ABSTRACT

FENGJIAO HU
Statistical Methods to Detect Differentially Methylated Regions with Next-Generation Sequencing Data
(Under the direction of Dr. VARGHESE GEORGE)

Researchers in genomics are increasingly interested in epigenetic factors such as DNA methylation because they play an important role in regulating gene expression without changes in the sequence of DNA. Abnormal DNA methylation is associated with many human diseases, including various types of cancer. We propose three different approaches to test for differentially methylated regions (DMRs) associated with complex traits, while accounting for correlations within and among CpG sites in the DMRs. One approach is a nonparametric method using a kernel distance statistic and the second one is a likelihood-based method using a binomial spatial scan statistic. Both of these approaches detect differential methylation regions between cases and controls along the genome. The kernel distance method uses the kernel function, while the binomial scan statistic approach uses a mixed effect model to incorporate correlations among CpG sites. Extensive simulations show that both approaches have excellent control of type I error, and both have reasonable statistical power. The binomial scan statistic approach appears to have higher power, while the kernel distance method is computationally faster. We also propose a third method under the Bayesian framework for comparing methylation rates when disease status is classified into ordinal multinomial categories (e.g., stages of cancer). The DMRs are detected using moving windows along the genome. Within each window, the Bayes factor is calculated to compare the two models corresponding to constant vs. monotonic methylation rates among the groups. As in the case of the scan

statistic approach, the correlations between the sites are incorporated using a mixed effect model. Results from extensive simulation indicate that the Bayesian method is statistically valid and reasonably powerful to detect DMRs associated with disease severity. The proposed methods are demonstrated using data from a chronic lymphocytic leukemia (CLL) study.

KEY WORDS: differentially methylated regions (DMRs), binomial scan statistic, kernel distance statistic, Bayes factor

STATISTICAL METHODS TO DETECT DIFFERENTIALLY METHYLATED REGIONS WITH NEXT-GENERATION SEQUENCING DATA

By

Fengjiao Hu

Submitted to the Faculty of the School of Graduate Studies

of Augusta University in partial fulfillment

of the Requirements of the Degree of

Doctor of Philosophy

July

2016

STATISTICAL METHODS TO DETECT DIFFERENTIALLY METHYLATED REGIONS WITH NEXT GENERATION SEQUENCING DATA

This dissertation is submitted by Fengjiao Hu and has been examined and approved by an appointed committee of the faculty of the Graduate School of Augusta University.

The signatures which appear below verify the fact that all required changes have been incorporated and that the dissertation has received final approval with reference to content, form and accuracy of presentation.

This dissertation is therefore in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____ _____

Date Major Advisor

_____

Departmental Chairperson

(Nursing Only)_____

Associate Dean for Graduate Programs

_____

Dean, School of Graduate Studies

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my committee chair, Dr. Varghese George, for his insightful guidance, constant encouragement and tremendous patience. Especially he continually and convincingly conveyed a spirit of adventure in regard to research, and instilled in me the qualities of being a good statistician and researcher. Without his guidance and persistent help, this dissertation would not have been possible.

I would also like to thank all my committee members, for their time and effort to serve on my committee. Especially I would like to thank Dr. Hongyan Xu, Dr. Duchwan Ryu and Dr. Santu Ghosh, not only for their lectures and detailed instructions in knowledge of statistical methods and genetics, but also for their constant support, scholarly advice, especially their timely help from methodology to data analysis, and their patience in editing this work.

I would like take this opportunity to thank my master thesis advisors, Dr. Robert E. Johnson (Virginia Commonwealth University) and Dr. Charles W. Champ (Georgia Southern University), not only for their guidance and instruction that help me build a foundation in statistics, but also for their unceasing encouragement and support to motivate me to achieve my goals in education and in life.

I also want to thank all the faculty, staff, and fellow graduates of the Department of Biostatistics & Epidemiology, especially Dr. Ramses F. Sadek from Cancer Center for their support.

Finally, I would like to thank my family for their support and love.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

CHAPTER 1

INTRODUCTION

Genome-wide association studies (GWAS), in which several hundred thousand to more than a million of single nucleotide polymorphisms (SNPs) are assayed in thousands of individuals, have identified hundreds of genetic variants associated with risk of a range of complex diseases including cancer, and have provided valuable insights into the complexities of their genetic architecture (Galvan, Ioannidis, and Dragani 2010, Hindorff et al. 2009, Manolio et al. 2009).

SNPs are notably a type of common genetic variation, which is a single base pair mutation at a specific locus of a gene's DNA sequence, consisting of two alleles, meaning within a population there are two commonly occurring base-pair possibility for a SNP location. Mutations are largely caused by extremely rare genetic variants that ultimately induce a detrimental change to protein function, which leads to the different phenotypes or traits, such as disease status (Bush and Moore 2012).

Genetic location associated with disease are called risk loci, which can only explain a small proportion of the disease risk, since each locus exerts a very small effect (Galvan, Ioannidis, and Dragani 2010). GWAS are generally based on the "common disease, common variants" assumption, where the most common genetic variants individually or in combination confer relatively small increments in risk (1.1 – 1.5 fold)

and only a small proportion of the phenotypic variation in population, are attributable to additive genetic factors (Hindorff et al. 2009). For example, only 5% of variation for human heights are attributed to additive genetic factors (Visscher 2008), For more complex traits, such as cancer, only <10% of phenotypic variation is explained by common variants (Frazer et al. 2009), and the variants identified through these studies have small effect size.

The difficulty with unexplained genetic variance is referred to as the "missing heritability" problem. Much of the speculation about missing heritability from GWAS has focused on the possible contribution of variants of low minor allele frequency (MAF), defined as roughly 0.5%<MAF<5%, or of rare variants (MAF<0.5%) (Manolio et al. 2009).

However, while many variants surely remain to be found, recently, people start to realize that for the majority of complex traits, such as cancer, the causes of the cellular changes are not only due to genetic factors, but also environmental factors and their interactions (Figure 1). For example, World Health Organization (2014) point out that, tobacco use is the cause of about 22% of cancer death; another 10% is due to obesity, a poor diet, lack of physical activity, and drinking alcohol. Other factors also have been involved, including infections, exposure to ionizing radiation, and environmental pollutants (Anand et al. 2008). For example, in the developing world, nearly 20% of cancers are due to infections such as hepatitis B, hepatitis C, and human papillomavirus (World Health Organization 2014). These factors interact together, which can alter the

phenotypes of mammalian cells, without changing the underlying DNA sequence. As a result, this causes increasing interest to researchers, in exploring non-genetic sources of phenotypic variation, including epigenetic changes.

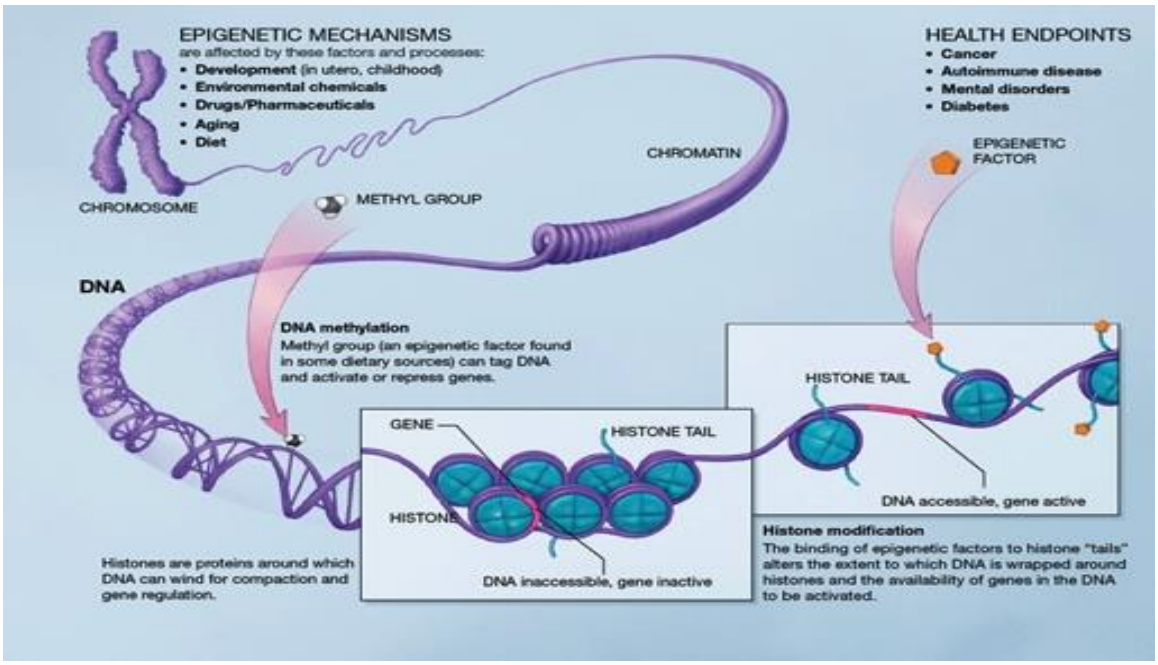

Figure 1: Epigenetics

1.1 DNA Methylation

Epigenetic state, defined as epigenetic inheritance, refers to a heritable change in the pattern of gene expression that is mediated by mechanisms other than alternations in the primary nucleotide sequence of a gene (Bird 2002, Russo, Martienssen, and Riggs 1996). As a result of a complex interplay of genetic and environmental factors, epigenome of a cell is highly dynamic to allow genetically identical cell to achieve diverse stable phenotypes in different organs (Bernstein, Meissner, and Lander 2007).

Epigenetic inheritance is the key to our understanding of the differences between growing, senescent and immortal cells, tumor and normal cells, various differentiated cells, and aging cells. Epigenetic templates that control gene expressions are transmitted to daughter cells independently of the DNA sequence. Normal cellular function relies on the maintenance of epigenomic homeostasis, but these stable patterns can sometimes become abnormal during fetal development, thereby cause pediatric cancer, or they can change during normal aging and contribute to common cancer risks in adults.

There are two main, inter-related types of epigenetic inheritance: DNA methylation (DNAm) and histone modification (Figure 1). The integration of DNA methylation with other epigenetic modification is clearly a complex process that depends on the collaboration of numerous components, many of which remain to be elucidated (Jin, Li, and Robertson 2011). Analysis of DNA methylation profiles will therefore enhance our understanding of the entire epigenome.

About 3-6% of all cytosines are methylated in normal human DNA (Esteller 2005). DNAm have been shown to be involved in cellular defense mechanisms, gene activation, embryonic development, gene imprint, allele inactivated, cell differentiation and development, as well as aging and cancerous (Bird 1986, Baylin 1997, Bestor and Tycko 1996).

In most cases, DNAm refers to hypermethylation of tumor-suppressor gene. Reactions using S-adenosyl-methionine as a methyl donor and catalyzed by enzymes called DNA methyltransferases (DNMTs) add a methyl group to the cytosine ring to form

5- methyl cytosine (5-mC) (Figure 2), which is also called “fifth base”, besides the four bases (adenine, guanine, cytosine, and thymine) that spell out the primary sequence of DNA.

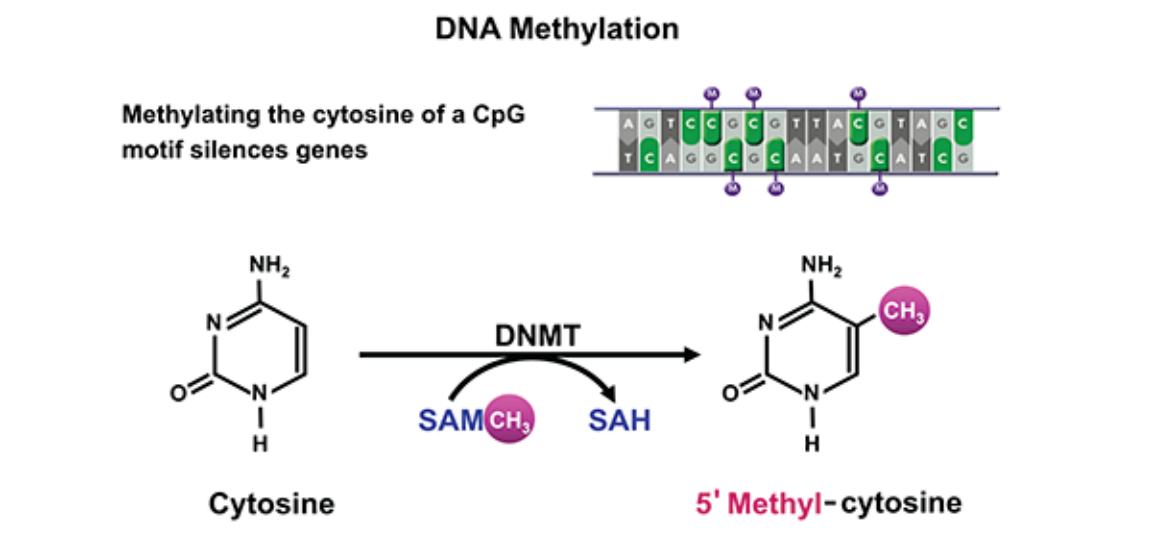

Figure 2: DNA Methylation

(http://pubs.niaaa.nih.gov/publications/arcr351/images/zakhari01.png)

DNAm variation at a single CpG site is known as a methylation variable position (MVP), which can be considered as the epigenetic equivalent of a SNP (Rakyan et al. 2004). In somatic cells, 5-mC occurs almost exclusively in the context of paired symmetrical methylation of a CpG site, in which a cytosine nucleotide is located next to a guanidine nucleotide.

Potentially methylated CpG dinucleotides are not randomly distributed in the human genome; instead, human genome consists of vast oceans of DNA sequence containing sparsely distributed but heavily methylated CpG dinucleotides punctuated by

short regions with unmethylated CpGs occurring at higher density, forming distinct islands in the genome (Bird et al. 1985). These CpG-rich regions are known as CpG islands, while the regions of low CpG density are defined as CpG oceans. Transition zones between CpG islands and CpG oceans are called CpG shores and display more tissue-specific variation in DNA methylation (Irizarry et al. 2009). In the bulk of the genome, about 80% of the CpG dinucleotides that are not associated with CpG islands are heavily methylated. In contrast, the dinucleotides in CpG islands, especially those associated with gene promoters, are usually unmethylated, whether or not the gene is being transcribed (Bird 2002).

The role of DNAm variation in complex disease has mainly been explored in the context of cancer, and has been discussed for more than 3 decades (Ehrlich and Wang 1981). The hypermethylation of tumor-suppressor genes, which is associated with their transcriptional silencing, is the key to the tumorigenic process, contributing to all of the typical hallmarks of a cancer cell that results from tumor-suppressor inactivation (Jones and Baylin 2002). Through further research, people realize that DNA methylation at CpG loci have important implications not only in cancer, but also other complex diseases (Kulis and Esteller 2010, Spisak et al. 2012).

DNA methylation has been found to not only be a marker for disease status, such as diagnosing cancer (Qureshi, Bashir, and Yaqinuddin 2010), but also can be used as a marker to differentiate disease severity, such as early and late stages in breast cancer (Klajic et al. 2013), ovarian cancer (Watts et al. 2008) and prostate cancer (Hoque 2009).

Most of DNA methylation have potential functions in inducing and suppressing cancer metastasis. Besides that, DNA methylation has been found to be associated with tumor size in colorectal cancer (Mitomi et al. 2010). Also patients with higher methylation further show more frequent recurrence compared to the low-methylation group, and shorter cancer related survival and recurrence-free survival (Mitomi et al. 2010). These findings show that it is very important to have better understanding of the epigenetics of cancer progression and metastasis.

1.2 NGS Technologies

Next-generation sequencing (NGS) technologies offer potential to accelerate epigenomic research substantially. With the development of technology, the traditional approach of isolating individual genes and studying them is being rapidly replaced by data sets generated from both individual laboratories and large consortia using new high-throughput technologies. High-throughput technologies now allow genome-scale mapping of DNA methylation and covalent modifications of histone proteins (Ren et al. 2000, Johnson et al. 2007, Lister et al. 2009). By using them, it is now feasible to interrogate various aspects of cellular processes, including sequence and structural variations and the transcriptome, epigenome, proteome and interactome (Ren et al. 2000).

1.2.1 Bisulfite Genomic Sequencing

Several large-scale analysis techniques are available that enable the survey of DNA methylation status at nucleotide resolution throughout the genome (Laird 2010, Cokus et al. 2008, Lister et al. 2008, Meissner et al. 2008, Pomraning, Smith, and Freitag

2009), including next-generating sequencing coupled with bisulfite treatment of DNA (Ren et al. 2000).

Key advantages of bisulfite sequencing are its comprehensive genomic coverage, high quantitative accuracy and excellent reproducibility. Steps in bisulfite sequencing are presented in Figure 3 (Lee et al. 2014). It starts with template preparation, consisting of building a library of nucleic acids (DNA or complementary DNA (cDNA)) and amplifying that library. Building libraries is to break genomic DNA into smaller sizes from which either fragment templates or mate-pair templates are created. Then emulsion PCR (emPCR) (Dressman et al. 2003) is used to ligate adaptors containing universal priming sites to the target ends, allowing complex genomes to be amplified with common PCR primers. Later, the DNA is separated into single strand and captured onto beads under condition that favor one DNA molecular per bead. After the successful amplification and enrichment of emPCR beads, the template is immobilized to a solid surface or support. The immobilization of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously. The nucleic acid is sequenced based on the library fragment and aligned to a reference genome, which means to find the position in the reference where the reads match with a minimum number of differences.

1.2.2 Bisulfite Microarrays

Bisulfite genomic sequencing remains as the gold standard for detection of DNA methylation – because such an approach can allow for a comprehensive assessment of a

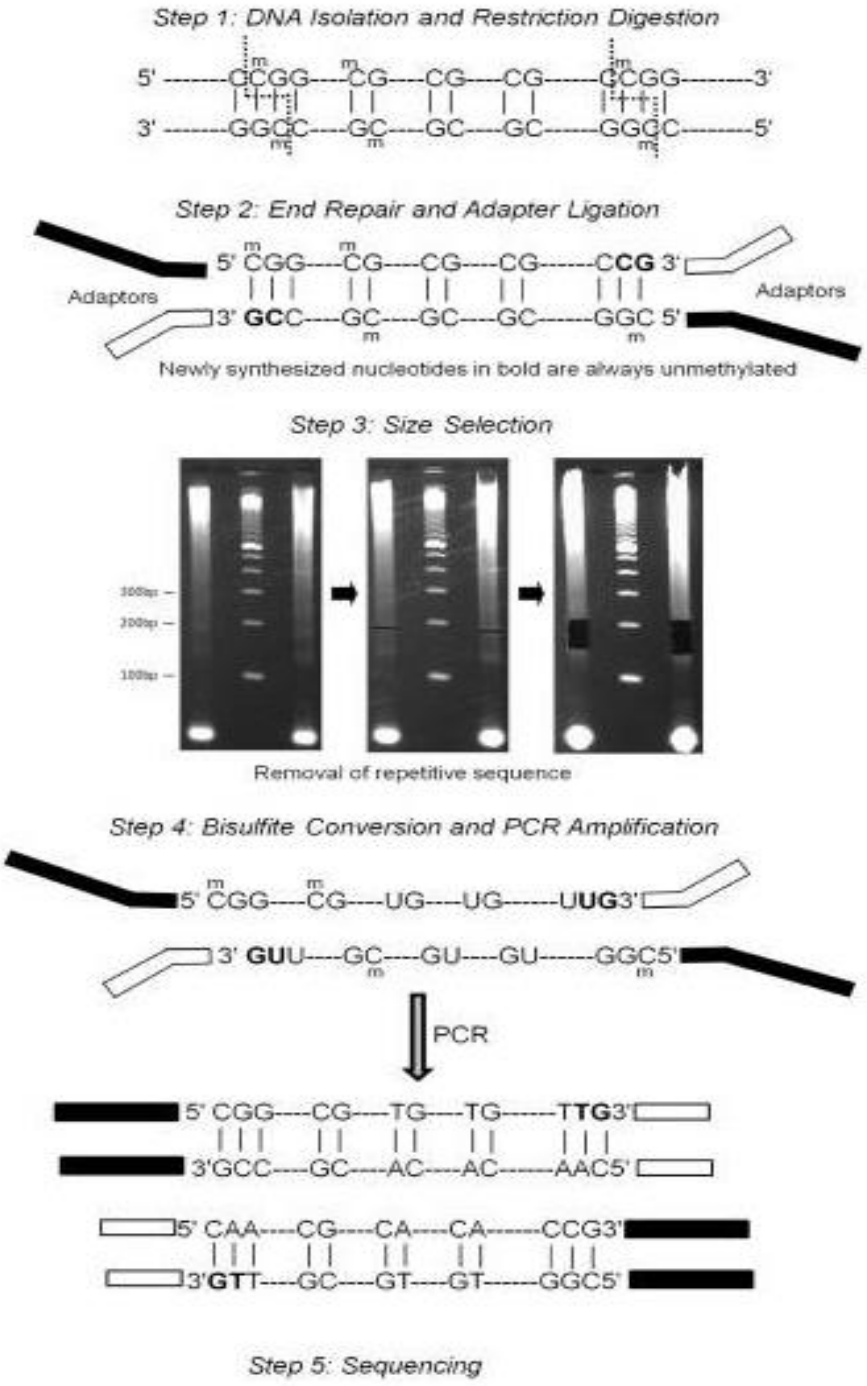

Figure 3: Key laboratory steps in bisulfite sequencing

small number of samples (Lister et al. 2009). Some biomarker research, however, requires effective high-throughput processing of hundreds of samples, for example, from clinical cohorts. The best compromise thus far in terms of reagent costs, time of labor, sample throughput and coverage may be the recently developed Illumina 450K Infinium Methylation BeadChip, which is by far the most widely used bisulfite microarray, and has been the focus of substantial bioinformatic methods development.

The Illumina 450K Infinium Methylation BeadChip allows researchers to assess the DNA methylation levels of close to half a million CpG sites distributed across the genome, which corresponds to 2% of all CpG sites of the human genome. It offers a good balance of genome-wide coverage, resolution (single base pair) and throughput (12 samples per chip and up to 96 samples per run). So it has enabled epigenome-wide association studies (EWAS) to explore the relationship between phenotypes and DNA methylation in large population-based studies (Rakyan et al. 2011).

However, the shortcoming of the chip technology is the high amount (about 500 nanograms) of required input DNA. Moreover, in contrast to sequencing-based methods, this approach does not address allele-specific and SNP-specific methylation and does not allow the discovery of variation beyond the probed loci.

1.3 Detecting DMRs with NGS Data

From Next-Generation sequencing, the numbers of molecules with a cytosine (methylated) and numbers of molecules with a uracil (unmethylated) are provided at over 2 million CpG sites in each individual. The total numbers of cytosines and uracils at each

CpG site are defined as the sequencing coverage, which varies at different CpG sites even for one individual. Thus the methylation rate is suggested for analysis instead of using methylation counts, which is calculated as the ratio of methylated counts over the sequencing coverage at each CpG site.

DNA methylation rates are continuous when measured across a large number of cells. They are susceptible to measurement error, and densely spaced across the genome (Jaffe, Feinberg, et al. 2012, Eckhardt et al. 2006, Irizarry et al. 2008). Besides that, methylation rates of DNAm vary strikingly across the genome, with strong local influences of base composition on single nucleotide variation (SNV) and regional effects of sequence (Hodgkinson and Eyre-Walker 2011). It has been shown that methylation rates at CpG sites are affected by those at nearby CpG sites, and have complicated correlation structures (Leek et al. 2010).

Based on these findings, recent research focus has been expanded to use patterns of methylation in neighboring CpG sites to detect differentially methylated regions (DMRs) in the genome based on methylation data from Next-Generation Sequencing. The advantage of detecting DMRs lies in the fact that it not only can account for the spatial dependence of CpG sites in DMR, but also may increase statistical power. That is because even though differences at any individual site may be small, if they are persistent across a region, statistical power to detect them may be greater for the region.

In addition, methylated rates have been shown to be strongly associated with some covariates, such as age (Bell et al. 2012, Teschendorff et al. 2010) and gender

(Kibriya et al. 2011, Liu et al. 2010), which can inflate or exaggerate the effect size estimate if not accounted for. As a result, it is very important to adjust for confounding and covariates when detecting DMRs that are associated with disease.

1.4 Dissertation Overview

The goal of this dissertation is to detect DMRs associated with disease status, and DMRs associated with disease severity, especially regions have the pattern of increased (decreased) methylation rates as the disease progress.

The proposed statistical methods to detect DMRs are based on methylation data from Next-Generation Sequencing. They are able to not only adjust for covariates and other confounding factors that potentially affect disease risk, but also account for correlation of methylation rates among CpG sites.

Before presenting our statistical methods, literature reviews are performed on current statistical methods for detecting DMRs, and presented in Chapter 2. Some current statistical methods for detecting DMRs are based on, detecting individual differentially methylated cytosines (DMCs) first, and then chaining the CpG sites based on a pre-defined distance. More recently, statistical methods have been proposed to detect DMRs with control of false discovery rate (FDR) using functional data analysis. So far, there are no existing packages for detecting DMRs that are associated with disease severity, however some possible statistical solutions are discussed in Chapter 2.

Chapter 3 and 4 propose two methods to detect DMRs—kernel distance method (KDM) and scan statistic method (SSM). Since the methylation rates have been shown to

have strong association with individual covariates, logistic regression of methylation rates are used to adjust for covariates, before calculating kernel distance and scan statistics.

Kernel distance statistic in Chapter 3 is calculated as a function of the difference in methylation rates between case and control groups at each CpG site, incorporating the correlation among CpG sites using a kernel function. The main advantage of KDM is that it is fast to compute, since it can be expressed as a quadratic kernel statistic. However, the calculation of kernel distance statistic strongly depends on the scaling factor, which represents the lengths of DMRs. Since the lengths of DMRs along the genome are varying and unknown, this might reduce the power of KDM.

SSM in Chapter 4 uses likelihood ratio to test the difference of methylation rates between two groups. This method uses moving windows along the genome, with multiple window sizes, which would help to reflect locations and lengths of the DMRs, and eventually increase the power compared to KDM. Since methylation rates are correlated with those at close-by CpG sites, the correlation are adjusted by a mixed-effect model. SSM also has advantages that able to account for unequal sequencing coverage at a particular CpG site across all individuals, by adjusting the methylation counts and sequencing coverage based on Xu et al. (2013).

A Bayesian method with Bayes factor (BFM) are proposed and presented in Chapter 5, to detect DMRs that associated with severity levels of diseases, and detect the regions with increasing (or decreasing) methylation rates as the disease severity increases

(or decreases). Patients are classified into groups, based on the disease severity (e.g. stages of cancer), and DMRs are detected by using the moving windows along the genome. Within each window, the Bayes factor is calculated to compare two models corresponding to constant vs. monotonic methylation rates among the groups. A mixed-effect model is used to not only incorporate the correlation of methylation rates between CpG sites in the region, but also to adjust clustering effect of observations at a CpG site for each individual.

Computer simulations were conducted to compare using KDM and SSM to detect DMRs associated with disease status. They also were conducted to study the behavior of using BFM to detect DMRs that were associated with disease severity. The simulation results are presented in Chapter 6, and the applicability of all these methods are demonstrated using a bisulfite sequencing dataset from a chronic lymphocytic leukemia (CLL) study, with results presented in Chapter 7. In Chapter 8 we present conclusions for future work.

Chapter 2

Literature Review

Statistical methods have been developed to detect differentially methylated regions (DMRs) that are associated with disease status. Some current statistical methods for detecting DMRs, detect individual differentially methylated cytosines (DMCs) first, and then chaining the CpG sites based on pre-defined distance.

Therefore, statistical methods that compare methylation rates at each CpG site, between case and control groups, are reviewed in Section 2.1. Here we mainly focus on the statistical methods developed for bisulfite sequencing data.

It is important to define the regions when detecting DMRs. The most straightforward approach to define the regions is to use pre-defined regions, such as CpG islands, CpG shores, promoters and introns. Many statistical methods are proposed to detect DMRs by testing whether the pre-defined candidate regions are differentially methylated. These statistical methods and packages are reviewed in Section 2.2.

Statistical methods based on pre-defined regions must be distinguished from those that can define regions of differential methylation. The latter is considerably more difficult because ensuring control of the false discovery rate (FDR) at the region level is not trivial; in particular, FDR control at the site-level does not give a direct way to region-level control when the region itself is also to be defined (Lun and Smyth 2014).

Some statistical methods and packages that can detect DMRs without pre-defined regions are reviewed in Section 2.3.

There is no existing package to detect DMRs that are associated with disease severity; some possible solutions for statistical methods are discussed in Section 2.4.

2.1 Statistical Methods to Detect Differentially Methylated CpG Sites

Many statistical methods have been proposed to detect CpG sites that are associated with disease status. They focus on comprehensive DNA methylation analysis of single base site, in order to find differentially methylated cytosines (DMCs) (Bock 2012).

One naïve approach of testing differential methylation between groups (e.g., cases and controls) is to use Pearson's chi-square test of independence with a 2 by 2 contingency table (methylated/unmethylated × case/control), which is based on the sum of the counts from NGS across subjects within a group for a given CpG site. However, this approach is problematic because it does not consider the differences of the sequencing coverage for different individuals. Without considering different sequencing coverage, the results would be biased since individuals with large sequencing coverage have undue influence on the test statistics.

In order to avoid the problems of unequal sequencing coverage for different individuals at one CpG site, methylation rate is used instead of methylated count alone. Methylation rates are calculated as the ratio of methylated counts over the sequencing

coverage at a CpG site for each individual. Using methylation rates also has advantage of taking into account of the between-subject variability in methylation levels, compared to using methylation counts.

Then Student's *t*-test can be applied to methylation rates, to test the differences of methylation rates between two groups. However, as typical for proportion data, methylation rates are restricted between 0 and 1, while the *t*-test is defined over $-\infty$ to ∞. Especially, in real data, a substantial proportion of samples with CpG sites have methylation proportions equal 0 and 1. The distribution of methylation rates is typically skewed, and displays substantial heteroscedasticity (Smithson and Verkuilen 2006), it tends to show smaller variances when located near the boundaries 0 and 1 as compared to the center of the unit interval. Considering all the above reasons, the normality assumption of the *t*-test may not hold for NGS methylation data.

In order to solve the issues with a normality assumption, an alternative choice is using the Mann-Whitney U test. This nonparametric approach has greater efficiency than the *t*-test on non-normal methylation data, but it assumes all the observations are independent. In addition, this method doesn't consider unequal sequencing coverage of every individual, at a CpG site.

Xu et al. (2013) proposed an adjusted chi-square statistic to detect differential methylated loci, by treating the NGS reads at a specific CpG site as a cluster within each individual, and then the problem becomes to compare two proportions in the presence of

clustered data. The methylation proportions are calculated based on adjusted methylation counts and coverage with design effects due to clustering.

The design effect used for adjustment is calculated based on Rao and Scott (1992), which is the ratio of estimated variances of methylation rates with clustering and without clustering, reflecting the variance inflation due to clustering. Here the estimated variance of methylation rates without clusters is based on a binomial distribution. This method has the advantage of no specific model assumption needed for the intra-cluster correlation, however, Xu et al. (2013) method cannot accommodate for other factors, such as confounders and covariates.

In order to adjust for covariates, logistic regression of methylation rates is considered, such as those used in the R package Methylkit (Akalin et al. 2012) to detect DMCs. The main advantage of logistic regression is that it allows for the inclusion of sample specific covariates, thus can adjust for confounding variables and covariates. This is very important for methylation data, since the methylation rates have been shown to be strongly associated with some individual covariates, such as age and gender.

Logistic regression calculates the log of the odds of methylation, which changes the range from 0 to 1 for methylation rates, to $-\infty$ to ∞ with the logit transformation. Then Wald test can be used to test association between methylation rates and disease status.

2.2 Statistical Methods to Detect DMRs with Pre-defined Regions

The R package Methylkit (Akalin et al. 2012) provides functionality to do either analysis on tiling windows across the genome or pre-defined regions of the genome. Within the windows or regions, logistic regression and Fisher's exact test can be conducted to detect the differences of methylation levels between cases and controls. A sliding linear model (SLIM) method was used to correct for multiple hypothesis testing (Wang, Tuominen, and Tsai 2011).

Instead of using fixed-length tiling windows, differential methylation analysis package (DMAP) for reduced representation bisulfite sequencing (RRBS) data, uses MspI-digested fragment of 40-220 bp lengths (Stockwell et al. 2014). This brings the advantages of being able to detect some DMRs that can't be detected with large moving windows. Especially for RRBS, only 2.5% of the genome is sequenced, the majority of windows will be empty or have partial inclusion of fragments. Further, if a small region is variably/differentially methylated between individuals, use of a 1000 bp or longer window might dilute the variation (Ehrlich and Lacey 2013). However, DMAP does not allow detection of DMCs, which need to be analyzed using MethylKit.

Both methods may result in inflated type I error rates when testing for group differences, by pooling reads among individuals. This is because methylation levels often vary significantly across individuals, as observed in cancer samples (Hansen et al. 2011). Also these methods have the disadvantage that they will unfortunately miss low-CpG-density DMRs that are abundant in the genome and critical for gene expression.

A whole genome DNA methylation analysis pipeline -- MethylSig (Park et al. 2014), uses the beta-binomial distribution to account for sequencing coverage and biological variation. It assumes that the observations are binomially distributed, conditional on the methylation proportion at a particular site, while the methylation proportion itself can vary across experimental units (e.g. patients), according to a beta distribution. This method also can incorporate local information across a chromosome to improve estimates of variances and/or methylation levels.

Recently, Ryu et al. (2016) developed a new statistical test based on functional data analysis, the generalized integrated functional test (GIFT), which tests for regional differences in methylation based on differences in the functional relationship between methylation rate and location of the CpG sites. In this method, subject-specific functional profiles are first estimated using wavelets, and the average profile within groups is calculated. An ANOVA-like test is used to compare groups for a region, by comparing the overall functional relationship to the average curve within each group. This method has the limitation that it mainly focuses on testing for differential methylation of a region, and other tools are needed to identify the candidate regions first.

2.3 Statistical Methods to Detect DMRs without Pre-defined Regions

Methods are being developed to detect DMRs without pre-defined regions, including bump-hunting techniques (Jaffe, Murakami, et al. 2012). This is a general method for bump detection. It introduces a bump hunting frame method that combines surrogate variable analysis (SVA) (Leek and Storey 2007), a statistical method for

modeling unexplained heterogeneity in genomic measurements, with regression modeling, smoothing techniques and modern multiple comparison approaches for detecting regions of interest based on high-throughput, genome-wide DNA methylation data. This method successfully exploits the correlation structure of methylation data to identify DMRs, and provides a genome-wide measure of uncertainty.

Besides bump-hunting techniques, some methods are developed specifically for detecting DMRs based on bisulfite sequencing data. For example, BSmooth (Hansen, Langmead, and Irizarry 2012), which has been reported as the first software package applicable to both methylated cytosine (mC) calling and DMR detection (Saito, Tsuji, and Mituyama 2014); also BiSeq, an algorithm developed by Hebestreit, Dugas, and Klein (2013) based on targeted bisulfite sequencing approaches, such as RRBS. Both methods use functional data analysis methods, where the functional relationship between methylation and location is modeled to estimate a subject-specific profile.

BSmooth is a pipeline to detect DMRs in whole genome bisulfite sequencing data, it basically relies on a local-likelihood smoother to smooth the methylation values sample-wise, which is appropriate to the slowly changing methylation levels over a region observed in the data. The group differences are tested via a test statistic that is similar to a *t*-test at each CpG site, DMRs are defined as adjacent CpG sites with absolute *t*-statistics above a pre-defined threshold with permutations for significance testing.

However, this method depends on the pre-defined threshold for absolute *t*-statistics, which would hinder automated analysis and possible lead to biased conclusion.

Moreover, fixed-length chaining criteria maybe problematic for detecting DMRs whose lengths range from hundreds of base pair as in small CpG islands, to millions of base pairs as in cancer aberrations (Hon et al. 2012).

In order to improve this, Li et al. (2013) created an optimized algorithm – eDMR, for detecting and annotating DMRs. It uses an adjustable spatial parameter for distance that bins the data into segments, in order to exam the spatial auto correlation of methylation data based on both the methylation changes and the *p*-values for each bin (Pedersen et al. 2012). Then they calculate the significance of the regions by combining the *p*-values of DMCs within the refined region. A false discovery rate (FDR) correction is also applied to correct for multiple hypothesis testing for the combined *p*-values.

BiSeq is a package that also uses FDR procedure to control the expected proportion of incorrectly rejected regions. The main advantage of BiSeq compared to BSmooth is power improvement by a hierarchical procedure. Also, it takes spatial dependence into account. It starts with using a beta-binomial model to account for biological variation between replicates, and then tests significance at each CpG site in all target regions for methylation differences, with a triangular kernel to capture the step-like changes observed in their data. The resulting *p*-values for each CpG site are transformed into normalized *z*-scores, and then the average is calculated for a given region, and compared to those obtained from resampling data. Eventually, the significant target regions are trimmed to the actually DMRs.

Model-based Analysis of Bisulfite Sequencing data -- MOABS (Sun et al. 2014) is another package that using beta-binomial model for replicated bisulfite sequencing (BS-seq) DNA methylation data. The advantage of MOABS is the introduction of a new metric, called credible methylation difference. Instead of using p-values, using credible methylation differences can not only indicate DMRs, but also directly measure the magnitude of the methylation differences. The MOABS authors suggested grouping differentially methylated sites into DMRs using a hidden Markov model or alternatively testing of pre-defined regions, but no specific details were given. Besides that, MOABS does not sacrifice resolution with low sequencing coverage, and it even has enough power to detect single CpG resolution differential methylation in low CpG density regulatory regions, with low-depth BS-seq experimental design (Sun et al. 2014).

The package – Dispersion shrinkage for sequencing data (DSS), also uses a beta-binomial model, to account for this hierarchy of variation between and within replicates (Feng, Conneely, and Wu 2014). A shrinkage approach is used to improve the performance when the number of replicates is low. However, DSS method can't control FDR when defining DMRs. These authors also considered the situation with only one sample in each group, and developed DSS-single (Wu et al. 2015). Since this method has difficulty to calculate variation with only one sample, information from neighboring CpG sites are used to estimate biological variation, by incorporating spatial correlation using smoothing procedure.

Another method, Spatial Clustering Method (SCM), developed by Yip et al. (2014) can detect DMRs using both methylation measurements and locations of CpG sites. However, they can't account for covariates. Also, this method doesn't consider unequal sequencing coverage from all individuals at each CpG site.

2.4 Statistical Methods to Detect DMRs associated with Disease Severity

Besides detecting DMRs with different methylation rates between cases and controls, it is also important to detect DMRs associated with multiple disease severity groups. That can help to have a better understanding of disease progression and to predict clinical aggressiveness.

However, most analysis are conducted by creating dichotomies based on biological subtypes, such as early and late cancer stages, and then detect DMRs by comparing the differences of DNA methylation rates between two groups (Klajic et al. 2013, Watts et al. 2008, Hoque 2009). Although such approaches are not incorrect, they may lose the information of the multiple disease status due to collapsing or ignoring of groups, and it cannot provide optimal features for the analysis.

In order to use multiple disease status, it is possible to run multiple testing for the association between DNA methylation and multiple group responses, using the methods for two groups. Although we can simply run analysis for all pair-wise comparisons and reduce them, it is not trivial when considering the regional nature of DMR, and would increase the multiple testing burden.

Another possible method is the generalized linear model that includes indicator variables for different levels of disease status. This method has the advantage that it can adjust for covariates. However analysts are often faced with noisy estimates of the category-specific regression coefficients, which can lead to unreasonable patterns in the regression coefficients corresponding to different levels of disease status, and eventually reduce the power (Dunson 2003).

In order to improve the efficacy of an overall test, one can take advantage of the fact that cancer develops through a series of stages, or different levels of disease severity in general, and develop statistical methods that can incorporate the ordering of disease status. However, the widely used trend test is not an ideal method, since it requires scores or weights for different levels of disease status, which are generally unknown.

Chapter 3

Kernel Distance Method to Detect DMRs Associated with Disease Status

3.1 Introduction

In order to detect differentially methylated regions (DMRs) along the genome, a kernel distance method (KDM) is proposed to detect the regions of CpG sites, which have different methylation rates between case and control groups.

Tango (1984) proposed a KDM to detect geographical clustering of disease. This method begins with a defined geographic area, taking all possible $n(n-1)/2$ pairs of n cases, and evaluates positive relationships of spatial distances between the members of a pair (Mantel 1967). If a cluster of diseased cases exists, however, the average pairwise distances among them can be drowned out by many pairs of random large distances. For this reason, it is very important to find a good measure of distances between two points.

Mantel (1967) and others suggested truncating larger distances, with a pre-defined critical distance. Alternatively, Tango (1984) used a nonlinear metric of distance that decreases more rapidly than linear, and also proposed a quadratic statistic

$$\mathbf{Q} = \mathbf{r}'\mathbf{A}\mathbf{r},$$

where $\mathbf{r}$ is a vector of relative frequencies, and $\mathbf{A}$ is a pre-defined matrix of a measure of closeness between two points. The matrix $\mathbf{A}$ is later referred to as the kernel matrix in

Schaid et al. (2013), which can serve as a smoother to create a plot. If this null hypothesis is rejected, showing evidence of true DMRs, peaks in smoothing plot would be observed, which will help to detect and locate DMRs.

Many researchers have tried to find an appropriate kernel matrix to measure the closeness. Tango (1984) first introduced the matrix **A** with (i, j) entry A_{ij} as

$$A_{ij} = e^{-d_{ij}},$$

where d_{ij} denotes the linear distance between two points i and j, and extended it to a more generalized form (Tango 1995),

$$A_{ij} = \begin{cases} e^{-d_{ij}/\tau}, & \left|d_{ij}/\tau\right| \leq 1 \\ 0, & \text{otherwise} \end{cases},$$

which is more natural in general and also relatively robust to the selection of the scaling parameter τ. However, considering the unclear interpretation of τ, Tango (2000) proposed another exponential threshold model,

$$A_{ij} = \begin{cases} e^{-4\left(\frac{d_{ij}}{\tau}\right)^2}, & \left|d_{ij}/\tau\right| \leq 1. \\ 0, & \text{otherwise} \end{cases}$$

The scale parameter τ is interpreted as a measure of cluster size, equal to the maximum allowed distance between cases in the same cluster. Cases further apart cannot be considered to be in the same cluster. Large values of τ are sensitive to large cluster, and small values of τ to small cluster. Schaid et al. (2013) used the tri-weight function

$$A_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{\tau}\right)^2\right)^3, & \left|\frac{d_{ij}}{\tau}\right| \leq 1 \\ 0, & \text{otherwise} \end{cases},$$

which has similar shape as a popular non-compact Gaussian function $A_{ij} = e^{-d_{ij}^2/\tau}$ with similar scaled distance.

KDM provided by Schaid et al. (2013) can be used to model methylation rates, which is calculated based on the difference of methylation rates between case and control groups for each CpG site, with the tri-weight kernel function to measure how the correlation of the methylation rates at different CpG sites depend on the distance between CpG sites. As we know, the correlations of methylation rates at different CpG sites in DMRs decrease as the distances of the two CpG sites increase.

Considering that confounding factors and covariates could affect methylation rates, the methylation rates are adjusted using logistic regression, before calculating the kernel distance statistic. First, we calculate the adjusted expected values of methylation counts at each CpG site for each individual, and then the sum of differences between observed and expected values of methylation counts (residuals) is used to calculate the adjusted methylation rates over all individuals at each CpG site for cases and controls.

Finally DMRs are detected by calculating kernel distance statistic based on the difference of adjusted methylation rates between case and control groups at each CpG site.

3.2 Kernel Distance Method to Detect DMRs

KDM is proposed here to detect DMRs, by adapting Schaid et al. (2013)'s method. Suppose m_{kij} is the count of the methylation molecules at CpG site j of individual i in group k, here $k = A$ for cases and $k = U$ for controls. We assume that

$$m_{kij} \sim B(c_{kij}, p_{kij}),$$

where c_{kij} is the coverage, and p_{kij} is the true methylation rate at CpG site j for individual i in group k, $k = A, U, i = 1,2, \dots, n_k, j = 1,2, \dots, s$.

Considering individual covariates, such as age and gender, we first use logistic regression to fit all data from both groups,

$$\log\left(\frac{p_{kij}}{1-p_{kij}}\right) = \log\left(\frac{m_{kij}}{c_{kij}-m_{kij}}\right) = \beta_0 + \beta_1 x_{ki}, \tag{3.1}$$

where x_{ki} represents the covariate value of individual i in group k. The fitted odds are calculated for methylation at CpG site j for individual i in group k, then are used to get the corresponding expected methylation rate

$$\hat{p}_{kij} = \frac{\hat{m}_{kij}}{\hat{c}_{kij}} = \frac{\exp(\hat{\beta}_0+\hat{\beta}_1 x_{ki})}{1+\exp(\hat{\beta}_0+\hat{\beta}_1 x_{ki})}, \tag{3.2}$$

with the difference of observed and expected methylated counts at CpG j for individual i in group k is calculated as adjusted methylation count

$$r_{kij} = m_{kij} - \hat{p}_{kij} c_{kij}. \tag{3.3}$$

with sums are calculated at CpG site j in case and control groups, respectively. That is,

$$r_{Aj} = \sum_{i=1}^{n_A} r_{Aij} \text{ and } r_{Uj} = \sum_{i=1}^{n_U} r_{Uij}.$$

Then the group effects for cases and controls are quantified as

$$\hat{\beta}_{Aj} = \frac{r_{Aj}}{c_{Aj}} \text{ and } \hat{\beta}_{Uj} = \frac{r_{Uj}}{c_{Uj}},$$

with

$$C_{Aj} = \sum_{i=1}^{n_A} c_{Aij} \text{ and } C_{Uj} = \sum_{i=1}^{n_U} c_{Uij}.$$

Eventually the difference between two groups

$$\delta_j = \hat{\beta}_{Aj} - \hat{\beta}_{Uj}$$

is calculated at each CpG site, and used in the quadratic statistic

$$\mathbf{Q} = \boldsymbol{\delta}'\mathbf{A}\boldsymbol{\delta},$$

where **A** is a pre-defined matrix representing the correlations of methylation rates between CpG sites.

Generally, the correlations of methylation decrease as the distances of the two CpG sites increase. Therefore the kernel matrix **A** should be based on a function that determines how rapidly the correlation decreases to 0 as the distance increases. Here we use the tri-weight function (Schaid et al. 2013),

$$A_{ij} = \begin{cases} \left(1 - \left(d'_{jl}\right)^2\right)^3, & d'_{jl} \leq 1 \\ 0, & \text{otherwise} \end{cases},$$

where $d'_{jl} = d_{jl}/\tau$ is a scaled distance based on scaling factor τ, and d_{jl} measures the distance between CpG site j and site l.

Since the lengths and numbers of DMRs are unknown and difficult to predict, and the lengths of DMRs are varying in the whole area under study, it is difficult to determine the value of scaling factor that represent the cluster size. When an appropriate size of

clusters cannot be predicted and many clusters are expected, it is common to repeat the procedure using different values of τ. Tango (2000) suggests allow τ to vary continuously from a small value near zero upwards until τ reaches about half the size of the whole study area. For convenience, we consider 10 values of τ as proposed by Schaid et al. (2013).

When a single test statistic is computed, the distribution of the kernel distance statistic can be well approximated by a scaled chi-square distribution (Tango 2012). However, the use of multiple scaling factors leads to the inappropriateness of an approximating scaled chi-square distribution to calculate *p*-values, and permutation method is required instead. In order to avoid multiple testing problems caused by multiple scaling factors, the minimum *p*-value across multiple scaling factors is used.

When the null hypothesis is rejected, the scaling factor that corresponding to the minimum p-value is recorded as the length of DMR, τ^*, and the corresponding kernel distance can be calculated as,

$$Q(\tau^*) = \sum_{j=1}^{m}\sum_{l=1}^{m}\left(A_{jl}(\tau^*)\delta_j\delta_l\right),$$

with percent contribution to $Q(\tau^*)$ at each CpG site calculated as $U_j(\tau^*)/Q(\tau^*)$, where

$$U_j(\tau^*) = \sum_{l=1}^{m}\left(A_{jl}(\tau^*)\delta_j\delta_l\right).$$

Then the distribution of methylation rates can be plotted based on the percent contribution $U_j(\tau^*)/Q(\tau^*)$ versus CpG site j, which can give us a graphical view of potential DMRs.

One advantage of KDM is able to use logistic regression to adjust for covariates that would affect the association between methylation and disease status, and can account for the correlation among CpG sites by using the kernel function, when detecting DMRs. The tri-weight kernel function in our proposed method can model the decreased correlation of methylation rates among CpG sites, as the distances increased.

This method also has the advantage of very fast computing time (Schaid et al. 2013), however, the power of KDM is reduced since it strongly depends on the pre-defined scaling parameter τ to reflect the unknown value of cluster size. If the value of τ is not close to the actual size, it might become difficult to detect real DMRs.

A method that will allow choosing the value of a parameter to reflect the scale at which clustering occurs, would be a good alternative in terms of achieving higher statistical power.

Chapter 4

Scan Statistic Method to Detect DMRs

4.1 Introduction

Scan statistic method (SSM) is an alternative to KDM in order to detect DMRs associated with disease status. SSM is a likelihood-based method, calculates the likelihood ratio to test the difference of methylation rates between two groups. This method uses moving windows along the genome, with multiple window sizes, allowing more accurate evaluation of the location and sizes of DMRs.

SSM is first studied by Naus (1965) to detect clusters in a point process in the one-dimensional setting. He applied the idea of maximum frequency to the case of ungrouped data and proposed a 'scan' test to test the null hypothesis of a purely random Poisson process, which assumes the probability of cases follows a uniform distribution across different locations.

However, it is well known that methylation data do not follow a uniform distribution, therefore, a reasonable modification is needed to take into account more accurate underlying distributions of methylation data. Kulldorff (1997) developed a likelihood-based SSM, which was extended to detect genetic variants by Ionita-Laza et al. (2012) by considering the Bernoulli distribution of variants at each position for every individual. The scan statistic is calculated based on the likelihood ratio of the frequencies

of variants carried among cases and controls within a window versus outside the window, with moving windows along the whole genome. Then the maximum scan statistic over the windows of all possible window sizes is defined as the global statistic. However, the approach considered by Ionita-Laza et al. (2012) can't be adapted for methylation data, since methylated counts at each CpG site for every individual follow a binomial distribution instead, conditional on the sequencing coverage.

We propose a new approach based on scan statistic for methylation data, which not only uses a binomial distribution for the methylation counts, but also allows the correlation of methylation rates between CpG sites. A mixed-effect logistic regression model is used for accounting for correlation of methylation rates between CpG sites, and estimating methylation counts. Besides that, the mixed-effect model can also allow for individual covariates such as age and gender.

The differences between the observed and expected methylated counts (residuals) are calculated based on the mixed-effect regression logistic model, and adjusted by "design effect", which is calculated at each CpG site as proposed by Xu et al. (2013). Considering unequal sequencing coverage of all individuals in the case and control group at each CpG site, the inflation factor for variance inflation, which is called "design effect", is calculated by treating NGS reads at each CpG site as clusters within each individual. These clusters are a natural result of the experimental design and the nature of the binomial data being measured on each subject within each group. The sequencing

coverage at each CpG site are also adjusted by the design effect, and in turn used to calculate the scan statistic.

SSM has an advantage of being able to detect DMRs with more than two groups with multinomial group responses. The details of SSM with multinomial responses are presented in Section 4.3.

4.2 Scan Statistic Method for Case-control Studies

In this Section, we will develop SSM to detect DMRs based on the difference in methylation rates between two groups (cases and controls). Since methylation rate at each CpG site are affected by those of close-by CpG sites, the correlation of methylation rates is adjusted by a mixed-effect model first.

4.2.1 Adjusting for correlation between CpG sites

Suppose m_{kij} is the count of the methylation molecules at CpG site j of individual i in group k, here $k = A$ for cases and $k = U$ for controls. We assume that $m_{kij} \sim B(c_{kij}, p_{kij})$, where c_{kij} is the coverage, and p_{kij} is the true methylation rate at CpG site j for individual i in group k, $k = A, U, i = 1,2, \dots, n_k, j = 1,2, \dots, s$.

To account for the correlation of methylation rates for nearby CpG sites, a random slope and intercept logistic regression model is considered to model methylation counts at each CpG site for every individual. A random slope and intercept logistic regression has the following form,

$$\log\left(\frac{p_{kij}}{1-p_{kij}}\right) = \log\left(\frac{m_{kij}}{c_{kij}-m_{kij}}\right) = \beta_0 + \beta_1 s_j + \beta_2 x_{ki} + \nu_{0ki} + \nu_{1ki} s_j, \quad (4.1)$$

where s_j represents the distance of CpG site j from the start point. In the mixed-effect model setting, the random effect $\boldsymbol{\nu}_{ki} = \begin{pmatrix} \nu_{0ki} \\ \nu_{1ki} \end{pmatrix}$ is assumed to vary independently across individuals, with $\boldsymbol{\nu_{ki}} \sim N\left(0, \begin{pmatrix} \sigma^2_{\nu_{0ki}} & \sigma_{\nu_{0ki}}\sigma_{\nu_{1ki}} \\ \sigma_{\nu_{0ki}}\sigma_{\nu_{1ki}} & \sigma^2_{\nu_{1ki}} \end{pmatrix}\right)$.

By adding x_{ki} in the mixed-effect model (4.1), we can also adjust for the covariates as in (3.1). The fitted odds of methylation counts can be calculated for CpG site j of individual i in group k, and can be used to get the corresponding adjusted expected methylation rate $\hat{p}_{kij}$ as in (3.2). The difference of observed and expected methylated counts at CpG j for individual i in group k is calculated as residual $r_{kij} = m_{kij} - \hat{p}_{kij}c_{kij}$ as in (3.3), $k = A, U, i = 1,2, \dots, n_k, j = 1,2, \dots, s$.

Considering the unequal sequencing coverage among individuals in a group at each CpG site, NGS reads at each CpG site within an individual is treated as a cluster, and the clustered data analysis method used by Xu et al. (2013) is adopted in our method. To calculate the design effect of clustered data, we first calculate the adjusted methylation counts r_{Aj} and r_{Uj} at CpG site j in case and control groups, respectively, ignoring the clustering within individuals. That is,

$$r_{Aj} = \sum_{i=1}^{n_A} r_{Aij} \text{ and } r_{Uj} = \sum_{i=1}^{n_U} r_{Uij}.$$

Then the group effects are quantified as,

$$\hat{\beta}_{Aj} = \frac{r_{Aj}}{c_{Aj}} \text{ and } \hat{\beta}_{Uj} = \frac{r_{Uj}}{c_{Uj}},$$

with

$$C_{Aj} = \sum_{i=1}^{n_A} c_{Aij} \text{ and } C_{Uj} = \sum_{i=1}^{n_U} c_{Uij}.$$

The variances of the group effects are given by

$$\hat{V}\left(\hat{\beta}_{Aj}\right) = \frac{n_A \sum_{i=1}^{n_A} \left(r_{Aij} - c_{Aij}\hat{\beta}_{Aj}\right)^2}{(n_A - 1)C_{Aj}^2} \text{ and } \hat{V}\left(\hat{\beta}_{Uj}\right) = \frac{n_U \sum_{i=1}^{n_U} \left(r_{Uij} - c_{Uij}\hat{\beta}_{Uj}\right)^2}{(n_U - 1)C_{Uj}^2}.$$

However, without clustering, the variances of the group effects from a binomial distribution would be

$$\hat{V}_B\left(\hat{\beta}_{Aj}\right) = \frac{\hat{\beta}_{Aj}(1-\hat{\beta}_{Aj})}{c_{Aj}} \text{ and } \hat{V}_B\left(\hat{\beta}_{Uj}\right) = \frac{\hat{\beta}_{Uj}(1-\hat{\beta}_{Uj})}{c_{Uj}}.$$

The design effects are defined as,

$$d_{Aj} = \frac{\hat{V}(\hat{\beta}_{Aj})}{\hat{V}_B(\hat{\beta}_{Aj})} \text{ and } d_{Uj} = \frac{\hat{V}(\hat{\beta}_{Uj})}{\hat{V}_B(\hat{\beta}_{Uj})}. \quad (4.2)$$

The design effects are then used to calculate sum of adjusted methylation counts and sequencing coverage in cases and controls as

$$\tilde{r}_{Aj} = \left\lceil \frac{r_{Aj}}{d_{Aj}} \right\rceil \text{ and } \tilde{r}_{Uj} = \left\lceil \frac{r_{Uj}}{d_{Uj}} \right\rceil \quad (4.3)$$

$$\tilde{C}_{Aj} = \left\lceil \frac{C_{Aj}}{d_{Aj}} \right\rceil \text{ and } \tilde{C}_{Uj} = \left\lceil \frac{C_{Uj}}{d_{Uj}} \right\rceil. \quad (4.4)$$

4.2.2 Binomial Scan Statistic

We assume $\tilde{r}_{Aj} \sim B\ (\tilde{C}_{Aj}, p_A)$ and $\tilde{r}_{Uj} \sim B\ (\tilde{C}_{Uj}, p_U)$, where $p_A - p_U$ is the true difference in methylation rates between cases and controls. Considering $\tilde{r}_{kj} \sim B\left(\tilde{C}_{kj}, p_k\right)$, then the likelihood of $\tilde{r}_{kj}$ is given by

$$f\left(\tilde{r}_{kj}\right) = \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} p_k^{\tilde{r}_{kj}} (1 - p_k)^{\tilde{C}_{kj} - \tilde{r}_{kj}}$$

$$= \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} \, exp\left(\tilde{r}_{kj} \log\left(\frac{p_k}{1-p_k}\right) + \tilde{C}_{kj} \log(1-p_k) \right).$$

For a specific window, after adjusting for correlation between CpG sites by using the mixed-effect model, $(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks})$ for the s consecutive CpG sites are assumed to be independent. Then the joint likelihood of adjusted methylation counts over consecutive s CpG sites in the defined region for group k is the product of the likelihoods of the s CpG site, which can be expressed as,

$$f(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks}) = \prod_{j=1}^{s} \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} exp\left(\tilde{r}_{kj} \log\left(\frac{p_k}{1-p_k}\right) + \tilde{C}_{kj} \log(1-p_k) \right)$$

$$= \prod_{j=1}^{s} \binom{\tilde{C}_{kj}}{\tilde{r}_{kj}} exp\left\{ \sum_{j=1}^{s} \tilde{C}_{kj} \left(\frac{\sum_{j=1}^{s} \tilde{r}_{kj}}{\sum_{j=1}^{s} \tilde{C}_{kj}} \log\left(\frac{p_k}{1-p_k}\right) + \log(1-p_k) \right) \right\}.$$

From this likelihood, we can see that the distribution of adjusted methylated counts follow a one-parameter exponential family $\boldsymbol{y} = (\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks}) \sim EXP(\eta, \phi, T, B_e, a)$ where

$$T(\boldsymbol{y}) = T(\tilde{r}_{k1}, \tilde{r}_{k2}, \dots, \tilde{r}_{ks}) = \frac{\sum_{j=1}^{s} \tilde{r}_{kj}}{\sum_{j=1}^{s} \tilde{C}_{kj}}$$

$$\eta = log\left(\frac{p_k}{1-p_k}\right) \rightarrow p_k = \frac{\exp(\eta)}{1+\exp(\eta)}$$

$$B_e(\eta) = -\log(1-p_k) = \log(1+e^{\eta})$$

$$\phi = \frac{1}{\sum_{j=1}^{s} \tilde{C}_{kj}}$$

$$a(\phi) = 1$$

and the log-likelihood $l(\eta; \boldsymbol{y}) = (\eta T(\boldsymbol{y}) - B_e(\eta))/\phi$ after ignoring an additive constant that does not depend on η.

Based on this likelihood function, we can find the maximum likelihood estimator (MLE) of parameter η in the one-parameter exponential family $EXP(\eta, \phi, T, B_e, a)$ as $\hat{\eta} = g_e(T(\boldsymbol{y}))$, where $g_e(T) = (B_e')^{-1}(T) = \log(T) - \log(1-T)$ (Agarwal, Phillips, and Venkatasubramanian 2006).

Let η_A and η_U be the parameters for the data with two groups in the same specified region. In order to test the hypotheses $H_0: \eta_A = \eta_U$ versus $H_1: \eta_A \neq \eta_U$, the ratio of the likelihood under H_1 versus H_0 can be used as a test statistic. More conveniently, we can use the log of this likelihood ratio as our test statistic, which we can refer to as the scan statistic. It is given by

$$\Delta = \kappa(T_A, \Phi_A) + \kappa(T_U, \Phi_U) - \kappa(T, \Phi), \tag{4.5}$$

where $\kappa(x, y) = (x g_e(x) - B_e(g_e(x)))/y$ and $\frac{1}{\Phi} = \frac{1}{\Phi_A} + \frac{1}{\Phi_U}$, $T = b_A T_A + (1 - b_A) T_U$ with $b_A = \frac{1}{\Phi_A} / (\frac{1}{\Phi_A} + \frac{1}{\Phi_U})$. Here we have

$$\Phi_A = \frac{1}{\sum_{j=1}^{s} \tilde{C}_{Aj}} \text{ and } \Phi_U = \frac{1}{\sum_{j=1}^{s} \tilde{C}_{Uj}},$$

$$T_A = \frac{\sum_{j=1}^{s} \tilde{r}_{Aj}}{\sum_{j=1}^{s} \tilde{C}_{Aj}} \text{ and } T_U = \frac{\sum_{j=1}^{s} \tilde{r}_{Uj}}{\sum_{j=1}^{s} \tilde{C}_{Uj}}$$

for cases and controls.

The statistic Δ is called binomial scan statistic, and it is the measure of the strength of H_1 versus H_0. The larger Δ is, the more likely it is that H_1 is true. For each

moving window, the binomial scan statistic can be calculated using (4.5), which can further be decomposed as,

$$\Delta = \frac{T}{\Phi}\left(r_A \log\left(\frac{r_A}{b_A}\right) + \left(\frac{b_A}{T} - r_A\right)\log\left(1 - T\frac{r_A}{b_A}\right) + (1 - r_A)\log\left(\frac{1 - r_A}{1 - b_A}\right)\right.$$
$$\left. + \left(\frac{1 - b_A}{T} - 1 + r_A\right)\log\left(1 - T\frac{1 - r_A}{1 - b_A}\right)\right) - \frac{1 - T}{\Phi}\log(1 - T)$$

where $b_A = \frac{\sum_{j=1}^{S}\tilde{C}_{Aj}}{\sum_{j=1}^{S}\tilde{C}_{Aj}+\sum_{j=1}^{S}\tilde{C}_{Uj}}$, $r_A = \frac{\sum_{j=1}^{S}\tilde{r}_{Aj}}{\sum_{j=1}^{S}\tilde{r}_{Aj}+\sum_{j=1}^{S}\tilde{r}_{Uj}}$, $T = \frac{\sum_{j=1}^{S}\tilde{r}_{Aj}+\sum_{j=1}^{S}\tilde{r}_{Uj}}{\sum_{j=1}^{S}\tilde{C}_{Aj}+\sum_{j=1}^{S}\tilde{C}_{Uj}}$ and

$\Phi = \frac{1}{\sum_{j=1}^{S}\tilde{C}_{Aj}+\sum_{j=1}^{S}\tilde{C}_{Uj}}$.

4.3 Scan Statistic Method for Multinomial Responses

Here we consider scenario that the number of groups is greater than two, and the groups are classified based on multinomial responses, such as different types of cancer patients and control group. The goal is to test whether there is any significant difference of methylation rates among these groups.

Before testing the differences among groups, the methylation counts and sequencing coverage need to be adjusted. As in Section 4.2, first we use the mixed-effect logistic regression model (4.1) to adjust for relevant covariates and the correlation of methylation rates among CpG sites. As before, we allow different sequencing coverage for different individuals at each CpG site. The design effects in (4.2) are calculated based on Xu et al. (2013), and used to calculate adjusted methylation counts $\tilde{r}_{kj}$ and sequencing coverage $\tilde{C}_{kj}$ for group k at CpG site j, as in (4.3) and (4.4).

Assuming all CpG sites in a DMR for group k have same methylation rate p_k, with adjusted methylation counts $\tilde{r}_{kj} \sim B(\tilde{C}_{kj}, p_k)$. Let $\eta_k = \log\left(\frac{p_k}{1-p_k}\right)$ be the parameters for the logit transformation of methylation rates of group k, then we are testing the hypothesis

$$H_0: \eta_1 = \eta_2 = \cdots = \eta_K = \eta$$

$$H_1: \text{Not all } \eta_1, \eta_2, \ldots, \eta_K \text{ are equal}$$

Here the groups are assumed to be independent, and the ratio of the likelihood under H_1 versus H_0 can be used as a test statistic. More conveniently, we can use the log of this likelihood ratio as our test statistic, which we can refer to as the scan statistic. It is given by

$$\Delta = \sum_{k=1}^{K} \kappa(T_k, \Phi_k) - \kappa(T, \Phi), \qquad (4.6)$$

where $\kappa(x, y) = (x g_e(x) - B_e(g_e(x)))/y$ and $T_k = \frac{\sum_{j=1}^{S} \tilde{r}_{kj}}{\sum_{j=1}^{S} \tilde{C}_{kj}}$, $\Phi_k = \frac{1}{\sum_{j=1}^{S} \tilde{C}_{kj}}$.

Define $\Phi = \frac{1}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$, $T = \frac{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{r}_{kj}}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$, thus $\frac{1}{\Phi} = \sum_{k=1}^{K} \frac{1}{\Phi_k}$, $T = \sum_{k=1}^{K} b_k T_k$ with $b_k = \frac{\frac{1}{\Phi_k}}{\frac{1}{\Phi}} = \frac{\frac{1}{\Phi_k}}{\sum_{k=1}^{K}\frac{1}{\Phi_k}} = \frac{\sum_{j=1}^{S} \tilde{C}_{kj}}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$. Then we have the scan statistic (4.6) simplified to,

$$\Delta = \sum_{k=1}^{K} \frac{T}{\Phi}\left(r_k \log\left(\frac{r_k}{b_k}\right) + \left(\frac{b_k}{T} - r_k\right)\log\left(1 - T\frac{r_k}{b_k}\right)\right) - \frac{1-T}{\Phi}\log(1-T)$$

where $b_k = \frac{\sum_{j=1}^{S} \tilde{C}_{kj}}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$, $r_k = \frac{\sum_{j=1}^{S} \tilde{r}_{kj}}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{r}_{kj}}$ and $\Phi = \frac{1}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$, $T = \frac{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{r}_{kj}}{\sum_{k=1}^{K}\sum_{j=1}^{S} \tilde{C}_{kj}}$.

The scan statistic Δ is calculated for each window using moving windows with a variable window (VW) size approach across the whole genome. DMR is the window with the highest value of the scan statistic. For each window W of size w, the binomial scan statistic can be calculated, and the one with the highest value denoted by LR_w. Then the maximum of LR_w over all values of w is used as the global statistic,

$$LR = \max_{w} LR_w.$$

However, LR calculation is unstable if the frequency of methylated counts within a given window is 0, for either cases or controls, a pseudo-count of 1 is added to the adjusted methylated and unmethylated counts at each CpG site, these additions efficiently assume that the null hypothesis is true at all sites.

Since the distribution of the scan statistic is unknown, an approximate p-value for the window with the largest LR_w is calculated using the permutation method. For case-control studies, SSM is expected to have higher power than the KDM, since SSM using moving window with variable window size overcomes the difficult problem of determining the value of scaling factor τ in the KDM. The use of moving windows also can result more accurate region of DMRs. In addition, SSM accounts for within cluster correlation by using the clustering effect based on Xu et al. (2013), while KDM does not.

SSM also has the advantage that it can be used for more than two groups, while KDM can only be used for two groups since it is calculated based on the difference of methylation rates at each CpG site. Though SSM can handle multiple groups, it cannot account for the ordering of the group responses. If ordering needs to be considered, the

maximum likelihood estimate becomes extremely tedious to calculate. Therefore, alternative approaches need to be developed to handle ordered multiple groups.

Chapter 5

A Bayesian Approach to Detect DMRs Associated with Disease Severity

5.1 Introduction

The KDM in Chapter 3 can only be used only for testing differentially methylation in two groups, while SSM can be extended to multinomial group responses. Here we consider groups classified based on ordinal responses, such as different stages of cancer, or severity levels of diseases, and test the association between methylation rates and disease severity, and find DMRs having the pattern of increased methylation rates as the diseases progresses. To this end, we propose a statistical method that not only can adjust for correlation of methylation rates between and within CpG sites, but also can incorporate monotonicity in response.

Classical statistical inference under constrained parametric spaces has been addressed by many authors, among which Bartholomew (1959) presented one of the first tests for the binomial problems employing inequality constraints. He proposed a test of $H_0: p_1 = p_2 = \cdots = p_K$ against the simple order $H_1: p_1 \leq p_2 \leq \cdots \leq p_K$ with at least one strict inequality, here p_k $(k = 1,2, \ldots, K)$ represents the binomial proportion. Under H_0, the ML estimator of p_k is the overall sample proportion π_k. If the sample binomial proportion satisfies $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_K$, then the order-restricted ML estimator is $\hat{p}_k =$

π_k (Bartholomew 1959). However, sometimes the sample proportions may not satisfy the ordering $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_K$, in that case, calculation of the restricted maximum likelihood estimator (RMLE) is subject to arbitrary orderings in the parameters is often difficult, and requires specialized algorithms that are not easily generalizable (Dunson 2003).

In order to solve this problem, Robertson and Wegman (1978) proposed a likelihood ratio statistic for the inequality-constrained binomial problem as a special case of an LR test, which compares parameters for independent samples from a single-parameter exponential family distribution. Before calculating the test statistic, they used the pool adjacent violator algorithm (Ayer et al. 1955) to pool "out-of-order" categories for which $\pi_k > \pi_{k+1}$ until the resulting sample proportions are monotone increasing. The order-restricted ML estimators $\hat{p}_k$ are the adjusted sample proportions.

The idea of applying an isotonic transformation to the unconstrained parameter estimates, motivated Dunson (2003) to create a Bayesian alternative approach for this problem, which has been adapted here. He proposed to use Bayes factors for assessing ordered trends, which are calculated based on the output from Gibbs sampling. The samples from the order-constrained model are derived by transforming draws from an unconstrained posterior density using an isotonic regression transformation.

Here we propose a Bayesian method using Bayes factor (BFM) to detect the region with increasing (or decreasing) methylation rates as the disease severity increases (or decreases). Patients are classified into groups based on the disease severity (e.g.

stages of cancer), and DMRs are detected by using the moving windows along the genome. Within each window, the Bayes factor is calculated and is used to test the hypothesis of constant versus monotonic increase in methylation rates corresponding to severity of the disease. A linear mixed-effect model is used to incorporate the correlation of methylation rates between and within CpG sites in the region.

5.2 Bayesian Method to Detect DMRs

Suppose m_{kij} is the count of the methylation molecules at CpG site j of individual i in group k, $k = 1,2 \dots K$. We assume $m_{kij} \sim B\ (c_{kij}, p_{kij})$, where c_{kij} is the coverage, and p_{kij} is the true methylation rate at CpG site j for individual i in group k.

To allow the correlation of methylation rates of nearby CpG sites, we consider a mixed-effect model. The logit link function for the methylation rate p_{kij} is expressed by

$$logit(p_{kij}) = \mu_k + \nu_{0ki} + \nu_{1kij}, \tag{5.1}$$

with ν_{0ki} and ν_{1kij} are the random effects. The random effect $\nu_{0ki} \sim N(0, \sigma_{\nu_0}^2)$ is used to model the correlation of methylation rates within each CpG site, while the random effect $\boldsymbol{\nu_{1ki}} = (\nu_{1ki1}, \nu_{1ki2}, \dots, \nu_{1kim})^T \sim N(\boldsymbol{\mu_0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu_0} = (0,0 \dots 0)^T$ is used to model the correlation of methylation rates between CpG sites.

Here μ_k in (5.1) is the fixed effect for each group, measures the association between methylation rates and group responses. The strength and direction of the association is modeled by prior distribution $N(\mu_\mu, \sigma_\mu^2)$, which means the parameters of μ_μ

and σ_μ^2 control the distribution of μ_k, and implies that all of the methylation rates are draws from a common distribution.

With assigned hyperpriors $\mu_\mu \sim N(0,1000^2)$, $\sigma_\mu^2 \sim IG(1,100)$, $\sigma_{v_0}^2 \sim IG(1,100)$ and $\mathbf{\Sigma^{-1}} \sim wish(\mathbf{I_m}, m)$ for m CpG sites in the moving window, the posterior distribution of μ_k is based on the mixed-effect logistic model (5.1), and used to calculate the Bayes factor for comparing these two models,

$$M_0: \mu_1 = \mu_2 = \cdots = \mu_K$$

$M_1: \mu_1 \leq \mu_2 \leq \cdots \leq \mu_K$ with at least one strict inequality,

to see whether there is an ordered constraint of methylation rates corresponding to severity of the disease.

In order to calculate the Bayes factor, first we draw samples $\mu_1, \mu_2, \ldots, \mu_K$ from the posterior distribution by using Gibbs sampling.

After that, an isotonic transformation is used to transform $\mu_1, \mu_2 \ldots \mu_K$ into $\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_K$, with $\tilde{\mu}_1 \leq \tilde{\mu}_2 \leq \cdots \leq \tilde{\mu}_K$ (Dunson 2003), with the min-max formula for the isotonic formula is

$$\tilde{\mu}_k = g_k(\boldsymbol{\mu}) = \min_{t \in U_k} \max_{s \in L_k} \left(\frac{\mathbf{1}'_{t-s+1} \boldsymbol{V}^{-1}_{[s:t]} \boldsymbol{\mu}_{[s:t]}}{\mathbf{1}'_{t-s+1} \boldsymbol{V}^{-1}_{[s:t]} \mathbf{1}_{t-s+1}} \right) \text{ for } j = 1,2, \ldots, K, \qquad (5.2)$$

here $\boldsymbol{V}$ is estimated based on samples from the posterior density of $\boldsymbol{\mu}$.

Also samples $\mu_1^0, \mu_2^0, \ldots, \mu_K^0$ are drawn from the prior density and transformed into $\tilde{\mu}_1^0, \tilde{\mu}_2^0, \ldots, \tilde{\mu}_K^0$, with $\tilde{\mu}_1^0 \leq \tilde{\mu}_2^0 \leq \cdots \leq \tilde{\mu}_K^0$, by using isotonic transformation (5.2).

The Bayes factor for each window (with moving windows along the genome) is given by,

$$BF = \frac{P(M_1|data)/P(M_1)}{P(M_0|data)/P(M_0)} = \frac{P(\tilde{\mu}_K > \tilde{\mu}_1)/P(\tilde{\mu}_K^0 > \tilde{\mu}_1^0)}{P(\tilde{\mu}_K = \tilde{\mu}_1)/P(\tilde{\mu}_K^0 = \tilde{\mu}_1^0)}$$

The windows with highest value of the Bayes factor among all windows is used for evaluating DMRs. BFM can detect DMRs associated with disease severity, especially detects DMRs with increasing (or decreasing) methylation rates, as the disease severity increase (or decrease). This method uses a mixed-effect model to not only adjust for correlation of methylation rates between CpG sites within each moving window, but also correlations within CpG sites.

Chapter 6

Simulation

This chapter considers simulation studies to compare the scan statistic method (SSM) and the kernel distance method (KDM). They were compared by empirical type I error, empirical power and computational efficiency. Simulation studies were also conducted to study the behavior of the Bayesian method using Bayes factor (BFM), when considering four ordinal group responses.

6.1 Comparison of Scan Statistic and Kernel Distance Methods

The first focus for simulation was to compare SSM and KDM, with respect to type I error, power and computational efficiency. For simplicity, (i) we did not include any covariates, and (ii) we used equal sample sizes for cases and controls in the simulation models.

Data were generated first under the null hypothesis to evaluate the validity by assessing how well each method would control the type I error without being overly conservative. In other words, we expected our test to have the actual type I error just under the nominal significance level, but not too far below the nominal level. A too conservative test (actual type I error substantially lower than the nominal level) would usually be underpowered. After establishing the validity for each method, we generated

data under the alternate hypothesis, for various levels of departures from the null, to compare the power for a broad set of alternatives, at various significant levels.

6.1.1 Data Generation Procedure

For the power comparisons at various alternate hypotheses and various significant levels, we assumed that there was only one DMR in the simulated genomic region, and all CpG sites within the region were equally spaced.

We generated a sample that contained N cases and N controls, and assumed every individual had equally spaced m CpG sites in the simulated region, of which r consecutive CpG sites in the middle were in the DMR.

Methylation counts at each CpG site for every individual were generated from $B(c_{kij}, p_{kij})$, $i = 1,2,\dots,N$, $j = 1,2,\dots,m$, $k = A,U$. Here the sequencing coverage c_{kij} were allowed to vary by sampling the values of it from a normal distribution $N(30,13)$, and then rounded to a integer, with a minimum of 5 based on the real data analysis from Xu et al. (2013). The correlated methylation rates p_{kij} were simulated based on a two-step procedure, proposed by Lacey, Baribault, and Ehrlich (2013) in order to model the spatial dependence for the methylation rates of nearby CpG sites.

First, independent random samples X_{kij} were generated from Beta-distribution for CpG site j of individual i in group k. Under the null hypothesis, X_{kij} were generated as $X_{kij} \sim Beta(\alpha_U, \beta_U)$, $k = A,U$. Under the alternative hypothesis, X_{kij} were generated under the same distribution for CpG sites outside of DMR. Within the DMR under the

alternate hypothesis, X_{kij} were generated $X_{Aij} \sim Beta(\alpha_A, \beta_A)$, where $\alpha_A \neq \alpha_U$ and $\beta_A \neq \beta_U$, so that the methylation rates were different between cases and controls within the DMR. Based on the property of the Beta distribution, with fixed α_U, β_A and β_U, only the values of α_A were changed, with effect size defined as $d = \frac{\alpha_A}{\alpha_A+\beta_A} - \frac{\alpha_U}{\alpha_U+\beta_U}$.

For each individual in each group, the vector of independent random variables $\mathbf{X_{ki}}$ was later transformed into a vector of correlated random variables with correlated methylation rates $\mathbf{p_{ki}} = 1 - \Phi(C\Phi^{-1}(1 - \mathbf{X_{ki}}))$, where $\Phi(\cdot)$ denoted the cdf of the standard normal distribution function with Cholesky decomposition C of the correlation matrix $\Sigma = CC'$.

All diagonal elements of the correlation matrix Σ are 1s, and off-diagonal element (i, j) was defined as the correlation coefficient ρ divided by the distance between CpG sites i and j, in order to allow for the fact that correlation of methylation rates for two CpG sites decreases as the distance between them increases.

6.1.2 Parameters for Simulation

As presented in Table 1, simulations were conducted with type I errors of 0.05 and 0.01, total sample sizes 48 and 60 with equal sample sizes in each group, and regions of 24 and 30 CpG sites with 6 in the middle as the DMRs. We assumed correlation coefficients of $\rho = 0.7$ and $\rho = 0.5$ for methylation rates between adjacent CpG sites, and those among non-adjacent sites were scaled down by dividing ρ by the distances

between sites. We set $\alpha_U = 0.1$, $\beta_A = \beta_U = 0.9$, and used different values of α_A to get different effect sizes d_{DM}. The simulation parameters are given in Table 1.

type I error (α)	0.05, 0.01
sample size per group (N)	24, 30
total number of CpG sites (m)	24, 30
number of CpG sites in DMR (r)	6
correlation coefficient (ρ)	0.5, 0.7
effect size (d_{DM})	0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2

Table 1: Parameters for simulation to compare SSM and KDM

6.1.3 Simulation Results

In order to present the effect of the parameters on the power of SSM and KDM, plots of power versus different values of effect sizes under the type I error of 0.05 are presented in Figures 4 and 5, corresponding to the 24-sites and 30-sites regions, respectively. The powers for SSM and KDM are very close to the type I error when the effect size is 0. The plots also show that the powers for SSM and KDM increase as the effect sizes increase; they also increase as sample sizes increase. Besides that, the plots show that SSM has uniformly better power than KDM.

Similar simulation results corresponding to type I error of 0.01, with the results are presented in Figure 6 and Figure 7. These results also show that both methods control the type I error, and that SSM have better power than KDM. As the significance level is reduced, the powers are also reduced as expected.

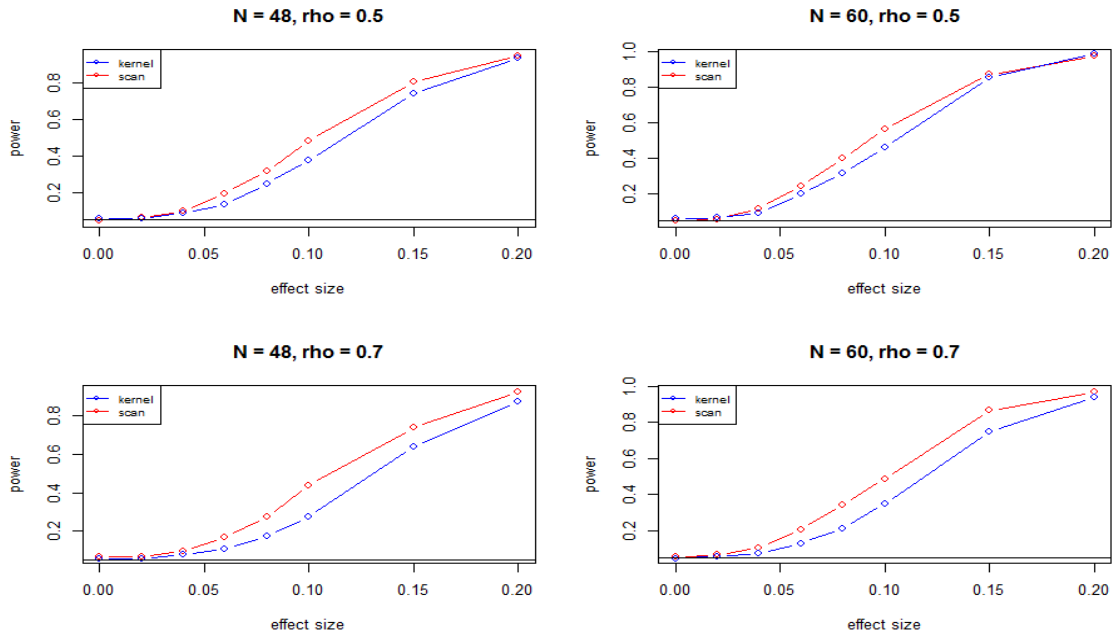

Figure 4: Power curves for SSM and KDM with 24 CpG sites, $\alpha = 0.05$

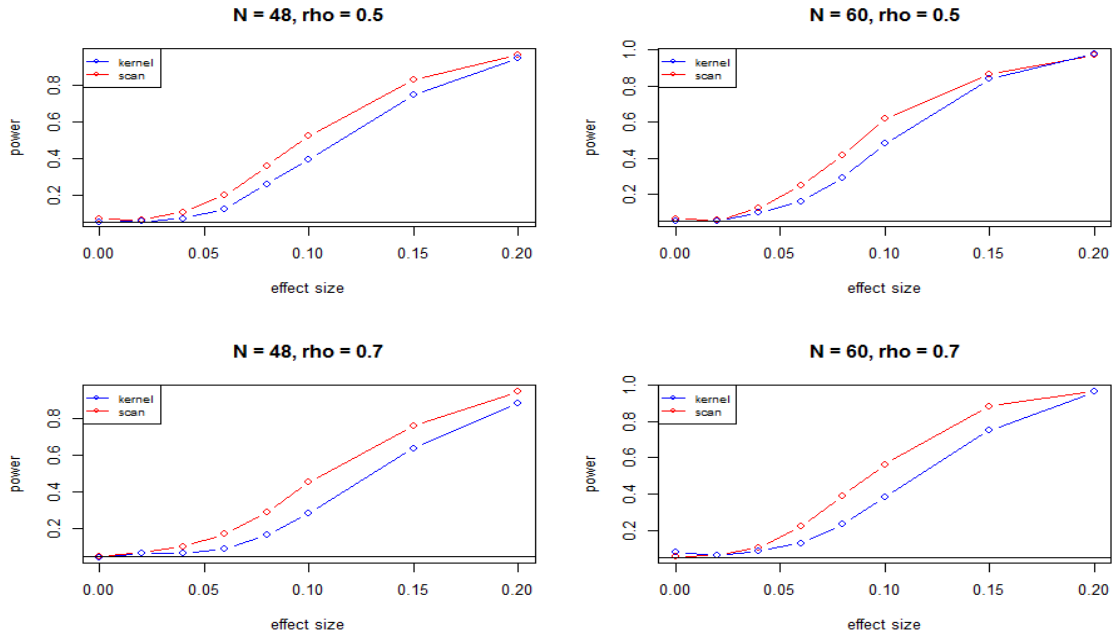

Figure 5: Power curves for SSM and KDM with 30 CpG sites, $\alpha = 0.05$

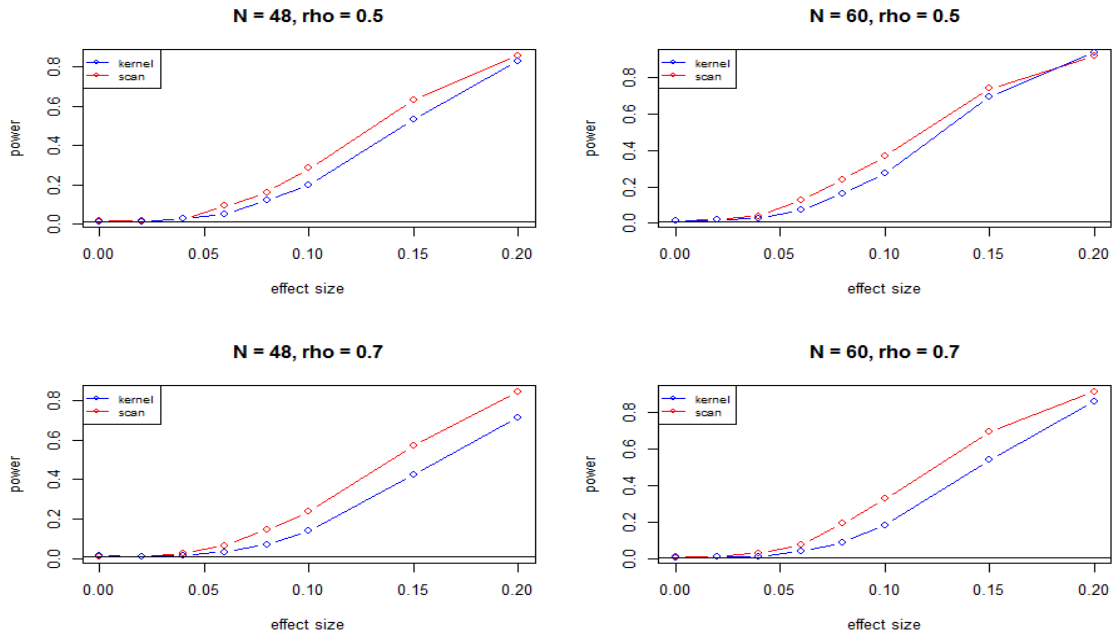

Figure 6: Power curves for SSM and KDM with 24 CpG sites, $\alpha = 0.01$

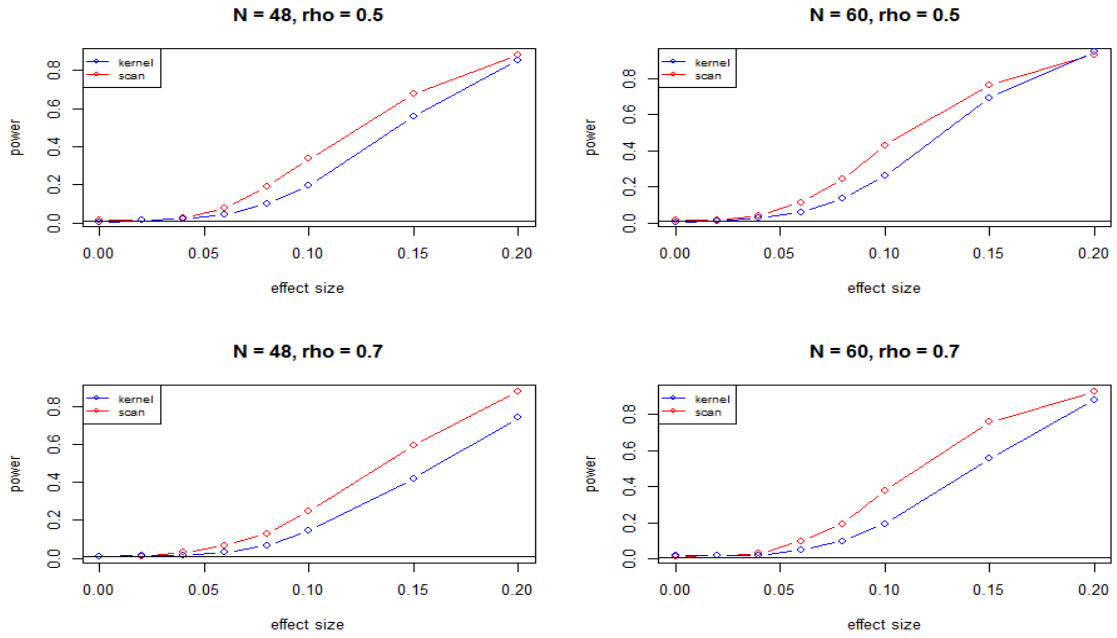

Figure 7: Power curves for SSM and KDM with 30 CpG sites, $\alpha = 0.01$

6.2 Simulation Study of Bayesian Method (BFM)

Simulation was also conducted to study the properties of BFM to detect DMRs. For simplicity, (i) we did not include any covariates, and (ii) we used equal sample sizes for each group in the simulation models. We assumed that there was only one DMR in the middle of simulated genomic region, and all CpG sites within the region were equally spaced.

6.2.1 Data Generation Procedure

We simulated data that contained K groups, with N individuals in each group, and assumed every individual had equally spaced m CpG sites in the simulated region, of which r ($< m$) consecutive CpG sites in the middle were in the DMR.

Methylation data were simulated at each CpG site of every sequence separately for every individual, under the fact that methylation rates were correlated between CpG sites, and methylation status at a CpG site were independent among different sequences as expected in NGS data. The data were simulated as described below:

1) First we generated random numbers for length and start point of each sequence, with a total of 100 sequences generated.

Suppose we generated random number a for the start point and c for the length of one sequence, then

2) We used vector $\mathbf{Y} = (Y_{kisa}, Y_{kis,a+1}, \dots, Y_{kis,a+c-1})$ to define the methylation status for sequence s of individual i in group k, and generated $\mathbf{Y}$ from multivariate Bernoulli distribution to consider the dependent feature.

The density of $P(\mathbf{Y}=\mathbf{y})=P(y_{kisa},y_{kis,a+1},\dots,y_{kis,a+c-1})$ of such a discrete random vector **Y** depends on 2^c probabilities, $p(0,0,\dots,0)$, $p(0,0,\dots,1)$, …, $p(1,1,\dots,1)$, specific to the different realization of **Y.** Considering the fact that if $(Y_1,Y_2,\dots,Y_S)$ follows multivariate Bernoulli distribution, the conditional distribution of $(Y_1,Y_2,\dots,Y_r)$ $(r<S)$ given the rest is also multivariate Bernoulli distribution (Dai, Ding, and Wahba 2013), We can incorporate an alternative parameterization that is able to avoid the high-dimensional parameters required in the multivariate Bernoulli distribution.

Because of the correlation of methylation rates between CpG sites, we treated methylation status Y_{kisj} at each CpG site j on one sequence was a branching process. We assumed that, for CpG site j, branching probabilities were same for each sequence of all individuals in group k. The branching probability p_{kj} (q_{kj}) was defined as the probability of methylated sequence read at CpG site j, conditional on methylated (unmethylated) sequence read at CpG site $j-1$ on the same sequence of same individual, and noted as $P\left(Y_{kisj}=1\middle|Y_{kis,j-1}=1\right)=p_{kj}$ and $P\left(Y_{kisj}=1\middle|Y_{kis,j-1}=0\right)=q_{kj}$.

Then the methylation status $(Y_{kisa},Y_{kis,a+1},\dots,Y_{kis,a+c-1})$ were generated as,

i) For the first CpG site of the sequence, the methylation status y_{kisa} was generated from Bernoulli distribution $Bern(m_a)$, with $m_a=(p_{ka}+q_{ka})/2$.

ii) The methylation status y_{kisj} for $j=a+1,\dots,a+c-1$ were generated with $y_{kisj}\sim Bern(p_{kj})$ if $y_{kis,j-1}=1$ or $y_{kisj}\sim Bern(q_{kj})$ if $y_{kis,j-1}=0$.

After all the sequence were generated at every CpG site for each individual, $\sum_s(y_{kisj} = 1)$ and $\sum_s(y_{kisj} = 0)$ were calculated, that were the number of methylated and unmethylated sequencing reads at CpG site j for individual i in group k. Then methylation count $m_{kij} = \sum_s(y_{kisj} = 1)$ and sequencing coverage $c_{kij} = \sum_s(y_{kisj} = 1) + \sum_s(y_{kisj} = 0)$ were calculated.

6.2.2 Simulation Parameters for the Bayesian Method

Simulation was conducted by using four groups of severity levels, with sample sizes of 50, and repeated for sample size of 100 in each group. The methylation data was simulated with a region of 24 CpG sites, and 6 of which (from site 10 to 15) were in the DMR.

Case I: The branching probabilities $P\left(Y_{kisj} = 1 \middle| Y_{kis,j-1} = 1\right) = p_{kj}$ were pre-defined, and were presented in Table 2, also we defined $q_{kj} = p_{kj} - 0.2$. The probabilities p_{kj} in Table 2 was defined to be symmetric, they increased from CpG site 1 to the middle of the DMR (CpG site 12 and 13), and then decreased.

	1	2	…	9	10	11	12	13	14	15	16	17	…	24
Group_1	0.44	0.46	…	0.60	0.62	0.64	0.66	0.66	0.64	0.62	0.60	0.58	…	0.44
Group_2	0.44	0.46	…	0.60	0.72	0.74	0.76	0.76	0.74	0.72	0.60	0.58	…	0.44
Group_3	0.44	0.46	…	0.60	0.82	0.84	0.86	0.86	0.84	0.82	0.60	0.58	…	0.44
Group_4	0.44	0.46	…	0.60	0.92	0.94	0.96	0.96	0.94	0.92	0.60	0.58	…	0.44

Table 2: conditional probabilities p_{kj} at each CpG site for simulation of BFM

Case II: p_{kj} in Table 2 reached maximum at the middle of the simulated genomic regions. In order to consider other possible variety, the simulation also conducted to generate maximum value of p_{kj} happening at any CpG site within simulated DMR (between sites 10 to 15), and varying for different individual. For each individual in every group, first we generated a random number r (between 10 to 15) for the location of the maximum value, and then branching probabilities p_{kj} were defined as increasing from 1 to r, and then decreasing from r to 15.

6.2.3 Simulation Results

Totally 1000 simulations were conducted, within each simulation, Bayes factor was calculated for each moving window, with window size of 6. They were calculated based on 3000 Gibbs samplers, with 1000 Gibbs samplers for burn in. The results of simulation of case I were presented in Figure 8 and Figure 9, while the results of simulation of case II were presented in Figure 10.

All the results show that the simulated DMR (window with CpG sites from 10 to 15), having maximum Bayes factor. This indicates that BFM can detect DMR. However, the Bayes factors are not symmetric, the windows on the right side have larger value, compares to corresponding ones on the left side. This might be caused by data generation, since the methylation status was generated based on that at previous CpG site of same sequence. The window with CpG sites from 10 to 15 was the simulated DMRs, with large methylation rates, this would cause the following CpG sites have large methylation rates, and would eventually have large Bayes factors.

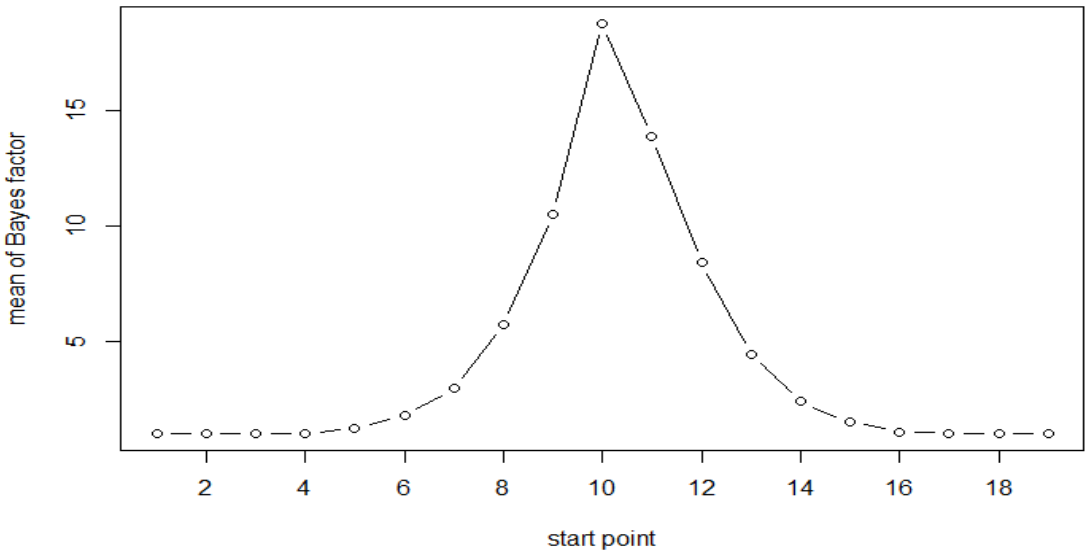

Figure 8: Mean of Bayes factors at each CpG site with $N = 50$ (Case I)

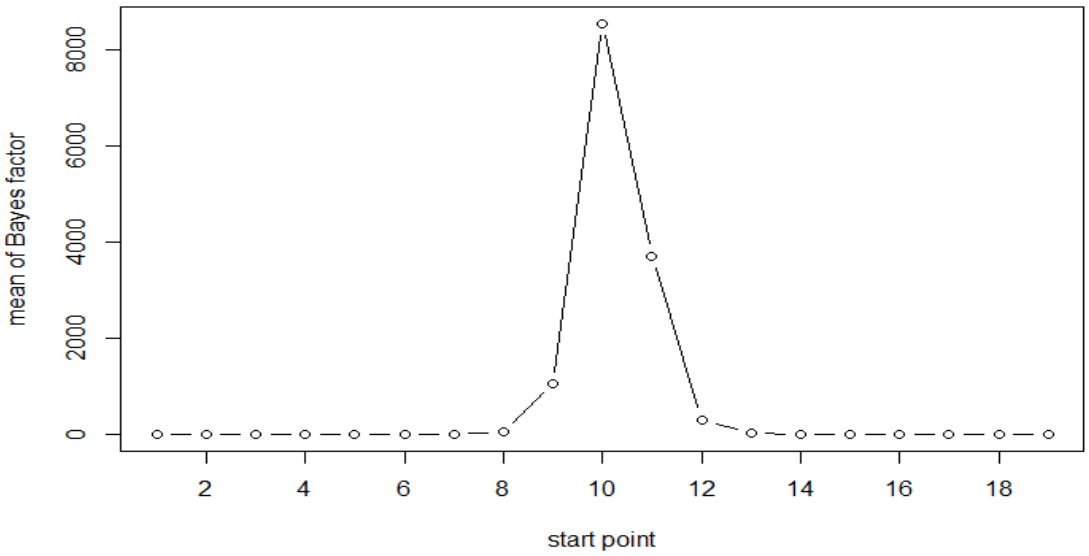

Figure 9: Mean of Bayes factors at each CpG site with $N = 100$ (Case I)

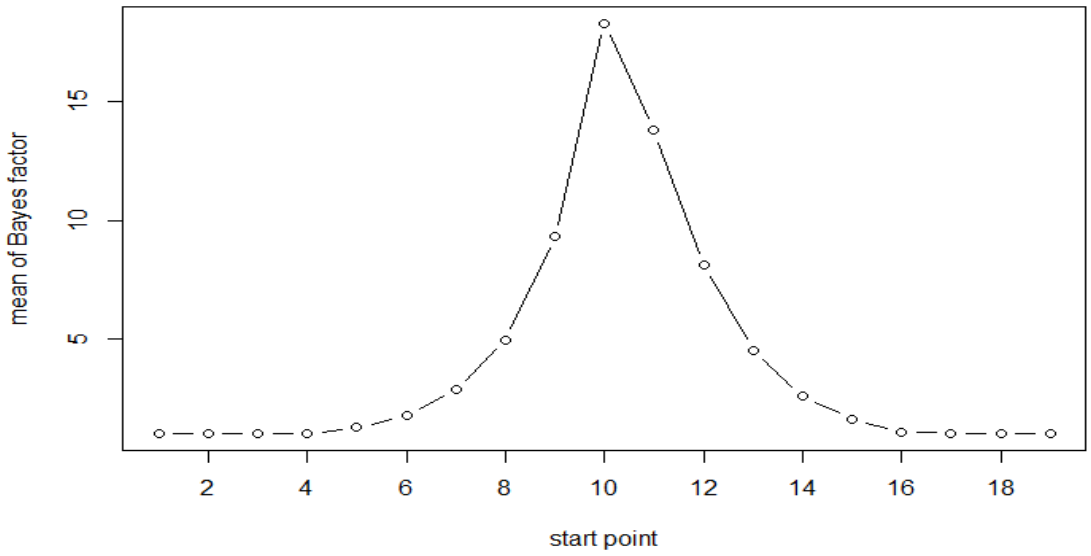

Figure 10: Mean of Bayes factors at each CpG site with $N = 50$ (Case II)

Comparing results presented in Figure 8 and Figure 9, Bayes factors increase as sample size increases. However, there is not much difference between the results in Figure 8 and Figure 10. This indicates that the shape of methylation rates within DMRs does not affect the ability of detecting DMRs using BFM.

start	end	N = 50(Case I)	N = 100(Case I)	N = 50 (Case II)
1	6	1.02	1.02	1.03
2	7	1.01	1.02	1.01
3	8	1.01	1.02	1.02
4	9	1.02	1.01	1.01
5	10	1.24	1.53	1.26
6	11	1.78	3.12	1.78
7	12	2.95	9.16	2.85
8	13	5.74	41.42	4.95
9	14	10.53	1052.07	9.31
10	15	18.79	8554.12	18.31

11	16	13.90	3718.77	13.79
12	17	8.44	306.07	8.12
13	18	4.43	21.91	4.50
14	19	2.40	5.66	2.60
15	20	1.52	2.22	1.60
16	21	1.07	1.11	1.07
17	22	1.03	1.04	1.02
18	23	1.01	1.03	1.02
19	24	1.03	1.03	1.01

Table 3: Mean Bayes factors at each CpG site based on simulation studies

CHAPTER 7

REAL DATA ANALYSIS

We next analyzed genome-wide methylation data from a study of chronic lymphocytic leukemia (CLL), which was the result of clonal expansion of malignant B cells. A B-cell lymphoma mainly of adults is heterogeneous disease (Chiorazzi, Rai, and Ferrarini 2005, Keating et al. 2003). It is clinically important to find heterogeneity of patients at the molecular level, which can help to design specific interventions for patients at different levels.

Over the last decade, research in CLL has resulted in multiple significant advances; such as identification of several molecular alternations with prognostic values. These include specific cytogenetic patterns (Dohner et al. 2000), mutational status of the immunoglobulin heavy chain variable gene (IgV_H) (Hamblin et al. 1999) and expression of CD38 (Hamblin et al. 2000). It has been found out that patients lacking the mutation have a poorer prognosis. Patients with lower levels of CD38 have slower disease progression (Damle et al. 1999, Hamblin et al. 1999).

Several research groups have demonstrated that DNA methylation of multiple promoter-associated CpG islands is common in CLL (Rahmatpanah et al. 2006, Kanduri et al. 2010, Martin-Subero et al. 2009). Detection of aberrant DNA methylation in CLL

could result in the development of an epigenetic classification of the disease with prognostic and therapeutic potential.

$CD19^+$ B cells from peripheral blood were collected from CLL samples and normal control subjects. All CLL samples were obtained from patients at the Ellis Fischel Cancer Center (EFCC), the GRU Cancer Center and the North Shore-LIJ Health System in compliance with the local Institutional Review Boards (Pei et al. 2012).

Illumina sequencing reads were generated for each sample, by using RRBS (Meissner et al. 2005). Totally 20 – 30 million reads were sequenced for each sample, and 63% to 75% were successfully mapped to either strand of the human genome (hg18) (Pei et al. 2012). The average sequencing depth per CpG was between 32x and 43x. Eventually RRBS provided counts of DNA molecules that were methylated and unmethylated at each CpG site, and overall methylation status of approximately 1.8 – 2.3 million CpG sites were determined consistently for each sample in the study (Pei et al. 2012).

Here genome-wide methylation data on 17,917 CpG sites of Chromosome 19 were analyzed, since Tong et al. (2010) pointed out that aberrant DNA methylation that associated with CLL, were located more frequently in chromosome 19.

7.1 Comparison of Scan Statistic and kernel Distance Methods

Based on CD38 levels, the samples were categorized as low- vs. high- risk, with 23 samples having CD38 levels ≤ 20 (low risk) and 17 samples having CD38 levels > 20 (high risk).

Here genome-wide methylation data on 17,917 CpG sites of Chromosome 19 were analyzed by both SSM and KDM to identify DMRs between high-risk and low-risk CLL samples. The percentage contributions of kernel distance statistic at each CpG site were plotted in Figure 11, while the detected DMRs from SSM were presented in Table 4. The peaks in Figure 11 showed the detected DMRs in Chromosome 19. Here all the values of kernel distance statistics were positive, because of quadratic expression of the kernel distance statistic.

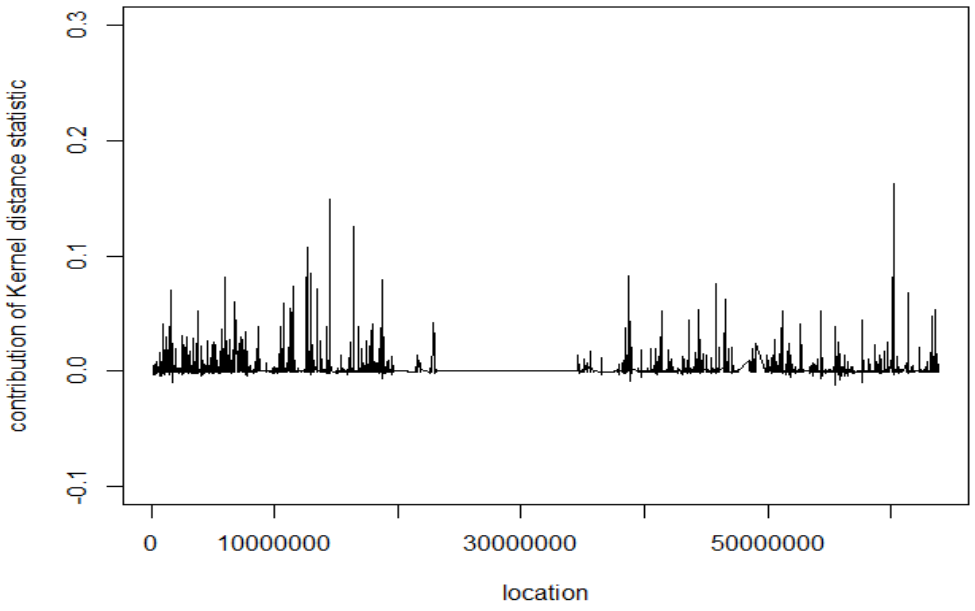

Figure 11: Contribution of kernel distance statistic at each CpG site for leukemia data

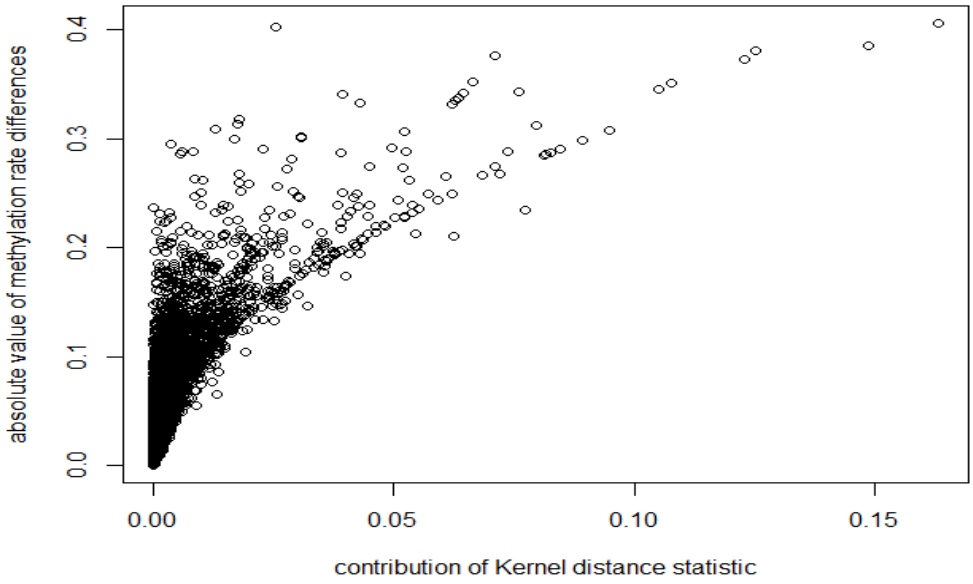

Figure 12: Contribution of kernel distance statistic versus methylation rates for leukemia data

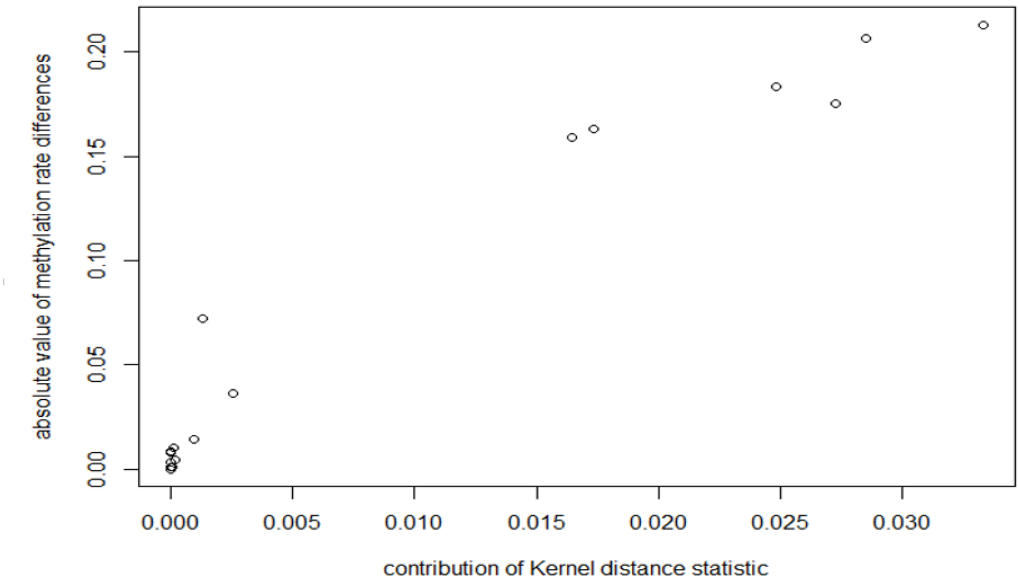

Figure 13: Contribution of kernel distance statistic versus methylation rates for simulation data

The wedge shape in Figure 12 showed that, a lot of CpG sites with small differences in methylation rates had very small contributions in the kernel distance statistic; and the CpG sites with large contributions in the kernel distance statistic were differentially methylated. This indicated that KDM can detect DMRs. Figure 13 plotted the absolute differences of methylation rates versus percentage contribution of kernel distance statistic at each CpG site based on simulated data in Section 6.1. The similar shape in Figure 12 and Figure 13 showed that KDM can detect DMRs, especially the tri-weight kernel function can incorporate the correlation structure of methylation rates between CpG sites.

Start	End	Window size	Scan statistic	p-value
951756	960480	15	6432.20	0.001
5748848	5855704	35	175.22	0.024
5949493	6059920	15	1816.97	0.037
6222967	6325326	40	126.29	0.042
6695897	6704448	5	121.95	0.039
7049880	7149391	20	46.16	0.042
8306311	8416558	105	693.16	0.02
10078223	10091192	15	1156.08	0.049
10261108	10336402	75	1169.73	0.048
10366854	10374990	5	1400.52	0.011
10529295	10537824	5	656.14	0.033
11311211	11369166	35	468.53	0.046
11852835	11937174	15	185.69	0.012
12036638	12128243	10	335.82	0.019
13780707	13818691	30	990.65	0.035
15871811	15874720	5	71.10	0.033
16211533	16298141	10	896.80	0.008
16779596	16818698	5	45.09	0.049
17181376	17207209	20	1476.43	0.032
17483944	17492848	5	608.50	0.027
18358107	18358200	5	1306.14	0.018
18839769	18849925	40	888.01	0.037
19196863	19220558	10	809.88	0
20751241	20751405	10	13.86	0.016

21443528	21449542	5	82.91	0.042
35558112	35558143	5	5.76	0.014
37528315	37528707	10	8.34	0.035
37808618	37858100	10	67.30	0.019
38315030	38359639	10	415.61	0.012
38576223	38632218	20	107.39	0.001
38980210	39003767	20	48.17	0.043
39760398	39760441	5	7.09	0.022
40193224	40214045	25	314.69	0.046
40495154	40706271	40	284.10	0.033
40958295	40995281	15	578.96	0.028
41323151	41345137	5	826.49	0.027
42400872	42516823	25	86.90	0.023
42631539	42651999	10	13.67	0.022
43411447	43472750	70	161.49	0.023
44099619	44158078	10	255.15	0.003
45388832	45464209	30	505.70	0.007
45812107	45821840	5	1100.34	0.005
46555659	46595121	5	527.04	0.007
47040515	47078316	5	211.34	0.008
50778928	50793474	5	1155.35	0.021
51010992	51058089	10	602.50	0.029
51059619	51079866	15	1135.07	0.026
51409109	51427742	5	10.71	0.027
53821358	53829676	15	3089.85	0.001
53914126	53934314	20	1843.99	0.011
53946213	53983289	5	1898.17	0.033
54819984	54835037	5	601.06	0.035
54872826	54884388	10	804.62	0.033
55714472	55760862	15	27.40	0.047
55853400	55911789	30	26.58	0.046
56884684	56887726	5	322.31	0.002
58388434	58388478	5	8.08	0.006
58980127	59064230	15	93.10	0.007
59643525	59652071	5	929.55	0.002
59652664	59666539	15	513.52	0.011
60109922	60545979	205	410.16	0.04
60790219	60808074	15	776.25	0.015
61304533	61424810	55	26.85	0.013
61741700	61798595	20	269.87	0.048
62277420	62310019	5	12.99	0.019
63565854	63570870	10	840.16	0.035

Table 4: Results of SSM for CLL data

SSM totally detected 66 DMRs with varying window sizes, these were presented in Table 4. The results in Table 4 could match with the peaks in Figure 9, indicated that both SSM and KDM can identify DMRs.

The start and end positions in base pairs for each detected DMR were used in UCSC genome browser to find the genes in the regions. Among them, the apolipoprotein gene cluster (*APOC1*, *APOC2*, *APOE*) was detected, which has been published that they have tight linkage with a chronic lymphocytic leukemia-associated translocation breakpoint (Shaw et al. 1989). We also detected the genes *CATSPERD*, *PRR22*, *RFX2*, and *MILT1*, these were shown in published work to be associated with leukemia (Wallingford et al. 2015). For example, translocation and fusion of *MILT1* with myeloid lymphoid leukemia could result in potent oncogenic activity (Chin et al. 2012, Doty et al. 2002).

Several lines of evidence suggested that the transcription factor *CREB* (cyclic AMP response element binding protein) may have a role in the pathogenesis of AML and other cancers (Crans-Vargas et al. 2002, Mayr and Montminy 2001). In our data, replication factor *C3* was detected, whose expression had been published that has a direct correlation with *CREB* in human acute myeloid leukemia (AML) cell lines, as well as in the AML cells from the patients (Chae et al. 2015). It is suggested that *C3* may have a role in neoplastic myelopoiesis by promoting the G1/S progression. Another detected gene *LAIR1*, also had been published that has a correlation with *CREB* (Kang et al. 2015). A pathway starts with *LAIR1*, activates downstream *CREB* in AML cells, sustains the

survival and self-renewal of AML stem cells. As a result, inhibition of expression of the ITIM-containing receptor *LAIR1* does not affect normal hematopoiesis but abolishes leukemia development (Kang et al. 2015).

7.2 Comparison of Bayesian Method with Scan Statistic Method for Two Groups

Before checking the performance of BFM on four ordinal group responses, BFM was used on the samples divided into two groups based on CD38 levels. BFM and SSM were compared, by using moving window of sizes 10 or 20.

	BFM >2	SSM ($p < 0.05$)	common
total	183	181	67
PubMed	42	41	18

Table 5: Comparison of BFM and SSM for window size of 10 ($p < 0.05$)

Using moving window of size 10, which was 10 CpG sites in each moving window, totally 181 genes in DMRs were detected by SSM with *p*-value < 0.05, and a total of 183 genes in DMRs were detected by BFM with value greater than 2. Among them 41 from SSM and 42 from BFM were found in PubMed that had been published that were leukemia associated, and 18 were detected by both methods (Table 5). They were *ACP5* (French et al. 2008), *ATF5* (Wang et al. 2014), *BIRC8* (Glodkowska-Mrowka et al. 2014), *C3* (Chae et al. 2015), *CARD8* (Xu et al. 2009), *CEACAM8* (Lasa et al. 2008), *CERS1* (Camgoz et al. 2013), *CKM* (Caldow, Digby, and Cameron-Smith 2015), *CRTC1* (Tang et al. 2015), *IL4l1* (Carbonnelle-Puscian et al. 2009), *LAIR1* (Kang et al. 2015), *MAP1S*(Haimovici et al. 2014), *NFIX* (O'Connor et al. 2015), *PDE4C* (Moon et al.

2002), *PLEKHG2* (Runne and Chen 2013), *PLVAP* (Rantakari et al. 2015), *RFX1* (Chen et al. 2000), *ZNF331* (McHale et al. 2009).

C3 and *LAIR1* genes were both detected, which were shown related to acute myeloid leukemia (Chae et al. 2015, Kang et al. 2015). Actually both *C3* and *LAIR1* genes connect with the transcription factor *CREB* (cyclic AMP response element binding protein), which has a role in the pathogenesis of AML and other cancers (Crans-Vargas et al. 2002, Mayr and Montminy 2001).

	BFM>4	SSM (p <0.01)	common
total	43	51	4
PubMed	9	8	1

Table 6: Comparison of BFM and SSM for window size of 10 (p <0.01)

	BFM>3	SSM (p<0.05)	common
Total	152	137	35
PubMed	35	36	8

Table 7: Comparison of BFM and SSM for window size of 20 (p <0.05)

	BFM>5	SSM (p<0.01)	common
total	30	30	6
PubMed	8	9	1

Table 8: Comparison of BFM and SSM for window size of 20 (p <0.01)

51 genes were detected by SSM with *p*-value < 0.01, and compared with 43 genes detected by BFM with cut point of 4 (Table 6), 9 genes from BFM and 8 genes from SSM were published that were associated with leukemia, but only 1 gene in common. Results with a moving window of size 20 were presented in Table 7 and Table 8. These results indicated that BFM could detect some genes that can't be detected by SSM

7.3 Bayesian Method for Ordinal Group Responses

In order to test whether the methylation rates increase as the CD38 level increases, the samples were classified into four risk groups based on CD38 level, with 5 samples in group 1 (Normal group), 24 samples in group 2 with CD38 $\leq$ 20, 9 samples in group 3 with 20$<$ CD38$\leq$50, and 8 sample in group 4 with CD38 > 50. Here moving windows with size of 10 were used for analysis.

Totally 789 windows had been detected as significant, by using $BF > \frac{0.95}{0.05} = 19$ as the cutoff for the significance (Dunson 2003). The start and end positions in base pairs for each detected DMR were used in UCSC genome browser to find the genes in the regions, and eventually found 125 genes. Among them 35 were published in PubMed that were associated with leukemia. Some of them were not detected when only considering two groups, they were *BRD4* (Stewart et al. 2013), *ELL* (Muto et al. 2015), *ERCC1* (Kong et al. 2012), *ERCC2* (Liu et al. 2014), *GDF15* (Secchiero et al. 2006), *JUND* (Gazon et al. 2012), *POLD1* (Sincennes et al. 2016), *PRDX2* (Agrawal-Singh et al. 2012), *RANBP3* (Hakata, Yamada, and Shida 2003), *SPIB* (Talby et al. 2006) and *TSPAN16* (Juric et al. 2007).

CHAPTER 8

DISCUSSION

Results from simulation and analysis of CLL data indicate that all three methods, SSM, KDM, and BFM are valid approaches to detect DMRs. All three methods detect DMRs, while adjusting for covariates with logistic regression and the correlation between CpG sites.

The tri-weight function used in KDM, incorporates the fact that the correlation decrease as the distances of two CpG sites increase, while SSM and BFM use a mixed-effect model to incorporate the correlation structure.

Although compound symmetric assumption used in SSM couldn't represent actual correlation structure, the sandwich estimate of the fixed effects is appropriate even when the correlation structure is mis-specified, with some trade off of the flexibility for robustness of inference. Here our simulation results also show that the mixed-effect model is able to adjust for correlation, when the simulated correlations decrease as the distances between CpG site increase.

BFM takes advantage of flexibility in Bayesian model, uses multivariate normal distributed random variables in the model to incorporate correlation structure, with inverse Wishart distribution as the prior for the correlation matrix. However, since the correlation structure is very complicated for methylation data, it might not the best

statistical model for the correlation structure. Some other better statistical methods can be considered to improve the robustness of the method for detecting DMRs.

Both SSM and KDM had reasonable power and good control of type I error, when detecting DMRs between cases and controls. SSM has better power compared to KDM, that was not only because SSM is a likelihood based method, while KDM is a non-parametric method; but also because SSM used moving window with multiple window sizes solved the difficulty of determining the value of τ in KDM. However the use of moving windows, with a mixed-effect model for adjusting correlation of methylation rates, caused SSM longer computation time.

The uncertainty of τ not only leads to disadvantages in terms of power for KDM, but also it caused KDM to only give rough regions of DMRs, since the results of DMRs were based on the plot of percentage contribution, which were calculated based on kernel distance statistic from only one value of τ. In reality, the lengths of DMRs ranged from hundreds of base pair as in small CpG islands, to millions of base pairs in cancer aberrations.

KDM doesn't have power as good as SSM, also because KDM is not able to adjust for unequal sequencing coverage for all individuals at each CpG site, while SSM calculates design effects based on Xu et al. (2013), which is used to adjust sequencing coverage and methylation counts. Some statistical methods that could adjust for unequal sequencing coverage should be proposed in the future, while using KDM. One possible solution is using logistic regression on methylation rates at each CpG site for every

individual, that might able to adjust for unequal sequencing coverage, since the methylation rates are calculated by considering the methylation data at each CpG site as binomial distributed. Another possible solution is using a mixed-effect logistic model with random intercept to adjust for the within cluster correlation, while treating methylation data at each CpG site as a cluster.

SSM has the advantage that it can be used for more than two groups, while KDM can only be used for two groups since it is calculated based on the differences of methylation rates. But SSM still has a limitation that it cannot consider the ordering of the group responses. When ordering needs to be considered, the maximum likelihood estimate is very difficult based on the constrained space. In that case, BFM should be considered. BFM is a valid approach to detect DMRs when considering ordinal group responses, especially to detect DMRs with methylation rates increasing (or decreasing) as disease severity increases.

Besides taking care of ordering of group responses, BFM also has an advantage over SSM by allowing for heterogeneity of effect across CpG sites, by modeling the methylation rates with a prior. On the other hand, SSM pools information across variants in a region, essentially assuming that each CpG site in the region has same methylation rates.

However, BFM assumes that methylation rates of CpG sites within each moving window are independent of those outside of the window, while SSM adjusts the correlation along whole genome. The comparison between BFM and SSM for CLL data with two groups

showed that there were few genes that detected by both methods, but BFM detected some DMRs, those SSM didn't detect. The results also show that it was very difficult to decide the best cut-point of the Bayes factor values to make decisions regarding DMRs. These were because results presented in Chapter 7 were based on how many genes were published from PubMed, which was not a good criteria for comparing two methods, since the approach of searching through PubMed was subjective. It would be helpful if there were a database with published and proven list of genes associated with diseases, which could be used to make comparisons when developing statistical methods.

BFM and SSM used a moving window to help decide the location and length of DMRs. But practically, it was very difficult to know the exact length of DMRs, this limitation was very common in statistical genetics, not only for detecting DMRs, but also for detecting rare variants (Schaid et al. 2013). Using cross validation or bootstrapping might help determine the window sizes. Some other methods for example, using genes and promoters can be used instead of moving windows, along with BFM and SSM to detect DMRs.

All our methods were only focused on DNA methylation data. However, large-scale cancer genomics projects such as TCGA (The Cancer Genome Atlas Research Network) are currently generating multiple layers of genomics data for early tumor, including DNA copy number, methylation, and mRNA expression. Statistical methods for integrating analysis and systematic modeling all these genomics data deserve more attention.

References

Agarwal, Deepak, Jeff M. Phillips, and Suresh Venkatasubramanian. 2006. The hunting of the bump: on maximizing statistical discrepancy. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. Miami, Florida: Society for Industrial and Applied Mathematics.

Agrawal-Singh, S., F. Isken, K. Agelopoulos, H. U. Klein, N. H. Thoennissen, G. Koehler, A. Hascher, N. Baumer, W. E. Berdel, C. Thiede, G. Ehninger, A. Becker, P. Schlenke, Y. Wang, M. McClelland, U. Krug, S. Koschmieder, T. Buchner, D. Y. Yu, S. V. Singh, K. Hansen, H. Serve, M. Dugas, and C. Muller-Tidow. 2012. "Genome-wide analysis of histone H3 acetylation patterns in AML identifies PRDX2 as an epigenetically silenced tumor suppressor gene." *Blood* no. 119 (10):2346-57. doi: 10.1182/blood-2011-06-358705.

Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, and C. E. Mason. 2012. "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles." *Genome Biol* no. 13 (10):R87. doi: 10.1186/gb-2012-13-10-r87.

Anand, P., A. B. Kunnumakkara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal. 2008. "Cancer is a preventable disease that requires major lifestyle changes." *Pharm Res* no. 25 (9):2097-116. doi: 10.1007/s11095-008-9661-9.

Ayer, Miriam, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. 1955. "An Empirical Distribution Function for Sampling with Incomplete Information." *The Annals of Mathematical Statistics* no. 26 (4):641-647.

Bartholomew, D. J. 1959. "A TEST OF HOMOGENEITY FOR ORDERED ALTERNATIVES." *Biometrika* no. 46 (1-2):36-48. doi: 10.1093/biomet/46.1-2.36.

Baylin, S. B. 1997. "Tying it all together: epigenetics, genetics, cell cycle, and cancer." *Science* no. 277 (5334):1948-9.

Bell, J. T., P. C. Tsai, T. P. Yang, R. Pidsley, J. Nisbet, D. Glass, M. Mangino, G. Zhai, F. Zhang, A. Valdes, S. Y. Shin, E. L. Dempster, R. M. Murray, E. Grundberg, A. K. Hedman, A. Nica, K. S. Small, Ther Consortium Mu, E. T. Dermitzakis, M. I. McCarthy, J. Mill, T. D. Spector, and P. Deloukas. 2012. "Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population." *PLoS Genet* no. 8 (4):e1002629. doi: 10.1371/journal.pgen.1002629.

Bernstein, B. E., A. Meissner, and E. S. Lander. 2007. "The mammalian epigenome." *Cell* no. 128 (4):669-81. doi: 10.1016/j.cell.2007.01.033.

Bestor, T. H., and B. Tycko. 1996. "Creation of genomic methylation patterns." *Nat Genet* no. 12 (4):363-7. doi: 10.1038/ng0496-363.

Bird, A. 2002. "DNA methylation patterns and epigenetic memory." *Genes Dev* no. 16 (1):6-21. doi: 10.1101/gad.947102.

Bird, A. P. 1986. "CpG-rich islands and the function of DNA methylation." *Nature* no. 321 (6067):209-13. doi: 10.1038/321209a0.

Bird, A., M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. 1985. "A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA." *Cell* no. 40 (1):91-9.

Bock, C. 2012. "Analysing and interpreting DNA methylation data." *Nat Rev Genet* no. 13 (10):705-19. doi: 10.1038/nrg3273.

Bush, W. S., and J. H. Moore. 2012. "Chapter 11: Genome-wide association studies." *PLoS Comput Biol* no. 8 (12):e1002822. doi: 10.1371/journal.pcbi.1002822.

Caldow, M. K., M. R. Digby, and D. Cameron-Smith. 2015. "Short communication: Bovine-derived proteins activate STAT3 in human skeletal muscle in vitro." *J Dairy Sci* no. 98 (5):3016-9. doi: 10.3168/jds.2014-9035.

Camgoz, A., E. B. Gencer, A. U. Ural, and Y. Baran. 2013. "Mechanisms responsible for nilotinib resistance in human chronic myeloid leukemia cells and reversal of resistance." *Leuk Lymphoma* no. 54 (6):1279-87. doi: 10.3109/10428194.2012.737919.

Carbonnelle-Puscian, A., C. Copie-Bergman, M. Baia, N. Martin-Garcia, Y. Allory, C. Haioun, A. Cremades, I. Abd-Alsamad, J. P. Farcet, P. Gaulard, F. Castellano, and V. Molinier-Frenkel. 2009. "The novel immunosuppressive enzyme IL4I1 is expressed by neoplastic cells of several B-cell lymphomas and by tumor-associated macrophages." *Leukemia* no. 23 (5):952-60. doi: 10.1038/leu.2008.380.

Chae, H. D., B. Mitton, N. J. Lacayo, and K. M. Sakamoto. 2015. "Replication factor C3 is a CREB target gene that regulates cell cycle progression through the modulation of chromatin loading of PCNA." *Leukemia* no. 29 (6):1379-89. doi: 10.1038/leu.2014.350.

Chen, L., L. Smith, M. R. Johnson, K. Wang, R. B. Diasio, and J. B. Smith. 2000. "Activation of protein kinase C induces nuclear translocation of RFX1 and down-regulates c-myc via an intron 1 X box in undifferentiated leukemia HL-60 cells." *J Biol Chem* no. 275 (41):32227-33. doi: 10.1074/jbc.M002645200.

Chin, L. K., C. Y. Cheah, P. M. Michael, R. N. MacKinnon, and L. J. Campbell. 2012. "11q23 rearrangement and duplication of MLLT1-MLL gene fusion in therapy-related acute myeloid leukemia." *Leuk Lymphoma* no. 53 (10):2066-8. doi: 10.3109/10428194.2012.666663.

Chiorazzi, N., K. R. Rai, and M. Ferrarini. 2005. "Chronic lymphocytic leukemia." *N Engl J Med* no. 352 (8):804-15. doi: 10.1056/NEJMra041720.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini, and S. E. Jacobsen. 2008. "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." *Nature* no. 452 (7184):215-9. doi: 10.1038/nature06745.

Crans-Vargas, H. N., E. M. Landaw, S. Bhatia, G. Sandusky, T. B. Moore, and K. M. Sakamoto. 2002. "Expression of cyclic adenosine monophosphate response-element binding protein in acute leukemia." *Blood* no. 99 (7):2617-9.

Dai, B., S. L. Ding, and G. Wahba. 2013. "Multivariate Bernoulli distribution." *Bernoulli* no. 19 (4):1465-1483. doi: 10.3150/12-Bejsp10.

Damle, R. N., T. Wasil, F. Fais, F. Ghiotto, A. Valetto, S. L. Allen, A. Buchbinder, D. Budman, K. Dittmar, J. Kolitz, S. M. Lichtman, P. Schulman, V. P. Vinciguerra, K. R. Rai, M. Ferrarini, and N. Chiorazzi. 1999. "Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia." *Blood* no. 94 (6):1840-7.

Dohner, H., S. Stilgenbauer, A. Benner, E. Leupolt, A. Krober, L. Bullinger, K. Dohner, M. Bentz, and P. Lichter. 2000. "Genomic aberrations and survival in chronic lymphocytic leukemia." *N Engl J Med* no. 343 (26):1910-6. doi: 10.1056/NEJM200012283432602.

Doty, R. T., G. J. Vanasse, C. M. Disteche, and D. M. Willerford. 2002. "The leukemia-associated gene Mllt1/ENL: characterization of a murine homolog and demonstration of an essential role in embryonic development." *Blood Cells Mol Dis* no. 28 (3):407-17.

Dressman, D., H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. 2003. "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations." *Proc Natl Acad Sci U S A* no. 100 (15):8817-22. doi: 10.1073/pnas.1133470100.

Dunson, D. B. 2003. "Bayesian inference on order-constrained parameters in generalized linear models." *Biometrics* no. 59 (2):286-295. doi: Doi 10.1111/1541-0420.00035.

Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck. 2006. "DNA methylation profiling of human chromosomes 6, 20 and 22." *Nat Genet* no. 38 (12):1378-85. doi: 10.1038/ng1909.

Ehrlich, M., and M. Lacey. 2013. "DNA methylation and differentiation: silencing, upregulation and modulation of gene expression." *Epigenomics* no. 5 (5):553-68. doi: 10.2217/epi.13.43.

Ehrlich, M., and R. Y. Wang. 1981. "5-Methylcytosine in eukaryotic DNA." *Science* no. 212 (4501):1350-7.

Esteller, M. 2005. "Aberrant DNA methylation as a cancer-inducing mechanism." *Annu Rev Pharmacol Toxicol* no. 45:629-56. doi: 10.1146/annurev.pharmtox.45.120403.095832.

Feng, H., K. N. Conneely, and H. Wu. 2014. "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data." *Nucleic Acids Res* no. 42 (8):e69. doi: 10.1093/nar/gku154.

Frazer, K. A., S. S. Murray, N. J. Schork, and E. J. Topol. 2009. "Human genetic variation and its contribution to complex traits." *Nat Rev Genet* no. 10 (4):241-51. doi: 10.1038/nrg2554.

French, D., L. H. Hamilton, L. A. Mattano, Jr., H. N. Sather, M. Devidas, J. B. Nachman, M. V. Relling, and Group Children's Oncology. 2008. "A PAI-1 (SERPINE1) polymorphism predicts osteonecrosis in children with acute lymphoblastic leukemia: a report from the Children's Oncology Group." *Blood* no. 111 (9):4496-9. doi: 10.1182/blood-2007-11-123885.

Galvan, A., J. P. Ioannidis, and T. A. Dragani. 2010. "Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer." *Trends Genet* no. 26 (3):132-41. doi: 10.1016/j.tig.2009.12.008.

Gazon, H., I. Lemasson, N. Polakowski, R. Cesaire, M. Matsuoka, B. Barbeau, J. M. Mesnard, and J. M. Peloponese, Jr. 2012. "Human T-cell leukemia virus type 1 (HTLV-1) bZIP factor requires cellular transcription factor JunD to upregulate HTLV-1 antisense transcription from the 3' long terminal repeat." *J Virol* no. 86 (17):9070-8. doi: 10.1128/JVI.00661-12.

Glodkowska-Mrowka, E., I. Solarska, P. Mrowka, K. Bajorek, J. Niesiobedzka-Krezel, I. Seferynska, K. Borg, and T. Stoklosa. 2014. "Differential expression of BIRC family genes in chronic myeloid leukaemia--BIRC3 and BIRC8 as potential new candidates to identify disease progression." *Br J Haematol* no. 164 (5):740-2. doi: 10.1111/bjh.12663.

Haimovici, A., D. Brigger, B. E. Torbett, M. F. Fey, and M. P. Tschan. 2014. "Induction of the autophagy-associated gene MAP1S via PU.1 supports APL differentiation." *Leuk Res* no. 38 (9):1041-7. doi: 10.1016/j.leukres.2014.06.010.

Hakata, Y., M. Yamada, and H. Shida. 2003. "A multifunctional domain in human CRM1 (exportin 1) mediates RanBP3 binding and multimerization of human T-cell leukemia virus type 1 Rex protein." *Mol Cell Biol* no. 23 (23):8751-61.

Hamblin, T. J., Z. Davis, A. Gardiner, D. G. Oscier, and F. K. Stevenson. 1999. "Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia." *Blood* no. 94 (6):1848-54.

Hamblin, T. J., J. A. Orchard, A. Gardiner, D. G. Oscier, Z. Davis, and F. K. Stevenson. 2000. "Immunoglobulin V genes and CD38 expression in CLL." *Blood* no. 95 (7):2455-7.

Hansen, K. D., B. Langmead, and R. A. Irizarry. 2012. "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions." *Genome Biol* no. 13 (10):R83. doi: 10.1186/gb-2012-13-10-r83.

Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg. 2011. "Increased methylation variation in epigenetic domains across cancer types." *Nat Genet* no. 43 (8):768-75. doi: 10.1038/ng.865.

Hebestreit, K., M. Dugas, and H. U. Klein. 2013. "Detection of significantly differentially methylated regions in targeted bisulfite sequencing data." *Bioinformatics* no. 29 (13):1647-53. doi: 10.1093/bioinformatics/btt263.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. 2009. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." *Proc Natl Acad Sci U S A* no. 106 (23):9362-7. doi: 10.1073/pnas.0903103106.

Hodgkinson, A., and A. Eyre-Walker. 2011. "Variation in the mutation rate across mammalian genomes." *Nat Rev Genet* no. 12 (11):756-66. doi: 10.1038/nrg3098.

Hon, G. C., R. D. Hawkins, O. L. Caballero, C. Lo, R. Lister, M. Pelizzola, A. Valsesia, Z. Ye, S. Kuan, L. E. Edsall, A. A. Camargo, B. J. Stevenson, J. R. Ecker, V. Bafna, R. L. Strausberg, A. J. Simpson, and B. Ren. 2012. "Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer." *Genome Res* no. 22 (2):246-58. doi: 10.1101/gr.125872.111.

Hoque, M. O. 2009. "DNA methylation changes in prostate cancer: current developments and future clinical implementation." *Expert Rev Mol Diagn* no. 9 (3):243-57. doi: 10.1586/erm.09.10.

Ionita-Laza, I., V. Makarov, Arra Autism Sequencing Consortium, and J. D. Buxbaum. 2012. "Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets." *Am J Hum Genet* no. 90 (6):1002-13. doi: 10.1016/j.ajhg.2012.04.010.

Irizarry, R. A., C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddeloh, B. Wen, and A. P. Feinberg. 2008. "Comprehensive high-throughput arrays for relative methylation (CHARM)." *Genome Res* no. 18 (5):780-90. doi: 10.1101/gr.7301508.

Irizarry, R. A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan, and A. P. Feinberg. 2009. "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores." *Nat Genet* no. 41 (2):178-86. doi: 10.1038/ng.298.

Jaffe, A. E., A. P. Feinberg, R. A. Irizarry, and J. T. Leek. 2012. "Significance analysis and statistical dissection of variably methylated regions." *Biostatistics* no. 13 (1):166-78. doi: 10.1093/biostatistics/kxr013.

Jaffe, A. E., P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry. 2012. "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies." *Int J Epidemiol* no. 41 (1):200-9. doi: 10.1093/ije/dyr238.

Jin, B., Y. Li, and K. D. Robertson. 2011. "DNA methylation: superior or subordinate in the epigenetic hierarchy?" *Genes Cancer* no. 2 (6):607-17. doi: 10.1177/1947601910393957.

Johnson, D. S., A. Mortazavi, R. M. Myers, and B. Wold. 2007. "Genome-wide mapping of in vivo protein-DNA interactions." *Science* no. 316 (5830):1497-502. doi: 10.1126/science.1141319.

Jones, P. A., and S. B. Baylin. 2002. "The fundamental role of epigenetic events in cancer." *Nat Rev Genet* no. 3 (6):415-28. doi: 10.1038/nrg816.

Juric, D., N. J. Lacayo, M. C. Ramsey, J. Racevskis, P. H. Wiernik, J. M. Rowe, A. H. Goldstone, P. J. O'Dwyer, E. Paietta, and B. I. Sikic. 2007. "Differential gene expression patterns and interaction networks in BCR-ABL-positive and -negative adult acute lymphoblastic leukemias." *J Clin Oncol* no. 25 (11):1341-9. doi: 10.1200/JCO.2006.09.3534.

Kanduri, M., N. Cahill, H. Goransson, C. Enstrom, F. Ryan, A. Isaksson, and R. Rosenquist. 2010. "Differential genome-wide array-based methylation profiles in prognostic subsets of chronic lymphocytic leukemia." *Blood* no. 115 (2):296-305. doi: 10.1182/blood-2009-07-232868.

Kang, X., Z. Lu, C. Cui, M. Deng, Y. Fan, B. Dong, X. Han, F. Xie, J. W. Tyner, J. E. Coligan, R. H. Collins, X. Xiao, M. J. You, and C. C. Zhang. 2015. "The ITIM-containing receptor LAIR1 is essential for acute myeloid leukaemia development." *Nat Cell Biol* no. 17 (5):665-77. doi: 10.1038/ncb3158.

Keating, M. J., N. Chiorazzi, B. Messmer, R. N. Damle, S. L. Allen, K. R. Rai, M. Ferrarini, and T. J. Kipps. 2003. "Biology and treatment of chronic lymphocytic leukemia." *Hematology Am Soc Hematol Educ Program*:153-75.

Kibriya, M. G., M. Raza, F. Jasmine, S. Roy, R. Paul-Brutus, R. Rahaman, C. Dodsworth, M. Rakibuz-Zaman, M. Kamal, and H. Ahsan. 2011. "A genome-wide DNA methylation study in colorectal carcinoma." *BMC Med Genomics* no. 4:50. doi: 10.1186/1755-8794-4-50.

Klajic, J., T. Fleischer, E. Dejeux, H. Edvardsen, F. Warnberg, I. Bukholm, P. E. Lonning, H. Solvang, A. L. Borresen-Dale, J. Tost, and V. N. Kristensen. 2013. "Quantitative DNA methylation analyses reveal stage dependent DNA methylation and association to clinico-pathological factors in breast tumors." *BMC Cancer* no. 13:456. doi: 10.1186/1471-2407-13-456.

Kong, J. H., Y. C. Mun, S. Kim, H. S. Choi, Y. K. Kim, H. J. Kim, J. H. Moon, S. K. Sohn, S. H. Kim, C. W. Jung, and D. H. Dennis Kim. 2012. "Polymorphisms of ERCC1 genotype associated with response to imatinib therapy in chronic phase

chronic myeloid leukemia." *Int J Hematol* no. 96 (3):327-33. doi: 10.1007/s12185-012-1142-6.

Kulis, M., and M. Esteller. 2010. "DNA methylation and cancer." *Adv Genet* no. 70:27-56. doi: 10.1016/B978-0-12-380866-0.60002-2.

Kulldorff, M. 1997. "A spatial scan statistic." *Communications in Statistics-Theory and Methods* no. 26 (6):1481-1496. doi: Doi 10.1080/03610929708831995.

Lacey, M. R., C. Baribault, and M. Ehrlich. 2013. "Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments." *Stat Appl Genet Mol Biol* no. 12 (6):723-42. doi: 10.1515/sagmb-2013-0027.

Laird, P. W. 2010. "Principles and challenges of genomewide DNA methylation analysis." *Nat Rev Genet* no. 11 (3):191-203. doi: 10.1038/nrg2732.

Lasa, A., E. Serrano, M. Carricondo, M. J. Carnicer, S. Brunet, I. Badell, J. Sierra, A. Aventin, and J. F. Nomdedeu. 2008. "High expression of CEACAM6 and CEACAM8 mRNA in acute lymphoblastic leukemias." *Ann Hematol* no. 87 (3):205-11. doi: 10.1007/s00277-007-0388-1.

Lee, Y. K., S. Jin, S. Duan, Y. C. Lim, D. P. Ng, X. M. Lin, GSh Yeo, and C. Ding. 2014. "Improved reduced representation bisulfite sequencing for epigenomic profiling of clinical samples." *Biol Proced Online* no. 16 (1):1. doi: 10.1186/1480-9222-16-1.

Leek, J. T., R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry. 2010. "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nat Rev Genet* no. 11 (10):733-9. doi: 10.1038/nrg2825.

Leek, J. T., and J. D. Storey. 2007. "Capturing heterogeneity in gene expression studies by surrogate variable analysis." *PLoS Genet* no. 3 (9):1724-35. doi: 10.1371/journal.pgen.0030161.

Li, S., F. E. Garrett-Bakelman, A. Akalin, P. Zumbo, R. Levine, B. L. To, I. D. Lewis, A. L. Brown, R. J. D'Andrea, A. Melnick, and C. E. Mason. 2013. "An optimized algorithm for detecting and annotating regional differential methylation." *BMC Bioinformatics* no. 14 Suppl 5:S10. doi: 10.1186/1471-2105-14-S5-S10.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker. 2008. "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." *Cell* no. 133 (3):523-36. doi: 10.1016/j.cell.2008.03.029.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker. 2009. "Human DNA methylomes at base resolution show widespread epigenomic differences." *Nature* no. 462 (7271):315-22. doi: 10.1038/nature08514.

Liu, D., D. Wu, H. Li, and M. Dong. 2014. "The effect of XPD/ERCC2 Lys751Gln polymorphism on acute leukemia risk: a systematic review and meta-analysis." *Gene* no. 538 (2):209-16. doi: 10.1016/j.gene.2014.01.049.

Liu, J., M. Morgan, K. Hutchison, and V. D. Calhoun. 2010. "A study of the influence of sex on genome wide methylation." *PLoS One* no. 5 (4):e10028. doi: 10.1371/journal.pone.0010028.

Lun, A. T., and G. K. Smyth. 2014. "De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly." *Nucleic Acids Res* no. 42 (11):e95. doi: 10.1093/nar/gku351.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. "Finding the missing heritability of complex diseases." *Nature* no. 461 (7265):747-53. doi: 10.1038/nature08494.

Mantel, N. 1967. "The detection of disease clustering and a generalized regression approach." *Cancer Res* no. 27 (2):209-20.

Martin-Subero, J. I., O. Ammerpohl, M. Bibikova, E. Wickham-Garcia, X. Agirre, S. Alvarez, M. Bruggemann, S. Bug, M. J. Calasanz, M. Deckert, M. Dreyling, M. Q. Du, J. Durig, M. J. Dyer, J. B. Fan, S. Gesk, M. L. Hansmann, L. Harder, S. Hartmann, W. Klapper, R. Kuppers, M. Montesinos-Rongen, I. Nagel, C. Pott, J. Richter, J. Roman-Gomez, M. Seifert, H. Stein, J. Suela, L. Trumper, I. Vater, F. Prosper, C. Haferlach, J. Cruz Cigudosa, and R. Siebert. 2009. "A comprehensive microarray-based DNA methylation study of 367 hematological neoplasms." *PLoS One* no. 4 (9):e6986. doi: 10.1371/journal.pone.0006986.

Mayr, B., and M. Montminy. 2001. "Transcriptional regulation by the phosphorylation-dependent factor CREB." *Nat Rev Mol Cell Biol* no. 2 (8):599-609. doi: 10.1038/35085068.

McHale, C. M., L. Zhang, Q. Lan, G. Li, A. E. Hubbard, M. S. Forrest, R. Vermeulen, J. Chen, M. Shen, S. M. Rappaport, S. Yin, M. T. Smith, and N. Rothman. 2009. "Changes in the peripheral blood transcriptome associated with occupational benzene exposure identified by cross-comparison on two microarray platforms." *Genomics* no. 93 (4):343-9. doi: 10.1016/j.ygeno.2008.12.006.

Meissner, A., A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. 2005. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." *Nucleic Acids Res* no. 33 (18):5868-77. doi: 10.1093/nar/gki901.

Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander.

2008. "Genome-scale DNA methylation maps of pluripotent and differentiated cells." *Nature* no. 454 (7205):766-70. doi: 10.1038/nature07107.

Mitomi, H., N. Fukui, N. Tanaka, H. Kanazawa, T. Saito, T. Matsuoka, and T. Yao. 2010. "Aberrant p16((INK4a)) methylation is a frequent event in colorectal cancers: prognostic value and relation to mRNA expression and immunoreactivity." *J Cancer Res Clin Oncol* no. 136 (2):323-31. doi: 10.1007/s00432-009-0688-z.

Moon, E., R. Lee, R. Near, L. Weintraub, S. Wolda, and A. Lerner. 2002. "Inhibition of PDE3B augments PDE4 inhibitor-induced apoptosis in a subset of patients with chronic lymphocytic leukemia." *Clin Cancer Res* no. 8 (2):589-95.

Muto, T., M. Takeuchi, A. Yamazaki, Y. Sugita, S. Tsukamoto, S. Sakai, Y. Takeda, N. Mimura, C. Ohwada, E. Sakaida, N. Aotsuka, T. Iseki, and C. Nakaseko. 2015. "Efficacy of myeloablative allogeneic hematopoietic stem cell transplantation in adult patients with MLL-ELL-positive acute myeloid leukemia." *Int J Hematol* no. 102 (1):86-92. doi: 10.1007/s12185-015-1779-z.

Naus, Joseph I. 1965. "The Distribution of the Size of the Maximum Cluster of Points on a Line." *Journal of the American Statistical Association* no. 60 (310):532-538. doi: 10.2307/2282688.

O'Connor, C., J. Campos, J. M. Osinski, R. M. Gronostajski, A. M. Michie, and K. Keeshan. 2015. "Nfix expression critically modulates early B lymphopoiesis and myelopoiesis." *PLoS One* no. 10 (3):e0120102. doi: 10.1371/journal.pone.0120102.

Park, Y., M. E. Figueroa, L. S. Rozek, and M. A. Sartor. 2014. "MethylSig: a whole genome DNA methylation analysis pipeline." *Bioinformatics* no. 30 (17):2414-22. doi: 10.1093/bioinformatics/btu339.

Pedersen, B. S., D. A. Schwartz, I. V. Yang, and K. J. Kechris. 2012. "Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values." *Bioinformatics* no. 28 (22):2986-8. doi: 10.1093/bioinformatics/bts545.

Pei, L., J. H. Choi, J. Liu, E. J. Lee, B. McCarthy, J. M. Wilson, E. Speir, F. Awan, H. Tae, G. Arthur, J. L. Schnabel, K. H. Taylor, X. Wang, D. Xu, H. F. Ding, D. H. Munn, C. Caldwell, and H. Shi. 2012. "Genome-wide DNA methylation analysis reveals novel epigenetic changes in chronic lymphocytic leukemia." *Epigenetics* no. 7 (6):567-78. doi: 10.4161/epi.20237.

Pomraning, K. R., K. M. Smith, and M. Freitag. 2009. "Genome-wide high throughput analysis of DNA methylation in eukaryotes." *Methods* no. 47 (3):142-50. doi: 10.1016/j.ymeth.2008.09.022.

Qureshi, S. A., M. U. Bashir, and A. Yaqinuddin. 2010. "Utility of DNA methylation markers for diagnosing cancer." *Int J Surg* no. 8 (3):194-8. doi: 10.1016/j.ijsu.2010.02.001.

Rahmatpanah, F. B., S. Carstens, J. Guo, O. Sjahputera, K. H. Taylor, D. Duff, H. Shi, J. W. Davis, S. I. Hooshmand, R. Chitma-Matsiga, and C. W. Caldwell. 2006.

"Differential DNA methylation patterns of small B-cell lymphoma subclasses with different clinical behavior." *Leukemia* no. 20 (10):1855-62. doi: 10.1038/sj.leu.2404345.

Rakyan, V. K., T. A. Down, D. J. Balding, and S. Beck. 2011. "Epigenome-wide association studies for common human diseases." *Nat Rev Genet* no. 12 (8):529-41. doi: 10.1038/nrg3000.

Rakyan, V. K., T. Hildmann, K. L. Novik, J. Lewin, J. Tost, A. V. Cox, T. D. Andrews, K. L. Howe, T. Otto, A. Olek, J. Fischer, I. G. Gut, K. Berlin, and S. Beck. 2004. "DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project." *PLoS Biol* no. 2 (12):e405. doi: 10.1371/journal.pbio.0020405.

Rantakari, P., K. Auvinen, N. Jappinen, M. Kapraali, J. Valtonen, M. Karikoski, H. Gerke, E. Khuda I. Iftakhar, J. Keuschnigg, E. Umemoto, K. Tohya, M. Miyasaka, K. Elima, S. Jalkanen, and M. Salmi. 2015. "The endothelial protein PLVAP in lymphatics controls the entry of lymphocytes and antigens into lymph nodes." *Nat Immunol* no. 16 (4):386-96. doi: 10.1038/ni.3101.

Rao, J. N., and A. J. Scott. 1992. "A simple method for the analysis of clustered binary data." *Biometrics* no. 48 (2):577-85.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. 2000. "Genome-wide location and function of DNA binding proteins." *Science* no. 290 (5500):2306-9. doi: 10.1126/science.290.5500.2306.

Robertson, T., and E. J. Wegman. 1978. "Likelihood Ratio Tests for Order Restrictions in Exponential Families." *Annals of Statistics* no. 6 (3):485-505. doi: Doi 10.1214/Aos/1176344195.

Runne, C., and S. Chen. 2013. "PLEKHG2 promotes heterotrimeric G protein betagamma-stimulated lymphocyte migration via Rac and Cdc42 activation and actin polymerization." *Mol Cell Biol* no. 33 (21):4294-307. doi: 10.1128/MCB.00879-13.

Russo, V. E. A., Robert A. Martienssen, and Arthur D. Riggs. 1996. *Epigenetic mechanisms of gene regulation, Cold Spring Harbor monograph series,*. Plainview, N.Y.: Cold Spring Harbor Laboratory Press.

Ryu, D., H. Xu, V. George, S. Su, X. Wang, H. Shi, and R. H. Podolsky. 2016. "Differential methylation tests of regulatory regions." *Stat Appl Genet Mol Biol*. doi: 10.1515/sagmb-2015-0037.

Saito, Y., J. Tsuji, and T. Mituyama. 2014. "Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions." *Nucleic Acids Res* no. 42 (6):e45. doi: 10.1093/nar/gkt1373.

Schaid, D. J., J. P. Sinnwell, S. K. McDonnell, and S. N. Thibodeau. 2013. "Detecting genomic clustering of risk variants from sequence data: cases versus controls." *Hum Genet* no. 132 (11):1301-9. doi: 10.1007/s00439-013-1335-y.

Secchiero, P., E. Barbarotto, M. Tiribelli, C. Zerbinati, M. G. di Iasio, A. Gonelli, F. Cavazzini, D. Campioni, R. Fanin, A. Cuneo, and G. Zauli. 2006. "Functional integrity of the p53-mediated apoptotic pathway induced by the nongenotoxic agent nutlin-3 in B-cell chronic lymphocytic leukemia (B-CLL)." *Blood* no. 107 (10):4122-9. doi: 10.1182/blood-2005-11-4465.

Shaw, D. J., H. G. Harley, J. D. Brook, and T. W. McKeithan. 1989. "Long-range restriction map of a region of human chromosome 19 containing the apolipoprotein genes, a CLL-associated translocation breakpoint, and two polymorphic MluI sites." *Hum Genet* no. 83 (1):71-4.

Sincennes, M. C., M. Humbert, B. Grondin, V. Lisi, D. F. Veiga, A. Haman, C. Cazaux, N. Mashtalir, B. Affar el, A. Verreault, and T. Hoang. 2016. "The LMO2 oncogene regulates DNA replication in hematopoietic cells." *Proc Natl Acad Sci U S A* no. 113 (5):1393-8. doi: 10.1073/pnas.1515071113.

Smithson, M., and J. Verkuilen. 2006. "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables." *Psychol Methods* no. 11 (1):54-71. doi: 10.1037/1082-989X.11.1.54.

Spisak, S., A. Kalmar, O. Galamb, B. Wichmann, F. Sipos, B. Peterfia, I. Csabai, I. Kovalszky, S. Semsey, Z. Tulassay, and B. Molnar. 2012. "Genome-wide screening of genes regulated by DNA methylation in colon cancer development." *PLoS One* no. 7 (10):e46215. doi: 10.1371/journal.pone.0046215.

Stewart, H. J., G. A. Horne, S. Bastow, and T. J. Chevassut. 2013. "BRD4 associates with p53 in DNMT3A-mutated leukemia cells and is implicated in apoptosis by the bromodomain inhibitor JQ1." *Cancer Med* no. 2 (6):826-35. doi: 10.1002/cam4.146.

Stockwell, P. A., A. Chatterjee, E. J. Rodger, and I. M. Morison. 2014. "DMAP: differential methylation analysis package for RRBS and WGBS data." *Bioinformatics* no. 30 (13):1814-22. doi: 10.1093/bioinformatics/btu126.

Sun, D., Y. Xi, B. Rodriguez, H. J. Park, P. Tong, M. Meong, M. A. Goodell, and W. Li. 2014. "MOABS: model based analysis of bisulfite sequencing data." *Genome Biol* no. 15 (2):R38. doi: 10.1186/gb-2014-15-2-r38.

Talby, L., H. Chambost, M. C. Roubaud, C. N'Guyen, M. Milili, B. Loriod, C. Fossat, C. Picard, J. Gabert, P. Chiappetta, G. Michel, and C. Schiff. 2006. "The chemosensitivity to therapy of childhood early B acute lymphoblastic leukemia could be determined by the combined expression of CD34, SPI-B and BCR genes." *Leuk Res* no. 30 (6):665-76. doi: 10.1016/j.leukres.2005.10.007.

Tang, H. M., W. W. Gao, C. P. Chan, Y. Cheng, J. J. Deng, K. S. Yuen, H. Iha, and D. Y. Jin. 2015. "SIRT1 Suppresses Human T-Cell Leukemia Virus Type 1 Transcription." *J Virol* no. 89 (16):8623-31. doi: 10.1128/JVI.01229-15.

Tango, T. 1984. "The Detection of Disease Clustering in Time." *Biometrics* no. 40 (1):15-26. doi: Doi 10.2307/2530740.

Tango, T. 1995. "A class of tests for detecting 'general' and 'focused' clustering of rare diseases." *Stat Med* no. 14 (21-22):2323-34.

Tango, T. 2000. "A test for spatial disease clustering adjusted for multiple testing." *Statistics in Medicine* no. 19 (2):191-204. doi: Doi 10.1002/(Sici)1097-0258(20000130)19:2<191::Aid-Sim281>3.0.Co;2-Q.

Tango, Toshiro. 2012. *Statistical Methods for Disease Clustering*: Springer Publishing Company, Incorporated.

Teschendorff, A. E., U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, D. A. Savage, E. Mueller-Holzner, C. Marth, G. Kocjan, S. A. Gayther, A. Jones, S. Beck, W. Wagner, P. W. Laird, I. J. Jacobs, and M. Widschwendter. 2010. "Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer." *Genome Res* no. 20 (4):440-6. doi: 10.1101/gr.103606.109.

Tong, W. G., W. G. Wierda, E. Lin, S. Q. Kuang, B. N. Bekele, Z. Estrov, Y. Wei, H. Yang, M. J. Keating, and G. Garcia-Manero. 2010. "Genome-wide DNA methylation profiling of chronic lymphocytic leukemia allows identification of epigenetically repressed molecular pathways with clinical impact." *Epigenetics* no. 5 (6):499-508.

Visscher, P. M. 2008. "Sizing up human height variation." *Nat Genet* no. 40 (5):489-90. doi: 10.1038/ng0508-489.

Wallingford, M. C., R. Filkins, D. Adams, M. Walentuk, A. M. Salicioni, P. E. Visconti, and J. Mager. 2015. "Identification of a novel isoform of the leukemia-associated MLLT1 (ENL/LTG19) protein." *Gene Expr Patterns* no. 17 (1):11-5. doi: 10.1016/j.gep.2014.11.003.

Wang, H. Q., L. K. Tuominen, and C. J. Tsai. 2011. "SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures." *Bioinformatics* no. 27 (2):225-31. doi: 10.1093/bioinformatics/btq650.

Wang, T., D. Qian, M. Hu, L. Li, L. Zhang, H. Chen, R. Yang, and B. Wang. 2014. "Human cytomegalovirus inhibits apoptosis by regulating the activating transcription factor 5 signaling pathway in human malignant glioma cells." *Oncol Lett* no. 8 (3):1051-1057. doi: 10.3892/ol.2014.2264.

Watts, G. S., B. W. Futscher, N. Holtan, K. Degeest, F. E. Domann, and S. L. Rose. 2008. "DNA methylation changes in ovarian cancer are cumulative with disease progression and identify tumor stage." *BMC Med Genomics* no. 1:47. doi: 10.1186/1755-8794-1-47.

World Health Organization. *Cancer Fact Sheet* 2014. Available from http://www.who.int/mediacentre/factsheets/fs297/en/.

Wu, H., T. Xu, H. Feng, L. Chen, B. Li, B. Yao, Z. Qin, P. Jin, and K. N. Conneely. 2015. "Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates." *Nucleic Acids Res* no. 43 (21):e141. doi: 10.1093/nar/gkv715.

Xu, H., R. H. Podolsky, D. Ryu, X. Wang, S. Su, H. Shi, and V. George. 2013. "A method to detect differentially methylated loci with next-generation sequencing." *Genet Epidemiol* no. 37 (4):377-82. doi: 10.1002/gepi.21726.

Xu, W. L., L. L. Zhou, Q. Y. Chen, C. Chen, L. L. Fang, X. J. Fang, and H. L. Shen. 2009. "[Effect of YB-1 gene knockdown on human leukemia cell line K562/A02]." *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* no. 26 (4):400-5.

Yip, W. K., H. Fier, D. L. DeMeo, M. Aryee, N. Laird, and C. Lange. 2014. "A novel method for detecting association between DNA methylation and diseases using spatial information." *Genet Epidemiol* no. 38 (8):714-21. doi: 10.1002/gepi.21851.