase and its associated transcription factors to the promoter region of the DNA strand. Then the RNA polymerase moves down the DNA strand from 3'end to 5' end and adds the complementary RNA nucleotides at the same time. Once the RNA polymerase encounters the terminator, the transcription step stops and a precursor messenger RNA (pre-mRNA) molecule falls off the DNA template.

In eukaryotes, a pre-mRNA is actually a chain of exons (coding segments) and introns (non-coding segments), which cannot be used directly to the subsequent protein production. In fact, the pre-mRNA has to be modified before becomes a mature mRNA. During the modification process, the introns are cut off from the pre-mRNA and the remaining exons are spliced together to form a mature mRNA. This modification is also known as RNA splicing. Finally, the mRNA with a poly-A tail added at its 3' end is ready to enter the cytoplasm through the nucleus pores.

Every mRNA consists of three parts: 5' untranslated region (5'UTR), protein coding region or open reading frame (ORF) and 3' untranslated region (3'UTR). The genetic information for protein synthesis is located in the ORF part. mRNA acts as an information carrier during the gene expression process. After it is exported to the cytoplasm, mRNA is used as a template and translated by ribosome according to genetic code. The genetic code interprets the nucleotide sequence within mRNA by mapping 64 three-nucleotide units, named as codons, to 20 amino acids. In other words, the mRNA implies an amino acid sequence. To form the sequence of amino acids indicated by mRNA, a variety of transfer RNA (tRNA) molecules are required. Each tRNA molecule has two ends, one of which carries an amino acid and the other carries an anticodon complementary to the codon of mRNA indicating the same amino acid.

The translation step involves the mRNA, ribosome, tRNA molecules and different

amino acids. After mRNA arrives at ribosome, the ribosome starts to find the start codon AUG. After the specific tRNA carrying the amino acid Met binds to the start codon, the ribosome shifts to the second codon on the mRNA. As the ribosome moves along the mRNA, the corresponding amino acids are chained together and the successive tRNA molecules are released. This process will not stop until the ribosome meets a stop codon. The translation step terminates with a chain of amino acids. The resulting amino acid sequence finally folds into a three-dimensional structure to form a synthesis protein. The whole process of gene expression in eukaryotes is depicted in Figure 1.1.

1.2.1 Differential Gene Expression

The gene expression process is controlled by the cell. The cell regulates the amount, timing and location of gene expression based on its needs. Every somatic cell within our body contains the same genome, but the liver cells are quite different from the heart cells with respect to both shapes and functions. It implies that different subsets of genes are expressed in these two types of cells. Moreover, during the period that a baby grows up to an adult, the number of cells of every organ increases at the same time. The cell growth has to involve the increasing expression of the related genes.

Besides, differential gene expression also explains the differences between a cancer cell and a normal cell. Cancer is actually a disease of abnormal cell growth. If a normal cell transforms into a cancer cell, the genes that regulate cell growth and differentiation must be changed. Thus, analyzing the differentially expressed genes in cancer cells can provide a better understanding of the causes of cancer and assist accurate clinical diagnosis. Furthermore, two patients suffering from the same type of cancer can have different symptoms, because their cancer cells may contain different subsets of

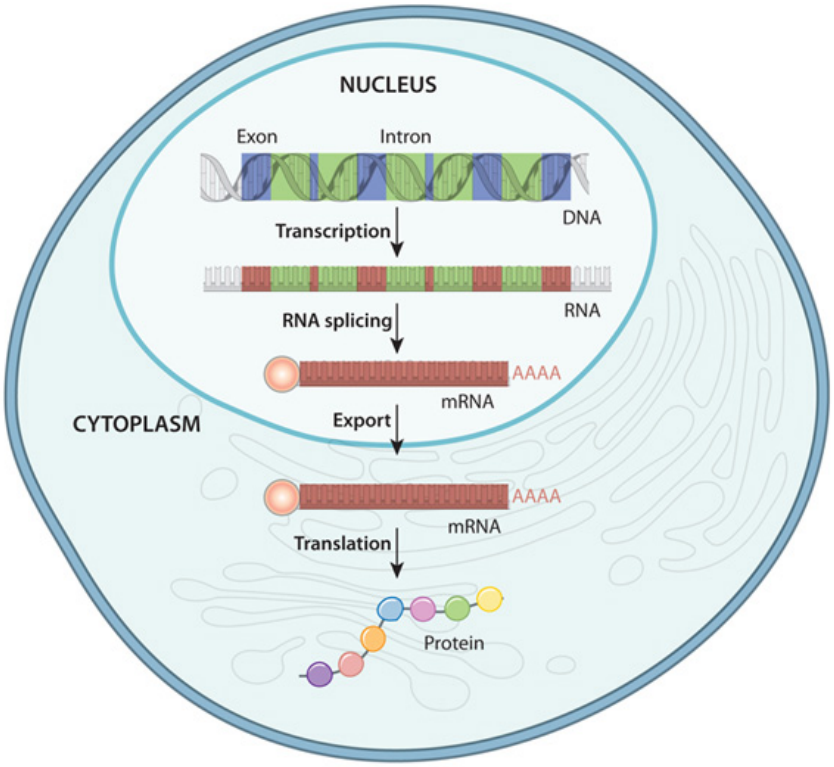


Figure 1.1: Gene expression in a eukaryote cell. The DNA sequence of a gene is firstly transcribed to pre-mRNA. After splicing the exons to form the mature mRNA, the poly-A tail is added. The mRNA is then exported to the cytoplasm and translated by ribosome to construct an amino acid sequence, which finally folds into the protein molecule. The figure comes from `http://www.nature.com/scitable/topicpage/gene-expression-14121669`.

genes expressed or different amounts of expression of the same sets of genes. It involves the studies of cancer subtypes. For instance, about 35 subtypes of non-hodgkin lymphoma (NHL) are confirmed at the current moment according to the reports at Cancer.Net (`http://www.cancer.net`). Thus, during the cancer diagnosis, from which subtype a patient suffers should also be examined so that the appropriate treatment can be correctly selected.

Overall, learning the differentially expressed genes is of great importance for inter-

preting functional elements of the genome, understanding the development of diseases, and improving clinical diagnosis and treatment selection.

1.2.2 RNA Sequencing

To analyze the differential gene expression between two types of tissues, mRNA instead of DNA should be sequenced and compared because only mRNA contains the protein coding instructions. Currently, the next generation RNA sequencing (RNA-Seq) is a widely used technology. RNA-Seq outperforms other existing technologies due to its high throughput, time and cost efficiency, high accuracy, and reproducibility [76, 49]. Figure 1.2 illustrates the key steps in the RNA-Seq technique.

In brief, an RNA-Seq experiment starts from the preparation of cDNA library. cDNA is made from an mRNA through reverse transcription. After the first cDNA strand is synthesized, the RNA strand is removed from the cDNA-mRNA hybrid and a complementary cDNA strand is generated to synthesize a double-strand cDNA. The double-strand cDNA is then cut into small pieces through RNA fragmentation or DNA fragmentation. The resulting cDNA fragments are subsequently ligated to sequencing adapters and sequenced by high-throughput DNA sequencing approaches. After this process, millions of short reads (30-400 bp in length) can be obtained. Given the reference genome or transcriptome, these reads are aligned and pooled into regions of interest, such as genes and exons. The number of short reads mapped to each region is counted to quantify the expression levels.

During the process of an RNA-Seq experiment, a high-throughput (or next generation) DNA sequencing (NGS) technique is required. Over the recent decades, several NGS platforms have been commercially developed and applied. These NGS platforms

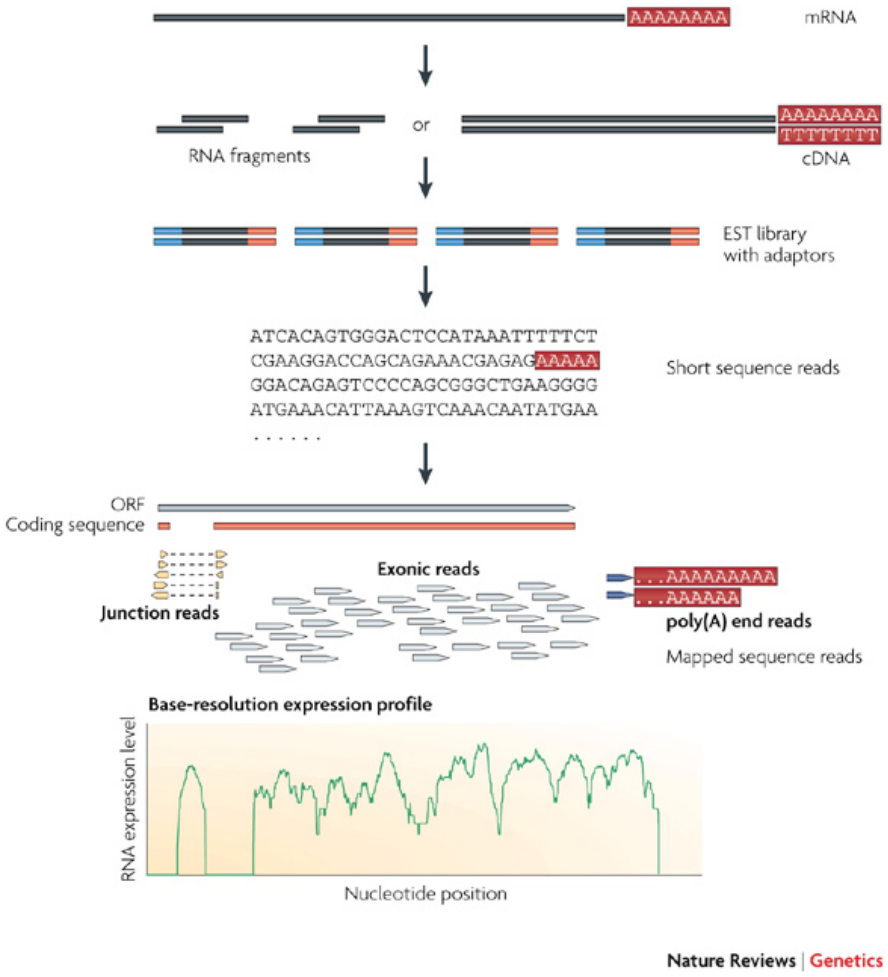


Figure 1.2: The schematic of an RNA-Seq experiments [76].

employ three general principles to sequence a DNA sample: sample preparation, immobilization and sequencing [52]. These three steps are briefly introduced below with the Genome Analyzer by Solexa/Illumina platform as an example, which is currently the most widely used NGS platform (see Figure 1.3).

The sample preparation step adds specific adapter sequences to the ends of each DNA fragment. These DNA fragments ligated by adapters constitute the adapter library. Once the adapter library is prepared, it is immobilized on the surface of a

flow cell for amplification. The Illumina flow cells are planar, fluidic devices that have primers immobilized on the inner surfaces, whose sequences correspond to the adapters. The amplification is then performed by a process called Bridge-PCR. In the Bridge-PCR process, an immobilized sequencing library fragment hybridizes with a primer on the flow cell surface to form a bridge structure. The resulting arched molecule acts as a template for a DNA extension reaction, which results in a bridge double-strand DNA. The two strands are then freed by using a denaturing reagent. Bridge-PCR produces clusters of thousands of DNA molecules on each flow cell lane by repeating the reagent flush cycles of denaturation, annealing, extension and wash. Each resulting cluster corresponds to a single DNA fragment and is dense enough to be detected in the sequencing step. The Illumina Genome Analyzer employs a sequencing-by-synthesis approach to sequence each cluster. In this step, four distinct fluorescent nucleotides are used to recognize the bases of the cluster of DNA fragments through polymerase synthesis. Starting from the hybridized sequencing primer, one base is extended at each time and the nucleotide fluorescence is shown in the imaging system simultaneously. Compared to other NGS platforms, the Illumina Genome Analyzer can generate millions of reads in 36 - 300 bp length with less time and cost [42]. Moreover, this technology creates few sequencing errors when the sequence reads are short [52], which especially suits an RNA-Seq experiment.

1.3 Epigenetics

Epigenetics studies the biological mechanisms that turn genes on or off but do not involve changes on the nucleotide sequence. Through the entire lifetime of a living organism, epigenetic changes occur naturally and modify the activation of each gene

regularly according to different biological functions. However, epigenetic modifications can also occur in response to other factors, such as lifestyle, climate changes, foods, and ages.

Epigenetic changes play an important role in evolution. Studies have shown that expression of the genes that are modified by epigenetic marks can be passed down to the future generations. For example, the experiment on mice carried out by [14] indicated that the descendants inherited the fearful memories of the first generation mice. Moreover, epigenetic traits vary rapidly in response to the environmental changes. Therefore, a living organism may adapt to different environment conditions merely through the epigenetic modifications that control the gene expression.

Three systems are commonly believed to initiate and sustain epigenetic modifications: DNA methylation and hydroxymethylation, histone modification and non-coding RNA (ncRNA)-associated silencing [17]. DNA methylation is a biochemical process that adds a methyl group (CH3) to the cytosine or adenine DNA nucleotides. Methylation of adenine only occurs in prokaryote organisms, while methylation of cytosine can be observed in almost all living organisms. DNA methylation may stably modify the expression of genes and inhibit transcription. Histones are proteins that act as spools around which DNA can wind such that the long DNA molecules are packaged inside the small cell nuclei. Histone modifications alter the gene expression through changing the way that the corresponding DNA wraps around the histones. ncRNA is transcribed from DNA but is not translated into a protein. ncRNAs regulate gene expression by involving in the formation of compact chromatin, histone modification, or DNA methylation. Figure 1.4 describes the mechanisms of DNA methylation and histone modification.

Epigenetic changes are not only of significance in both individual development and

evolution, but also responsible for a variety of diseases when we pay more attention to their roles in human health. Abnormal epigenetic modifications can result in incorrect activation or silencing of genes, which may lead to various cancers, mental disorders, immune disorders, and other syndromes. In our study, we focus on the relationships between cytosine methylation and the development of colorectal cancer.

1.3.1 Cytosine Methylation

Methylation of cytosine (or cytosine methylation) is currently the best-studied epigenetic modification since it plays a crucial role in many essential biological processes, such as embryonic growth, X chromosome inactivation, genomic imprinting, regulation of gene expression and the development of cancer [62]. Cytosine methylation means adding a methyl group at the fifth carbon position of the cytosine ring. According to its specific structure, methylated cytosines are often called 5-methylcytosines or 5mC for short.

In general, there are two types of cytosine methylation differentiated by the sequence contexts that 5mCs locate in: CpG methylation and non-CpG methylation. Here, CpG indicates a nucleotide C linked with a nucleotide G by phosphate along the DNA sequence. It was demonstrated that most CpGs (over 70%) in human genome are methylated [18]. In particular, the CpG-rich regions, which are also called CpG islands (or CGIs), attracted the most discussion since a majority of genes have their promoters overlapped by CGIs [65]. It is generally believed that methylation of the CGIs within the promoter regions has an inhibitory effect on transcription process and might be a key factor in the development of cancers [5, 62]. Furthermore, recent studies have also shown that CpG methylation in other regions, such as CpG island shores

and gene bodies, can distinguish between different tissue cells and between normal and cancer cells [15, 34].

On the other hand, non-CpG methylation involves cytosine methylation occurring in CHG and CHH sequence contexts, where H represents nucleotide A, C, or T. Non-CpG methylation was primarily observed and studied in plant cells [71]. For example, an abundance of CHH methylation was found genome-wide in Arabidopsis and other flowering plants [56]. However, in mammal genomes, non-CpG methylation only exists in specific tissues or at some restricted development stages. It was shown in a previous study that non-CpG methylation was enriched in human embryonic stem cells but disappeared during the cellular differentiation process [41]. Moreover, the detailed function of non-CpG methylation in the gene regulatory process is still unclear. As a result, non-CpG methylation is seldom discussed while studying the association between DNA methylation in human genome and diseases.

Besides, recent studies have proven that 5mCs in mammalian genomes can be oxidized to form 5-hydroxymethylcytosine (or 5hmC). Current experimental data suggests its critical role in brain development and cancer [57]; however, the precise biological function of 5hmCs is still unclear. In wet-lab experiments, it brings about more challenges to distinguish between 5mC and 5hmC because of their identical reactions towards the bisulfite treatment (see the following section). Moreover, the occurrences of 5hmC are substantially less than 5mCs in human genomes. Therefore, the impact of 5hmCs is always ignored in current studies.

1.3.2 Bisulfite Sequencing

Sodium bisulfite treatment is a widely-used approach to determine the 5mCs in a DNA sequence. After the treatment of sodium bisulfite, unmethylated Cs will be converted into uracils (Us) while the 5mCs keep unchanged. In the subsequent polymerase chain reaction (PCR) amplification step, the Us will be converted to thymines (Ts) and the 5mCs remain as Cs. Such different reactions of unmethylated Cs and 5mCs allow us to distinguish them by simply comparing the DNA sequences before and after bisulfite treatment [38].

We note that it is both expensive and time-consuming to establish the DNA sequence twice in practical research. A general solution is to compare the bisulfite treated DNA sequence with the corresponding reference genome sequence that is published online with an open access. In recent decades, next generation sequencing (NGS) technologies have been rapidly developed and widely used because of their advantages in saving time and cost. By using NGS technologies to obtain the bisulfite treated DNA sequence, which defines BS-Seq, the methylation states can be analyzed in genome-wide scale at the single base-pair resolution. Furthermore, applying BS-Seq in whole genome enable the detection of methylation states of all Cs regardless of their sequence contexts. The general process of BS-Seq is described in Figure 1.5. The reads generated from BS-Seq are also called BS reads.

In recent years, an increasing number of alignment tools have been developed for mapping the whole-genome bisulfite sequencing (WGBS) data, such as BS Seeker [12], Bismark [37], Last [23] and so on. Based on the alignments reported by these mapping tools, the methylation status of a single cytosine in the reference genome can be measured. Particularly, a C in reference genome (or a genome C) is said to have a *methy-*

lated call if it is mapped by a C in some read (or a read C) in an alignment; whereas if a genome C is mapped by a read T in an alignment, it has an *unmethylated call*. In most studies in analyzing DNA methylation, the methylation status of a single genome C is usually quantified by the *average methylation level*, which is defined as the number of methylated calls divided by the total number of methylated and unmethylated calls. In fact, since a sample in a practical WGBS experiment contains DNA molecules from distinct cells, the methylation status at a single C site is actually defined by its *methylation level*, which is the fraction of calls having methylation at that site within the sample. According to the assumption that the methylated Cs should follow a binomial distribution, the average methylation level is usually regarded as an unbiased estimate of the true methylation level at the single site. In the following article, we refer average methylation level to methylation level equivalently if no special conditions are mentioned.

1.3.3 Differentially Methylated Regions

The same CpG locus may have differential methylation states in different tissues, which attempts to control the activation of the gene in response to distinct functional requirements. While aberrant methylation occurs at single CpG loci, it might give rise to a variety of human diseases. Such negative impacts of abnormal CpG methylation have been demonstrated by numbers of studies. The methylation of CpGs within the promoter of gene DAPK1 was proved to inhibit DAPK1 expression and further cause the chromic lymphocytic leukemia (CLL) [61]. Besides, two methylated cytosines in the promoter of gene SLC12A6 were also shown to have association with bipolar disorder [51].

It does make sense to learn the methylation states of single CpG positions; however, it is not convincible enough. Analysis of single CpG sites is very sensitive to the influence of SNPs and sequencing errors, and hence might result in incorrect judgement of methylation states [50]. In fact, the functions of methylation are always associated with genomic regions, such as the gene promoter regions, gene bodies, etc. [34]. In particular, CpG islands (CGIs) that overlap these genomic regions are paid the most attentions in the studies of differential methylation states between two groups of cells. In human somatic cells, most CGIs of the gene promoters are normally unmethylated, while the CGIs in gene bodies are methylated in a tissue-specific manner. Methylated CGIs in the gene promoters may lead to long-term silencing and play a critical role in various cancers, such as breast and colon cancers. When CGIs in gene bodies are aberrantly methylated, the cancer-causing somatic and germline mutations might occur. In order to discover the comprehensive maps between methylation of the whole human genome and various types of diseases, the differentially methylated regions (DMRs) are primarily identified in most studies [16, 30].

A DNA region is considered as a DMR if the DNA methylation levels (defined in the previous section) within this region are consistently and statistical-significantly different between a pair of sample groups. There are several types of DMRs depending on the objectives of studies. In our work, we aim at identifying DMRs between cancer and normal samples and DMRs between different tissues. In regard of the length of a DMR, it might be ranged from a single CpG to millions of bases. The DMRs containing only one CpG are generally referred as differentially methylated CpGs (or DMCs).

Moreover, a DMR is directional: if the methylation levels of cancer group are significantly higher than those of normal group within the same region, the DMR is usually called *hyper-methylated* region; whereas if the methylation levels of cancer group are

significantly lower than those of normal group, the DMR is termed as *hypo-methylated* region. In regard of human colorectal cancer, it has been shown that the CGIs at gene promoters are mostly hyper-methylated but globally the cancer genome shows large hypo-methylated blocks [30, 3].

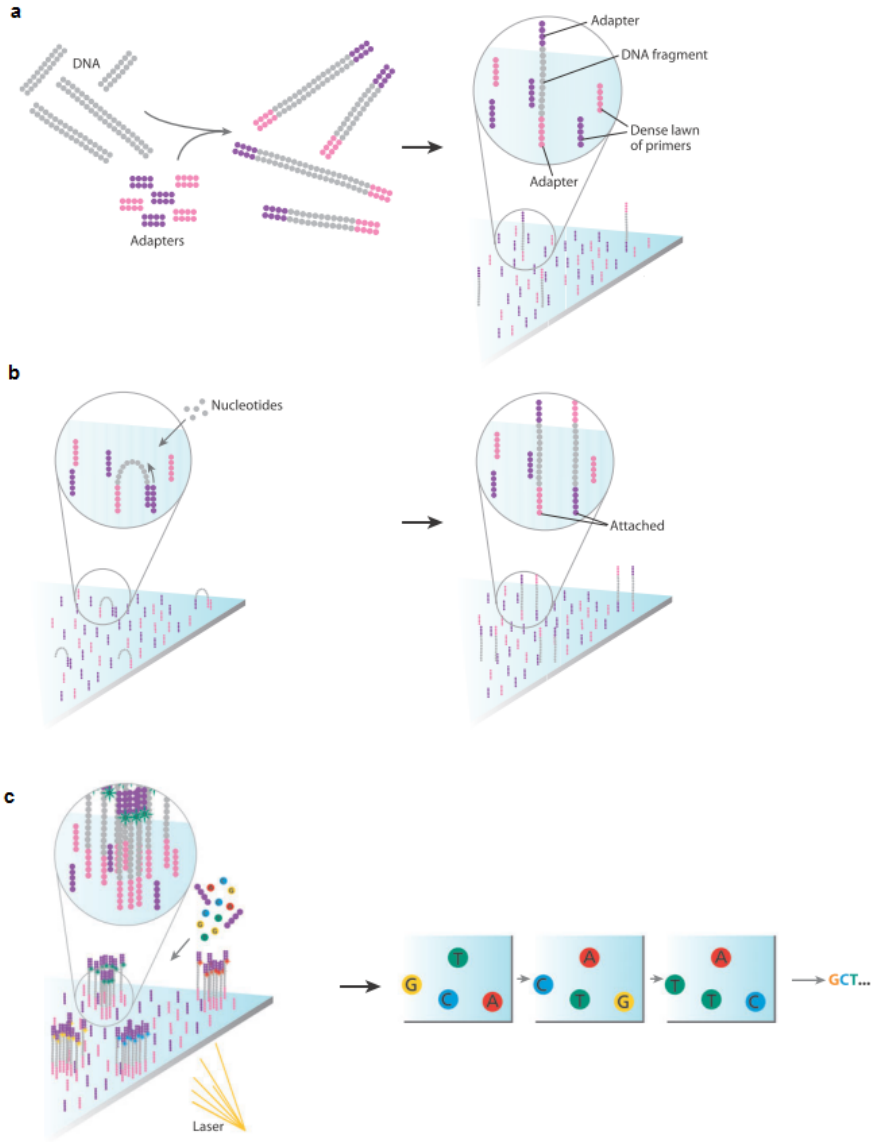


Figure 1.3: The sequencing approach of Illumina platform. Figure is taken from [45].

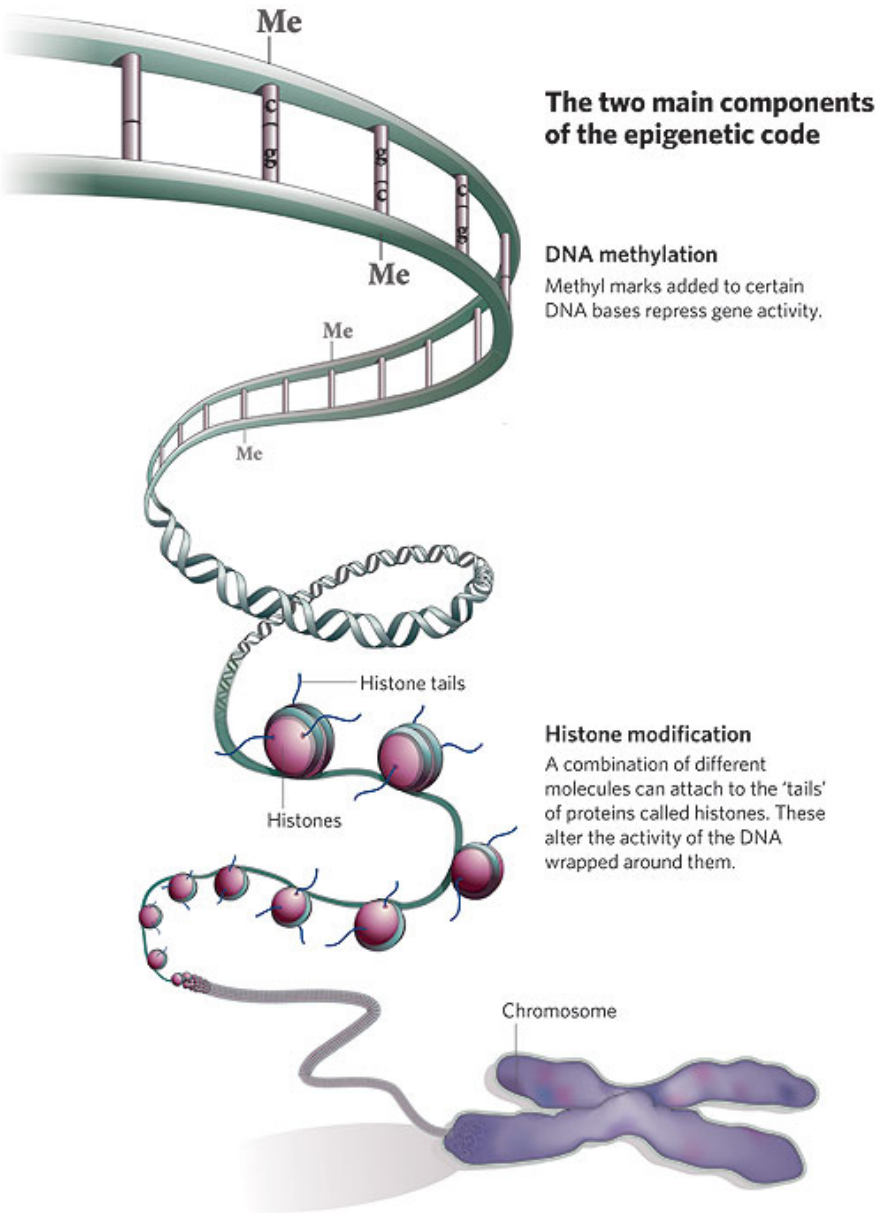


Figure 1.4: DNA methylation and histone modification. DNA methylation can repress gene activity. Histone modification occurs at the tails of histone proteins and may change the way that DNA wraps around them. Figure is downloaded from `http://www.precisionnutrition.com/epigenetics-feast-famine-and-fatness`.

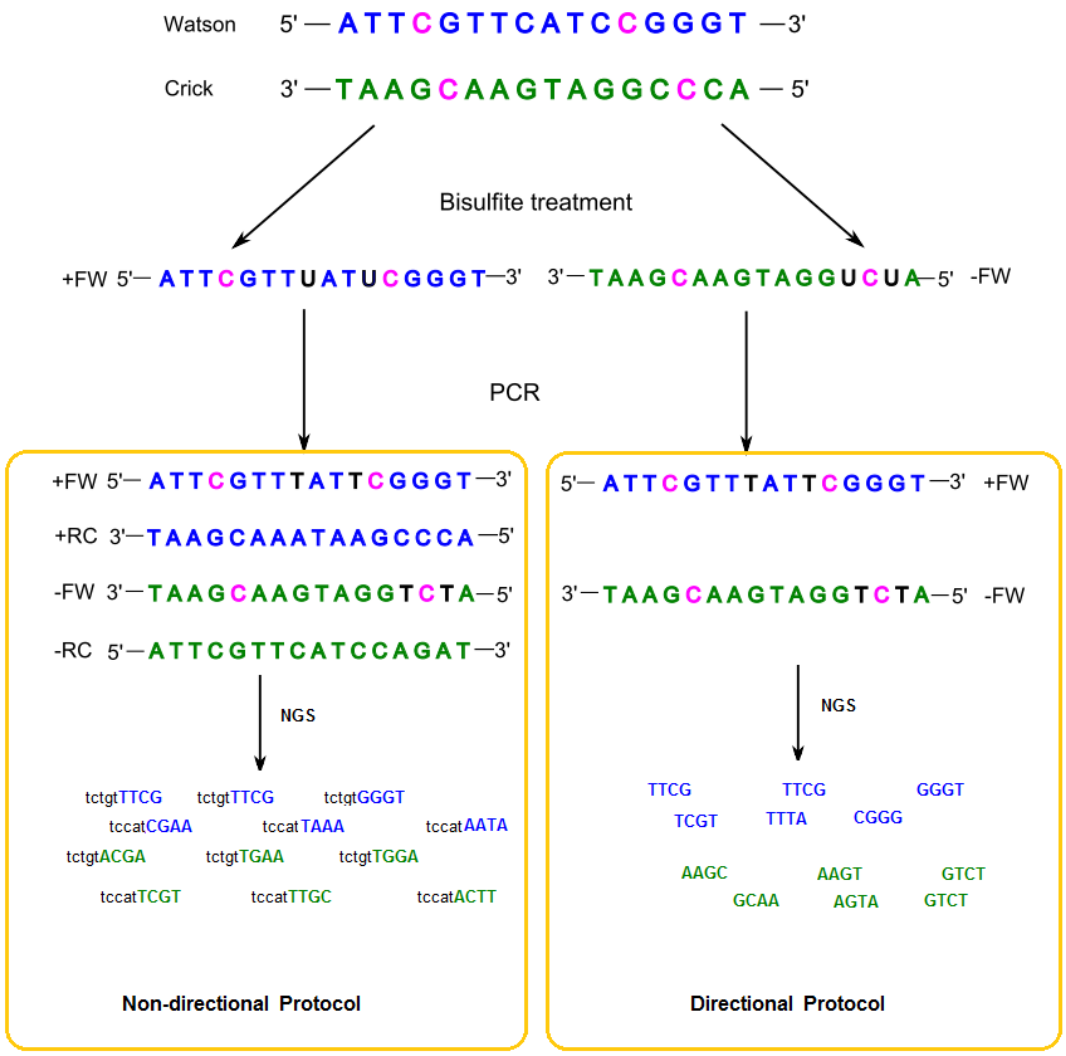


Figure 1.5: Bisulfite sequencing (BS-Seq). BS-Seq combines bisulfite treatment and next generation sequencing technologies. Bisulfite treatment that converts unmethylated Cs to Us but keeps 5mCs (in rose color) unchanged. During the PCR process, Us are converted to Ts while 5mCs remain as Cs. With respect to the PCR procedure, two types of amplification protocols exist: directional and non-directional protocols. Non-directional protocols replicate the bisulfite converted strands first, and then generate reads with tags. In contrast, reads produced from directional protocols are only from either the forward converted strand or the reverse converted strand.

Chapter 2

l_1-Penalized Model-based Clustering for RNA-Seq Count Data

2.1 Introduction

Based on the resulting RNA-Seq count data from NGS, researchers can use statistical methods to perform various downstream analysis, such as identifying differentially expressed (DE) genes (e.g., [2, 63]) and classification or clustering of samples (e.g., [4, 78]) and genes (e.g., [69]). Note that clustering can be performed on either genes or samples, or both simultaneously. In this thesis, we focus on the clustering of samples based on the variation of gene expression profiles across different tissues or subtypes of cancer cells.

Studies on clustering different RNA-Seq samples are very significant to characterize themselves across different organisms or different phases during pathogenic development. It could be the premise for the diagnosis of atypical clinical and histopathological presentation. Although a lot of articles contribute to the clustering analysis with the use

of the microarray data (e.g., [70], [59], [80] and [55]), which was the most popular technology to quantify gene expression before the advent of RNA-Seq, there are fewer methods proposed for the latter. Besides, the RNA-Seq count data has different characteristics from microarray data in numerous ways. Compared with microarray data, RNA-Seq has more accurate digital signals inherently than the analogue measure of microarray; RNA-Seq is more flexible to detect new types of isoforms while microarray will fail when non-specific and off-target probe-array binding occurs; RNA-Seq count data consists of non-negative integers and is normally modeled by Poisson or NB distribution while microarray data is in real domain and usually modeled by normal distribution. The first two differences make RNA-Seq supersede microarray and the last difference require researchers to develop new analysis tools specifically for RNA-Seq. However, RNA-Seq count data encounters one same challenge as that in microarray: Due to the typically large number of genes compared to number of samples, this problem falls into the "large p small n" paradigm that has attracted a lot of attention recently in biostatistics ([27, 44, 75, 8, 31]).

In recent years, several methods have been developed to cluster samples by using RNA-Seq data. In 2008, a Bayesian method was proposed in [4] to compute the dissimilarity matrix in the clustering analysis of small RNA cloning data from sequencing technology. Two years later, it was suggested to transform the data for stabilizing the variance and then compute the squared Euclidean distance using these transformed data [2]. In the same year, M. D. Robinson and his colleagues provided a clustering function using 500 features with highest variance in their edgeR software package [63]. In 2011, D. M. Witten modeled the RNA-seq count data with a Poisson log-linear model and computed the dissimilarity matrix by a modified log-likelihood using the power transformed data [78]. These methods mentioned above are distance-based methods.

Model-based clustering is a good alternative approach which applies a finite mixture model to represent each cluster by each model component. Moreover, in a model-based clustering method, variable selection can be conducted simultaneously by putting a penalty on the mixture model. The finite mixture model has been deeply studied and widely applied in many areas, such as astronomy, biology, genetics, economics and so on [47]. For all high-throughput biotechnologies, such as RNA-Seq and microarrays, it is a key point to select variables. Among large amount of genes that are measured, usually only a small number of genes are of interest. In order to exclude non-informative genes, some dimension reduction methods are applied prior to clustering. D. Ghosh and A. M. Chinnaiyan (2002) proposed a principal component analysis approach to cluster genes based on microarray data [25]. However, the separation between dimension reduction and clustering could damage the clustering structure of the original data ([80] and [58]). In contrast, an advantage of model-based method is to select variables while conducting sample clustering by putting a penalty on it. A. Khalili and J. Chen (2007) developed the penalized model-based method for mixture of regression models [36]. They provided strict proof for the asymptotic properties of regression coefficients estimated by EM algorithm inside the model. In the clustering analysis of microarray data, a l_1 penalized model-based clustering with a finite Normal mixture model was proposed and studied for clustering samples based on microarray data. The EM algorithm was also applied to estimate parameters for this l_1 penalized Normal mixture model. Additionally, in [55] a modified BIC was proposed to determine the number of clusters and the amount of penalization. For this penalized finite Normal mixture model, there are closed-forms for estimators of parameters in the M-step during the EM process. However for Poisson and NB mixture model, especially for the NB model, there are no closed-forms for estimators of some parameters and optimization methods are needed

to find the numerical solutions in the M-step. Now, we propose a new penalized model-based approach to cluster samples according to the differential gene expression profiles across distinct tissues or subtypes of cancer cells. We also compare our method with three competing methods: PoiClaClu [78], edgeR [63] and DESeq [2] in both simulation and four real data sets. Before we present our method and discuss the experiment results, we need to review these three competing approaches first.

2.1.1 Competitive Methods Reviews

To better present the differences between our method and the three competing methods, a brief introduction of each method will be provided. Generally, these three methods adopt the similar strategy, which starts from calculating a specific dissimilarity matrix to build a hierarchical tree and then does clustering by cutting this tree in the target level.

PoiClaClu

PoiClaClu [78] is available in the CRAN repository. In this approach, for gene p in the sample j under biological condition k the read count X_{pj} is modeled by a Poisson log linear distribution:

$$X_{pj}|j \in C_k \sim \text{Poisson}(N_{pj}d_{kp}),$$

where $N_{pj} = s_j g_p$ and $C_k = \{j : \text{sample } j \text{ is in the cluster } k\}$. s_j is interpreted as the size factor and estimated by $\hat{s}_j = \frac{\sum_p X_{pj}}{\sum_{p,j} X_{pj}}$ and g_p is the total read counts of gene p over all samples. The parameter d_{kp} indicates fluctuation around the base line for gene p in the biological condition k. Then the maximum likelihood estimation (MLE) is applied to estimate N_{pj} as $\hat{N_{pj}} = \frac{\sum_p X_{pj} \sum_j X_{pj}}{\sum_{p,j} X_{pj}}$. As the estimate of g_p is defined by $\hat{g}_p = \sum_j X_{pj}$,

$\hat{N_{pj}}$ can be represented by $\hat{s}_j\hat{g}_p$. d_{kp} is subsequently estimated by MLE: $\hat{d}_{kp} = \frac{\sum_{j\in C_k,p} X_{pj}}{\sum_{j\in C_k,p} \hat{N}_{pj}}$. Additionally, a Gamma($\beta,\beta$) prior is given on d_{kp}, whose rate and shape parameters are both β, and then the estimator of d_{kp} is computed by the posterior mean: $\hat{d}_{kp} = \frac{\sum_{j\in C_k,p} X_{pj}+\beta}{\sum_{j\in C_k,p} \hat{N}_{pj}+\beta}$.

To define the distance between sample j' and sample j, the author tests the null hypothesis $H_0 : d_{pj'} = d_{pj} = 1$. A modified version of log-likelihood ratio statistic is proposed to represent this distance:

$$\sum_j (\hat{N}_{pj} + \hat{N}_{pj'} - \hat{N}_{pj}\hat{d}_{pj} - \hat{N}_{pj'}\hat{d}_{pj'} + X_{pj}\log\hat{d}_{pj} + X_{pj}\log\hat{d}_{pj'}).$$

This so called modified version is from the Bayesian estimate of d_{pj}. Accordingly, the dissimilarity matrix is captured and the hierarchical tree is built to conduct clustering.

The author also proposes a remedy for the defect in Poisson model when it is used to model the data with biological replicates. The overdispersed data will be transformed by a power transformation: $X_{pj}^{\alpha} \to X'_{pj}$, where $\alpha \in (0, 1]$ is determined by a goodness of fit test for Poisson model. It is presented that this power transformation performs well on moderately overdispersed data. However, the effect of this power transformation on the parameters' estimating process needs to be discussed carefully. For example, the estimate of size factor $\hat{s}_j$ used in their method will be changed after the power transformation. From our view, no sound explanation can be carried out for such change.

DESeq

DESeq [2] is available in the Bioconductor repository. The authors model the read count by NB distribution: $X_{pj}|j \in C_k \sim \text{NB}(\mu_{pj}, \sigma^2_{pj})$, where $\mu_{pj} = g_{kp}s_j$. s_j is the size factor which is estimated by median ratio method (presented in the methods section). To

get the estimate of g_{kp}, the average of X_{pj} is computed over samples j that correspond to the condition k. In order to cluster samples, a variance stabilizing transformation (VST) is presented on the normalized data $y_{pj} = \frac{X_{pj}}{s_j}$. The motivation is to make variable variances of X_{ij} for different genes homoscedastic. The VST function in the paper is given by

$$\tau(x) = \int^x \frac{dg}{\sqrt{w(g)}},$$

where $w(g)$ is the variance-mean dependence. Then the authors squared Euclidean distances between samples based on the VST data.

EdgeR

EdgeR [63] is available in the Bioconductor repository. With regard to clustering RNA-Seq samples, the authors propose a straightforward method. Euclidean distances are computed between each pair of samples based on the selected genes with the highest variance. The default number of selected genes is 500.

2.1.2 Our Contribution

As mentioned in the previous sections, all the competing methods are distance-based methods. The benefit of such methods is the ability to visualize the clustering results clearly by the related techniques, such as hierarchical clustering or multidimensional scaling. However, they lack a probabilistic interpretation for clustering and provide no statistically sound way of determining the number of clusters. To tackle these issues, the model-based clustering [47] is a popular statistical approach. This method allows soft allocation of samples to clusters. Moreover, with a well-defined likelihood, many criteria such as the Akaike Information Criterion (AIC) [1], the Bayesian Information

Criterion (BIC) [67] and Extended Bayesian Information Criterion (EBIC) ([9] and [10]), have been developed to perform model selection.

Another important issue related to the clustering of RNA-Seq count data is gene selection. Such data have large dimension P (the number of genes) and small sample size n and generally a lot of genes are noise in the sense they are not differentially expressed across different clusters. Accordingly, it is natural to perform gene selection when we conduct clustering. Besides that the set of selected genes can be of interest in itself, this will potentially improve clustering accuracy. To our knowledge, gene selection for RNA-Seq count data is not currently available for clustering, except to heuristically select those genes with larger variances in a preprocessing step [63]. We will apply the penalized model-based method to perform gene selection and clustering simultaneously. To determine the number of clusters and the number of important genes, we use the Bayesian information criterion (BIC) as in [55]. To model the RNA-Seq count data, Poisson model can work quite well if no biological replicates exist [46]; otherwise, the negative binomial (NB) model ([63, 2]) can be applied because biological replicates may give rise to over-dispersion (i.e., the variance is larger than the mean). In this chapter, we focus on NB mixture model of which the Poisson mixture model is a special case. Our proposed method can often get better clustering results. The proposed method can be executed in the R package PMixClus available at `https://github.com/TianYe00/PMixClus.git`.

2.2 Methods

Here we mainly introduce the NB mixture model with l_1 penalty for clustering RNA-Seq samples. Additionally the l_1 penalized Poisson mixture model will also be pre-

sented as an alternative. To estimate parameters, we utilize the expectation maximization (EM) algorithm. Then a proper initialization strategy of parameters will be proposed for EM algorithm.

2.2.1 Model

Suppose that the read count data $\boldsymbol{x}$ contains n samples (rows) and P genes (columns). Let $\boldsymbol{x_j} = (x_{j1}, \ldots, x_{jP})$ denote the read counts of P genes in sample j for $j = 1, \ldots, n$. For mathematical simplicity, it is assumed that all genes are independent and $\boldsymbol{x_j}$ follows a finite mixture distribution $\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j}; \boldsymbol{\psi_{jk}})$, where f_k is the discrete distribution for kth cluster with parameter vector $\boldsymbol{\psi_{jk}}$, and π_k is the mixing proportion satisfying $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. The log-likelihood of the mixture model is

$$\log L(\boldsymbol{\Theta}) = \sum_{j=1}^{n} \log\{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x_j}; \boldsymbol{\psi_{jk}})\}, \tag{2.1}$$

where $\boldsymbol{\Theta} = \{(\pi_k, \boldsymbol{\psi_{jk}}) : k = 1, \ldots, K; j = 1, \ldots, n\}$. Denote $\mu_{jkp} = E(x_{jp})$ if sample j belongs to cluster k. To facilitate gene selection, one key development here is to decompose μ_{jkp} as $\mu_{jkp} = s_j \gamma_p \theta_{kp}$ subject to identifiability constraint $\sum_{j=1}^{n} s_j = n$, where s_j is the size factor for sample j (sequencing depth for sample j), γ_p is the average read counts of gene p over all samples and θ_{kp} represents cluster-specific effect for gene p. If $\theta_{kp} = 1$ for all k, it means that gene p is not differentially expressed across clusters and should be treated as noise variables. This observation suggests that we can perform gene selection by shrinking many of the θ_{kp} towards 1. Thus we propose to use the

following LASSO penalty [72]:

$$P_\lambda(\boldsymbol{\Theta}) = \lambda \sum_{k=1}^{K} \sum_{p=1}^{P} |\log \theta_{kp}| \ ,$$

where $\lambda > 0$ is a tuning parameter. Accordingly, the penalized log-likelihood function for the mixture model is defined as,

$$\log L_P(\boldsymbol{\Theta}) = \log L(\boldsymbol{\Theta}) - P_\lambda(\boldsymbol{\Theta}) \ .$$

NB mixture model

We now introduce the method based on the NB distribution. We assume $x_{jp}|j \in C_k \sim \text{NB}(\mu_{jkp}, \phi_p)$, where NB(.,.) is the negative binomial distribution with mean $\mu_{jkp} = s_j \gamma_p \theta_{kp}$ and variance $\mu_{jkp} + \phi_p \mu_{jkp}^2$, C_k contains the samples for cluster k, and ϕ_p is the gene-specific dispersion parameter. When $\phi_p = 0$, this reduces to the Poisson model. Thus the parameter set $\boldsymbol{\psi_{jk}}$ in the NB mixture model (2.1) is $\boldsymbol{\psi_{jk}} = \{(\mu_{jkp}, \phi_p) : p = 1, \ldots, P\} = \{(s_j, \gamma_p, \theta_{kp}, \phi_p) : p = 1, \ldots, P\}$.

Poisson mixture model

The Poisson mixture model is considered as a special case when the target data set is believed to have no overdispersion. We assume $x_{jp}|j \in C_k \sim \text{Poisson}(\mu_{jkp})$ with mean and variance $\mu_{jkp} = s_j \gamma_p \theta_{kp}$. Therefore, the parameter set $\boldsymbol{\psi_{jk}}$ in the Poisson mixture model (2.1) is $\boldsymbol{\psi_{jk}} = \{(\mu_{jkp}, \phi_p) : p = 1, \ldots, P\} = \{(s_j, \gamma_p, \theta_{kp}) : p = 1, \ldots, P\}$.

The EM algorithm is applied to find the optimizer of the penalized likelihood. The details are presented in the following sections.

2.2.2 Initialization Strategy

Many articles have already demonstrated the importance of the initialization strategy when the mixture model is fitted by EM algorithm [68, 47]. In order to estimate the size factor s_j, many approaches have been developed (e.g. [6, 2, 40]). Here we introduce three of them.

- **Total count.** This method simply estimates s_j as $\hat{s}_j = \frac{n \sum_p x_{jp}}{\sum_{j,p} x_{jp}}$.

- **Median ratio.** S. Anders and W. Huber [2] propose this method to compute $\hat{s}_j$ as,

$$\hat{s}_j = \underset{p}{\text{median}}\{\frac{x_{jp}}{(\prod_{j=1}^{n} x_{jp})^{\frac{1}{n}}}\}.$$

  It means that they compute the ratios of read counts to the geometric mean of read counts for each gene and then calculate the median of these resulted ratios over all genes.

- **Quantile.** J. Bullard and others [6] propose this method to estimate s_j by $\hat{s}_j = \frac{q_j}{\sum_j q_j}$, where q_j is the third quantile of read counts for sample j.

The total count method is seldom used since the bias is easily caused by a few large counts. Whereas median ratio method and quantile method perform more robust on such data sets. For our simulation, all these three methods can perform well since we want to focus our study on clustering. Therefore we apply median ratio method to estimate s_j in our approach. For the NB mixture model, we fix s_j when running the EM algorithm to improve computation efficiency due to the complicated model; while for Poisson mixture model, we provide an alternative option that s_j can be updated inside the EM algorithm.

We use the K-means method to get initial class labels and the starting values of θ_{kp} and γ_p are obtained by simple moment estimator. For the initial values of ϕ_p, J. Lu and others proposed a dispersion estimator for the over-dispersed log-linear model by applying the goodness-of-fit statistic in the analysis of SAGE data [43] . J. Li and others applied a similar idea to estimate the transformation for data exhibiting over-dispersion using the Poisson goodness-of-fit statistic [40] . Following this idea, we use the NB goodness-of-fit statistic to obtain the starting values of ϕ_p. Generally, the larger read count leads to lower dispersion [2]. Accordingly, we divide genes into M groups according to the mean counts of genes and then estimate the dispersion parameters for each group of genes. In the mth group, the goodness-of-fit statistics is

$$\mathrm{GOF}_{mp} = \sum_j \frac{(x_{jp} - \hat{\mu}_{k(j)p})^2}{\hat{\mu}_{k(j)p}(1 + \phi_m \hat{\mu}_{k(j)p})} \ ,$$

where $k(j)$ denotes the cluster identity for sample j. Since x_{jp}'s are independently NB distributed, the approximate distribution of GOF_{mp} is χ^2 with $(n-1)(P/M-1)$ degrees of freedom. In order to get rid of the outliers, we set $S_m = \{p : \mathrm{GOF}_{mp} \text{ in } (\epsilon, 1-\epsilon) \text{ quantile of all } \mathrm{GOF}_{mp}\}$, where $\epsilon \in (0, \frac{1}{2})$ is a fixed constant. Then $\hat{\phi}_m$ is estimated by

$$\sum_{p \in S_m} \mathrm{GOF}_{mp} = (1-2\epsilon)(n-1)(P/M-1) \ .$$

We set $\epsilon = 0.25$ and divide genes into 10 groups. This initialization strategy borrows information from different genes within a group to deal with the problem of small sample size typical for RNA-seq data.

2.2.3 EM Algorithm

We use EM algorithm [13, 47, 21, 55, 69] to estimate the parameters. The complete-data penalized log-likelihood is given by

$$\log L_{c,P}(\boldsymbol{\Theta}) = \sum_{j=1}^{n}\sum_{k=1}^{K} z_{kj}\{\log \pi_k + \log f_k(\boldsymbol{x_j}; \boldsymbol{\psi_{jk}})\} - \lambda \sum_{k=1}^{K}\sum_{p=1}^{P} |\log \theta_{kp}|, \tag{2.2}$$

where $z_{kj} = 1$ if sample j belongs to cluster k and $z_{kj} = 0$ otherwise. z_{kj} is treated as missing data in the EM algorithm. The EM algorithm will be developed for penalized NB mixture model and penalized Poisson mixture model respectively.

EM algorithm for penalized NB mixture model

(a) E-step

We need to compute the conditional expectation of penalized log-likelihood (2.2) for the complete data with respect to z_{kj} given data $\boldsymbol{x}$. On the $(m+1)$th iteration, this conditional expectation is

$$\begin{aligned} Q_P(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(m)}) &= E_{\hat{\boldsymbol{\Theta}}^{(m)}}(\log L_{c,P}(\boldsymbol{\Theta})|\boldsymbol{x}) \\ &= \sum_k \sum_j \hat{\tau}_{kj}^{(m)}\{\log \pi_k + \log f_k(\boldsymbol{x_j}; \mu_{jk}, \phi)\} - \lambda \sum_{k=1}^{K}\sum_{p=1}^{P} |\log \theta_{kp}| \end{aligned} \tag{2.3}$$

where $\hat{\tau}_{kj}^{(m)}$ is the posterior probability that the sample j comes from kth cluster given the estimates of other parameters from previous iterations:

$$\hat{\tau}_{kj}^{(m)} = E_{\hat{\boldsymbol{\Theta}}^{(m)}}(z_{kj}|\boldsymbol{x}) = \frac{\hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j}; \hat{\mu}_{jk}^{(m)}, \hat{\phi}^{(m)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j}; \hat{\mu}_{jk}^{(m)}, \hat{\phi}^{(m)})} \quad .$$

(b) M-step On the $(m+1)$th M-step, we firstly get the estimator of π by maxi-

mizing the leading term of (2.3):

$$\hat{\pi}_k^{(m+1)} = \sum_j \hat{\tau}_{kj}^{(m)}/n, \qquad k = 1, ..., K.$$

To avoid unnecessary computational complication, we do not estimate π_k with the whole function (2.3) and [36] showed that this method still worked well. For the same reason, the size factor s_j in the NB mixture model is computed by median ratio method [2] in advance and keeps fixed in the EM algorithm. Then we maximize (2.3) with respect to θ_{kp}, γ_p and ϕ_p for $p = 1, \ldots, P$ and $k = 1, \ldots, K$. Since it is hard to jointly maximize over these parameters, we maximize each parameter in turn with others fixed. For the NB model, the maximizers cannot be found in closed form. To find the numerical solutions, we apply the Newton Raphson (NR) algorithm which is similar to that used in [26] to compute the maximizers. Here we present the first and second derivatives of (2.3) with respect to γ_p, θ_{kp} and ϕ_p for the NR algorithm.

Calculate the first and second derivatives of (2.3) with respect to γ_p for $p = 1, \ldots, P$,

$$\frac{\partial Q_P}{\partial \gamma_p} = \sum_k \sum_j \hat{\tau}_{kj}^{(m)} \left\{ -(\phi_p^{-1} + x_{jp}) \frac{\phi_p s_j \theta_{kp}}{1 + \phi_p s_j \gamma_p \theta_{kp}} + \frac{x_{jp}}{\gamma_p} \right\}$$

and

$$\frac{\partial^2 Q_P}{\partial \gamma_p^2} = \sum_k \sum_j \hat{\tau}_{kj}^{(m)} \left\{ (\phi_p^{-1} + x_{jp}) \frac{(\phi_p s_j \theta_{kp})^2}{(1 + \phi_p s_j \gamma_p \theta_{kp})^2} - \frac{x_{jp}}{\gamma_p^2} \right\} \quad .$$

Calculate the first and second derivatives of (2.3) with respect to θ_{kp} for $p = 1, \ldots, P$ and $k = 1, \ldots, K$,

$$\frac{\partial Q_P}{\partial \theta_{kp}} = \sum_j \hat{\tau}_{kj}^{(m)} \left\{ -(\phi_p^{-1} + x_{jp}) \frac{\phi_p s_j \gamma_p}{1 + \phi_p s_j \gamma_p \theta_{kp}} + \frac{x_{jp}}{\theta_{kp}} \right\} - \frac{\lambda \pi_k}{\theta_{kp}} \operatorname{sign}(\log \theta_{kp})$$

and

$$\frac{\partial^2 Q_P}{\partial \theta_{kp}^2} = \sum_j \hat{\tau}_{kj}^{(m)} \left\{ (\phi_p^{-1} + x_{jp}) \frac{(\phi_p s_j \gamma_p)^2}{(1 + \phi_p s_j \gamma_p \theta_{kp})^2} - \frac{x_{jp}}{\theta_{kp}^2} \right\} + \frac{\lambda \pi_k}{\theta_{kp}^2} \operatorname{sign}(\log \theta_{kp}) \ .$$

Calculate the first and second derivatives of (2.3) with respect to ϕ_p for $p = 1, \ldots, P$,

$$\begin{aligned} \frac{\partial Q_P}{\partial \phi_p} = \sum_j \sum_k \hat{\tau}_{kj}^{(m)} \Bigg\{ & -\frac{\Psi(\phi_p^{-1} + x_{jp})}{\phi_p^2} + \frac{\Psi(\phi_p^{-1})}{\phi_p^2} + \frac{\log(1 + \phi_p s_j \gamma_p \theta_{kp})}{\phi_p^2} \\ & - \frac{(1 + x_{jp}\phi_p) s_j \gamma_p \theta_{kp}}{\phi_p + \phi_{kp}^2 s_j \gamma_p \theta_{kp}} + \frac{x_{jp}}{\phi_p} \Bigg\} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 Q_P}{\partial \phi_p^2} = \sum_j \sum_k \hat{\tau}_{kj}^{(m)} \Bigg\{ & \frac{\Phi(\phi_p^{-1} + x_{jp}) - \Phi(\phi_p^{-1})}{\phi_p^4} \\ & + \frac{2(\Psi(\phi_p^{-1} + x_{jp}) - \Psi(\phi_p^{-1}) - \log(1 + \phi_p s_j \gamma_p \theta_{kp}))}{\phi_p^3} \\ & + \frac{s_j \gamma_p \theta_{kp}}{\phi_p^2 (1 + \phi_p s_j \gamma_p \theta_{kp})} + \frac{(x_{jp} \phi_p^2 s_j \gamma_p \theta_{kp} + 2\phi_p s_j \gamma_p \theta_{kp} + 1) s_j \gamma_p \theta_{kp}}{(\phi_p + \phi_p^2 s_j \gamma_p \theta_{kp})^2} - \frac{x_{kp}}{\phi_p^2} \Bigg\} \ , \end{aligned}$$

where Ψ is the digamma function and Φ is the trigamma function.

Before implementing NR algorithm, we set reasonable boundaries for their domains such that in each continuous subdomains the function Q_P is continuous and twice differentiable. It is motivated by the improvement of computational efficiency and tackling the discontinuity of $\frac{\partial Q_P}{\partial \theta_{kp}}$ when $\theta_{kp} = 1$. Then we apply the bounded NR in the specified subdomain which contains the maximum. J. J. Goeman proposes a similar idea to deal with l_1 penalized likelihood optimization [26] .

On the $(m + 1)$th M-step, we compute $\hat{\theta}_{jp}^{(m+1)}$ in the subdomains: $(0, 1)$, 1 and

$(1, \max_j\{\frac{x_{jp}}{(\gamma_p * s_j)}\}]$. We assume v is a positive minimum, say $1e-7$, and let $t_1 = 0$, $t_2 = 1$ and $t_3 = \max\{\boldsymbol{x_p}/(\gamma_p^{(i)} * \boldsymbol{s})\}$ be the boundary points of subdomains. The algorithm goes through the following process.

1. If $\frac{\partial Q_p}{\partial \theta_{kp}}(\theta_{kp} = v) \leq 0$, then $\hat{\theta}_{kp}^{(m+1)} = v$. This can happen in the special case that $x_{jp} = 0$ for all $j \in C_k$.

2. Otherwise, if $\text{sign}(\frac{\partial Q_p}{\partial \theta_{kp}}(\theta_{kp} = t_2 - v)) = \text{sign}(\frac{\partial Q_p}{\partial \theta_{kp}}(\theta_{kp} = t_3 + v))$, then $\hat{\theta}_{kp}^{(m+1)} = 1$.

3. Otherwise, if $\text{sign}(\frac{\partial Q_p}{\partial \theta_{kp}}(\theta_{kp} = t_1 + v)) \neq \text{sign}(\frac{\partial Q_p}{\partial \theta_{kp}}(\theta_{kp} = t_2 - v))$, we compute the $\hat{\theta}_{kp}^{(m+1)}$ with NR in the subdomain $(0, 1)$; otherwise we compute the $\hat{\theta}_{kp}^{(m+1)}$ with the bounded NR in the subdomain $(1, t_3)$.

After $\hat{\theta}_{jp}^{(m+1)}$ is obtained, we compute $\hat{\phi}_p^{(m+1)}$ in the subdomain (v, t_p). t_p can be decided by the empirical estimate of dispersion from raw data. We suggest to use the 90 percentile of estimated dispersions but maximum since there may be some outliers that show extremely large values. Then On the $(m+1)$th M-step, we compute $\hat{\phi}_p^{(m+1)}$ as:

1. if $\text{sign}(\frac{\partial Q_p}{\partial \phi_p}(\phi_p = v)) = \text{sign}(\frac{\partial Q_p}{\partial \phi_p}(\phi_p = t_p)) < 0$, then $\hat{\phi}_p^{(m+1)} = v$;

2. otherwise, if $\text{sign}(\frac{\partial Q_p}{\partial \phi_p}(\phi_p = v)) = \text{sign}(\frac{\partial Q_p}{\partial \phi_p}(\phi_p = t_p)) > 0$, then $\hat{\phi}_p^{(m+1)} = t_p$;

3. Otherwise, $\hat{\phi}_p^{(m+1)}$ is computed by the bounded NR method.

To calculate $\hat{\theta}_{jp}^{(m+1)}$ or $\hat{\phi}_p^{(m+1)}$ in the last step of the computation process, the bounded NR method is applied in the domain (t_a, t_b) as following. For convenience, we use δ to substitute $\hat{\theta}_{jp}^{(m)}$ or $\hat{\phi}_p^{(m)}$ and use δ^{new} to substitute $\hat{\theta}_{jp}^{(m+1)}$ or $\hat{\phi}_p^{(m+1)}$.

1. Apply NR to update the parameter δ to δ^{new}.

2. If δ^{new} goes out of range (t_a, t_b) and the times of out-of-range are less than 2,

$$\delta^{new} = \begin{cases} t_a + v & \text{if } \delta^{new} < t_a \\ t_b - v & \text{if } \delta^{new} > t_b \end{cases}$$

3. If δ^{new} goes out of range (t_a, t_b) and the times of out-of-range are equal to 2,

   (a) if $\text{sign}(\frac{\partial Q_p}{\partial \delta_{kp}}(\theta_{kp} = t_a + v)) = \text{sign}(\frac{\partial Q_p}{\partial \delta_{kp}}(\delta_{kp} = (t_a + t_b)/2))$, then $t_a = (t_a + t_b)/2$;

   (b) otherwise, $t_b = (t_a + t_b)/2$;

   then update δ^{new}:

$$\delta^{new} = \begin{cases} t_a + v & \text{if } \delta^{new} < t_a \\ t_b - v & \text{if } \delta^{new} > t_b \end{cases}$$

   and reset the times of out-of-range to 0.

4. Go to step 1 until convergence.

By confining the NR into some subdomains, some drawbacks of the NR algorithm in nonconvex optimization are overcome. For our complex model, the NR algorithm may diverge in the whole real domain when the initial value is not very good (Figure 2.1). To tackle this, we confine the NR algorithm into a subdomain and shrink this subdomain when it performs divergence.

EM algorithm for penalized Poisson mixture model

(a) E-step

On the $(m + 1)$th iteration, the conditional expectation of penalized log-likelihood

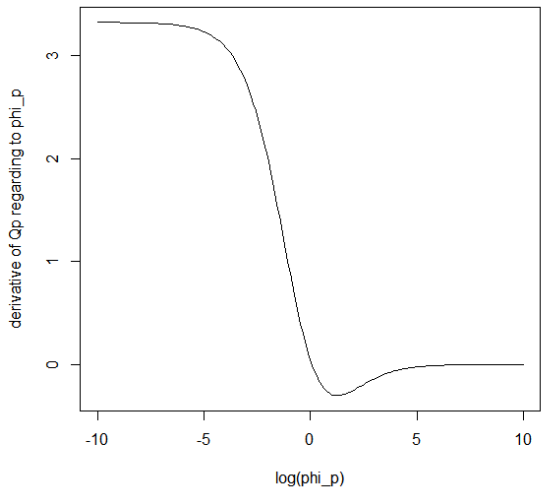


Figure 2.1: One case when update ϕ_p with the NR algorithm. The NR algorithm will be divergent if the initial value of ϕ_p is not so good.

(2.2) for the complete data with respect to z_{kj} given data $\boldsymbol{x}$ is

$$\begin{aligned} Q_P(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}^{(m)}) &= E_{\hat{\boldsymbol{\Theta}}^{(m)}}(\log L_{c,P}(\boldsymbol{\Theta})|\boldsymbol{x}) \\ &= \sum_k \sum_j \hat{\tau}_{kj}^{(m)} \{\log \pi_k + \log f_k(\boldsymbol{x_j}; \mu_{jk})\} - \lambda \sum_{k=1}^{K} \sum_{p=1}^{P} |\log \theta_{kp}| \ , \quad (2.4) \end{aligned}$$

where $\hat{\tau}_{kj}^{(m)}$ is the posterior probability that the sample j comes from kth cluster given the estimates of other parameters from previous iterations:

$$\hat{\tau}_{kj}^{(m)} = E_{\hat{\boldsymbol{\Theta}}^{(m)}}(z_{kj}|\boldsymbol{x}) = \frac{\hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j}; \hat{\mu}_{jk}^{(m)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(m)} f_k(\boldsymbol{x_j}; \hat{\mu}_{jk}^{(m)})} \ .$$

(b) M-step

On the $(m+1)$th iteration, we firstly get the estimate of π_k, which has the same form with that in the penalized NB mixture model. Due to simpler structure than NB mixture

model, the computation of size factor s_j inside EM algorithm is affordable. To compute the maximum estimate of s_j for $j = 1, \ldots, n$, we firstly solve $\frac{\partial Q_P}{\partial s_j} = 0$ and hence get

$$s_j = \frac{\sum_k \hat{\tau}_{kj}^{(m)} \sum_p x_{jp}}{\sum_k \tau_{kj} \sum_p \theta_{kp}\gamma_P - \sum_k \hat{\tau}_{kh}^{(m)}[\sum_p \theta_{kp}\gamma_p - \sum_p x_{hp}/s_h]} \ ,$$

where h is the index such that $\sum_k \hat{\tau}_{kh}^{(m)} \sum_p \hat{\theta}_{kp}^{(m)} \hat{\gamma}_P^{(m)}$ is minimum. Since $\sum_k \hat{\tau}_{kj}^{(m)} = 1, \quad j = 1, \ldots, n$,

$$s_j = \frac{\sum_p x_{jp}}{\sum_k \tau_{kj} \sum_p \theta_{kp}\gamma_P - \sum_k \hat{\tau}_{kh}^{(m)}[\sum_p \theta_{kp}\gamma_p - \sum_p x_{hp}/s_h]} \ . \tag{2.5}$$

Since $\sum_{j=1}^{n} s_j = n$, then

$$\sum_{j \neq h} \frac{\sum_p x_{jp}}{\sum_k \hat{\tau}_{kj}^{(m)} \sum_p \theta_{kp}\gamma_P - \sum_k \hat{\tau}_{kh}^{(m)}[\sum_p \theta_{kp}\gamma_p - \sum_p x_{hp}/s_h]} + s_h = n \ . \tag{2.6}$$

Although there is no closed-form estimator for s_j, we can always find the unique solution of s_h in the domain $(0, n]$ by solving (2.6) and subsequently obtain the value of s_j for $j \neq h$ by (2.5). The detailed proof will be provided in the end of this section.

By solving the equation $\frac{\partial Q_P}{\partial \gamma_p} = 0$, we can get the closed forms of γ_p for $p = 1, \ldots, P$:

$$\gamma_p = \frac{\sum_j x_{jp}}{\sum_k \sum_j \hat{\tau}_{kj}^{(m)} \theta_{kp} s_j} \ .$$

To compute the MLEs of θ_{kp} for $k = 1, \ldots, K$ and $p = 1, \ldots, P$, we still need to solve $\frac{\partial Q_P}{\partial \theta_{kp}} = 0$ and get

$$\theta_{kp} = \frac{\sum_j \hat{\tau}_{kj}^{(m)} x_{jp} - \lambda \pi_k \operatorname{sign}(\log \theta_{kp})}{\sum_j \hat{\tau}_{kj}^{(m)} \gamma_p s_j} \ . \tag{2.7}$$

Then transform (2.7) as:

$$\text{sign}(\log\theta_{kp})\{|\theta_{kp}-1|+\frac{\lambda\pi_k}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}\}=\frac{\sum_j\hat{\tau}_{kj}^{(m)}X_{jp}}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}-1 \quad . \tag{2.8}$$

From equation (2.8), we can easily obtain the relation:

$$\text{sign}(\log\theta_{kp})=\text{sign}(\frac{\sum_j\hat{\tau}_{kj}^{(m)}x_{jp}}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}-1) \quad .$$

According to this relation, we can solve (2.8):

$$|\theta_{kp}-1|+\frac{\lambda\pi_k}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}=|\frac{\sum_j\hat{\tau}_{kj}^{(m)}x_{jp}}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}-1| \quad .$$

As a result, the closed forms of θ_{kp} for $k=1,\ldots,K$ and $p=1,\ldots,P$ are

$$\theta_{kp}=\text{sign}(\log\theta_{kp})(|\frac{\sum_j\hat{\tau}_{kj}^{(m)}x_{jp}}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j}-1|-\frac{\lambda\pi_k}{\sum_j\hat{\tau}_{kj}^{(m)}\gamma_p s_j})_+ +1 \quad ,$$

where $(.)_+$ is the indicator function of positive real domain.

Theorem 1. *There exists an unique s_h in the interval $(0,n]$ that satisfies equation (2.6), where all other arguments are nonnegative constants and $\sum_k\hat{\tau}_{kj}^{(m)}=1, \quad j=1,\ldots,n.$*

Proof. For the sake of succinct inference, we firstly denote:

$$A_j=\sum_p x_{jp} \quad ,$$

$$B_j=\sum_k(\hat{\tau}_{kj}^{(m)}-\hat{\tau}_{kh}^{(m)})\sum_p\theta_{kp}\gamma_P \quad ,$$

$$C = \sum_{p} x_{hp}$$

and

$$D = \sum_{j \neq h} \frac{A_j}{B_j} \ .$$

Accordingly, equation (2.6) can be transformed to,

$$s_h + D - n - \sum_{j \neq h} \frac{A_j C}{B_j^2 s_h + B_j C} = 0. \tag{2.9}$$

Since h is the index such that $\sum_k \tau_{kh}^{(m)} \sum_p \theta_{kp} \gamma_P$ is minimum, then $B_j \geq 0$. If $B_j = 0$ for $j \neq h$, the (2.6) degenerates into a first-order linear equation and the solution can be easily computed as

$$s_h = \frac{n \sum_p x_{hp}}{\sum_{j \neq h} \sum_p x_{jp} + \sum_p x_{hp}} \ .$$

If $B_j = 0$ for some $j \neq h$ but not all, the (2.9) can be transformed as

$$s_h + D - n - \sum_{j \in G_1} \frac{A_j C}{B_j^2 s_h + B_j C} + \sum_{j \in G_2} \frac{A_j s_h}{C} = 0 \ , \tag{2.10}$$

where $G_1 = \{j : j \neq h \text{ and } B_j \neq 0\}$ and $G_2 = \{j : j \neq h \text{ and } B_j = 0\}$. Then we need to consider the nontrivial case when $B_j > 0$ for $j \neq h$ or $B_j = 0$ for some $j \neq h$ but not all. When s_h closes to 0, the left hand sides of (2.9) and (2.10) will close to $-n$. When $s_h = n$, the left side of 2.9 equals to

$$D - \sum_{j \neq 1} \frac{A_j C}{B_j^2 n + B_j C} \ . \tag{2.11}$$

Since $A_j > 0$ and $B_j > 0$ for $j \neq h$ and $C > 0$, the formula (2.11) is positive. Similarly, we can easily prove the left hand side of (2.10) is positive. Obviousy the left hand sides

of (2.9) and (2.10) are both increasing function. Consequently, there is an unique s_h in $(0, n]$ to satisfy (2.6). □

2.2.4 Model Selection and Hybrid-Hierarchical Tree

For penalized model-based clustering, it is important to determine the number of clusters K and the regularization parameter λ. Several useful selection criteria can be applied in the model-based clustering, such as AIC, BIC and some modified versions. Here we mainly use BIC ([55]) which is defined as

$$BIC = -2 \log L(\hat{\Theta}) + \log(n) d_e \ ,$$

where $d_e = K + 2P + KP - 1 - q$ is the effective number of parameters and $q = \#\{(k, p) : \hat{\theta}_{kp} = 1\}$. Besides, in our software PMixClus, we provide an alternative selection criteria EBIC [9, 10] that is defined as

$$EBIC = -2 \log L(\hat{\Theta}) + (\log(n) + \log(P)) d_e \ .$$

EBIC puts a heavier penalty on d_e and hence prefer to choose the model with more genes excluded. For the simulation and real data experiments in this thesis we only apply BIC to select model.

In many cases, we may want to visualize the hierarchical clustering structure. For this purpose, we use the hybrid-hierarchical (HH) tree guided by the result from penalized model-based clustering. The HH tree applies agglomerative clustering to the set of clusters obtained from model-based clustering. Thus it produces a partial hierarchical clustering containing only clusters coarser than the output from model-based cluster-

ing. The method was first proposed in [35, 74] and later summarized and extended in [81]. Since one objective of clustering is to identify subtypes of cells, the partial hierarchical tree can be used to investigate how the subtypes organize themselves into coarser groups. Hence, we adopt the method to build a tree for the clusters reported by our penalized mixture model.

Let K_0 be the number of clusters selected by penalized model-based clustering. When building the HH tree, in the ith merging step, we have $K_0 - i + 1$ clusters $C_1, \ldots, C_{K_0-i+1}$. The distance between two clusters, C_a and C_b is defined by

$$D(C_a, C_b) = \log \frac{\prod_{j \in C_a} f_a(\boldsymbol{x}_j; \hat{\mu}_{ja}, \boldsymbol{\phi}) \prod_{j \in C_b} f_b(\boldsymbol{x}_j; \hat{\mu}_{jb}, \boldsymbol{\phi})}{\prod_{j \in C_c} f_c(\boldsymbol{x}_j; \hat{\mu}_{jc}, \boldsymbol{\phi})} \ ,$$

where $C_c = C_a \cup C_b$ and $\hat{\mu}_{ja}$, $\hat{\mu}_{jb}$ and $\hat{\mu}_{jc}$ are the MLE based on observations from cluster C_a, C_b and C_c, respectively. The values of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_P)$ above assume the estimated values from the penalized model-based clustering. The HH tree is mainly used as a heuristic-based visualization tool to see how the clusters can be further grouped.

2.3 Results

We compare the results of our proposed method with those of PoiClaClu [78], edgeR [63] and DESeq [2] and evaluate the performances in terms of clustering accuracy and gene selection. PoiClaClu measures the distance using the power transformed sequencing data based on the Poisson log-linear model. edgeR models the read counts with NB distribution and proposes a method to compute the distance matrix based on the 500 selected genes that have the largest dispersion across all samples. DESeq uses variance stabilizing transformation of count data based on the NB model and then computes the

pairwise squared Euclidean distances.

2.3.1 Simulation Study

Simulation setup

In each data set, the read count $x_{jp}|j \in C_k$ for gene p in sample j belonging to cluster k follows the distribution NB($s_j\gamma_p\theta_{kp}, \phi_p$). The RNA-seq count data $\boldsymbol{x}$ with $P = 10000$ genes are generated from two clusters, each of which contains 10 samples. The size factor s_j is generated from Unif(0.5, 1.7) and γ_p from Exp(1/100). Among the 10000 genes, the first 3000 genes are differentially expressed between clusters. For some constant $z > 1$, in the first cluster, θ_{kp} is set to be z and $1/z$ for the first 1500 genes and the next 1500 genes, respectively. Similarly, for the second cluster, θ_{kp} is set to be $1/z$ and z, for the first 1500 genes and the next 1500 genes respectively. For the remaining 7000 genes, we set $\theta_{kp} = 1$. By the descriptions above, z represents the level of fluctuation between different clusters and we consider two values $z = e^{0.2}$ and $z = e^{0.5}$ in our simulations. For the dispersion parameter ϕ_p, we test four values: $\phi_p = 0.01$, $\phi_p = 0.1$ and $\phi_p = 0.5$, and $\phi_p = 1/(100 + \gamma_p)$, the last of which is similar to the setup used in [2] and [69]. For each setup, 50 data sets are generated.

Evaluation of clustering

We compare the clustering performances of PoiClaClu, edgeR, DESeq, and our proposed method PMixClus. Here we assume the true number of clusters is known, since the other three methods do not suggest a value for the number of clusters. In other words, for these competing methods we get the group labels by cutting the hierarchical tree at the true level. Furthermore, to allow for a fair comparison, we estimate the

size factor s_j by median ratio method in all algorithms. To assess the clustering performance, we use the rand index (RI) [60], which measures the similarity between the true clusters and the estimated clusters. Suppose that $C = \{C_1, \ldots, C_c\}$ denotes the true clusters and $S = \{S_1, \ldots, S_s\}$ denotes the estimated clusters, then RI is defined as

$$\text{RI} = \frac{a+b}{\binom{n}{2}} \ ,$$

where a represents the number of pairs of samples that are in the same cluster in C and in the same cluster in S and b represents the number of pairs of samples that are in the different cluster in C and in the different cluster in S. The higher the RI value is, the more accurate is the estimated clustering.

When $z = e^{0.5}$ (larger differences between clusters), all methods can correctly assign the samples to the two clusters for all settings of dispersion parameter ϕ_p. For $z = e^{0.2}$, we compare the RI values of different methods in Table 2.1. All methods except for edgeR could achieve correct clustering results when dispersion value is not too high (last three settings of ϕ_p). When the over-dispersion is high ($\phi_p = 0.5$), PMixClus performs best among all methods.

Table 2.1: Mean values of RIs and standard errors over 50 simulated data sets with the level of fluctuation $z = e^{0.2}$.

ϕ_p	PMixClus	PoiClaClu	edgeR	DESeq
0.5	0.990(0.074)	0.760(0.204)	0.502(0.094)	0.851(0.233)
0.1	1(0)	1(0)	0.490(0.060)	1(0)
0.01	1(0)	1(0)	0.751(0.258)	1(0)
$1/(100+\gamma_p)$	1(0)	1(0)	0.671(0.242)	1(0)

Evaluation of gene selection

To evaluate the performance of gene selection for the proposed method, we report noise features exclusion rate (NER), informative features exclusion rate (IER) and accuracy (ACC). NER is the ratio of the number of noise features excluded by a method to the number of true noise features. IER is the ratio of the number of informative features excluded to the number of true informative features. ACC is the proportion of true noise features and true informative features correctly found among all features. According to these measures, a variable selection method can be expected to achieve good performance if it gets large NER and ACC and small IER.

In Table 2.2, we summarize the NERs, IERs and ACCs of our proposed method on simulated data sets with different settings of z and ϕ_p. We obtain a good balance of NER and IER when the over-dispersion was not too high. When over-dispersion is high, many informative features are falsely excluded. However this is not unexpected since it is hard for any algorithm to distinguish between differences in expression caused by different clusters and caused by over-dispersion. Note that we can still obtain the relative high RI values shown in Table 2.1 even with high dispersion.

Dispersion parameter estimation

Dispersion parameters have a great effect on gene selection and clustering. However there are some difficulties that prevent us from obtaining accurate estimates of dispersion. Firstly, as a result of high costs of the experiment, the RNA-seq data normally has low sample size and hence it is challenging to estimate the dispersion accurately. Secondly, the fluctuation among different clusters may give rise to larger estimated values of dispersion for the DE genes. Thirdly, for the penalized mixture model, over-

estimation can result from shrinking the mean parameters. We use a robust initial value of ϕ_p as explained in Section 2 that borrows information from multiple genes. Additionally, we impose the penalty on the log-transformed θ_{kp} so that the shrinkage can be reduced when differences of DE genes among clusters enlarge.

To illustrate the estimation of dispersion, we generate one data set from each simulation setup and plot the estimates of the dispersion parameter ϕ_p in Figure 2.2. It is worth noting that estimates obtained by our method come closer to the true value of ϕ_p when $\log(\gamma_p)$ increases and the fluctuation decreases. When the true dispersion is large ((a), (b), (e) and (f) in Fig. 2.2), the proposed method can get more accurate estimates.

Evaluation of model selection

To select proper models for simulated data sets, we examine $K = 1, 2, 3, 4$ and a grid of values for λ. We apply BIC to obtain the optimal combination of λ and K. Table 2.3 shows that BIC can select the correct K in most cases ($K = 4$ never selected) except when the differences in both clusters and dispersion are large. In the latter case, this is likely due to that differences in expression caused by dispersion can be confounded with differences between cluster. However when BIC falsely selected $K = 3$, we could still build the correct HH tree in most such cases. For example, when $z = e^{0.5}$ and $\phi_p = 0.5$ for all p, BIC selects $K = 3$ in 46 simulated data sets but we can obtain correct clustering results by cutting HH trees at $K = 2$.

2.3.2 Application to Real Data

We study the performances of PMixClus and the other three competing methods on four real data sets: Liver and Kidney [46], MAQC-2 [6], Yeast [54] and Cervical Can-

cer [79]. There are only technical replicates in the Liver and Kidney and MAQC-2 data sets. The Yeast data set has both technical and biological replicates while the Cervical Cancer has only biological replicates. Additionally, the MAQC-2 data set was generated from the MicroArray Quality Control consortium and hence we can use the data set from real time reverse-transcription PCR (qRT-PCR) as the gold standard to identify the DE genes.

Introduction of real data sets

Liver and Kidney compared the expression of 22925 genes between a liver sample and a kidney sample from a human male. Seven technical replicates were generated for each sample. We extracted five replicates, which had the same library preparation (at the 3 pM concentration) for each sample. We focused on the 18228 genes whose total gene counts over all samples are not less than 5. The data set can be downloaded from a supplementary file in [46].

MAQC-2 is the mRNA-seq data set related to MicroArray Quality Control Project, comparing two types of biological samples (Brain and UHR). There were seven technical replicates with one specific library preparation for each biological sample. A subset of genes (around one thousand) were assayed by qRT-PCR ([7]). Based on the fold changes of genes in qRT-PCR data, we selected 188 genes from the subset, including 141 DE genes (fold change > 2) and 47 non-DE genes (fold change < 0.2). Then we replicated the 47 non-DE genes for five times so that the ratio of DE genes to non-DE genes is more reasonable. The sequencing data set can be downloaded from `http://bowtie-bio.sourceforge.net/recount` [22] and the qRT-PCR data can be downloaded from `www.ncbi.nlm.nih.gov/geo` with GEO Accession GSE5350.

Yeast is the RNA-seq data set, comparing the replicates of *Saccharomyces cere-*

visiae cultures. Three replicates were tested under each of two library preparation, oligo(dT) (dT) and random hexamers (RH). Specifically, there was one original replicate, one technical replicate and one biological replicate under each library preparation protocol. We focused on the 6710 genes whose total gene counts over all samples are at least 3. The data set can be downloaded from a supplementary file in [2].

Cervical Cancer is the microRNA (miRNA), 18-30 nucleotides in length, sequencing data set which were used to compare cervical cancer tissues and normal tissues. This data set included 29 samples from each of cervical cancer tissues and 29 from each of normal tissues with 714 miRNA. Among cervical cancer tissue samples, there are 21 squamous cell carcinomas (SCC), 6 adenocarcinomas (ADS) and 2 unclassified. We excluded two unclassified samples from analysis. We focused on the 636 genes whose total gene counts over all samples are at least 5. The data set can be downloaded from a supplementary file in [79].

Clustering and model selection

For Liver and Kidney and MAQC-2 data sets, all of the algorithms output the correct clustering when the number of clusters is specified to be $K = 2$. Furthermore, for the proposed PMixClus, $K = 2$ is indeed selected by BIC. Additionally, since there are only technical replicates in these two experiments, our proposed penalized Poisson mixture model can also perform correct clustering results and select $K = 2$ by BIC.

The clustering of Yeast and Cervical Cancer data sets is more challenging. For Yeast data set, the six samples fall into two known clusters (dT and RH). PMixClus and PoiClaClu obtain the same clustering results with $K = 5$. With a hierarchical representation (using HH tree for our method), all samples are correctly clustered at level $K = 3$ except for a RH biological replicate (Figure 2.3). In contrast, clustering

from edgeR and DESeq looks worse based on the respective constructed trees.

For the Cervical Cancer data set, we again identify $K = 5$ by BIC and we build the HH tree with five leaves at the bottom. To make comparison with other methods, we cut all hierarchical trees constructed by different methods at level 5 (Figure 2.4). Visually, our method matches best with the known three clusters (color coded), followed by PoiClaClu. Figure 2.5 depicts the RI values of different methods when K varies from 2 to 5. The method PMixClus(HH) is based on the tree constructed by HH method, and the method PMixClus is based on applying our method with K fixed to a value between 2 and 5 when using the BIC to choose λ only. We also present the model-based clustering result when no gene selection is performed ($\lambda = 0$). From Figure 2.5 (b), it is clear that our proposed methods obtain better clustering results than the other three, even with no gene selection, except for the case of $K = 2$ when PoiClaClu is only inferior to PMixClus(HH). Notably, in this example there are 3 known clusters.

Gene selection

From Figure 2.5(b), we can tell that the results with gene selection (the red line and the blue line) are better than those without gene selection (the black line). In other words, the penalized model performs better than the standard model. In Figure 2.6, based on MAQC-2 data we display the ratios of correctly included genes to the total number of DE genes when the number of selected genes increases (obtained using different λ), as well as the ratio of correctly excluded genes to the total number of non-DE genes. The BIC selected a model which includes 298 genes (shown as solid triangle and solid circle in the figure).

2.4 Conclusion

In this work, we proposed the penalized model-based method to accomplish clustering analysis on RNA-seq count data. Typically these data sets have the characteristics of high dimension and low sample size. Moreover, many of the features are noninformative about the cluster and hence should be automatically excluded in order to increase the accuracy of clustering. The proposed method has the desired ability of performing clustering and gene selection simultaneously. In addition, model-based approach allows us to apply some statistical model selection criteria, such as BIC or EBIC, to determine the number of clusters.

The l_1 penalty is applied in the mixture model to select genes. However, it may penalize the large values excessively [83]. This bias can also lead to larger estimates of dispersion and further result in inaccuracy of the gene selection. Some other penalties may be applied to tackle this problem, such as hard thresholding penalty or SCAD penalty ([19]). Nevertheless, these penalties can complicate the numerical computation significantly. Another possible solution to improve the performance of variable solution is to penalize the dispersion parameters at the same time. However, cautious studies must be required, because it will give an impact back on the estimation of mean parameters and further affect the clustering results. Consequently, to design better penalties in the NB mixture model for clustering with fast numerical implementation will be our future research topics.

In addition, there are some discussions on the possibility that transform count data from RNA-Seq to continuous data ([84, 39]). Since there are many well designed methods, most of which are normal-based statistical approaches, for microarray data, it is a nature thought to apply them on RNA-Seq data. In order to fulfill this purpose, two

major hindrances have to be stepped over. Firstly, the RNA-Seq count data follow a skewed distribution. Secondly, variance for each gene usually increases as its expression level becomes higher. The most common method is to log-transform count data and then standardize them. However, these standardizing log-transformed data still have unequal variabilities, especially for low coverage genes, and cannot produce perfect normal distribution though they are less skewed. In 2014, Law et al. proposed the method 'voom' to incorporate a precision weight, in which the mean-variance trend was considered, into the succeeding transformation method after normalized count data with log-counts per million[39]. These ideas provide us an alternative option to cluster samples based on RNA-Seq data, for example we can transform the data first and then use the model from [55]. However, we should be careful when we do this. It is almost impossible to get a perfect transformation which can result in a normal distribution and equal variabilities. Accordingly, the resulting bias need to be further analyzed when we use normal-based methods on transformed count data. Additionally, since the cluster labels are unknown and sometimes even the number of clusters is unknown, we have to do transformation across all samples together. It results in that variability between different clusters will be shrunk when we try to stabilize the variance. Therefore, such transformations may destroy clustering structures.

Table 2.2: Mean values and standard errors of NERs, IERs and ACCs over 50 simulated data sets when we use the true $K = 2$ or select K by BIC.

z	ϕ_p	NER(K=2)	IER(K=2)	ACC(K=2)	NER(BIC)	IER(BIC)	ACC(BIC)
$e^{0.2}$	0.5	0.985(0.042)	0.946(0.035)	0.706(0.019)	0.985(0.010)	0.935(0.031)	0.709(0.004)
	0.1	0.999(0.000)	0.960(0.005)	0.712(0.001)	0.945(0.051)	0.600(0.209)	0.781(0.039)
	0.01	0.839(0.009)	0.162(0.013)	0.839(0.005)	0.817(0.078)	0.158(0.021)	0.824(0.049)
	$1/(100+\gamma_p)$	0.790(0.028)	0.181(0.022)	0.799(0.014)	0.790(0.028)	0.181(0.022)	0.799(0.014)
$e^{0.5}$	0.5	0.966(0.113)	0.882(0.269)	0.712(0.026)	0.752(0.168)	0.180(0.181)	0.772(0.095)
	0.1	0.771(0.026)	0.041(0.083)	0.828(0.008)	0.750(0.055)	0.029(0.004)	0.817(0.038)
	0.01	0.698(0.011)	0.072(0.008)	0.767(0.007)	0.698(0.011)	0.072(0.008)	0.767(0.007)
	$1/(100+\gamma_p)$	0.628(0.049)	0.084(0.092)	0.714(0.009)	0.621(0.013)	0.071(0.007)	0.714(0.008)

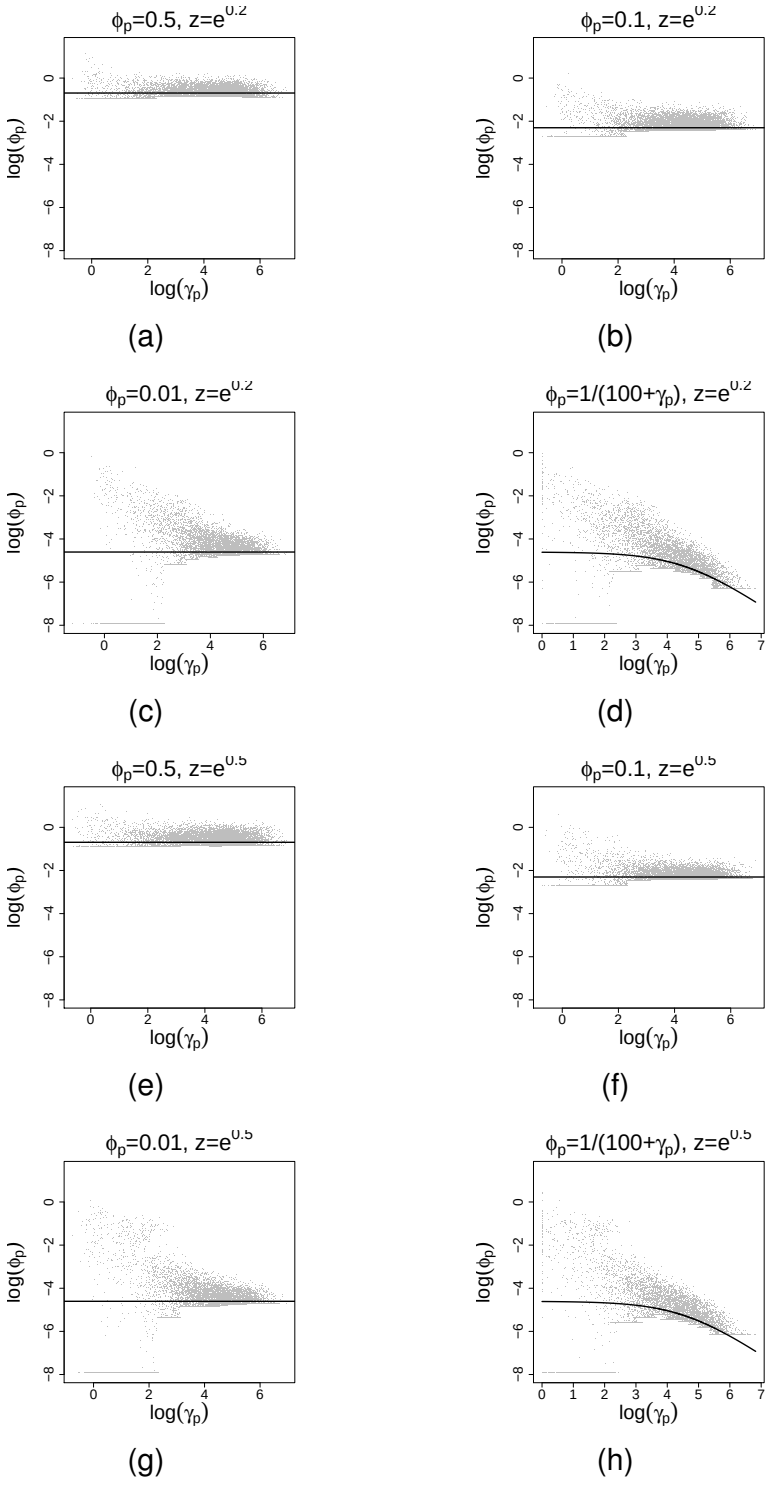


Figure 2.2: This figure displays the $\log(\phi_p)$ versus the true values of $\log(\gamma_p)$. The grey dots are the estimates from our proposed method with fixed $K = 2$. The black line represents the true dispersion value.

Table 2.3: Frequencies of the number of clusters K selected by BIC from 50 simulated data sets. The mean values of selected λ are also reported.

		$K = 1$		$K = 2$		$K = 3$	
z	ϕ_p	Freq	λ	Freq	λ	Freq	λ
$e^{0.2}$	0.5	10	0(0)	36	7.71(0.93)	4	7.39(0)
	0.1	1	0(0)	48	20.09(0)	1	14.07(1.11)
	0.01	0	–	45	20.09(0)	5	17.81(3.12)
	$1/(100 + \gamma_p)$	0	–	50	21.82(1.57)	0	-
$e^{0.5}$	0.5	2	0(0)	2	10.31(0)	46	4.84(0.96)
	0.1	1	0(0)	48	7.39(0)	1	7.39(0)
	0.01	0	–	50	14.39(0)	0	–
	$1/(100 + \gamma_p)$	1	0(0)	49	14.39(0)	0	–

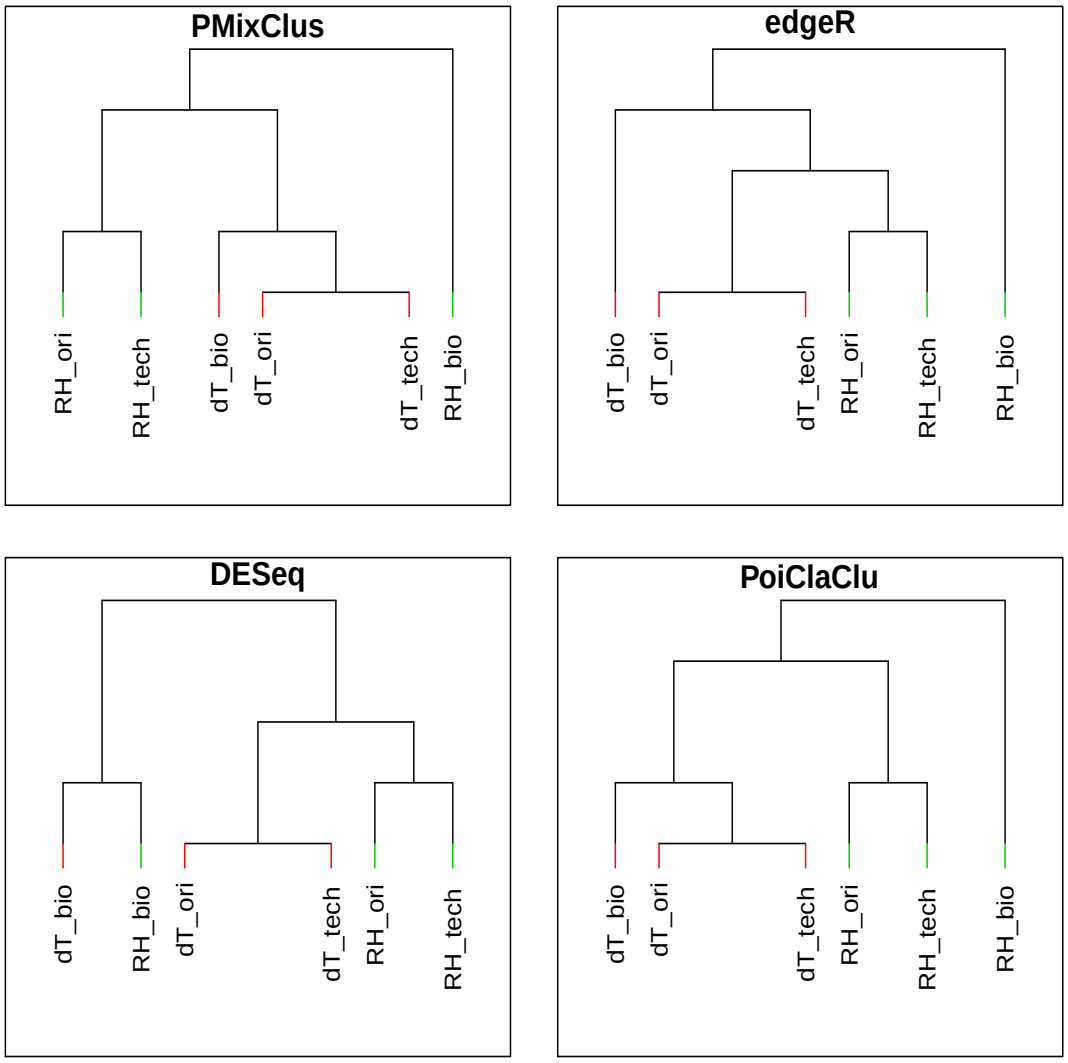


Figure 2.3: The hierarchical cluster dendrograms of Yeast data set. The dT samples and RH samples are in red and green respectively.

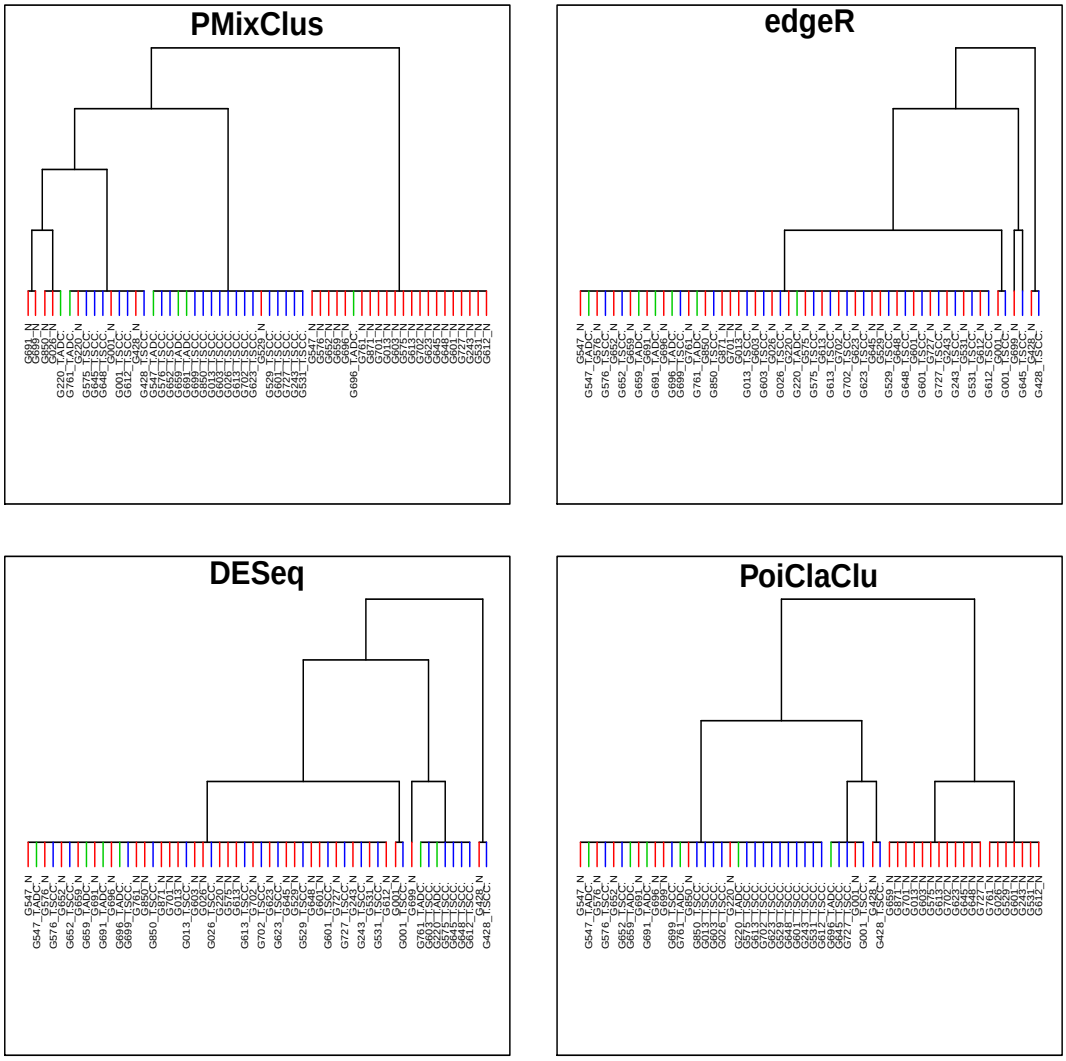


Figure 2.4: The hierarchical cluster dendrograms of Cervical Cancer data set. The dendrograms of PoiClaClu, edgeR and DESeq are from cutting the hierarchical tree at the fifth level. The ADC, SCC and normal samples are in green, blue and red respectively.

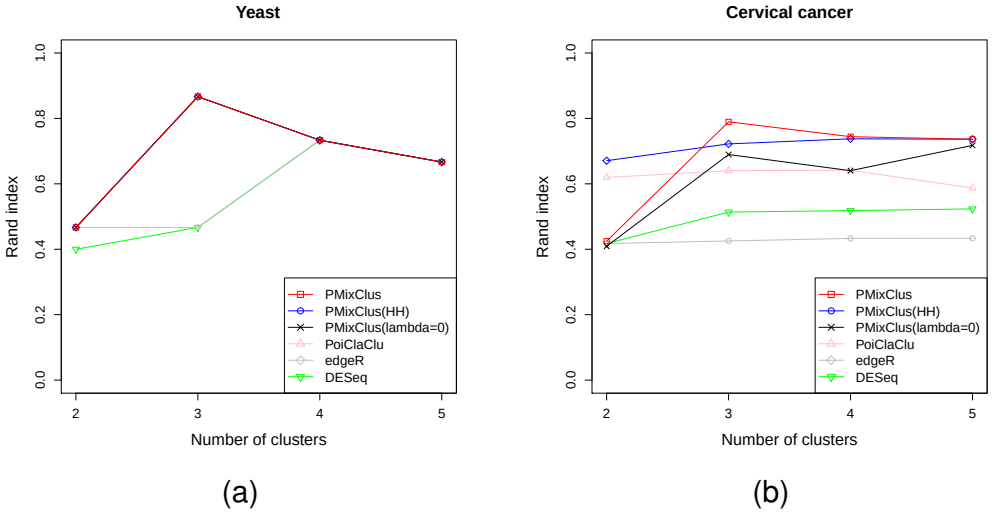


Figure 2.5: RIs by cutting the hierarchical tree at corresponding levels, except for PMixClus and PMixClus($\lambda = 0$). The curves for PMixClus and PMixClus($\lambda = 0$) show the RIs resulting from using different fixed number of clusters in our method.

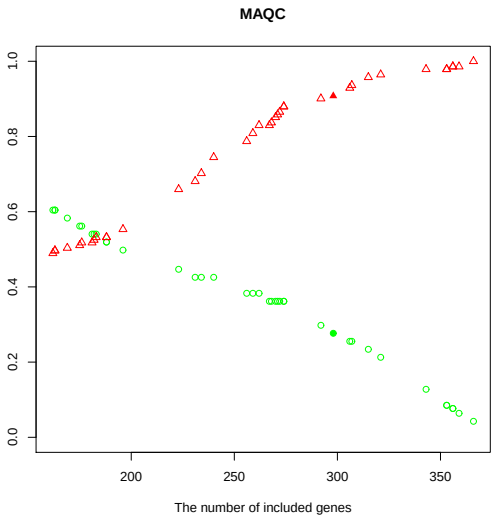


Figure 2.6: Plot for the the true positive ratio (number of correctly included DE genes/number of DE genes, red triangular) and the true negative ratio (number of correctly excluded non-DE genes/number of non-DE genes, green circle).

Chapter 3

Detection of DMRs Based on 3D Rank Clustering

3.1 Introduction

The analysis of the genome-wide methylation generally consists of two steps: aligning the BS-Seq reads onto the reference genome and detecting the DMRs. Here we focus on the detection of DMRs which is expected to give a clue and explanation for some aberrant cells, such as cancer cells. If take no account of cost, WGBS is a good method to analyze DMRs since it can present a panorama of whole genome at single resolute and hence provide all information needed for downstream analysis. In the real application, however, biologists have to reduce both the number of biological samples and the sequencing coverage per sample to make the experiment cost affordable [82]. Additionally, the incomplete bisulfite treatment and PCR lead to much more computationally complex data [48]. As a result, the low quality data is the first challenge for downstream analysis. The other challenge is to account for biological variance among

biological replicates. To detect the reproducible DMRs, which show the common features of the same biological group, the biological variance should be considered in the analysis method. Moreover, the biological variance in the cancer group is normally larger than that in the normal group [30, 66]. The Fisher's exact test p-value (FETP) [41] and BSmooth [29] are most popular methods for detection of DMRs based on the WGBS data. FETP has no treatment for low quality data and fails to consider the biological variance, whereas BSmooth utilize the local likelihood smoothing to increase the statistical power for low quality data and a signal-to-noise statistic to bring biological variance into account. However, the BSmooth assumes the biological variances are same for both cancer and normal groups. Here we will compare our method with BSmooth in simulation and real data experiments due to the more similarities between these two methods.

3.1.1 BSmooth Reviews

As we discussed in previous sections, WGBS enables a straight-forward quantification of methylation at very high resolution, but it fails to be widely applied due to the expensive genome-wide sequencing procedure, the complex output data and the lack of analysis tools [48]. By far, only a few analyzing approaches have been developed to identify DMRs with WGBS data. Among these methods, BSmooth is one of the most popular approaches and tackles most of the challenges. Here a brief review is presented.

BSmooth was the first analysis pipeline to detect DMRs from WGBS data. After aligning the WGBS reads onto the reference genome, BSmooth firstly smooth the methylation levels by using local likelihood smoothing in each sample. The smoothing window should be large enough such that at least 70 CpGs are included and it is at least

2k bases wide. For the local likelihood smoothing method, BSmooth approximates the logit function of methylation levels by second degree polynomial. In other words, BSmooth assumes the logit function of methylation levels is second degree smoothing and accordingly fits this function to reduce the bias caused by low quality data. After smoothing by moving the fixed-width window globally, BSmooth employs a statistical test (similar to t-test) to identify the differentially methylated CpG sites and form DMRs.

There are some drawbacks in BSmooth method. The second degree polynomial approximation of logit function may result in extrapolation: when one methylation level is predicted close to the boundary of one genomic region, it will be highly influenced by the continued slope from that region and usually predicted as 0 or 1 [32]. Additionally, this smoothing method may lead to the failure detection of the low-CpG-density DMRs or the small-size DMRs because methylation levels in such DMRs can be easily affected by surrounding CpGs. The other drawback is BSmooth does not consider the difference of biological variances between different biological groups.

3.1.2 Our Contribution

The WGBS technology provides the methylation information of all Cs in a genome-wide scale, regardless of their sequence contexts. Since methylation in most mammalian cells can only be found in CpG contexts, we focus discussion on detecting DMRs composed of differentially methylated CpGs purely. We note that the average sequencing coverage of WGBS data is always very low and the number of WGBS reads mapped to each CpG site is not uniformly distributed. As a consequence, it is impossible to tell the methylation status of a CpG site simply based on its methylated calls and

unmethylated calls.

In order to identify DMRs from WGBS data, it is necessary to have a proper estimation of the methylation status at each CpG site. Regarding this, the smoothing method is a good choice and has been discussed by many studies. The essential idea is to estimate the methylation level of a single CpG site based on its neighboring methylation information. However, it is not reasonable to treat the methylation level as a smoothly varied function of genomic distance, which is the fundamental assumption in BSmooth [29]. As we discussed in the previous sections, many small-size DMRs and low-CpG-density DMRs are very prone to failure in detection under such assumption. Inspired by the strategy in BiSeq [32], which is designed for reduced representation bisulfite sequencing (RRBS) data, we decided to define CpG clusters first and then to smooth methylation levels of the CpGs within each cluster. Different from BiSeq, however, our method defines a CpG cluster based on three criterions instead of the only distance condition in BiSeq. Experiments showed that our clustering approach is able to accurately determine the boundaries between genomic regions with different methylated directions.

After the CpG clusters are defined, the methylation levels within each single cluster are estimated by a modified local kernel smoothing method. To estimate the methylation of a single CpG site, the methylation levels of neighbor CpGs in the pre-clustered regions are utilized. By restricting the smoothing window in pre-clustered regions, our method can overcome the drawback introduced by BSmooth for the low-CpG-density DMRs or the small-size DMRs if such DMRs are prominent in the surrounding CpGs. Compared with normal kernel smoother, ours considers the influences of both the genomic distance between two CpGs and the sequencing coverage of each neighbor CpG site. In other words, the neighbor CpGs which are closer to the targeted CpG and have

higher quality will take more effects on that targeted CpG. Besides, we always put the heaviest weight to the targeted CpG to avoid the too large effect brought by some neighbor CpGs with extremely large coverage. Additionally, our smoothing method is not function based and hence there is no extrapolation found in BSmooth. Therefore, more information is remained and utilized by our smoothing method compared to the others.

Moreover, to provide a powerfully statistical analysis of methylation, we model the number of methylated calls at each CpG site to follow the beta-binomial distribution so that the biological variation among samples is considered. In the subsequent hypothesis test, the Wald test is applied at each single CpG site to identify whether it has significantly different methylation states between the biological groups. After all the CpGs in the pre-clustered regions are scanned and examined, we prune away non-differentially methylated CpGs (non-DMCs) at the edge of each cluster and merge the adjacent DMRs which have the same methylation direction.

To sum up, we have developed a novel approach to detect DMRs from WGBS data and implemented it into an R package named as DMReSearch. It can be freely downloaded at `https://github.com/TianYe00/DMReSearch.git`. DMReSearch consists of four main steps: three-dimensional rank clustering, modified local kernel smoothing, model building and testing, and boundaries defining. Figure 3.1 depicts the working scheme of our method. The following sections will discuss each step in detail.

3.2 Methods

3.2.1 Data Preparation

After mapping the WGBS reads onto the reference genome, we are able to count the methylated calls and unmethylated calls of every CpG in each sample directly from the alignment results. For convenience, we define the *methylation count* of a CpG to be the number of its methylated calls and its *total count* to be the sum of methylated calls and unmethylated calls. Accordingly, dividing the total count by the methylation count at a CpG site results in its methylation level. In addition, we say a CpG site is *covered* in a sample if its corresponding total count is non-zero. We also call the total count of a CpG as its *coverage*.

We note that there always exist some CpG sites that are not covered in every sample, which is believed as a result of the low sequencing coverage of WGBS technology. In case of the disturbance from non-informative CpGs, we perform a data filtration step: only the CpGs that are covered by an adequate number of alignments in most samples for each biological group are remained. The methylation counts, total counts (or coverages), the strand information and the corresponding genomic locations of all the remained CpGs are used as input data in the subsequent procedures.

3.2.2 Three Dimensional Rank Clustering

In DMReSearch, we propose an alternative approach to define CpG clusters. Like BiSeq, spacial distances between CpGs act as a key factor in clustering. Instead of simply grouping the close CpGs throughout each chromosome, we attempt to find the cluster centers first. Inspired by the clustering approach proposed in [64], our method

also considers the local density of each CpG site and its minimum distance to a higher-density CpG. However, different from [64], we introduce a methylation-specific criterion to find the cluster centers. Therefore, we have to discuss every CpG in three dimensions with respect to the three criterions during the clustering process. In the below, we give detailed discussion in the scenario of a single chromosome.

Suppose there are totally n CpGs remained after the filtration step. For the ith CpG, we denote its genomic location by l_i. Then, given any two different CpGs with their genomic locations l_i and l_j, the distance between them is $d_{ij} = |l_i - l_j|$. Based on these notations, we are ready to define the *local density* of the ith CpG as,

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \ .$$

Here, $\chi(d)$ is an indicator function of the set $\{d \leq 0\}$ and d_c is a self-defined cutoff distance. In other words, $\chi(d) = 1$ if $d \leq 0$ and $\chi(d) = 0$ otherwise. The local density of a CpG site actually counts its neighbor CpGs within a distance of d_c. By default, d_c =300bp in DMReSearch. The clustering dimension ρ_i is hence defined.

Once the local densities of all CpGs are computed, we define another clustering dimension δ_i. If the local density of the ith CpG is not the largest, δ_i is the minimum distance between the ith CpG and any other CpGs whose local densities are larger than ρ_i. Otherwise, δ_i is defined specially to be the maximum distance from the ith CpG to any other CpG sites. To summarize,

$$\delta_i = \begin{cases} \min\limits_{\{j:\rho_j > \rho_i\}} (d_{ij}) & \text{if } \rho_i \neq \rho_{max} \ , \\ \max\limits_j (d_{ij}) & \text{otherwise} \ , \end{cases}$$

where ρ_{max} means the maximum local density. This dimension factor is used to avoid that adjacent centers are too close.

The third dimension is defined according to the fluctuations of differences of the mean methylation levels at a CpG site in different biological groups. Suppose that we have two groups of samples. In particular, we define group 1 to be the control group containing normal samples and group 2 to be the case group (or cancer group) with respect to cancer samples. For the ith CpG, we first calculate its mean methylation level of all samples in each group and we denote these two mean methylation levels by $\bar{y}_{ik}$, $k = 1$ or 2. Then the difference between these two means can be easily obtained, which is $\lambda_i = \bar{y}_{i2} - \bar{y}_{i1}$. Note that λ_i has a range between -1 and 1. After the mean difference λ_i is computed for every CpG, for ith CpG we compute the standard deviation of all λ_j's that correspond to the nearest $2R$ CpGs around the ith CpG, i.e.,

$$\sigma_i = \mathrm{SD}(\{\lambda_j : j \in I_i\}), \quad I_i = \{j : |j - i| \leq R\}$$

where $SD(\Lambda)$ is a function to calculate the standard deviation of the values in set Λ and the value of R depends on the average density over all CpGs. Finally, we have the third dimensional variable ϕ_i defined as,

$$\phi_i = \exp(10 \cdot \sigma_i) \ .$$

The reasonable selection of cluster centers should be the CpG sites having larger ρ_i and δ_i but smaller ϕ_i. To quantify the selection criteria, we introduce a score γ_i for the

ith CpG that can be computed by the following formula,

$$\gamma_i = \frac{\rho_i \cdot \delta_i}{\phi_i} \quad .$$

Thus finding cluster centers is equivalent to choosing the largest scores. One straightforward way is to sort $\{\gamma_i, i = 1, \ldots, P\}$ in a decreasing order, where P is the total number of CpGs. Suppose that the ordered score set is $\Gamma' = \{\gamma_{(j)}, j = 1, \ldots, P\}$ and c_T cluster centers are required for the subsequent analysis. The selected cluster centers are exactly the CpGs locating at $\{l_{(j)}, j = 1, \ldots, c_T\}$ that correspond to the first c_T largest scores in Γ'.

After the cluster centers are found, we extend each center CpG in both forward and backward directions to form a candidate CpG cluster. Our extension strategy depends on the fact that CpGs locating within a distance of 500bp have highly correlated methylation levels [16]. Accordingly, we repeatedly extend a cluster by including non-center CpGs whose minimum distance to the existed CpG sites in the cluster is at most d_c. In order to reduce the cases of incorrect extension, we start from two adjacent cluster centers and make extension inside their interval region at the same time. For the rightmost (or leftmost) center CpG, the extension on the right (or left) works simply based on the initial distance condition. At last, we discard the CpG clusters that contain less than three CpGs.

Considering the large size of genomic data, we design an iterative procedure to define all potential CpG clusters. In brief, we repeat the above steps to find clusters from the remaining CpG sites that are not covered by any cluster until we cannot find any cluster containing at least three CpGs. The whole procedure of defining CpG clusters is summarized in Algorithm 1. In each iterative step, the default value of c_T is 2% of

the total number of remaining CpGs.

Our three-dimensional rank clustering approach works much better than clustering CpGs simply based on their spacial distances like [32]. To illustrate the advantages of our approach, we take a toy example shown in Figure 3.2. We extracted a part of real data corresponding to 63 CpG sites in a DNA segment on the human chromosome 22. The distance between any two adjacent CpGs is at most 300bp. We simulated a hypermethylated region consisting of 39 CpGs. If we make use of the clustering method merely based on the distances, all these CpG sites will be assigned into one cluster. Whereas, our clustering approach accurately separates the CpG sites into two clusters, one of which is exactly the DMR.

The toy example also illustrates the significance of the third dimension in our clustering approach, which considers the variance of methylation levels at each single CpG site. If we cluster the CpGs simply based on the two distance-related dimensions, we will obtain three clusters, one of which spans across both the DMR and non-differentially methylated (non-DM) region. Accordingly, we believe that the third dimension is crucial to improve the accuracy of defining the DMRs and provide more reliable clustering information for the downstream statistical analysis.

3.2.3 Modified Local Kernel Smoothing

As we discussed previously, it is too costly to design a perfect WGBS protocol that provides adequate samples and sufficient sequencing coverage. Therefore, the methylation levels obtained directly from the alignment of WGBS reads take too few information to suggest correct DMRs between different biological conditions. In practice, the foremost task of DMR detection with WGBS data is to improve the data quality. One well-

Algorithm 1 3*DRCluster*

Input: A set $\mathcal{D} = \{l_i, i = 1, 2, \ldots, n\}$ of genomic locations of CpGs, integers d_c and c_T.
Output: A list $\mathcal{R}$ of CpG clusters.

$\mathcal{F} \leftarrow \mathcal{D}$
$I_F = \{i : l_i \in \mathcal{F}\}$
repeat
  Calculate $\Gamma = \{\gamma_i : i \in I_F\}$.
  Rank Γ in a decreasing order $\rightarrow \Gamma' = \{\gamma_{(i)} : i \in I_F\}$.
  Select cluster centers $\mathcal{O} = \{l_{(i)} : i = 1, 2, \ldots, c_T\}$.
  Sort $\mathcal{O}$ increasingly $\rightarrow \mathcal{O} = \{O_j : j = 1, 2, \ldots, c_T\}$.
  Find two boundaries of each cluster I_j with center $O_j \in \mathcal{O}$.
  $\mathcal{I} \leftarrow \{I_j : j = 1, 2, \ldots, c_T\}$
  $\mathcal{D}_c \leftarrow \{l_i : l_i$ is covered by some cluster $I_j\}$.
  $\mathcal{F} \leftarrow \mathcal{F} \setminus \mathcal{D}_c$
  Delete I_j with less than 3 CpGs from $\mathcal{I}$
  $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{I}$
until $\mathcal{I} == \emptyset$
return $\mathcal{R}$

accepted technique depending on the spacial correlation of methylation levels (shown in [33, 16]) is to select a kernel smoother to estimate and adjust the methylation level of every single CpG site in a single sample. Briefly speaking, a smoothing procedure borrows the methylation information of neighboring CpGs to enhance the accuracy of the estimated methylation level at a specific CpG.

In our approach DMReSearch, we propose a modified local kernel smoother inspired by AKSmooth [11] and Nadaraya-Watson estimator [53, 77]. The most attractive characteristics of our kernel smoother are: (i) smoothing is performed within each single pre-clustered region; (ii) the kernel weights the real genomic distances between CpG sites; (iii) the coverage at each single CpG site is also considered as weights. Let us discuss the details below.

For the jth sample, we denote the methylation count and the total count of the ith CpG with m_{ij} and n_{ij}, respectively. We assume that each m_{ij} follows a binomial

distribution $m_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$, which agrees with the assumptions in many other studies [29, 32]. The success probability p_{ij} actually implies the true methylation level of the ith CpG, which is the proportion of calls having methylation at the ith CpG in the sample j. It is effortless to obtain an unbiased estimate of p_{ij} to be the fraction $y_{ij} = \frac{m_{ij}}{n_{ij}}$.

Furthermore, depending on the spacial correlation of methylation levels, we may suppose p_{ij} to be a sample-cluster-specific function $f_h(l_i)$ over location l_i with bandwidth h in the gth CpG cluster. We estimate $f_h(l_i)$ (i.e., p_{ij}) with the following smoother,

$$\hat{p}_{ij} = \hat{f}_h(l_i) = \frac{\sum_{t \in C_g} K_h(l_i, l_t) \cdot W_i(t) \cdot y_{ij}}{\sum_{t \in C_g} K_h(l_i, l_t) \cdot W_i(t)} \ ,$$

where $C_g = \{t : t\text{th CpG locates in the } g\text{th cluster}\}$,

$$K_h(l_i, l_t) = K_h\left(\frac{|l_t - l_i|}{h}\right) \ ,$$

and

$$W_i(t) = \begin{cases} \log(n_{tj} + 1) & \text{if } t \neq i \ , \\ \max_{t \in C_g} \log(n_{tj} + 1) & \text{if } t = i \ . \end{cases}$$

In the smoother $\hat{f}_h(l_i)$, the triangular kernel K_h is applied to assign different weights with respect to the genomic distances between the ith CpG and its neighboring CpGs in the same cluster. The closer the two CpGs locate, the higher weight is assigned. Comparatively, AKSmooth adopts an order-distance in its kernel function, which is defined according to the indices of CpG sites. The order-distance between the ith CpG and the tth CpG is then $|t-i|$, regardless of their practical genomic distance. However, it is not surprising that the kernel might blow up the influence of some CpGs that locates farther from the ith CpG than the tth CpG but have the same or smaller order-distance

as the tth CpG. Therefore, taking the real genomic distance into account may be more reasonable.

Besides, another component W_i of the smoother $\hat{f}_h(l_i)$ weighs the coverage information of each neighboring CpG. Since the CpG sites with higher coverages are more likely to give rise to more reliable estimation of their methylation levels, we allow the CpGs with higher coverages inside the neighborhood of the specific ith CpG to make more contributions to an accurate estimation of its methylation level. Therefore, we are able to utilize the given knowledge of methylation more sufficiently and hence provide the more comprehensive and precise information in the downstream analysis.

Compared to the global smoothing strategy in BSmooth, our cluster-based local smoothing approach is standout to define the boundaries of DMRs. We employ the same toy example in Figure 3.2 to illustrate this advantage of our method. According to the smoothing results of BSmooth (see Figure 3.2(b)), it is difficult to find out the clear boundaries of both the non-DM region and the DMR. We further note that BSmooth incorrectly increased the methylation levels of those CpGs near the right end of the non-DM region. It should be attributed to the basic assumption of BSmooth that the methylation levels vary smoothly along the whole DNA sequence. In contrast, the smoothing results of our method (see Figure 3.2(c)) present the boundaries of the two genomic regions clearly and correctly.

3.2.4 Model and Hypotheses Test

When we analyze the methylation in a single sample (see the previous discussion of smoothing), based on the binomial distribution the methylation level of a single CpG site can be simply estimated by the fraction of its methylation count over its total count.

However, it becomes insufficient to apply the same modeling approach when the differences of methylation states are discussed across distinct samples. In other words, the biological variance of methylation levels has to be considered for each single CpG site among all samples under different conditions. Therefore, we build a hierarchically structured model to take biological variance into account.

We assume that the methylation count m_{ij} of the ith CpG in sample j follows the binomial distribution:

$$m_{ij} \sim \text{Binomial}(n_{ij}, p_{ijk}) \ ,$$

where n_{ij} is the total count and p_{ijk} implies the true methylation level of the ith CpG in sample j in the biological group k, for all $i = 1, ..., P$, $j = 1, ...N$ and $k = 1, 2$. Furthermore, we note that beta distribution is widely recognized suitable to describe the methylation levels among distinct samples due to its flexibility and precision to present the overdispersion brought by biological replicates [20]. Therefore, we apply beta distribution to model the methylation level p_{ijk}:

$$p_{ijk} \sim \text{Beta}(\alpha_{ik}, \beta_{ik}) \ .$$

For the convenience of illustration, we substitute the parameters of the beta distribution with the mean $\mu_{ik} = \frac{\alpha_{ik}}{\alpha_{ik}+\beta_{ik}}$ and the dispersion $\phi_{ik} = \frac{1}{1+\alpha_{ik}+\beta ik}$. Accordingly, m_{ij} follows the Beta-Binomial distribution with parameters μ_{ik}, ϕ_{ik}:

$$m_{ij} \sim \text{Beta-Binomial}(\mu_{ik}(\frac{1}{\phi_{ik}} - 1), (1 - \mu_{ik})(\frac{1}{\phi_{ik}} - 1), n_{ij}) \ .$$

In the subsequent step, the parameters in the beta-binomial model are required to be estimated. We note that the structure of the beta-binomial distribution is too complex to

have any close form of the maximum likelihood estimator for either μ_{ik} or ϕ_{ik}. For the sake of efficient computation, we employ the method of moment to estimate μ_{ik}, that is,

$$\hat{\mu}_{ik} = \frac{\sum_{j \in G_k} \tilde{m}_{ij}}{\sum_{j \in G_k} n_{ij}} \ ,$$

where $\tilde{m}_{ij} = \hat{p}_{ijk} n_{ij}$ is the pseudo-count of methylation after smoothing, and $G_k = \{j : \text{sample } j \text{ belongs to the biological group } k\}$. We further notice that the low sample size in a WGBS experiment leads to an increasing number of challenges to accurately estimate the dispersion ϕ_{ik} of each single CpG. To tackle such difficulty, we presume that CpGs locating inside each pre-clustered genomic regions share the same dispersion parameter but differentiate the parameter in the control group from that in the cancer group. Under this assumption, let ϕ_{gk} denote all ϕ_{ik}'s in the CpG cluster g for the biological group k. Then the dispersion parameter ϕ_{gk} can be estimated by utilizing the beta-binomial goodness of fit statistic:

$$GOF_{ig} = \sum_{j \in G_k} \frac{\tilde{m}_{ij} - n_{ij}\hat{\mu}_{ik}}{n_{ij}\hat{\mu}_{ik}(1 - \hat{\mu}_{ik})[1 + (n_{ij} - 1)\phi_{gk}]} \ .$$

Since m_{ij} is independently beta-binomial distributed, $\sum_{i \in C_g} GOF_{ig}$ approximately follows the χ^2 distribution with $(N_k - 1)(N_g - 1)$ degrees of freedom, where N_k is the number of samples in the biological group k and N_g is the number of CpGs in cluster g. To avoid the damaging effects brought by outliers, we only consider the CpGs corresponding to the set $S_g = \{t : GOF_{tg} \text{ is } (\epsilon, 1 - \epsilon) \text{ quantile of all } GOF_{ig}\}$, where $\epsilon = 0.1$ is a fixed constant. Therefore we can have the estimator of ϕ_{gk} by solving the following equation:

$$\sum_{i \in S_g} GOF_{ig} = (1 - 2\epsilon)(N_k - 1)(N_g - 1) \ .$$

Note that our approach to parameter estimation is different from that of BSmooth. BSmooth makes no effort to distinguish the biological variation between the normal and cancer samples. In contrast, we discuss distinct dispersion parameters for two sample groups respectively, which is based on the observations that methylation levels in cancer samples usually have larger dispersion than the methylation levels in normal samples [30, 66].

Finally, we employ the hypothesis test H_{i0}: $\mu_{i1} = \mu_{i2}$ to tell whether a CpG is differentially methylated between two biological groups. We form the Wald test statistic w_i for the ith CpG in every pre-clustered region,

$$w_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{\widehat{var}(\hat{\mu}_{i1}) + \widehat{var}(\hat{\mu}_{i2})}} \ ,$$

where

$$\begin{aligned} \widehat{var}(\hat{\mu}_{ik}) &= \frac{\sum_{j \in G_k} \widehat{var}(\tilde{m}_{ij})}{(\sum_{j \in G_k} n_{ij})^2} \\ &= \frac{\sum_{j \in G_k} n_{ij} \hat{\mu}_{ik}(1 - \hat{\mu}_{ik})[1 + (n_{ij} - 1)\hat{\phi}_{ik}]}{(\sum_{j \in G_k} n_{ij})^2} \ . \end{aligned}$$

Based on the simulation data set, we make use of Q-Q plot to derive the distribution of our beta-binomial Wald test statistic w_i under the null hypothesis. According to this Q-Q plot and the density histogram of Wald test statistics, we find that w_i approximately follows the normal distribution (see Figure 3.3). The heavy tail in Q-Q plot is from non-DMCs which are falsely included in the DMR. Since the non-DMCs has much less dispersion than DMCs, the estimations of their dispersions are inaccurate when we try to borrow information of all CpGs in the same pre-cluster by goodness of fit statistics. However, Figure 3.4 and Figure 3.5 show that our method can still control

the Type I errors well. At last, we define the DMCs to be those CpGs whose p-values are less than a self-defined significance level.

3.2.5 DMRs Identification

Given all the DMCs in all pre-clustered regions, we firstly prune away the non-DMCs in the edge of each pre-clustered region. Secondly we examine the distances between any two adjacent DMCs. If the distance exceeds 300bp, the corresponding two DMCs will be assigned into two new clusters. Thirdly we check the direction of each DMC. If a switch of methylation directions exist between two adjacent DMCs, these two DMCs are separated into two clusters. Fourthly after all pre-clustered regions are trimmed and divided to new regions, we compute the distance between two neighboring clusters, which is actually the minimum distance between their boundary CpGs. If two clusters are close enough to each other, say at most 300bp, and have the same methylation direction, they will be merged into one cluster. At last, we throw away those tiny clusters with less than three CpGs or those clusters with not large enough proportion of DMCs (80% by default). The remaining clusters are thus reported as the candidate DMRs.

3.3 Results

We carried out both simulation and real data experiments to test the performance of our method DMReSearch. By now, we only compared DMReSearch with BSmooth, because BSmooth is the only software package to detect DMRs with smoothing based method and designed for WGBS data.

3.3.1 Simulation Study

Simulation Setup

Considering the complexity of WGBS data, we borrowed information from the data set experimented by Hansen et al. [30]. In their experiments, two biological groups were tested, each of which had three samples. We extracted the data corresponding to the CpGs locating at chr22. Among these CpG sites, we only remained those CpGs whose coverage was more than one in at least two samples of each biological group. We took these CpGs as the template to generate the simulation data set. As a result, we obtained the template data with 177,342 CpGs.

We firstly simulated the starting locations and ranges of the DMRs. In this step, we selected the CpGs whose local densities were at least three, from which the centers of DMRs were randomly chosen by following the Bernoulli distribution with the success probability 0.04. The length of a DMR was generated from the normal distribution Normal$(400, 400^2)$, but truncated by the upper bound 2000bp and the lower bound 10bp. Then starting from the center CpG of each DMR, the extension step was performed by including the neighboring CpGs that satisfied both of the two conditions: (i) the distances from the center should not exceed half of the DMR length; (ii) the minimum distance from the CpGs within the DMR should be at most 300bp. Subsequently, we merged the DMRs that intersected with each other or whose boundaries had a distance of less than 300bp in between. Moreover, we discarded those DMRs that containing less than three CpGs. Finally, we labeled the resulting DMRs as hypermethylation or hypo-methylation with equal probability.

Next, we simulated the methylation level of each CpG. On the one hand, we simulated the methylation levels for DMCs. For the ith CpG, if it located in a hyper-

methylated DMR, the methylation levels p_{ij2} of normal samples would be generated from the beta distribution Beta(a, b) with mean $\frac{a}{a+b}$ while the methylation levels p_{ij1} of the cancer samples would be the addition of p_{ij2} and the methylation difference Δp_i. If the ith CpG located in a hypo-methylated DMR, the methylation levels p_{ij2} of normal samples would be obtained according to Beta(b, a) with mean $\frac{b}{a+b}$ and the methylation levels p_{ij1} would be computed by subtracting Δp_i from p_{ij2}. With respect to CpGs inside DMRs, we simulated their methylation levels by different settings of both (a, b) and Δp_i that explains the distinct biological variances and data (or signal) qualities respectively. We used $a = 20$ and $b = 80$ to simulate the case of lower biological variance, while $a = 2$ and $b = 8$ for the case of higher biological variance. In addition, we generated the methylation differences Δp_i from the uniform distribution $Unif(0.15, 0.3)$ to indicate the data with weak signal, but $Unif(0.25, 0.4)$ for strong signal data. Particularly, if the simulated methylation levels exceeded the range between 0 and 1, they would be specially defined as 0.01 or 0.99 depending on their values.

On the other hand, the methylation levels of non-DMCs were derived by three steps. Firstly, we applied the DMReSearch to smooth the methylation levels in the template data set with 500bp window size. Then we computed the mean values $\bar{p}_i$ across the three samples in the control groups. Finally, we generated p_{ijk} from Beta$(100\bar{p}_i, 100(1-\bar{p}_i))$ if the corresponding data set had the lower biological variance; otherwise, Beta$(10\bar{p}_i, 10(1-\bar{p}_i))$ was applied for the case of higher biological variance.

After the methylation levels were assigned, we continued to simulate the total count and methylation count for each CpG. The simulated total count n_{ij} of the ith CpG in the jth sample was obtained by adding 5 or 15 to its true total count in the template data set. As a result, we were able to have two types of setup in terms of the average sequencing coverage: the low coverage 8× and the high coverage 18×. Once the total

counts were generated, the methylation counts of all the CpGs in the biological group k could be created by employing $m_{ij} \sim \text{Binomial}(n_{ij}, p_{ijk})$.

Note that we increased the values of total counts instead of utilizing the true values of the template data set. It should be attributed to the extremely low coverage (3×) of the template data set, which could cause larger biases of the simulated methylation counts that follow the binomial distribution. For example, if the total count is 1, the methylation count will be either 0 or 1 regardless of what the true methylation level is. We can also consider the more extreme situation when the total count is 0. In these two cases, the bias may be amplified when the sample size is very low (just 3 samples in each group).

To sum up, we totally generated eight different simulation setups by various combinations of coverages (8× or 18×), biological variances (lower or higher) and signal qualities (strong or weak).

Evaluation

For the sake of fair comparison, we used the same window width of 100bp in the smoothing process to run both BSmooth and DMReSearch. We compared the DMCs reported by each software tool with the true DMCs that we simulated. If a true DMC is correctly identified by a tool, it will be considered as a true positive (TP). If a true DMC is not identified by a tool, it will be treated as a false negative (FN). Similarly, a true negative (TN) and a false negative (FP) indicate a true non-DMC that is correctly reported as non-DM and that is falsely identified as DM, respectively. To evaluate the performance of either tool, we plotted the ROC curve for each of the eight simulation experiments. Figure 3.4 depicted the detection results of BSmooth and DMReSearch on the low-coverage data sets. We noticed that both software tools achieved more pre-

cise results when the biological variation was lower or the signal quality was stronger. In all four experiments, DMReSearch obviously achieved the more accurate detection with respect to higher true positive rate (TPR) and lower false positive rate (FPR). In particular, We noticed that DMReSearch outperformed BSmooth notably in the most disadvantageous situation where the data set had lower sequencing coverage, higher biological variance and weak signals (Figure 3.4(a)). We believed that the superior performance of DMReSearch was contributed by its cluster-based smoothing approach. We also found that BSmooth could hardly achieve higher TPR when FPR was required as about 2%. In contrast, our software DMReSearch could guarantee TPR over 60% in this case. When we compared their results in the experiments with high-coverage data sets (see Figure 3.5), we observed that both DMReSearch and BSmooth achieved approximately the same high accuracies. However, DMReSearch performed a little better than BSmooth, especially requiring smaller FPRs ($< 2\%$). Overall, the simulation study showed the strong ability of DMReSearch to accurately detect DMRs, even in the worst case when the data set had low sequencing coverage, the samples had high biological variation but the supporting signals of DMRs were quite weak.

3.3.2 Application to Real Data

To test our software package DMReSearch on real data, we downloaded the WGBS reads of the study of colorectal cancer from NCBI Sequence Read Archive (SRA). The access number of the study is SRP006774. There are totally two groups of WGBS reads: one is the normal colonic sample group and the other is the colorectal cancer group. Each group contains three experiments and each experiment has two runs. In addition, the six biological samples are also experimented by capture bisulfite sequenc-

ing that can achieve higher sequencing coverage at targeted genomic regions. After aligning the reads onto the reference human genome hg19, we computed the total count and methylation count of each CpG in each biological sample and extracted the data corresponding to chromosome 22 for comparison. On average, with respect to chromosome 22 the coverage of WGBS data is 2× and that of capture bisulfite sequencing is 16×.

In the subsequent analysis, we applied the default settings of arguments in DMReSearch and BSmooth. Specially, DMReSearch reported DMCs with the default significant level of 0.01 and BSmooth selected its top 0.05 quantile DMCs by default. At last, DMReSearch detected 5599 DMCs that composed 1235 DMRs, while BSmooth identified 3714 DMCs and 926 DMRs.

Assessment of Smoothing Methods

For the high coverage data set from capture bisulfite sequencing, we filtered out those CpGs with low coverage and only kept the CpGs whose total counts are more than 9 in at least two samples of either the cancer group or the normal group. To compare with low coverage WGBS data, we only used the CpGs that were also reported in the low coverage data. The methylation levels of the remained CpGs were regarded as the benchmarks to assess the smoothed methylation levels of the low coverage WGBS data set. We compared the smoothing tools in BSmooth and DMReSearch with their default arguments. For each data set, we computed the two mean values for every CpG of smoothed methylation levels across all samples in the normal group and the cancer group, respectively. We subsequently calculated the correlations between the mean values corresponding to the low coverage data set and those to the high coverage data set. At last, we employed the correlations to evaluate the performance of the smoothing

tools in both software packages. The higher the values of correlation, the more reliable the smoothed low coverage data from WGBS. BSmooth finally got the correlation 0.697 for the cancer group and 0.709 for the normal group. Comparatively, DMReSearch achieved the values of correlation 0.839 and 0.822 for the cancer and normal groups, respectively. Similarly, we also calculated the detailed correlation values for all six samples between these two sets of data and the results were established in Table 3.1. It was not surprising at all that the smoothing method of DMReSearch could provide more reliable estimated data for the downstream analysis to detect DMRs.

Tool	Cancer1	Cancer2	Cancer3	Normal1	Normal2	Normal3
DMReSearch	0.682	0.768	0.718	0.670	0.722	0.699
BSmooth	0.639	0.670	0.622	0.622	0.696	0.652

Table 3.1: Correlation values between the smoothed methylation levels of the WGBS data set and the raw methylation levels of the capture bisulfite sequencing data. Cancer$k$ represents the kth sample in the cancer group and Normal$k$ is the kth sample in the normal group.

Evaluation with Differentially Expressed Genes

As mentioned in the introduction section, there is a strong relationship between DMRs and DE genes. Therefore to evaluate the experimental results of real data, we adopted the studies of DE genes in colorectal cancer samples [24, 28]. We applied the online analysis tool GEO2R to compare the normal and cancer groups of samples in the accession of $GSE4183$. Finally, the top 250 differentially expressed (DE) genes in terms of false discovery rates were reported. We further found that six out of the 250 DE genes located on the chromosome 22, which are ZNRF3, TIMP3, FBLN1, TYMP, APOL1 and APOBEC3G. We focused our comparison on the genomic regions corresponding to these 6 DE genes.

After searching for the DMRs reported by both softwares, we found that DMReSearch could detect DMRs in four DE genes, whereas BSmooth could only identify DMRs in two DE genes. Moreover, neither BSmooth nor DMReSearch could find DMRs in gene APOL1 and gene APOBEC3G. The failure of detection should be attributed to the sparsity of CpGs inside these two genes and the small differences of methylation levels between the cancer samples and normal samples (see Figure 3.9).

When we paid attention to gene ZNRF3, we observed that DMReSearch and BSmooth reported quite distinct DMRs (see Figure 3.6). BSmooth detected one DMR (blue bars), while DMReSearch identified two DMRs (red bars). We specially examined the differences of the mean methylation levels between cancer and normal groups inside the only DMR reported by BSmooth and found them not convincible at all. Figure 3.6(b) showed the mean differences of the methylation levels at the three CpGs in this DMR. BSmooth reported it as a hyper-methylated DMR. However, the conclusion could hardly make any sense according to the differences that were approximately 0, +0.25, and −0.15. In contrast, the results of applying our software package DMReSearch showed strongly consistent differential methylation at all CpGs inside each detected DMR. We specially zoomed in the first DMR found by DMReSearch to illustrate its reasonable identification (see Figure 3.6(c)). Note that the only one CpG with different direction corresponded to a much smaller difference value of +0.1 compared to the rest four mean differences of −0.5 on average.

For the FBLN1 (see Figure 3.7), both DMReSearch and BSmooth were able to identify reasonable DMRs with respect to the differences of the mean methylation levels between two groups of samples. We believed that the distinct results of these two software tools were caused by their different modeling and testing strategies. Unfortunately, we were unable to tell which tool reported the correct DMRs due to the lack of

truth.

DMReSearch identified DMRs in the DE genes TYMP and TIMP3, where BSmooth failed to detect any DMR. It also suggested that DMReSearch is able to provide much more informative DMRs for further analysis of the associations between differential methylation and the development of cancer.

3.3.3 Application to Real Data II

In order to achieve a more comprehensive evaluation of the performance of DMReSearch in identifying DMRs, we took another comparison experiment with the Lister data [41]. Lister and his colleagues studied the differences between the cytosine methylation states of two human cell lines: H1 human embryonic stem cells and IMR90 fetal lung fibroblasts. In their study, each of the two cell lines was sequenced by WGBS technology and aligned to the human reference genome (hg18) by using the Bowtie algorithm. We obtained the aligned and summarized data from the Salk Institute website (`http://neomorph.salk.edu/human_methylome/data.html`), which has two replicates for each cell line. We extracted the data that corresponds to the chromosome 21 (chr21) and discarded the calls in non-CpG context. For the remaining calls in CpG context, we ignored the strand factor by summing the methylation calls and total calls on both the forward and the reverse strands, respectively. Note that the summing procedure cannot influence the final results because CpG contexts in both strands are symmetrically distributed. Then, we applied DMReSearch and BSmooth to analyze this set of data and evaluated the agreements between their output DMRs and DE genes. In our experiment, we found five DE genes on chr21 by using Cuffdiff [73] on the web-based platform Galaxy (`https://usegalaxy.org`).

We noted that the Lister Data owned a high sequencing coverage of about 16× on average, which is much higher than those of most other WGBS experiments. We extracted a low-coverage data set from the Lister data and treated the original data as a benchmark to test the performance of DMReSearch and BSmooth. In detail, we selected the alignment calls in CpG context corresponding to the forward strand of chr21. The resulting sub-data set had a sequencing coverage of about 9×. We subsequently run DMReSearch and BSmooth to detect DMRs by using both the low- and high-coverage data sets and their own default parameters. In the experiments on the lower-coverage data set, DMReSearch finally identified 5489 DMRs that covered 27383 DMCs, while BSmooth reported 975 DMRs and 7750 DMCs. For the high-coverage data experiment, DMReSearch identified 6413 DMRs which contained 31209 DMCs while BSmooth detected 990 DMRs covering 7776 DMCs.

Note that the low-coverage data set is a subset of the high-coverage data set. Thus, the true DMCs covered by the reads in the high-coverage data set should also have a chance to be called by the low-coverage data set. In other words, the results gotten from the low-coverage data set should be consistent with those from high-coverage data set. Accordingly, we examined the number of DMCs detected by each tool based on both the low-coverage data set and the higher-coverage data set. Among the 27383 DMCs found by DMReSearch with the low-coverage data set, 21752 DMCs were also detected by using the high-coverage data set. Comparatively, the number of common DMCs reported by BSmooth was 5428. In other words, about 70% DMCs found by BSmooth in the low-coverage data set were also identified by itself with the high-coverage data set; whereas, the common DMCs detected by DMReSearch in both data sets were almost 80% to its low-coverage results. It was more apparent to conclude that DMReSearch could achieve higher agreement between the results of experiments on low-coverage

and high-coverage data sets when we focused comparison on the DMRs detected within the five DE genes (see Figure 3.10 to Figure 3.14). Among these five DE gene regions, both software tools failed to detect any DMR with the low-coverage data in the forth region (Figure 3.13). Additionally, BSmooth could not find any DMRs with its default parameter settings in the third DE gene no matter which data set it used. However, DMReSearch was able to identify a large number of DMRs within this region. As shown in Figure 3.12, DMReSearch even obtained the highest consistency in DMR detection using both low- and high-coverage data. Inside each of the other three DE gene regions, DMReSearch could identify most of the DMRs that were found by running on the high-coverage data. In contrast, the results of BSmooth in these two types of data sets showed much more discrepancies.

Furthermore, we evaluated the smoothing performances of both tools by using the low-coverage data set. We calculated the correlation between their smoothed methylation levels from the low-coverage data set and the original ones from high-coverage data set. The values of correlation are summarized in Table 3.2. DMReSearch achieved much higher correlation scores than BSmooth, which implied that our cluster-based smoothing approach could retain more original features as improving the overall quality of the WGBS data. The better smoothing results could also benefit the accuracy of the downstream analysis to identify DMCs and DMRs, as shown above.

Tool	Lung1	Lung2	Embryonic1	Embryonic2
DMReSearch	0.898	0.897	0.737	0.744
BSmooth	0.605	0.615	0.301	0.355

Table 3.2: Correlation values between the smoothed methylation levels from the low-coverage data set and the raw methylation levels of the Lister WGBS data. *Lung*1 and *Lung*2 represent the two biological replicates of the human IMR90 fetal lung fibroblasts cell lines and *Embryonic*1 and *Embryonic*2 denote the two replicates in human embryonic stem cells.

In addition, we specially noticed that BSmooth defined an unreasonable boundary for one region (see Figure 3.15). It is obvious that the two CpG sites on leftmost side have much smaller mean differences of methylation levels compared to the other CpGs. Moreover, we could also easily observe that these two CpGs are not close to the others at all. As a result, it might be incorrect to include these two CpGs into a DMR, like what BSmooth did. In contrast, DMReSearch provided a much clear and reasonable boundary for the DMR detected by itself. This observation also illustrated the advantages of our clustering-before-smoothing strategy applied in DMReSearch.

In order to quantitatively interpret the reliability of each tool in the detection of DMRs by using the experimental results of high-coverage data set, we define a DMR reported by a tool to be an unreasonable DMR if it satisfies one of the following two criteria: (i) the average mean difference of the DMCs within a DMR is less than 0.05; (ii) more than 1/3 DMCs inside the DMR have the opposite direction to the reported direction of the DMR. Based on these two criteria, DMReSearch only reported 8 unreasonable DMRs among the total 6413 DMRs it identified. However, 281 out of the 990 DMRs found by BSmooth were unreasonable. It is not surprising at all that BSmooth obtained so many unreasonable output DMRs considering its poorer smoothing performance and fuzzy boundaries.

3.4 Conclusion

We presented a new method to detect DMRs based on WGBS data. Our method has been implemented into an R package called DMReSearch. The proposed approach builds on a three-dimensional rank clustering method followed by a modified local kernel smoothing strategy and Wald hypothesis test based on the beta-binomial distri-

bution. We carried out a series of simulation experiments and real data tests to evaluate the performance of DMReSearch. The experimental results showed that DMReSearch has a strong ability to identify more informative DMRs and achieve higher accuracies than BSmooth, especially for the data set with lower sequencing coverage, higher biological variation and weak signal quality. It also implies that applying DMReSearch can help reduce the cost of WGBS experiments because the software tool can obtain the satisfactory accuracies when analyzing low coverage and small sample size data.

DMReSearch is a user-friendly software package that can be executed on personal computer (PC) to detect DMRs in the whole genome scale. Limited by the memory space of a regular PC, running DMReSearch requires more time, most of which is used in the process of clustering. We will try to optimize the coding of DMReSearch and update the package for the computing servers with large memory spaces and multiple cores in order to improve the efficiency of our software tool.

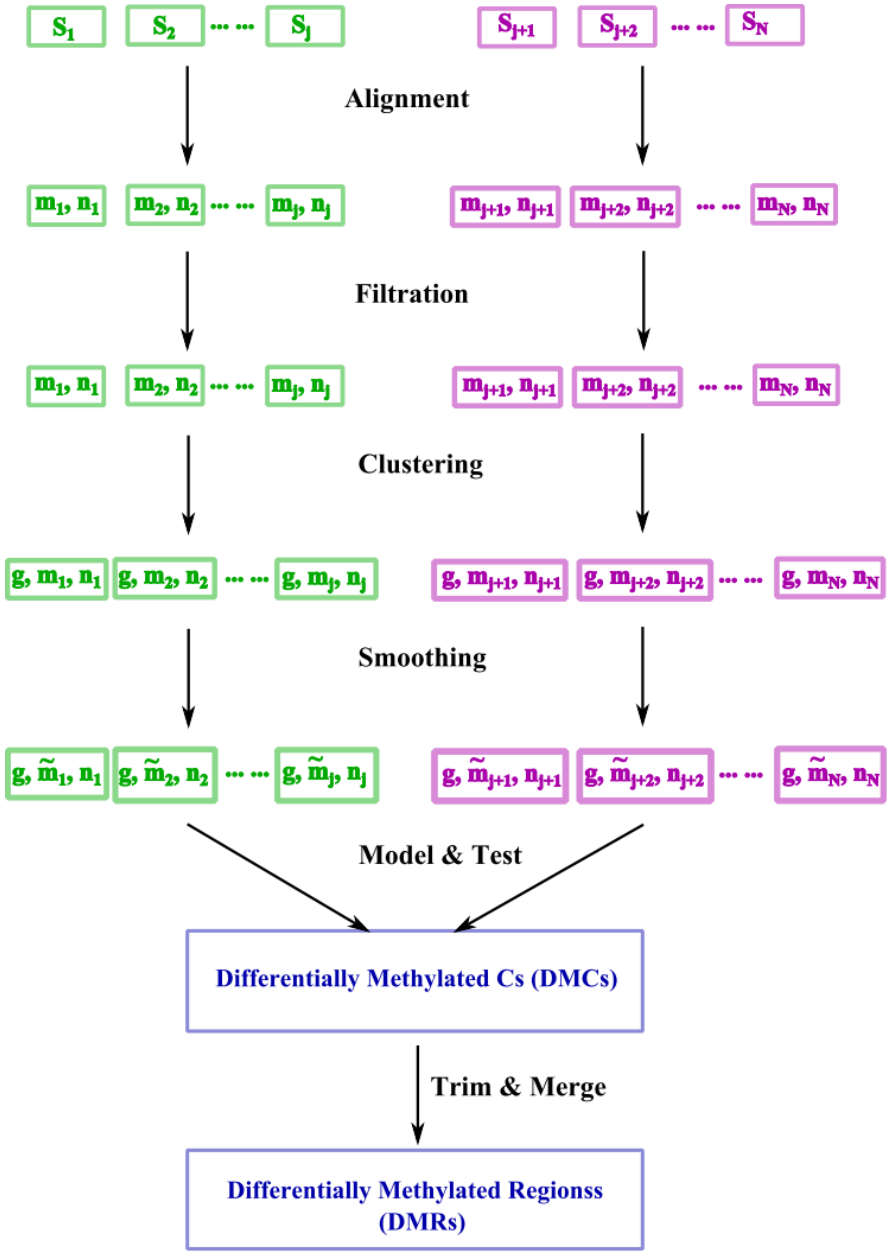
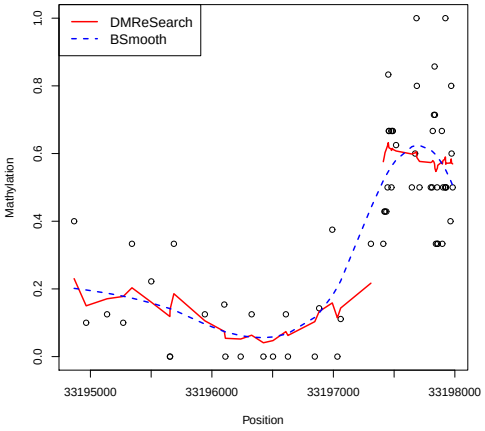
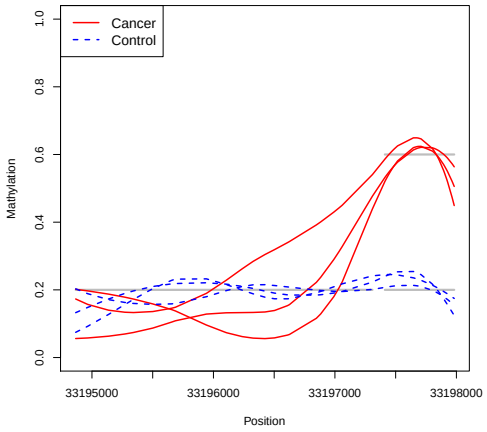


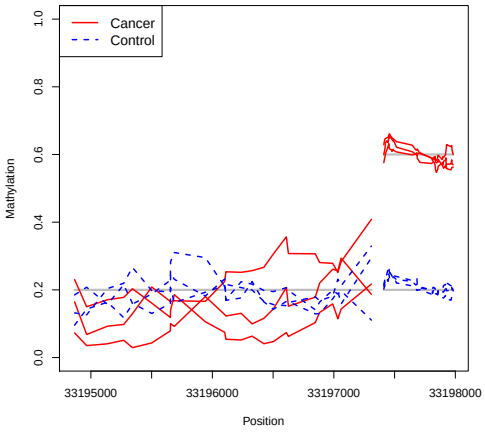
Figure 3.1: The schematic working procedures of DMReSearch. $S_1, \ldots, S_N$ represent the N samples in two condition groups. m_j and n_j indicate the methylation counts and total counts of Cs in each sample j after alignment. The clustering process results in the CpG groups g. $\tilde{m}_j$ denote the smoothed methylation counts.



(a)

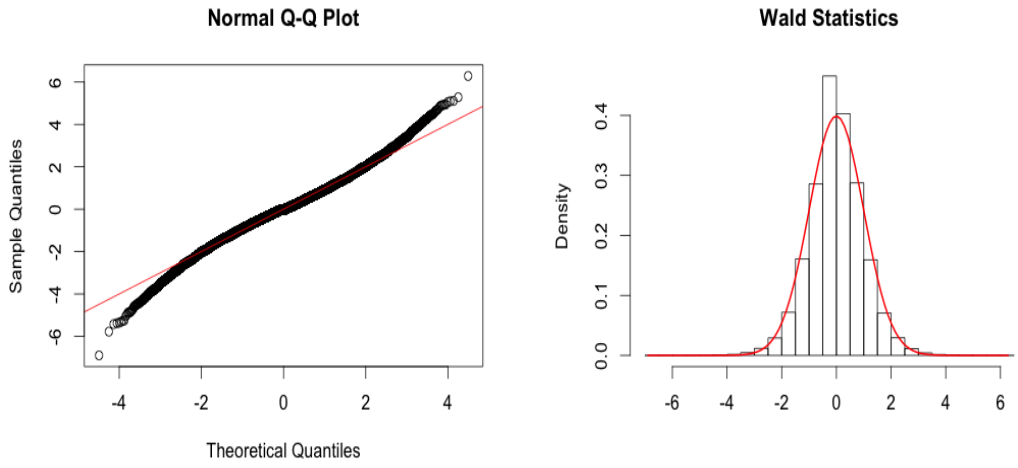


(b)



(c)

Figure 3.2: Toy example. (a) The smoothing performances of BSmooth and DMReSearch on a sample in cancer group. Each circle represents the methylation level of a CpG. (b) and (c) The smoothing results of BSmooth and DMReSearch on all six samples respectively. The grey lines imply the mean methylation levels of different groups in different regions. The mean methylation levels of two sample groups in non-DM region are exactly the same. For the DMR, the mean methylation level of cancer group is much higher than that of the normal group.



(a) (b)

Figure 3.3: Normal QQ plot of beta-binomial Wald statistics and histogram of Wald statistics. In Figure (a), each circle is a beta-binomial Wald statistic. In Figure (b), the red curve represents the standard normal distribution.

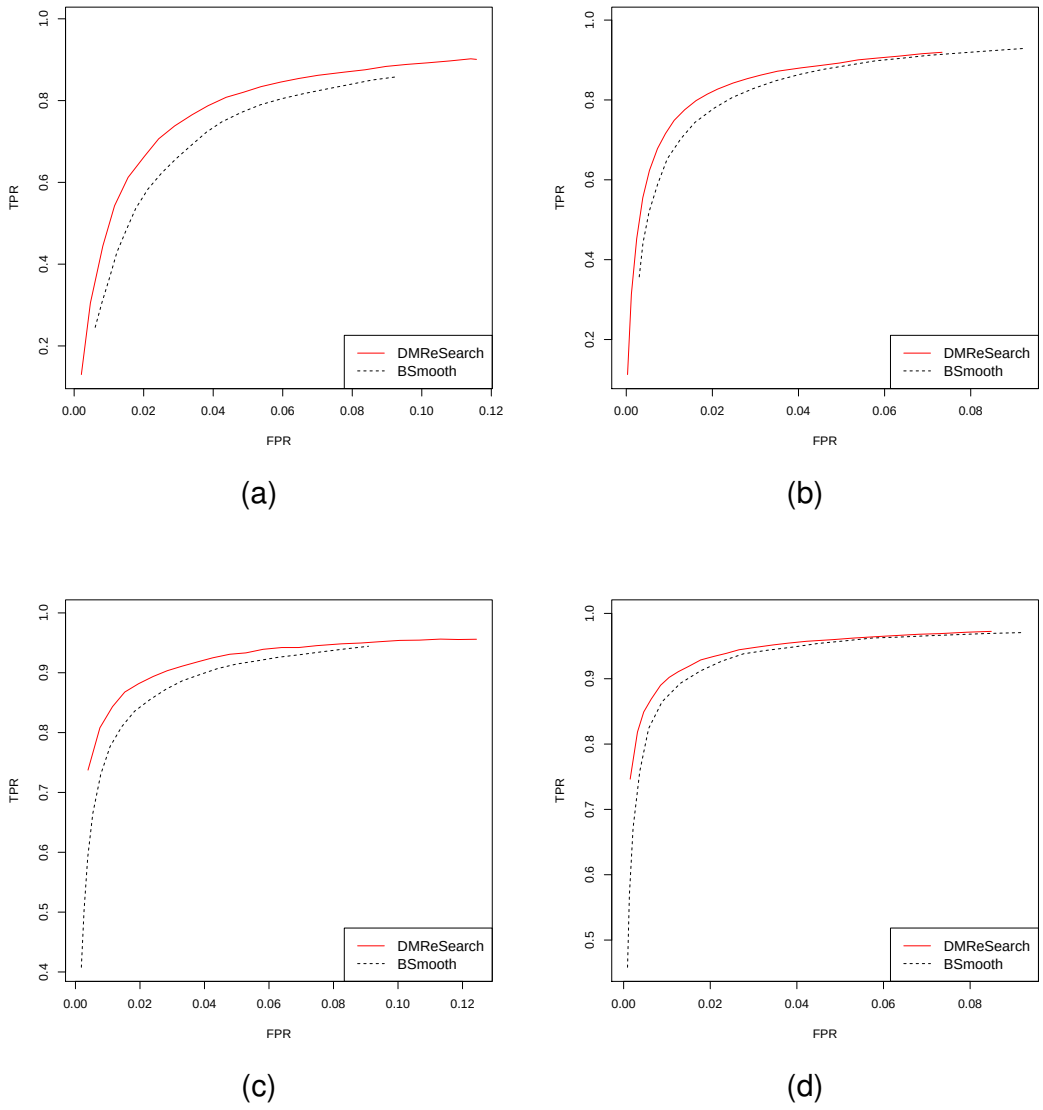


Figure 3.4: The ROC curves for low coverage setups. The first and second columns are for high and low biological variance setups respectively. The first and second rows are for weak and strong signal setups respectively.

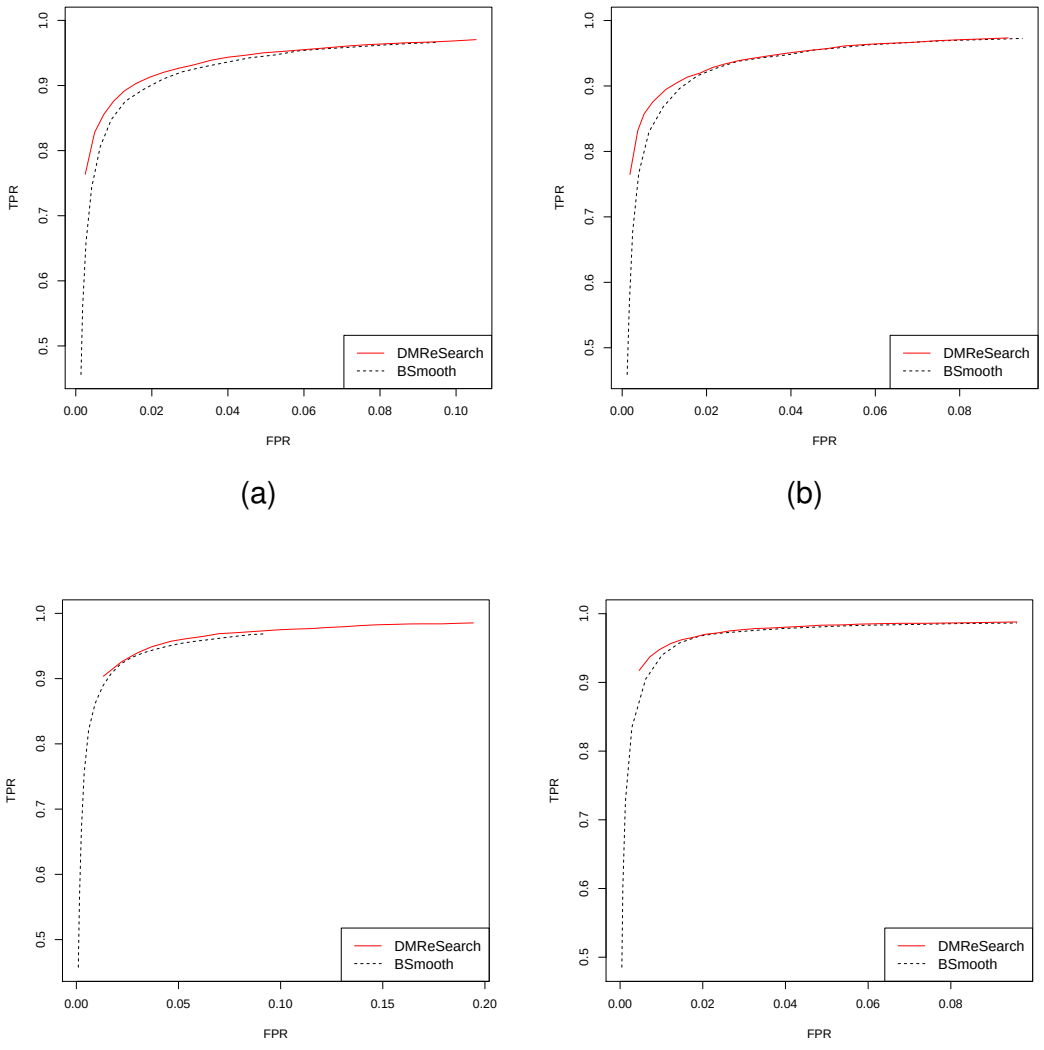


Figure 3.5: The ROC curves for high coverage setups. The first and second columns are for high and low biological variance setups respectively. The first and second rows are for weak and strong signal setups respectively.

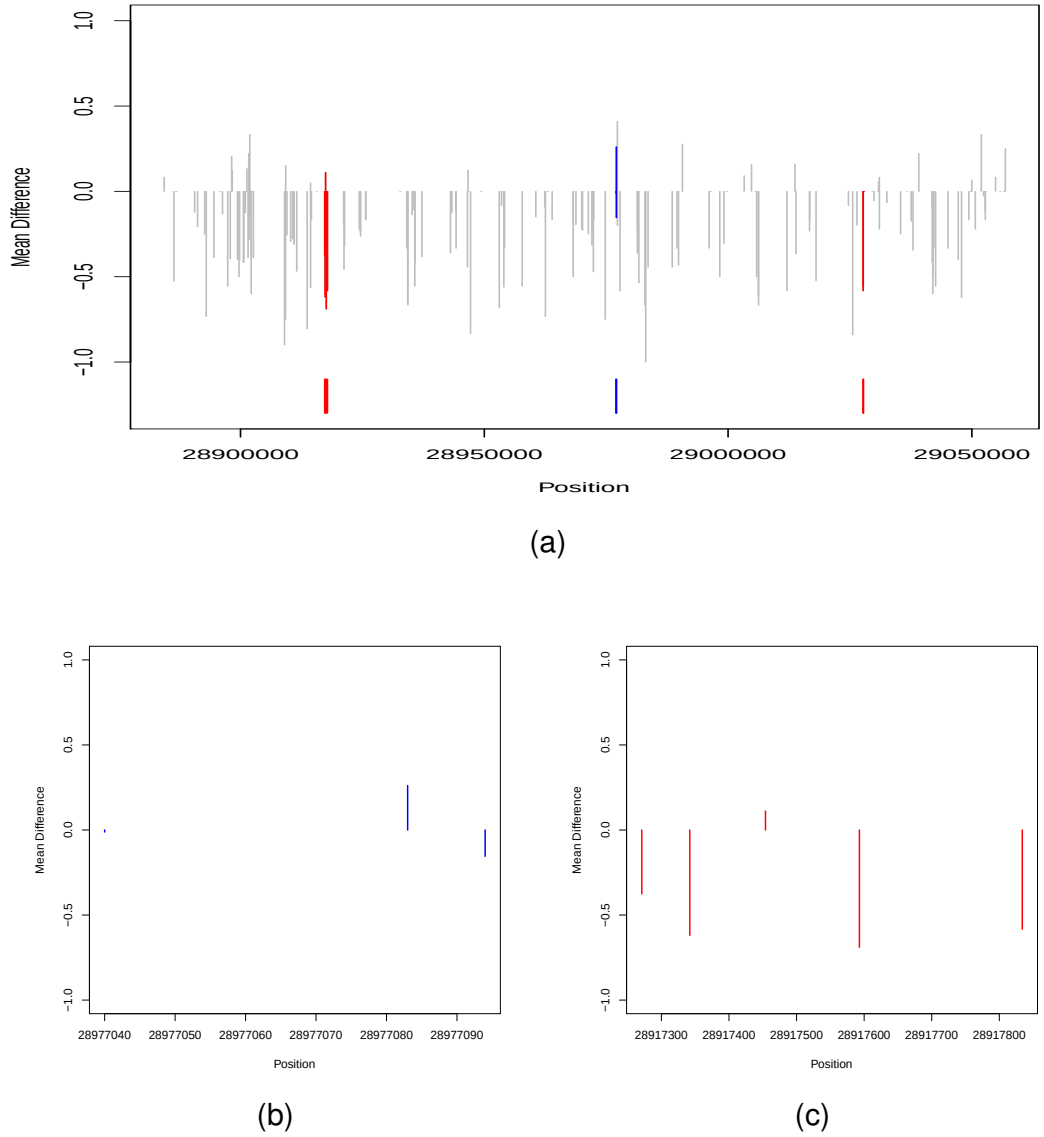


Figure 3.6: The DMRs found by DMReSearch and BSmooth in the gene ZNRF3. The y-axis is the mean of the methylation levels in cancer group minus the the mean of the methylation levels in normal group. The red color represents the DMRs found by DMReSearch and the blue color represents the DMRs found by BSmooth. The grey color indicates that the corresponding CpG is not identified as a DMC by either software. The figures in the second row are the zoomed-in DMRs found by BSmooth and DMReSearch. The colored segments in the bottom are to indicate the DMC locations.

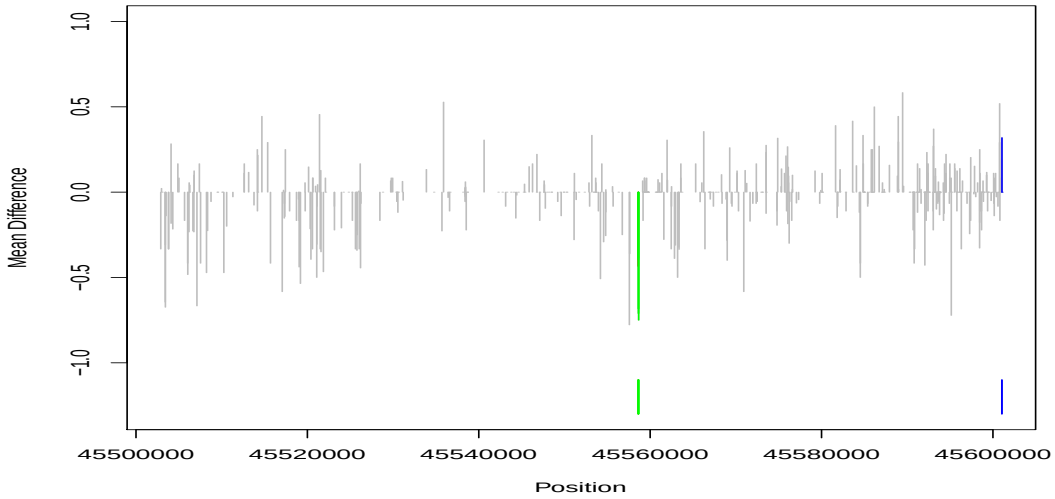
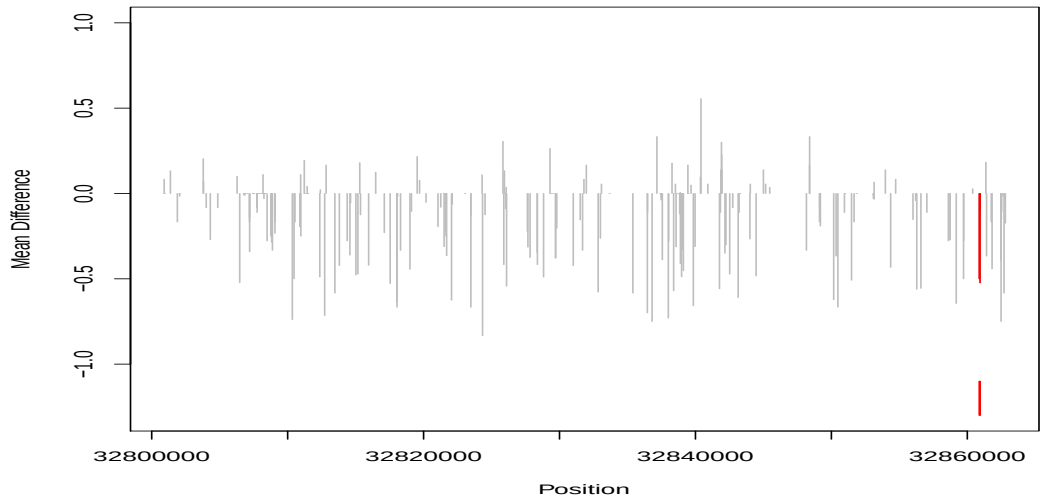
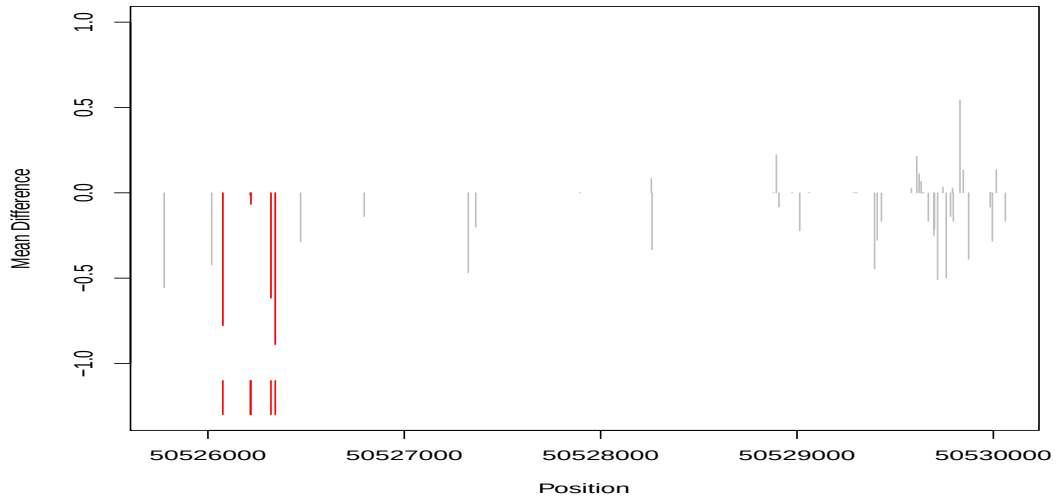


Figure 3.7: The DMRs found by DMReSearch and BSmooth in the gene FBLN1. The y-axis is the mean of the methylation levels in cancer group minus the the mean of the methylation levels in normal group. The blue color represents the DMRs found by BSmooth and the green color represents the DMRs found by both methods. The grey color indicates that the corresponding CpG is not identified as a DMC by either software. The colored segments in the bottom are to indicate the DMC locations.

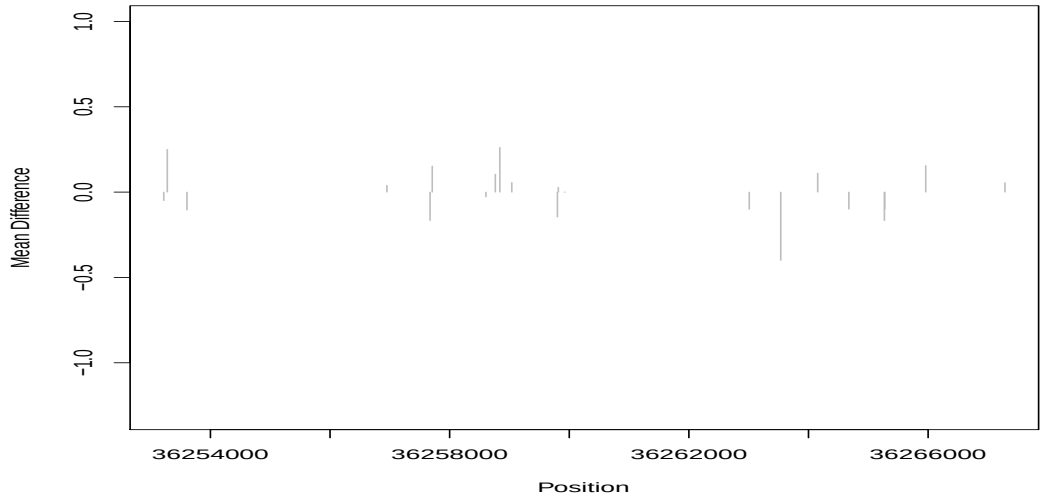


(a)

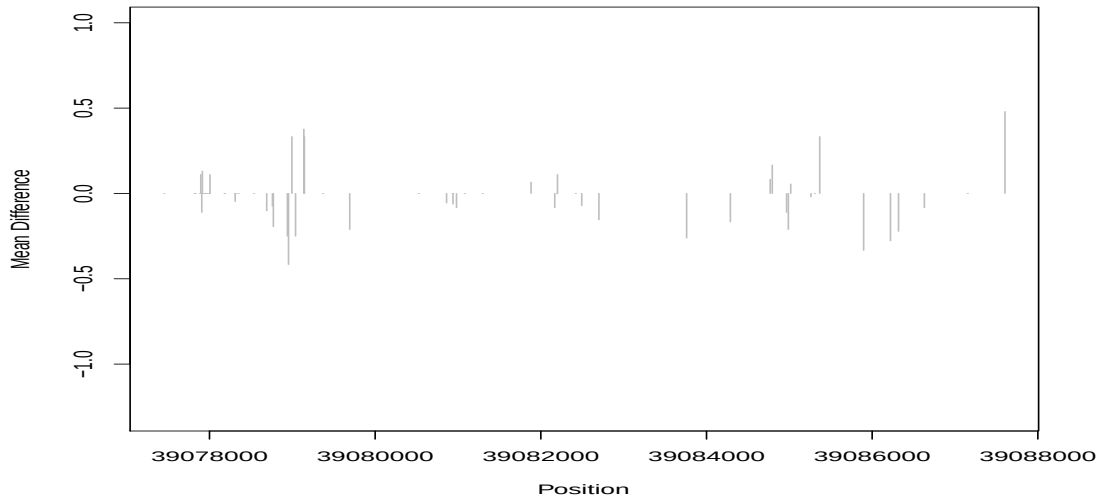


(b)

Figure 3.8: The DMRs found by DMReSearch in the gene TIMP3 and TYMP. The y-axis is the mean of the methylation levels in cancer group minus the the mean of the methylation levels in normal group. The red color represents the DMRs found by DMReSearch. The grey color indicates that the corresponding CpG is not identified as a DMC by either software. The colored segments in the bottom are to indicate the DMC locations.

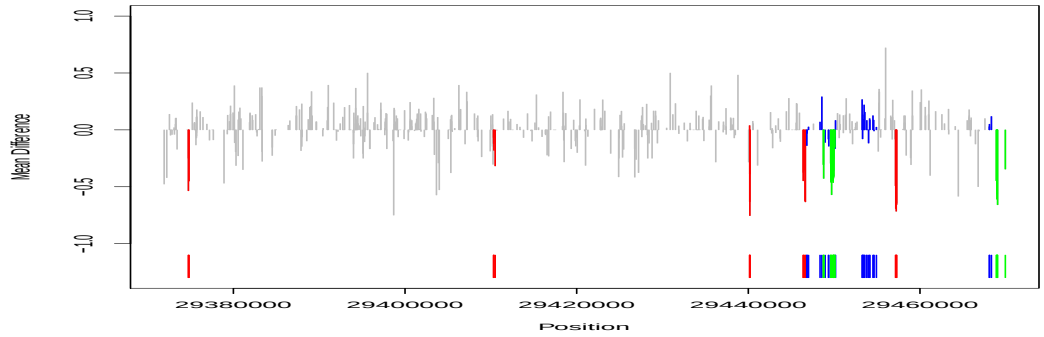


(a)

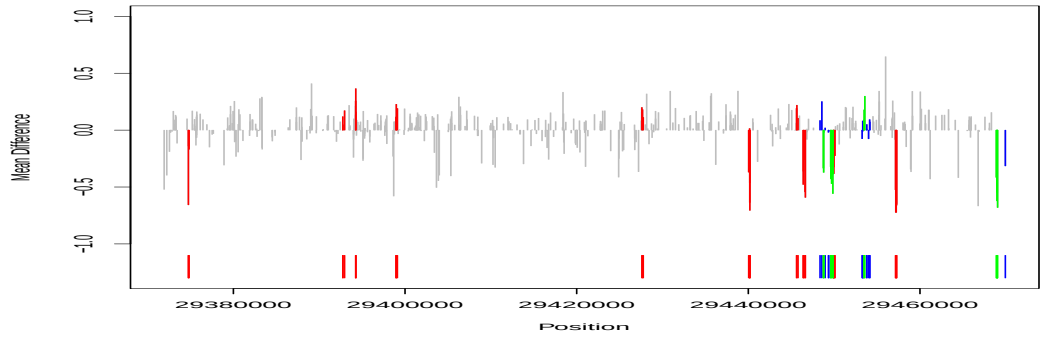


(b)

Figure 3.9: The mean difference plot in the gene APOL1 and APOBEC3G. The y-axis is the mean of the methylation levels in cancer group minus the the mean of the methylation levels in normal group. The grey color indicates that the corresponding CpG is not identified as a DMC by either software. No DMRs were detected by both software tools in these two genes.

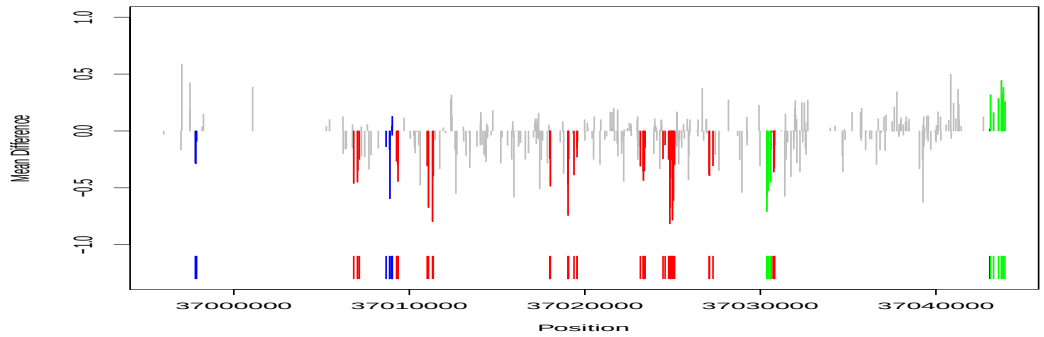


(a)

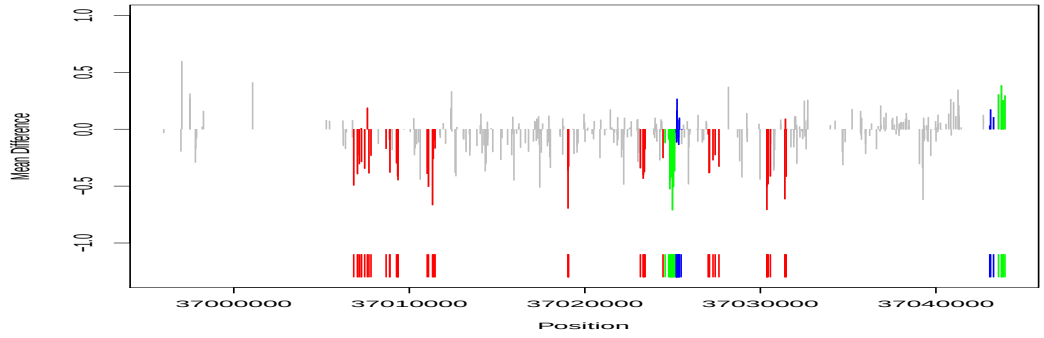


(b)

Figure 3.10: DMRs detection results within the first DE gene by using (a) low-coverage and (b) high-coverage data sets, respectively. The y-axis is the mean of the methylation levels in IMR90 fetal lung fibroblasts cells minus the the mean of the methylation levels in h1 human embryonic stem cells. The red color represents the DMRs found by DMReSearch. The DMRs found by BSmooth are marked by blue. The green color indicates the DMRs found by both two tool. The grey means CpGs that are not considered differentially methylated by both tools. The colored segments in the bottom are to indicate the DMC locations.

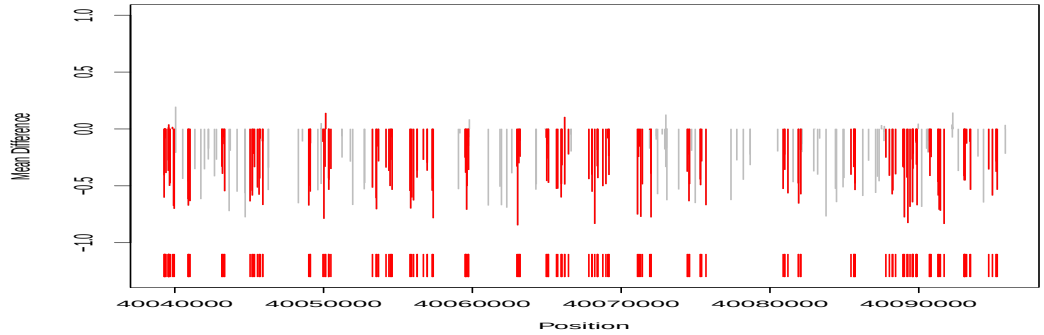


(a)

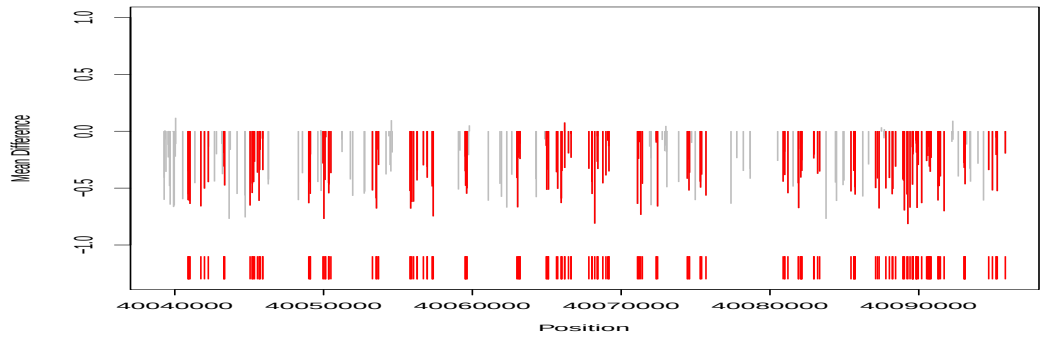


(b)

Figure 3.11: DMRs detection results inside the second DE gene by using (a) low-coverage and (b) high-coverage data sets, respectively. The color marks have the same meaning with those in Figure 3.10

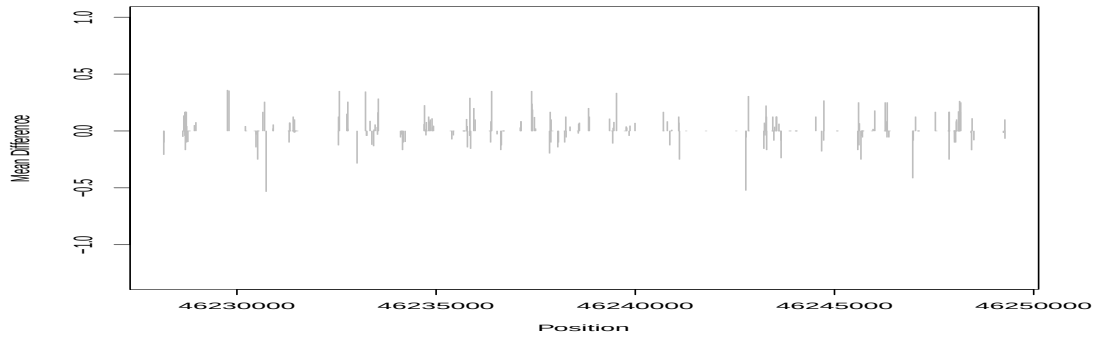


(a)

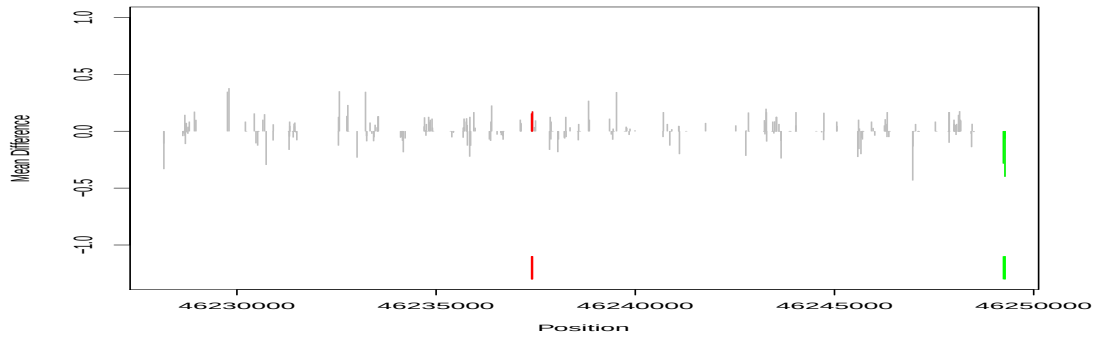


(b)

Figure 3.12: DMRs detection results corresponding to the third DE gene by using (a) low-coverage and (b) high-coverage data sets, respectively. The color marks have the same meaning with those in Figure 3.10

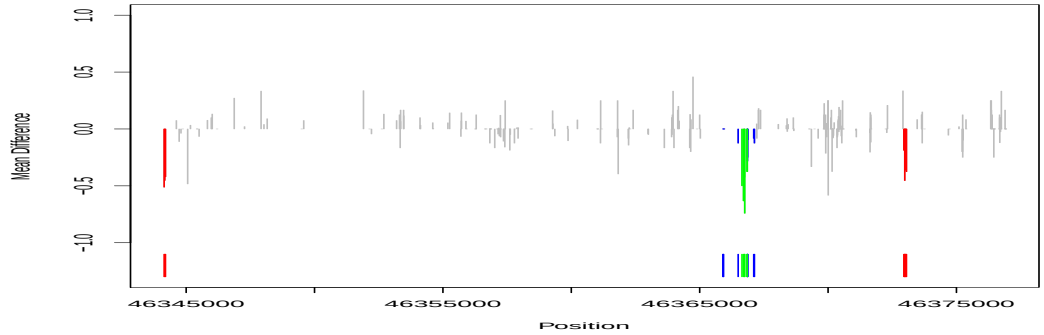


(a)

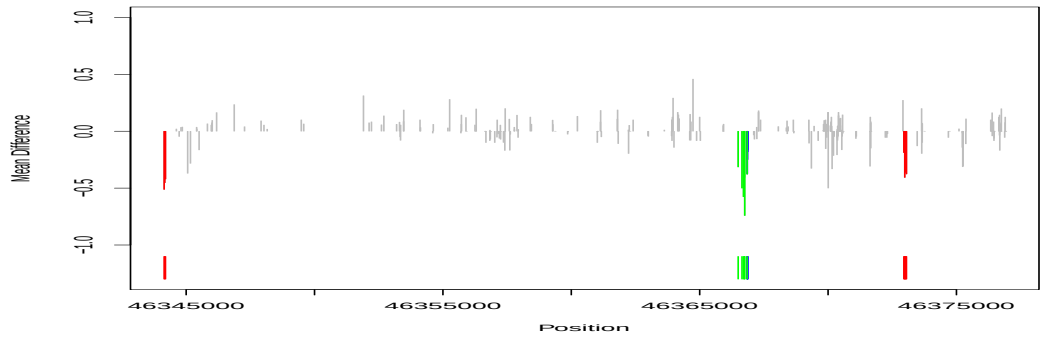


(b)

Figure 3.13: DMRs detection results corresponding to the forth DE gene by using (a) low-coverage and (b) high-coverage data sets, respectively. The color marks have the same meaning with those in Figure 3.10



(a)



(b)

Figure 3.14: DMRs detection results corresponding to the fifth DE gene by using (a) low-coverage and (b) high-coverage data sets, respectively. The color marks have the same meaning with those in Figure 3.10

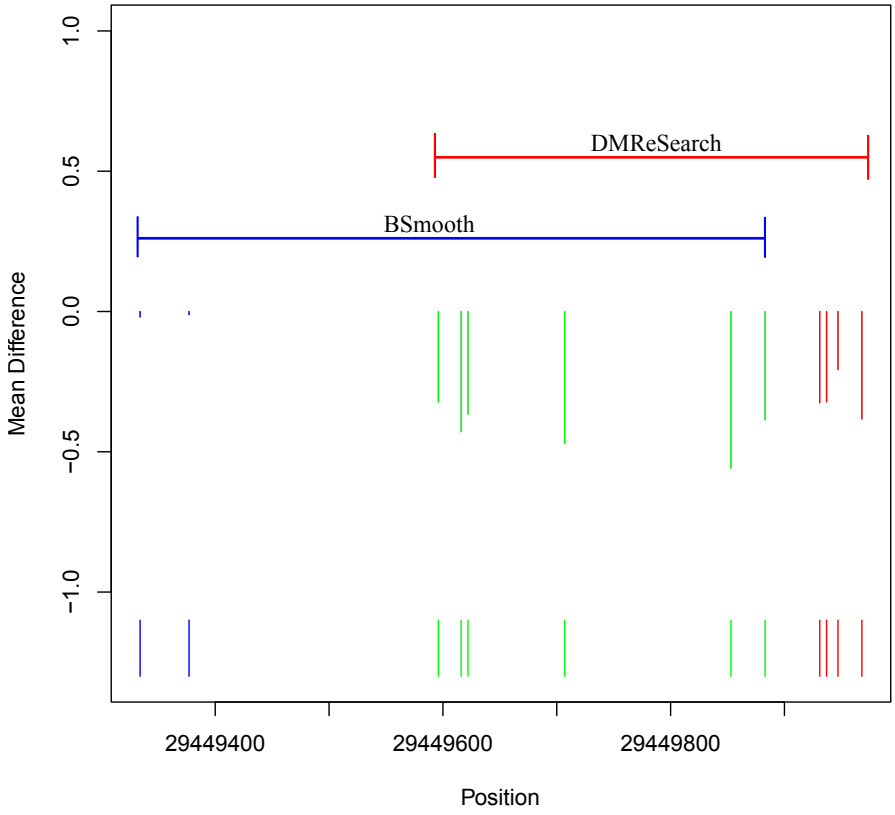


Figure 3.15: A zoomed-in region in the first DE gene. BSmooth included two unreasonable CpG sites into the DMR (blue). Comparatively, DMReSearch defined more reasonable boundaries of the DMR (red) by excluding these two CpGs and extending the other side of boundary.

Chapter 4

Conclusion

4.1 Summary

Clustering is to distribute observations that are cohesive and distinguishable from others into some homogeneous groups. It is a significant technological tool to analyze NGS data. In this thesis we mainly discussed two clustering problems for NGS data. One is to cluster samples based on RNA-Seq data by using the differentially expressed genes across different tissues or subtypes of cancer cells. The other is to detect the DMRs based on BS-Seq data with samples in two different biological conditions. Roughly speaking, these two issues need us to cluster the sequencing data in two different directions. We assume that the clustering of samples is vertical clustering and the detection of DMRs is horizontal clustering. Both of these two clustering problems encounter the same challenge from NGS data: low sample size and high dimension. Due to low sample size, it is difficult to accurately estimate parameters in the statistical models, especially for the estimation of biological variances. Inside both of the proposed clustering methods, we make the same strategy, which borrows information from the

correlated features, to estimate the biological variances. However these two problems have dramatic differences in response to the different directions and distinct types of NGS experiments.

In the vertical clustering, we need to conduct clustering based on the huge number of features, most of which are non-informative. Accordingly it is significant to select informative features that show differential expressions across different groups. However, the unknown group information of samples would inherently impede the variable selection. To tackle these difficulties, we propose the l_1 penalized model-based clustering method and develop the R package PMixClus. This method can simultaneously select variables and cluster samples. It can further determine the number of clusters through some model selection criteria, such as BIC or EBIC. We also provide numerical solutions for both Poisson and NB mixture models. Although Poisson model is not applicable for many RNA-Seq data sets that have biological replicates, it can perform well when there are only technological replicates or may be applied in other real world problems.

In the horizontal clustering, one challenge is to accurately identify DMRs that are surrounded by noises and the other one is the low coverage of CpGs because of the high cost of the WGBS experiments. We develop one software DMReSearch to solve these problems and accurately identify DMRs. We propose a three-dimensional rank clustering method to pre-cluster the CpGs and subsequently apply a modified local kernel smoothing approach in each pre-clustered region. Our results from simulation and real data experiments all show this pretreatment could not only set DMR boundaries more precisely but also improve the quality of low coverage data. Then Beta-Binomial distribution is used in order to take biological variances into consideration. After estimation of parameters, we utilize the Wald test to find DMCs and accordingly trim and merge

the pre-clustered regions to form final DMRs.

4.2 Future Works

For the proposed penalized model-based clustering, l_1 penalty is used to select variables automatically in order to simplify computation process; however the excessive penalty on large values may result in larger estimates of dispersion and further make variable selection inaccurate. Although we design a new parametrization and accordingly penalize log-transform of θ_{kp} to reduce the shrinkage of large values, it cannot eliminate this bias thoroughly. One possible solution for this limitation is to utilize other penalties, such as hard thresholding penalty or SCAD penalty ([19]). Besides, penalizing the dispersion parameters simultaneously may be another choice. Both of these solutions can complicate the model structure and further challenge the numerical computation, especially for NB model. We plan to develop better penalties to select variables in NB mixture model and the corresponding efficient numerical solutions. Additionally, some articles [40, 78] propose the power transformation methods in order to get rid of the overdispersion such that the Poisson model can be applied even for RNA-Seq data sets that involve biological replicates. We also tried the power transformation method in [78] to transform data for our proposed Poisson model, but the results from both simulation and real data were not ideal (worse than those without transformation). Therefore, we want to seek for a new transformation method that is more appropriate for Poisson mixture model since the numerical computation in Poisson mixture model is much easier than that in NB mixture model.

A shortcoming of DMReSearch is that it is slower than some other softwares, like BSmooth. Therefore we will improve the computational efficiency by utilizing large

memory equipment and parallel computation. During we study this topic, we are wondering if there is one special case that the methylation levels of all CpGs in one true DMR perform small differences between two biological conditions, which cannot be treated as DMCs for any statistical test, but their differences are highly consistent across them. We think that if the number of such CpGs is large enough, the corresponding region will be highly probable to be DMR. However, to our best knowledge, there is no method to consider this situation. One possible solution is to put a hypothesis test on the whole region. Actually BiSeq [32] applies a group hypothesis test on each region, but it is still required that the p-value of each CpG inside every detected DMR should be statistically significant. Another difficulty to tackle this problem is strong dependence on the accuracy of the pre-clustering process. If the pre-clustered regions are not satisfactory, the group test will increase false positives and false negatives. Due to the lack of DMR knowledge, it is hard to define accurate boundaries for each pre-clustered region. It will be our future work to find some ways to set more accurate boundaries, especially for those significant regions. Furthermore, we will also try to design a reasonable and sensitive group test specifically for DMR detection with WGBS data.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Caski, editors, *In Proc. of the Second International Symposium on Information Theory*, pages 267–281. Budapest: Akademiai Kiado, 1973.

[2] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[3] Benjamin P. Berman, Daniel J Weisenberger, Joseph F. Aman, Toshinori Hinoue, Zachary Ramjan, Yaping Liu, Houtan Noushmehr, Christopher P. E. Lange, Cornelis M van Djk, Rob A. E. M. Tollenaar, David Van Den Berg, and Peter W. Laird. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nucler lamina-associated domains. *Nature Genetics*, 44(1):40–46, November 2011.

[4] P. Berninger, D. Gaidatzis, E. van Nimvegen, and M. Zavolan. Computational analysis of small RNA cloning data. *Methods*, 44:13–21, 2008.

[5] Adrian P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, May 1986.

[6] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(94), 2010.

[7] R. D. Canales, Y. Luo, J. C. Willey, B. Austermiller, C. C. Barbacioru, C. Boysen, K. Hunkapiller, R. V. Jensen, C. R. Knight, K. Y. Lee, Y. Ma, B. Maqsodi, A. Papallo, E. H. Peters, K. Poulter, P. L. Ruppel, R. R. Samaha, L. Shi, W. Yang, L. Zhang, and F. M. Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology*, 24(9):1115–1122, 2006.

[8] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14(2):244–258, 2013.

[9] J. Chen and Z. Chen. Extended Bayesian information criterion for model selection with large model space. *Biometrika*, 94:759–771, 2008.

[10] J. Chen and Z. Chen. Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica*, 22:555–574, 2012.

[11] Junfang Chen, Pavlo Lutsik, Ruslan Akulenko, Jörn Walter, and Volkhard Helms. AKSmooth: enhancing low-coverage bisulfite sequencing data via kernel-based smoothing. *Journal of Bioinformatics and Computational Biology*, 12(6):1442005, December 2014.

[12] Pao-Yang Chen, Shawn J. Cokus, and Matteo Pellegrini. Bs seeker: precise mapping for bisulfite sequencing. *BMC Bioinformatics*, 11:203, April 2010.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[14] Brian G Dias and Kerry J Ressler. Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nature Neuroscience*, 17(1):89–96, January 2014.

[15] Akiko Doi, In-Hyun Park, Bo Wen, Peter Murakami, Martin J. Aryee, Rafael Irizarry, Brian Herb, Christine Ladd-Acosta, Junsung Rho, Sabine Loewer, Justine Miller, Thorsten Schlaeger, George Q Daley, and Andrew P. Feinberg. Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature Genetics*, 41(12):1350–1353, December 2009.

[16] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K. Rakyan, John Attwood, Matthias Burger, John Burton, Tony V. Cox, Rob Davies, Thomas A. Down, Carolina Haefliger, Roger Horton, Kevin Howe, David K. Jackson, Jan Kunde, Christoph Koenig, Jennifer Liddle, David Niblett, Thomas Otto, Roger Pettett, Stefanie Seemann, Christian Thompson, Tony West, Jane Rogers, Alex Olek, Kurt Berlin, and Stephan Beck. Dna methylation profiling of human chromosomes 6, 20, and 22. *Nature Genetics*, 38(12):1378–1385, December 2006.

[17] Gerda Egger, Gangning Liang, Ana Aparicio, and Peter A. Jones. Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, 429(6990):457–463, May 2004.

[18] Melanie Ehrlich, Miguel A. Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett,

Kenneth C. Kuo, Roy A. McCune, and Charles Gehrke. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Research*, 10(8):2709–2721, April 1982.

[19] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[20] Silvia L. Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, August 2010.

[21] C. Fraly. Algorithms for model-based Gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.

[22] A. C. Frazee, B. Langmead, and J. T. Leek. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(449), 2011.

[23] Martin C. Frith, Royta Mori, and Kiyoshi Asai. A mostly traditional approach improves alignment of bisulfite-converted DNA. *Nucleic Acids Research*, 40(13):e100, March 2012.

[24] Orsolya Galamb, Balazs Gyorffy, Ferenc Sipos, Sandor Spisak, Anna Maria Nemeth, Pal Miheller, Zsolt Tulssay, Elek Dinya, and Bela Molnar. Inflammation, adenoma and cancer: objective classification of colon biopsy specimens with gene expression signature. *Disease Markers*, 25(1):1–16, 2008.

[25] D. Ghosh and A. M. Chinnaiyan. Mixture modeling of gene expression data from mi- croarray experiments. *Bioinformatics*, 18(2):275–286, 2002.

[26] J. J. Goeman. L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, 52:70–84, 2010.

[27] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804, 2010.

[28] Balazs Gyorffy, Bela Molnar, Hermann Lage, Zoltan Szallasi, and Aron C. Eklund. Evaluation of macroarry preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLos ONE*, 4(5):e5645, May 2009.

[29] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13:R83, Octorber 2012.

[30] Kasper Daniel Hansen, Winston Timp, H'echor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G. McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A. Irizarry, and Andrew P. Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, August 2011.

[31] N. Hao and H. H. Zhang. Interaction screening for ultra-high dimensional data. *Journal of the American Statistical Association*, 109:1285–1301, 2014.

[32] Katja Hebestreit, Martin Dugas, and Hans-Ulrich Klein. Detection of significantly differentially methylated regions in targeted bisulfite sequencing. *Bioinforamtics*, 29(13):1647–1653, July 2013.

[33] Rafael A. Irizarry, Christine Ladd-Acosta, Benilton Carvalho, Hao Wu, Sheri A. Brandenburg, Jeffrey A. Jeddelon, Bo Wen, and Andrew P. Feinberg. Compre-

hensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, 18(5):780–790, May 2008.

[34] Peter A. Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews, Genetics*, 13(7):484–492, May 2012.

[35] G. Karypis, E.-H. Han, and V. Kumar. Chameleon: Hierarchical clustering using dynamic modeling. *Computers*, 32(8):68–75, 1999.

[36] A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102:1025–1038, 2007.

[37] Felix Krueger and Simon Andrews. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572, June 2011.

[38] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon Andrews. DNA methylation analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–151, January 2012.

[39] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 15(29), 2014.

[40] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, 13(3):523–538, 2012.

[41] Ryan Lister, Mattia Pelizzola, Robert H. Dowen, R. David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R. Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo,

Lee Edsall, Jessica Antosiewicz-Bourget, Ron Stewart, Victor Ruotti, A. Harvey Millar, James A. Thomson, Bing Ren, and Joseph Ecker. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322, November 2009.

[42] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11, 2012.

[43] J. Lu, J. K. Tomfohr, and T. B. Kepler. Identifying differential expression in multiple sage libraries: an overdispersed log-linear model approach. *BMC Bioinformatics*, 6(165), 2005.

[44] S. Ma, J. Huang, and X. Song. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*, 12(4):763–775, 2011.

[45] E. R. Mardis. Next-Generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9:387–402, 2008.

[46] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509–1517, 2008.

[47] G. J. McLachlan and D. Peel. *Finite Mixture Models*. New York:Wiley, 2000.

[48] Klaas Mensaert, Simon Denil, Geert Trooskers, Wim Van Criekinge, Oliver Thas, and Tim De Meyer. Next-generation technologies and data analytical approaches for epigenomics. *Environmental and molecular mutagenetics*, 55(3):155–170, April 2014.

[49] M. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11:31–46, 2010.

[50] Karin B. Michels, Alexandra M. Binder, Sarah Dedeurwaerder, Charles B. Epstein, John M. Greally, Ivo Gut, E. Andres Houseman, Benedetta Izzi, Karl T. Kelsey, Alexander Meissner, Aleksandar Milosavljevic, Kimberly D. Siegmund, Christoph Bock, and Rafael A. Irizarry. Recommedations for the design and analysis of epigenome-wide association studies. *Nature Methods*, 10(10):949–955, October 2013.

[51] Dirk Moser, Savira Ekawardhani, Robert Kumsta, Haukur Palmason, Christoph Bock, Zoi Athanassiadou, Klaus-Peter Lesch, and Jobst Meyer. Functional analysis of a potassium-chloride co-transporter 3 (SLC12A6) promoter polymorphism leading to an additional DNA methylation site. *Neuropsychopharmacology*, 34(2):458–467, January 2009.

[52] S. Myllykangas, J. Buenrostro, and H. P. Ji. Overview of sequencing technology platforms. *Bioinformatics for High Throughput Sequencing*, pages 11–25, 2012.

[53] E.A. Nadaraya. On estimating regression. *Theory of Probability and Its Applications*, 9(1):141–142, 1964.

[54] U. Nagalakshmi, Z. Wong, K. Waern, C Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 302:1344–1349, 2008.

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

[56] Mattia Pelizzola and Joseph R. Ecker. The dna methylome. *FEBS Letters*, 585(13):1944–2000, July 2011.

[57] Gerd P. Pfeifer, Swati Kadam, and Seung-Gi Jin. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetic Chromatin*, 6(1):10, May 2013.

[58] Raftery and A. E. Discussion of "bayesian clustering with variable and transformation selection" by liu et al. *Bayesian Statistics*, 7:266–271, 2003.

[59] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M Reich, E. Latulippe, J Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lender, and T. Golub. Multiclass cancer diagnosis using tumor gene expression signature. *PNAS*, 9:3273–3975, 1998.

[60] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[61] Apama Raval, Stephan M. Tanner, John C. Byrd, Elizabeth B. Angerman, James D. Perko, Shih-shih Chen, Björn Hackanson, Michael R. Grever, David M. Lucas, Jennifer J. Matkovic, Thomas S. Lin, Thomas J. Kipps, Fiona Murray, Dennis Weisenburger, Warren Sanger, Jane Lynch, Patrice Watson, Mary Jansen, Yuko Yoshinaga, Richard Rosenquist, Pieter J. de Jong, Penny Coggill, Stephan Beck, Henry Lynch, Albert de la Chapelle, and Christoph Plass. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell*, 129(5):879–890, June 2007.

[62] Keith D. Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, August 2005.

[63] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expresion data. *Biostatistics*, 26:139–140, 2010.

[64] Alex Rodriguez and Alessandro Laio. Clusting by fast search and find of density peaks. *Science*, 344(6191):1492–1496, June 2014.

[65] Serge Saxonov, Paul Berg, and Douglas L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5):1412–1417, January 2006.

[66] Till Schoofs, Christian Rohde, Katja Hebestreit, Hans-Ulrich Klein, Stefanie Göllner, Isabell Schulze, Mads Lerdrup, Nikolaj Dietrich, Shuchi Agrawal-Singh, Anika Witten, Monika Stoll, Eva Lengfelder, Wolf-Karsten Hofmann, Peter Schlenke, Thomas Büchner, Klaus Hansen, Wolfgang E. Berdel, Frank Rosenbauer, Martin Dugas, and Carsten Müller-Tidow. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood*, 121(1):178–187, January 2013.

[67] G. Schwartz. Estimating the dimensions of a model. *Annals of Statistics*, 6:461–464, 1978.

[68] W. Seidel, K. Mosler, and M. Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52:481–487, 2000.

[69] Y. Si, P. Liu, P. Li, and P. B. Thomas. Model-based clustering for RNA-seq data. *Bioinformatics*, 30:197–205, 2014.

[70] P. T. Spellman, G. Sherlock, V. R. Iyer, M. Zhang, K. Anders, M. B. Eisen, P. O. Broun, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3975, 1998.

[71] Miho M. Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews, Genetics*, 9(6):465–476, June 2008.

[72] R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

[73] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven J. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010.

[74] S. Vaithyanathan and B Dom. Model-based hierarchical clustering. In *In Proc. 16th Conf. Uncertainty in Artificial Intelligence*, pages 599–608. UAI, 2000.

[75] L. Wang, J. Zhou, and A. Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68(2):353–360, 2012.

[76] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[77] Geoffrey S. Watson. Smooth regression analysis. *The Indian Journal of Statistics*, 26(4):359–372, December 1964.

[78] D. M. Witten. Classification and clustering of sequencing data using a Poisson model. *Annals of Applied Statistics*, 5:2493–2518, 2011.

[79] D. M. Witten, R. Tibshirani, S. Gu, A. Fire, and W. Lui. Ultra-high through-put sequencing-based small rna discovery and discrete statistical biomarker analysis in a collection of cervical tumous and matched controls. *BMC Biology*, 8(58), 2010.

[80] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17:763–774, 2001.

[81] S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003.

[82] Michael J. Ziller, Kasper D. Hansen, Alexander Meissner, and Martin J. Aryee. Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing. *Nature Methods*, 12(3):230–232, March 2015.

[83] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[84] I. Zwiener, B. Frisch, and H. Binder. Transforming rna-seq data to improve the performance of prognostic gene signatures. *PLoS ONE*, 9(1):e85150, 2014.

Publication

1. Tian, Y., Sun, R. and Lian, H. l_1 Penalized model-based clustering for RNA-Seq count data. Submitted.

2. Sun, R., Tian, Y. and Chen, X. TAMeBS: a sensitive bisulfite-sequencing read mapping tool for DNA methylation analysis. Accepted by IEEE BIBM 2014.

3. Lai, P., Tian, Y. and Lian, H. Estimation and variable selection for generalised partially linear single-index models. Journal of Nonparametric Statistics. 2014; 26(1): 171-185.

4. Hu, Y., Tian, Y. and Lian, H. Letter to the Editor. The Annals of Applied Statistics. 2013; 7(2): 1244-1246.