Summer 2017

A functional data analytic approach for region level differential DNA methylation detection

Mohamed Salem F. Milad

Department: Mathematics and Statistics

A FUNCTIONAL DATA ANALYTIC APPROACH FOR REGION LEVEL DIFFERENTIAL DNA METHYLATION DETECTION

by

MOHAMED SALEM F. MILAD

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

2017

Approved

Dr. Gayla R. Olbricht, Advisor

Dr. V.A. Samaranayake

Dr. Robert Paige

Dr. Ronald L. Frank

Dr. R.W. Doerge

PUBLICATION DISSERTATION OPTION

This dissertation has been prepared using the publication option. This dissertation contains the following two manuscripts for publication:

PAPER I: pages 21 - 44 “Testing Differentially Methylated Regions through Functional Principal Component Analysis”, Section 2 will be submitted to the journal *Epigenetics*.

PAPER II: pages 45 - 69 “Smoothed Functional Principal Component Analysis for Detecting Differentially Methylated Regions”, Section 3 will be submitted to the journal *Bioinformatics*.

ABSTRACT

DNA methylation is an epigenetic modification that can alter gene expression without a DNA sequence change. The role of DNA methylation in biological processes and human health is important to understand, with many studies identifying associations between specific methylation patterns and diseases such as cancer. In mammals, DNA methylation almost always occurs when a methyl group attaches to a cytosine followed by a guanine (i.e. CpG dinucleotides) on the DNA sequence. Many statistical methods have been developed to test for a difference in DNA methylation levels between groups (e.g. healthy vs disease) at individual cytosines. Site level testing is often followed by a post hoc aggregation procedure that explores regional differences. Although analyzing CpGs individually provides useful information, there are both biological and statistical reasons to test entire genomic regions for differential methylation. The individual loci may be noisy but the overall regions tend to be informative. Also, the biological function of regions is better studied and are more correlated to gene expression, so the interpretation of results will be more meaningful for region-level tests. This study focuses on developing two techniques, functional principal component analysis (FPCA) and smoothed functional principal component analysis (SFPCA), to identify differentially methylated regions (DMRs) that will enable discovery of epigenomic structural variations in NGS data. Using real and simulated data, the performance of these novel approaches are compared with an alternative method (M3D) for region level testing.

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Dr. Gayla R. Olbricht for her guidance, support, and encouragement during the course of my Ph.D. research. Her critical thinking improved my research skills and prepared me for future challenges. It was a great honor for me to be a part of her research group. I am particularly thankful to Dr. V.A. Samaranayake for his inestimable support and scientific advice. I am deeply grateful to him. Special thanks are due to my committee members Dr. Robert Paige, Dr. Ronald L. Frank, and Dr. R.W. Doerge for their support and guidance throughout the completion of my Ph.D. program.

I gratefully acknowledge the Ministry of Higher Education – Libya for financial support during my pursuit of graduate studies at Missouri S&T.

Finally, I wish to thank my entire family, my parents, my sisters, and my brothers. Special thanks are due to my wife for her patience and support.

.

TABLE OF CONTENTS

Page

PAPER

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

1.1. BASICS OF GENETICS

The field of genetics involves the study of heredity and genetic variation, which includes investigating the properties of genes. Genes are sections of deoxyribonucleic acid (DNA) located inside each cell of an organism that encode proteins and play a role in determining the nature of living organisms [1]. Organisms inherit phenotypic traits or characteristics based on the genes transmitted by their parents. For example, the products of sexual reproduction often resemble their parents because they have inherited half of their genetic material from each parent. Investigating the function of genes at the molecular level is part of field a known as molecular genetics, which combines genetics with molecular biology. The Central Dogma of Molecular Biology[2] offers a way to understand how genes are converted to functional information. The Central Dogma describes how genes are transcribed to messenger RNA (the transcriptome) that is translated to proteins, which mediate most of the cell's biochemical functions (the proteome). Thus, molecular genetics involves not only inheritance but how genes are expressed, which controls how much of specific proteins are produced at the cellular level and can ultimately affect phenotypes. Many phenotypic traits are complex in nature as they may be determined by multiple genes and also influenced by the environment [1, 2].

1.2. GENOMICS

The field of genomics involves the study of the complete set of deoxyribonucleic acid (DNA) in an organism (i.e., the genome) [2]. The mapping of genomes for particular

organisms enables better understanding of the location of genes and their functions on a large scale. The human genome is the complete sequence of genetic information in humans, which is stored in each cell's nucleus and mitochondria. Genetic information is encoded in the DNA molecule and is stored on structures called chromosomes (Figure 1.1). DNA is double-stranded and is comprised of millions of nucleotides, organic molecules that function as subunits and are composed of a nitrogen nucleobase (i.e., guanine (G), adenine (A), thymine (T), and cytosine (C)), a five-carbon sugar, and at least one phosphate group [3]. DNA can be annotated into important substructures such as protein-coding genes and non-coding sequences (Figure 1.1). These main structures can be further annotated into substructures such as exons, introns, CpG islands, and promoter regions that play specific roles in different molecular processes [4].

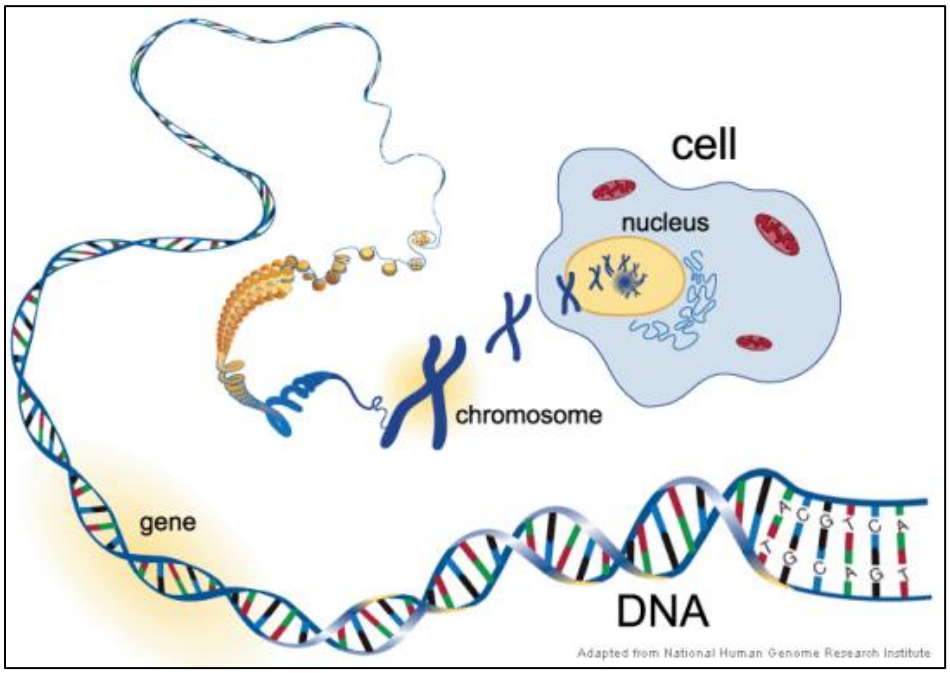

Figure 1.1. Location and Structure of the DNA Molecule in the Human Genome. Figure obtained from [5].

The first genome-wide DNA sequence in humans, with a total of about three billion nucleotide positions, was completed by the Human Genome Project [6] in 2001. DNA sequencing determines the order of all the nucleotides for any particular organism. This sequence information was initially used in organism specific genome projects, combined with computational methods and domain expertise, to map the location of genes and other substructures, to be used as reference for future studies of that organism. With advances in technology, all of the genetic information from any individual can now be revealed using whole-genome sequencing [7]. Such genome-wide studies are powerful tools for exploring genetic contributions to phenotypic variation and have the potential to allow for important health advances, such as personalized medicine, in the future.

1.3. EPIGENETICS

Epigenetics refers to heritable changes in genetic activity and expression that take place without any change in the DNA sequence. The word "epigenetics" comes from the Latin "epi," which means "above" or "on top of" the genetic information. Epigenetics encompasses all the information that is contained in the cell and expressed for more than one cell generation, as the DNA sequence remains stable [8]. In other words, epigenetics is "the study of mitotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [9]. DNA methylation and histone modifications, which involve the addition of chemical marks to the DNA or histone proteins, are two key epigenetic mechanisms (Figure 1.2) [10]. DNA methylation occurs when a methyl (Me) group attaches to a cytosine (C) base on the DNA molecule. Histone modifications occur when certain chemical groups (e.g. methyl, acetyl) attach to the tails of histone proteins. Epigenetic

modifications help to regulate gene expression [11-13] and epigenetic aberrations correlate with cancer [14, 15] and other diseases [16-17]. Environmental factors can affect epigenetic mechanisms. An advantage of this environmental influence is that drugs can be formulated to modify epigenetic patterns in cancer cells [18].

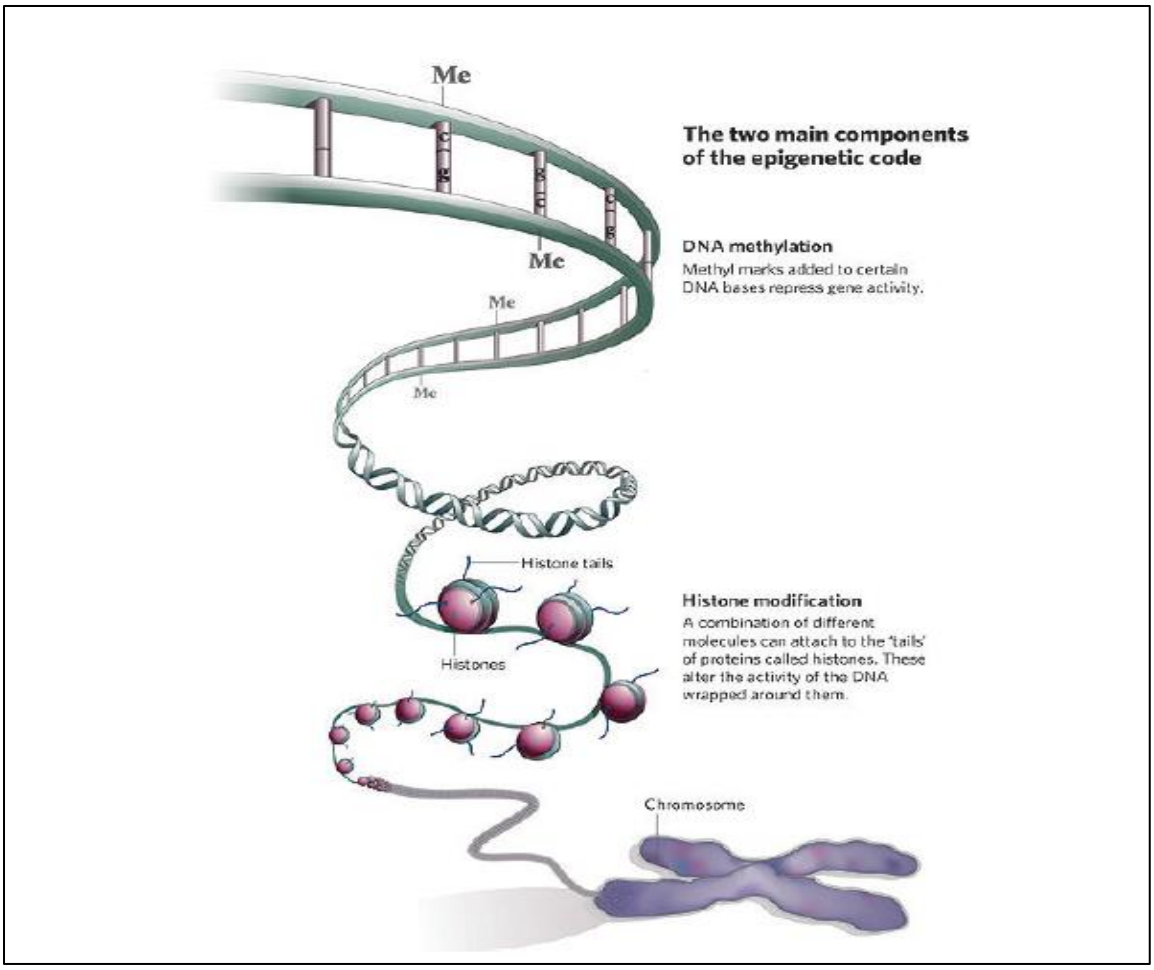

Figure 1.2. Two key epigenetic modifications: (1) DNA methylation and (2) histone modification. Figure obtained from Qiu (2006) [19].

1.4. DNA METHYLATION

The four DNA nucleotides (adenine (A), cytosine (C), guanine (G), and thymine (T)) can be categorized into two classes, the pyrimidine-based nucleotides (C and T) and the purine-based nucleotides (G and A). On the double stranded DNA (dsDNA) a purine on one strand pairs with a pyrimidine on the other strand. Called complementary base pairing, this always occurs where A is paired with T while G is paired with C. The A/T pairing is secured by two hydrogen bonds, but the G/C pairing is mediated by three, creating a stronger bond [20]. The length of the dsDNA is measured by the number of nucleotide base pairs (bp), which ranges from a few thousand (i.e., kilo base pairs (kbp)) in single-celled organisms to several million (i.e., mega base pairs (Mbp)) per molecule for complex organisms [21].

DNA methylation occurs in most organisms, but not in the budding yeast, *Saccharomyces cerevisiae*, and the nematode worm, *Caenorhabditis elegans*. It is also limited to embryonic development in the fruit fly, *Drosophila melanogaster* [22, 23]. In mammals, DNA methylation usually takes place when cytosine (C) is followed by guanine (G) in the $5' - 3'$ direction of the DNA sequence. This is denoted as CG or CpG, the latter notation showing that cytosine and guanine are connected by a phosphate on one of the DNA strands [24]. In plants, DNA is methylated in three sequence contexts: CG, CHG and CHH (where H = A, T or C). At least three DNA methylation pathways exist in plants and each pathway appears to methylate cytosines in different sequence contexts [25].

In mammals, DNA methylation occurs as the result of a family of de novo DNA methyltransferase enzymes (DNMT3) and is maintained during DNA replication by a maintenance DNA methyltransferase (DNMT1) [26]. Plants also have methyltransferase

enzymes, some similar to DNMT1 and other unique to plants [25]. At some locations, known as CpG islands, the number of CpG sites in relation to the CG content in a sequence of a particular length is higher than expected. CpG islands occur upstream of many genes and are usually unmethylated [25]. Recent research suggests that the relationship between genetic variation, DNA methylation, and expression is complex [27].

DNA methylation plays a key role in many biological processes, including genomic imprinting, X-chromosome inactivation, embryonic development, and the silencing of transposable elements [22, 28-31]. In plants, DNA methylation is essential for genome stability and plant development [28, 31]. In humans, specific DNA methylation patterns have been associated with the development of cancer [14]. An overall loss of DNA methylation (hypomethylation) that occurs with a gain in methylation (hypermethylation) at the CpG islands in promoter regions is often found in cancer cells [14, 17].

1.5. NEXT-GENERATION SEQUENCING TECHNOLOGY

The introduction of next-generation sequencing (NGS) technology (also known as high-throughput sequencing) in the 2000s enabled researchers to conduct genome-wide investigations of many different molecular level phenomena, including gene expression and epigenetic modifications such as DNA methylation. NGS is a high-throughput technology that allows cost-effective processing of millions of sequencing reads in parallel [32]. Although several companies manufacture NGS technologies (e.g. Illumina, Roche 454, Life Technologies), there is a set of general processing steps shared between them even though their specific technical details may differ (Figure 1.3) [33].

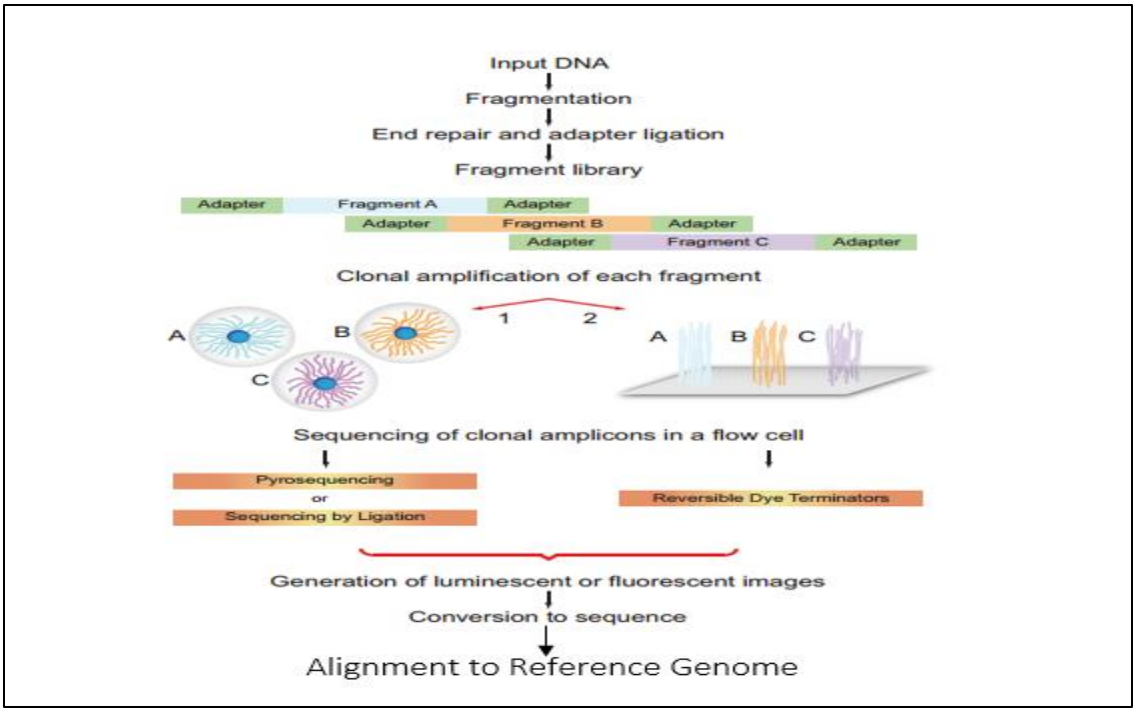

Figure 1.3 Next generation sequencing processing steps for platforms requiring clonally amplified templates (Roche 454, Illumina, and Life Technologies). Input DNA is converted to a sequencing library by fragmentation, end repair, and ligation to platform specific oligonucleotide adapters. Individual library fragments are clonally amplified by either (1) water in oil bead– based emulsion PCR (Roche 454 and Life Technologies) or (2) solid surface bridge amplification (Illumina). Flow cell sequencing of clonal templates generates luminescent or fluorescent images that are algorithmically processed into sequence reads. These reads are then aligned to a reference genome and evaluated based on the biological mechanism being investigated. Figure modified from Voelkerding et al. (2010) [33].

NGS enables the cost-efficient generation of large sequencing data sets based on whole genomes at single-base resolution. Today, NGS is used for variant detection by resequencing (personnel genomes), transcriptome analysis (RNA-seq), and the discovery of epigenetic variations (DNA methylation) [34]. NGS methods offer advantages for such large-scale studies over the traditional Sanger sequencing developed in 1977. The NGS high-throughput platforms have a higher coverage giving a more reliable and accurate result compared to those obtained via Sanger Sequencing technology [34]. Coverage is one of the common measures of the amount of sequence data generated and it refers to the

average number of times each base in the genome is sequenced [32]. In addition, the next-generation sequencing platforms are able to detect methylation levels at individual cytosines due to their higher accuracy and sensitivity, making them suitable for epigenomic investigations.

1.6. GENOME-WIDE METHYLATION PROFILING APPROACHES

DNA methylation can be investigated at a genome-wide level using a variety of technologies, most notably microarrays and next-generation sequencing (NGS). The focus of this work is on NGS technologies as they can cover cytosine sites across the entire genome and not just a pre-chosen subset covered by microarrays. NGS entails a series of steps as described in section 1.5, including library preparation, amplification, sequencing, imaging, and alignment, resulting in millions of sequencing reads per run [33]. An improvement over microarray technologies, next-generation technologies can cover a wide breadth of the genome including repetitive elements [35]. Along with these newer sequencing technologies, novel approaches have been developed to obtain genome-wide profiles of DNA methylation. Some require bisulfite-converted genomic DNA for template preparation, such as MethylC-seq and RRBS (reduced representation bisulfite sequencing) [35, 36]; some rely on the enrichment of methylated DNA, such as MeDIP-seq (methylated DNA immunoprecipitation sequencing) and MBD-seq (methylated DNA binding domain sequencing) [37, 38]; and some use methylation-sensitive characteristics of restriction enzymes to digest genomic DNA [38]. Each method has advantages and disadvantages with regard to covered regions, sequencing depth, accuracy, and cost. For example, MeDIP-seq and MBD-seq cannot investigate at a single-base resolution, but they can reflect high-to-

medium methylation of DNA sequences covering broader regions [22]. In contrast, whole genome bisulfite-based methods such as methylC-seq provide measurements at single cytosines and are considered the gold standard, but the cost is still too high for many smaller-scale labs to utilize this technique [23]. As a compromise, RRBS combines the use of restriction enzymes with bisulfite sequencing and NGS to obtain methylation levels at individual cytosines in a subset of the genome with high CpG content. This reduces the cost at the expense of losing information in some regions. The focus on the methods developed in this work are on the bisulfite based methods, described in more details in next section.

1.7. BISULFITE SEQUENCING-BASED METHODS TO PROFILE DNA METHYLATION

Using NGS to quantify DNA methylation at the single based level relies on a technique called bisulfite sequencing [39]. This technique utilize a process called bisulfite conversion of genomic DNA, in which the DNA molecules undergo a bisulfite treatment that allows methylated and unmethylated cytosines to be differentiated at single-base resolution. Using this method, unmethylated cytosines are converted to uracils (which will be read as thymines by the DNA polymerase), leaving methylated cytosines unmodified [39]. When the bisulfite-treated DNA is amplified by a polymerase chain reaction (PCR), it yields products in which unmethylated cytosines appear as thymines (Figure 1.4). Therefore, when combined with NGS, it is possible to infer the number of cytosine and thymine reads at a specific genomic position of a known cytosine site. This results in a count of the number of methylated reads and unmethylated reads at single-base resolution which could be converted to a methylation percentages at that position.

1.8. REDUCED REPRESENTATION BISULFITE SEQUENCING

Bisulfite sequencing can be combined with NGS for whole genome studies using methods such as BS-seq or methylC-seq [40]. Although these are considered the gold standard, they are often cost prohibitive, especially for large genomes, large sample sizes, and small labs. A smaller-scale method developed by Meissner et al. [36] also employs bisulfite-converted DNA and provides insights into a subset of the methylome [41]. Meissner et al. [36] pioneered the reduced representation bisulfite sequencing (RRBS) approach, which is more feasible for case-control studies in humans with large sample sizes [42] and for use in smaller labs. RRBS digests genomic DNA with a methylation-insensitive restriction enzyme, where fragments of a specific length are used to filter the most informative genomic subset. Then, a bisulfite conversion of the fragments is undertaken to establish DNA methylation levels [45], which ultimately provides DNA methylation patterns in the chosen segments of the genome. These restricted fragments often cover core promoters and CpG islands [41], which contain key regulatory parts of the genome. Altogether, RRBS comprises only ~1% of the underlying whole genome [41]. DNA quantities as little as 10-300 ηg are sufficient to produce accurate DNA methylation levels with RRBS [43]. Therefore, RRBS is suitable for many clinical samples (e.g., tumors) that only supply a small quantity of genomic input DNA material. The following section describes selected steps in preparing an RRBS library (see Figure 1.4).

The first step is to isolate the genomic DNA. Using highly purified genomic input DNA is mandatory when generating a high-quality RRBS library [43]. Otherwise, contaminated DNA molecules might interact with the restriction enzymes, which affect the bisulfite conversion [43]. The second step is the digestion reaction and fragmentation.

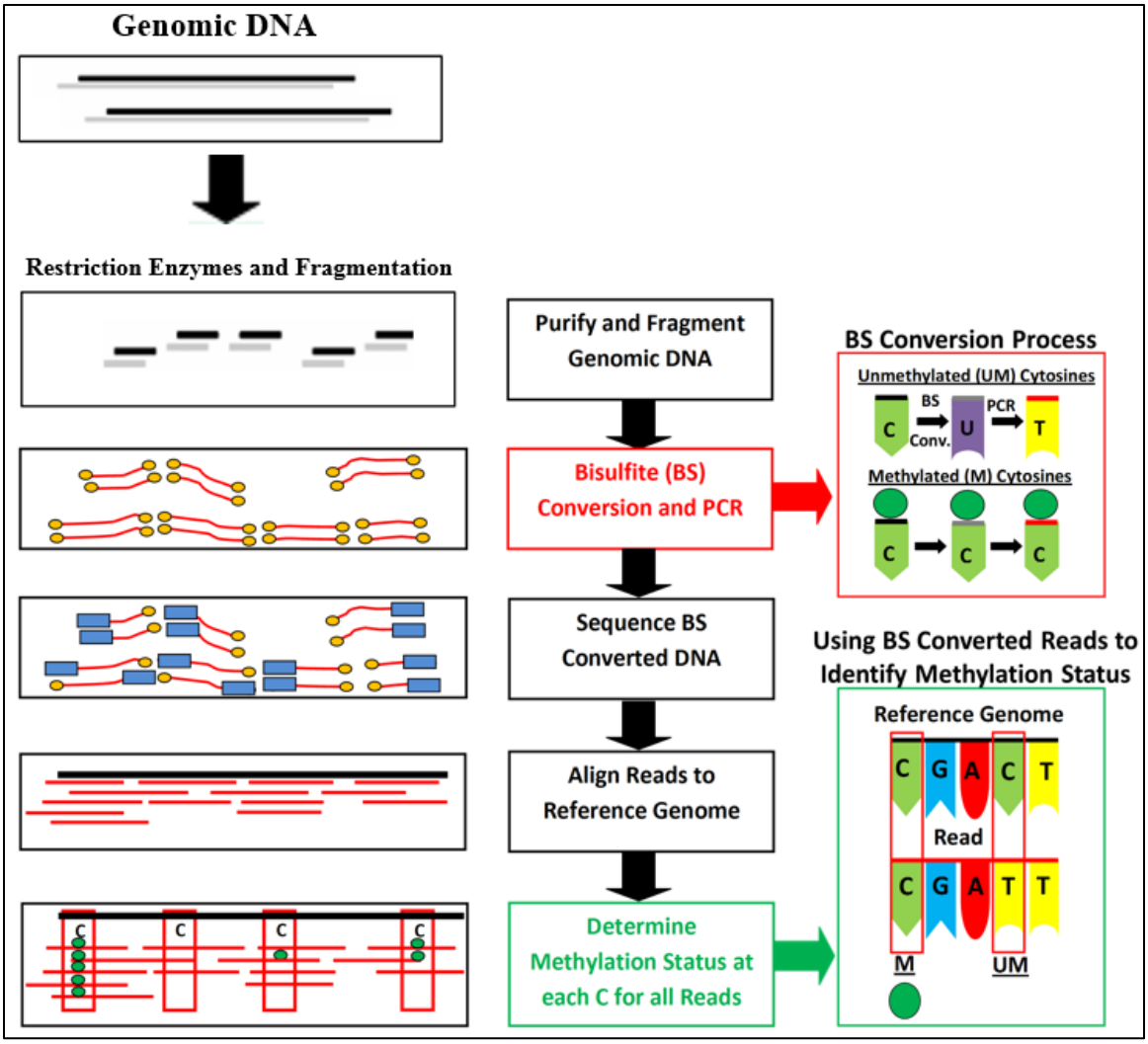

Figure 1.4. Workflow of a RRBS Library Preparation. Image Courtesy of Olbricht (2006) [44].

Two commercially available enzymes, MspI and TaqI, [43] can be used since they are insensitive to CpG DNA methylation, and will thus not bias the methylation measurements. However, only MspI produces fragments that contain CpG dinucleotides, at both ends, which is important to aid in capturing CpG dense regions while reducing the genomic space. One disadvantage of MspI is that a methylated cytosine in the first position of the restriction motif C↓CGG hampers the digestion reaction [43].

Some intermediate steps are then taken to prepare the fragments remaining after the restriction enzyme digestion for sequencing. Methylation adapters are ligated to double stranded sheared DNA fragments so that the fragments can be hybridized to the flow cell for sequencing [45].

Using RRBS libraries, both single-end and paired-end sequencing can be conducted, but adapters must be methylated cytosines to maintain compatibility with the subsequent bisulfite conversion. One advantage of paired-end sequencing is that it improves the mapping efficiency by fostering unique alignments. However, it can also produce inaccurate DNA methylation levels because overlapping pairs produce redundant DNA methylation information [46]. Before sequencing, fragments are size selected where fragments that do not meet a minimum length are filtered out before the remaining fragments are bisulfite converted. It has been shown through in silico analyses that a size selection for fragments of 40-220 bp that contains the MspI restriction motif C↓CGG covers the preparation of most promoter sequences and CpG islands [47].

In the third main step, the digested and size-selected fragments are bisulfite converted as described in section 1.7 and then amplified by PCR. Step four involves sequencing the bisulfite converted fragments using a NGS platform. At this time, RRBS has only been performed on Illumina platforms [46]. In step five, sequenced reads are aligned to a reference genome. Finally, in step six, the methylation status of each cytosine is determined for all reads and summarized, resulting in a count of the number of methylated and the number of unmethylated reads at each cytosine sequenced. Note that in mammals, typically only CpG sites are summarized, but all cytosines are of interest in plant

1.9. LITERATURE REVIEW FOR STATISTICAL METHODS

Table 1.1 provides information about various statistical methodologies used to discover individual differentially methylated CpG sites and differentially methylated regions for bisulfite-based NGS technology. Note that some methods were utilized for whole genome bisulfite sequencing studies such as BS-seq [48], while other were only tested on RRBS data. Both methods yield data with counts of both methylated and unmethylated reads at each cytosine site where data are available. At times, researchers may want to relate individual CpG sites to a particular phenotype [49]. Early BS-seq studies typically profiled cell lines, but did not collect replicates, and used the Fisher's exact test (FET) to define differentially methylated cytosine sites between phenotypic conditions [50]. This strategy may be adequate for comparing cell lines, but overall, the use of FET should be avoided because FET does not take into account inherent biological variations. As such, when using FET in a two-condition comparison, the data must be condensed to account for each condition, meaning that any within-condition variability is disregarded. Because this process underestimates variability and magnifies differences, the false positive rate is much higher. Similarly, using a binomial distribution (e.g., within a logistic regression framework, such as methylKit [51]) does not enable an accurate estimation of biological variability without using an over-dispersion term. Therefore, the optimal statistical model for measuring replicated BS-seq [48] DNA methylation is a beta-binomial. Based on the methylation proportion at any given site, the observations are binomially distributed, while the methylation proportion can vary across experimental units (e.g., patients), in a beta distribution. This is advantage can be seen in the latest versions of the BiSeq [48] and methylSig [52] methods, which employ beta-binomial assumptions.

Table 1.1 Statistical Methods to Detect Differentially Methylated Loci or Regions

Methods	Designed for	Determines regions or uses predefined	Accounts for covariates	Statistical elements
BSmooth	BS-seq	Determines	No	Bump hunting on smoothed t-score
BiSeq	BS-seq	Determines	Yes	Beta-binomial (Wald test)
MethylKit	BS-seq	Predefined	Yes	Logistic regression
MethylSig	BS-seq	Determines	No	Beta-binomial (Likelihood-Ratio Test)
MAGI	BS-seq	Predefined	No	FET and logistic regression
M3D	RRBS	Predefined	Yes	Kernel-based

Although site level tests can be informative, differentially methylated regions (DMRs) have a greater ability to predict phenotypes [53]. Another advantage of using DMRs is that although differences at any given site may be small and noisy, variations across a region can often be more easily detected since neighboring methylation levels are typically highly correlated [54]. However, methods operating on predefined regions differ from those that define regions of differential methylation after site level testing (i.e., the regions are not known in advance). Although the false discovery rate (FDR) needs to be controlled across the regions tested in both cases, this is nontrivial when the regions are not known in advance, making this strategy much more difficult for controlling the false positives. When the region is undefined prior testing, it is impossible to extrapolate the region-level FDR control from the site-level tests [53]. As such, the best approach is to use predefined regions which can be defined based on annotation regions (e.g., CpG islands, CpG shores, exons, or introns), or defined based on non-annotation regions that are defined based on CpG density. These non-annotation based CpG clusters can be defined as follows: (1) CpG sites that covered at least 75% of samples are defined as frequently covered CpG sites, and (2) a maximum distance of 100 base pairs to the nearest neighbor within a region is accepted. The predefined regions are limited to regions with at least 20 frequently covered CpG sites [50].

At the region-level, there are several statistical methods that can be employed, such as methylSig [52], methylKit [51], and others. Of the approaches that determine the regions after site-level testing, BSmooth is a widely cited package that looks for runs of smoothed, absolute t-like scores beyond a threshold. However, this approach does not contain a permutation strategy to control region-level FDR. Because DNA methylation levels usually have a strong spatial correlation, if such correlations could be accounted for in a region level testing procedure, the statistical power of that approach would increase greatly. Interestingly, M3D [55] has proposed such a nonparametric statistical test that would detect DMRs from predefined regions based on CpG density, while also accounting for spatial correlation. This method uses a radial basis function (RBF) kernel function to derive the Maximum Mean Discrepancy (MMD) between the data sets to assess the homogeneity of the underlying methylation distribution. MAGI [56] characterizes testing regions using existing annotation information, assuming spatial homogeneity across regions, but does not adjust for spatial correlations between individual cytosine sites. There are two versions of MAGI for site level and region level tests. Each version has the option of a FET if replicates are not available or a logistic regression when replication is present. The first version ($MAGI_c$) tests for differences between methylation levels at individual cytosine sites within each annotated region. The second version ($MAGI_g$) is comprised of two steps: (1) using an a priori threshold to classify each cytosine as either methylated or unmethylated, and, (2) performing a single FET or logistic regression on the resulting data for each region, with the assumption that the resulting data are binomially distributed [56].

This dissertation focuses on developing methods for DMR testing over predefined regions based on CpG density. Functional data analysis techniques are employed to more

fully utilize the nature of correlated methylation levels over genomic regions. Two techniques, functional principal component analysis (FPCA) and smoothed functional principal component analysis (SFPCA), are proposed to identify differentially methylated regions (DMRs) that will enable discovery of epigenomic structural variations in NGS data. The performance of these novel approaches are compared with the only other method (M3D) that investigates shape changes over a predefined region.

1.10. INTRODUCTION TO FUNCTIONAL DATA ANALYSIS

The main idea of functional data analysis (FDA) [57] involves analyzing data that can be represented as curves or functions. Typically, a trajectory of data is collected on one or more individuals of the form (t_i, y_i) where y_i represents the quantity of interest at time or positiont_i. Although these observations are collected at discrete points, the idea behind FDA is that there is underlying function $x(t)$ that is smooth such that data at sequential points y_{i-1}, y_i, y_{i+1} are linked to each other in some way and likely to exhibit similarity. This smoothness property is important for functional data, as otherwise wise it could be treated as multivariate data. One or more derivatives of the function $x(t)$ is assumed to exist due to the smoothness, where $D^m x(t)$ indicates the derivative of order m at argument t. Studying these derivatives of the function allows the exploration of properties such as velocity and acceleration.

To estimate the function $x(t)$ and certain number of its derivatives, the discrete data y_i are typically utilized [57]. However, the data observed may not be smooth due noise or measurement error. When the signal-to-noise ratio is low or sparsely sampled, it is helpful to have data from a random sample of individual records so that information can

be drawn from similar trajectories to obtain a more stable estimate of a specific curve. In functional data analysis, the goal is to represent experimental data collected over time or space with a series linear combinations of basis functions that are mathematically independent [60].

A function $X(t)$ can be represented as $X(t) = \sum_{j=1}^{K} C_j \phi_j(t) = \boldsymbol{\phi}(\boldsymbol{t})\boldsymbol{C}$ with K known basis functions $\phi_j(t)$ and $C_1, \dots, C_K$ are coefficients to be estimated. ϕ_j are a set of basis functions that are mathematically independent and have the property that they can approximate any function well by taking a linear combination of a sufficient number K of these functions. Commonly used basis functions include the Fourier basis for periodic data and the B-spline basis for non-periodic data [57].

The Fourier basis utilizes basis functions that represent sine and cosine functions of increasing frequency. The Fourier basis expansion for periodic data is given as: $\hat{x}(t) = c_0 + c_1 \sin(\omega t) + c_2 \cos(\omega t) + c_3 \sin(2\omega t) + c_4 \cos(2\omega t) + \cdots$.

The system is defined by basis functions: $\{\phi_1(t) = \sin(\omega t), \phi_2(t) = \cos(\omega t), \dots, \phi_{2k-1}(t) = \sin(k\omega t), \phi_{2k}(t) = \sin(k\omega t)\}$, where $\phi_0(t) = 1$, the constant $\omega = \frac{2\pi}{P}$, and P is defined as the period of the first sine/cosine pair. If the values of t_j are equally spaced on the interval and the period of the function is equal to the length of the interval, then the basis is orthogonal and computing the coefficients becomes easier, especially in situation when the Fast Fourier Transformation (FFT) can be applied. However, newer methods such as B-spline or wavelets can match or exceed this computational efficiency [57].

The B-spline basis utilizes polynomial segments that are joined end-to-end such that the segments are constrained to be smooth where they join. The points at which the

segments join are called knots. Over each segment, a spline is a polynomial of order m (order = degree+1). Polynomials in neighboring segments are required to have matching derivatives up to order m-2 to impose smoothness. The number of basis functions will be uniquely defined by the sum of the B-spline order and the number of interior knots. Without interior knots, the spline becomes a simple polynomial. Note that with increasing order the approximation of the function and its derivatives improve such that by order four, the fit is very good [57]. B-Spline basis functions also have nice computational properties since the inner product matrix of K basis functions is band structured. They are used when data are not known to be periodic.

1.11. FDA FOR METHYLATION DATA

Since the methylation levels are often strongly correlated between CpG sites within a region, it is natural to represent RRBS data as functional data that can be represented as a linear combination of basis functions over a genomic region. This functional representation allows the investigation of dominant modes of variation in the data. One approach for this is functional principal component analysis (FPCA), where statistics are calculated that summarize key features of the functions describing the curve of methylation levels over a defined region and test for differences between conditions. In contrast, most methylation studies to date either use statistics based on single CpG sites or summarize single CpG values within a region of interest. In this work, statistical methods based on FDA are developed for methylation data. Using simulated data, the utilities of FDA in the context of methylation are explored and compared to other methods used in the literature.

1.12. SUMMARY

The main purpose of this dissertation is to develop novel applications of functional data analysis (FDA) procedures for DNA methylation data from RRBS studies. Specifically, these FDA methods will enable testing for differentially methylated region (DMRs) and pinpoint genomic regions that are likely to be biologically meaningful. Testing differentially methylated regions through functional data analysis is described in this dissertation for two papers: (1) Testing differentially methylated regions through functional principal component analysis and (2) Smoothed functional principal component analysis for detecting differentially methylated regions.

The first paper develops functional principal component analysis (FPCA) based on Fourier and B-spline basis functions that successfully tests for differentially methylated regions (DMRs) between two groups (e.g., case and controls) in RRBS data. An empirical comparison, using a simulation based on real data, shows a significant increase in true positive rates for detecting DMRs for the FPCA method in comparison with the M3D approach. The FPCA method also shows considerable robustness with respect to coverage depth and replication number.

The second paper develops a smoothed FPCA (SFPCA) for detecting DMRs by combining a goodness-of-fit detecting with a roughness penalty to maintain the advantages of basis expansion while improving the smoothness. The SFPCA method compares region level differences in the average SFPCA scores between the cases and the controls. The SFPCA scores take into account all information across all CpG sites in a genomic region, capture summary information about dominant modes of variation in the methylation profiles, and improve smoothness of estimated functional principal component curves. In

comparison to the M3D method, the SFPCA technique had substantially higher true positive rates and was robust in relation to coverage depth and replications, using a simulation based on real data.

PAPER

I. TESTING DIFFERENTIALLY METHYLATED REGIONS THROUGH FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

M. Milad and Dr. Gayla R. Olbricht*

Department of Mathematics and Statistics, Missouri University of Science and Technology

*Corresponding Author. OlbrichtG@mst.edu

ABSTRACT

DNA methylation is an epigenetic modification that plays an important role in many biological processes and diseases. Several statistical methods have been proposed to test for DNA methylation differences between conditions at individual cytosine sites, followed by a post hoc aggregation procedure to explore regional differences. While there are benefits to analyzing CpGs individually, there are both biological and statistical reasons to test entire genomic regions for differential methylation. Variability in methylation levels measured by next-generation sequencing (NGS) is often observed across CpG sites in a genomic region. Evaluating meaningful changes in regional level methylation profiles between conditions over noisy site level measurements is often difficult to implement with parametric models. To overcome these limitations, this study develops a nonparametric approach, based on functional principal component analysis (FPCA), to detect predefined differentially methylated regions (DMR). The performance of this approach is compared with an alternative method (M3D), using real and simulated data.

Keywords: functional principal component; epigenomics; DNA methylation; next-generation sequencing

1. INTRODUCTION

DNA methylation is an epigenetic modification involved in gene silencing and tissue differentiation [1]. The role of DNA methylation in human health has been heavily researched in cancer studies as specific methylation patterns are associated with cancer [2]. Methylation can alter the function of genes by adding a methyl (CH_3) group to DNA at cytosine sites [3]. In mammals, DNA methylation almost always occurs when a methyl group attaches to a cytosine (C) when followed by a guanine (G) on the DNA sequence (i.e., CpG dinucleotides) [3]. A number of biological processes in mammals (e.g., the silencing of transposable elements, gene expression regulation, genomic imprinting, and X-chromosome inactivation) involve methylation [4]. Although the methylation of CpG locations in promoter regions is linked to gene silencing, recent research indicates that CpG methylation within genes bodies' correlates with gene expression in a more complex manner [5].

To obtain quantitative methylation data with base pair resolution across the genome, a bisulfite treatment of DNA is followed by next-generation sequencing (NGS). The bisulfite treatment transforms unmethylated cytosine (C) nucleotides into uracils (U), which amplify as thymine (T) during a polymerase chain reaction **(**PCR**)** [6] while methylated cytosines remain unchanged. The bisulfite treated sample is then sequenced via NGS to obtain a library of sequencing reads. After sequencing reads derived from bisulfite treated DNA are aligned to a reference genome, the methylation status of a cytosine in the reference can be assessed by observing the aligned reads that overlap it. This means that when a C in a bisulfite-treated read overlaps a cytosine in the reference, the reference cytosine is methylated for that read [7]. However, if a T in a bisulfite treated read overlaps

cytosine in the reference, then the reference cytosine is unmethylated for that read [7]. This approach can be applied to the whole genome using methylC-seq [8] or BS-seq [9] methods. However, such studies are often costly for organisms with large genome sizes or for case-control studies where large sample sizes are needed. An alternative way to pair bisulfite sequencing with NGS, called reduced representation bisulfite sequencing (RRBS) [10], focuses on capturing an informative subset of the genome. RRBS utilizes restriction enzymes, such as MspI or TaqI, to cleave at CCGG loci so as to select an informative subset of short reads to sequence [7] This process allows for more accurate and specific results, with greater coverage of CpG-dense regions, including promotors, CpG islands, and repetitive sequences. It reduces the numbers of nucleotides to be sequenced to 1% of the genome and thus has a lower cost than sequencing all cytosines genome wide [7]. Many statistical issues are shared between whole genome methods and RRBS, but the following discussion is in the context of RRBS for illustrative purposes.

An essential issue in DNA methylation analysis is identifying genomic loci or regions with varying methylation levels related to distinct biological conditions. The individual loci may be noisy (especially in heterochromatin) but the overall regions tend to be informative [11]. Region-level conclusions are also often more meaningful biologically, making it desirable to consider summarizing information across individual loci in a region. Recently, new statistical methods and software tools have been created to identify differentially methylated regions (DMRs) from RRBS data [12-14]. Most of these methods search for DMRs by first testing each cytosine site, then applying a post hoc aggregation procedure. Post-hoc aggregation reflects the fact that it is unknown which regions are of interest before testing, thus a procedure is needed to control the type I error rate while also

letting the data guide the search for locations of informative regions. One of the first methods developed, BSmooth [15], uses a smoothing process across the genome within each sample to improve the accuracy when estimating the methylation level for any single CpG site. This smoothing process is beneficial since methylation levels of neighboring cytosines are known to be highly correlated [15]. BSmooth distinguishes differentially methylated regions by combining neighboring differentially methylated cytosines (DMCs), which are found using a t-statistic approach, with either a quantile or direct t-statistic cutoff [12]. A majority of the newer methods, such as BiSeq [16] and methylSig [17], also use local smoothing, along with a beta binomial model of methylation at individual cytosine sites; these two methods then combine the results of tests at individual loci to compute a measure of significance for an estimated DMR. Another method called MethylKit [13] uses annotation to provide a statistical test that pools the sequencing reads across an annotated unit (e.g. gene) by group. The MethylKit approach is able to test at both the site level and for predefined regions based on annotation. With multiple samples, a logistic regression with a binary predictor corresponding to condition status is used, which can be expressed as a binomial-based test [13].

In contrast, the MAGI [18] method tests directly for DMRs instead of computing measures of significance for each region based on tests of individual cytosine sites. MAGI assumes that methylation homogeneity exists within a predefined region, so no adjustments are made for spatial correlations between cytosine sites. Methylation levels at each cytosine site for each sample are labeled with a binary representation documenting whether or not they exceed a specified decision boundary, with those exceeding the boundary declared methylated and those falling below declared unmethylated. A Fisher's Exact Test (FET) is

then performed over each predefined region, counting the number of cytosine sites that have changed states between groups [18]. A logistic regression is utilized in place of FET when replicates are available.

A newer alternative approach, M3D [19], relies on the Maximum Mean Methylation Discrepancy (M3D) method to assess changes in the shapes of methylation profiles within the local predefined regions being tested. It applies a machine learning technique called Maximum Mean Discrepancy (MMD) [20] to test the homogeneity in underlying methylation-generating distributions. The method uses a radial basis function (RBF) kernel to construct the MMD between data sets under different conditions in each region being tested; this number is modified based on changes in coverage profiles. The M3D statistics are compared to a null distribution of observed M3D statistics between replicate pairs [19]. It has been suggested that the shape of the methylation profile is a crucial factor in predicting gene expression, supporting the notion of a functional role for the methylation pattern [21]. In a review of the literature, it appears that only the M3D method considers differences in the shape of the methylation profile over the region. The advantage of the M3D method is based on a number of factors. First, the method is sensitive to spatially correlated changes in methylation profiles. Second, the method explicitly accounts for difference in coverage profiles between conditions. Thirdly, the method models inter-replicate variability along the whole genome.

Building on the strengths of M3D, this research explores the use of functional data analysis (FDA) techniques to characterize additional properties of the curve shapes of methylation profiles in genomic regions. Since previous studies have indicated the importance of methylation profile shape in predicting gene expression the FDA techniques

could be advantageous in detecting the profiles that M3D is unable to find. Specifically, in this research, a nonparametric approach based on functional principal component analysis (FPCA) is introduced to detect differential methylation regions (DMRs) from predefined regions, which explicitly accounts for adjusting spatial correlations between cytosine sites. FPCA allows investigation of dominant modes of variation in the RRBS data using the eigenfunctions of the methylation profile covariance function. This method can be employed to test for changes in shape of methylation profiles across regions, as opposed to testing only at individual cytosine sites. This study compares the performance of FPCA to the only other existing method (M3D) that tests for region level shape differences using a simulation based on real RRBS data.

2. METHODS

The computational procedure in this section follows the approach of DE-FPCA [22], an approach that was developed for gene expression studies. DE-FPCA uses functional principal component analysis (FPCA) to decompose gene expression profiles and summarize differences in profiles between groups by using a test statistics based on functional principal component scores that enables differential expression testing. This idea enables finding differences in shape change by representing the expression profile of a gene by a functional curve, called a gene expression function, and this approach is especially powerful in detecting alternative splicing [22]. Although functional data analysis techniques appear to offer many advantages to genomic studies, such methods have not been explored for analyzing DNA methylation data. FPCA has a natural application to aid in solving the issue of testing for differentially methylated regions (DMRs), but the data

collected and defining of regions differs from that of gene expression data and thus the following formulation is needed in the DNA methylation context.

In this study, the methylation profile across CpG sites is decomposed within a predefined region by using functional principal components and calculating the FPC scores to test for DMRs between two groups (e.g., cases and control) of samples. The FPCA scores related to an eigenfunction are computed for all observed methylation profiles in each genomic region and can indicate eigenfunctions with large variation between two groups. The methylation profile function is defined as follows. Let t be the genomic position of CpG site within a predefined genomic region and T be the length of the genomic region being considered. Assume that random samples from two different conditions are collected and sequenced via RBBS. There are n_A case samples and n_B controls samples. Let $x_i(t)$ denote the methylation level for the CpG site at genomic position t for the ith case sample with $y_i(t)$ defined similarly for the ith control sample. Thus $x_i(t)$ and $y_i(t)$ are empirical methylation functions [22].

First, as a brief review of functional principal component analysis (FPCA) [23], consider the following. Let $X(t)$ be a centered, square-integrable function, describing the methylation level of CpG sites over the predefined region. Let ϕ_1,ϕ_2, … be the orthonormal eigenfunctions of the covariance function of $X(t)$. By the Karhunen Loève theorem [23], the centered process in the eigenbasis functions can be expressed as $X(t) = \sum_{k=1}^{\infty} \xi_k \phi_k(t)$, where $\xi_k = \int X(t)\,\phi_k(t)\,dt$ is the principal component coefficient associated with the kth eigenfunction $\phi_k(t)$, with $E(\xi_k) = 0$, $Var(\xi_k) = \lambda_k$, and $E(\xi_k \xi_l) = 0$ for $k \neq l$. The covariance function $R(s,t)$ can be written as $R(s,t) = Cov\big(X(s), X(t)\big) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\ \phi_k(t)$ [22].

The first eigenfunction ϕ_1 represents the principal mode of variation in $X(t)$ in that $\phi_1(t)$ maximizes the variance of $\xi = \int X(t)\,\phi(t)\,dt$ where $Var(\xi) = Var[\int X(t)\,\phi(t)\,dt\,] = \int\int \phi(s)\,R(s,t)\,\phi(t)\,dsdt$ [22]. This $\phi_1(t)$ represents the dominant mode of variation in methylation levels over the region. Similarly, ϕ_k is the function that maximizes $Var(\xi)$ in the functional space that is orthogonal to $\phi_1, \ldots, \phi_{k-1}$.

Using the above information, the eigenfunctions $\{\phi_1, \phi_2, \ldots\}$ should satisfy

$$\int R(s,t)\phi_k(s)ds = \lambda_k\phi_k(t), \qquad (1)$$

where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ for any integer $k \geq 1$. The eigenfunctions $\{\phi_1, \phi_2, ..\}$ can be found by solving equation (1) [22].

2.1. PERFORMING FPCA ON DNA METHYLATION DATA

In the context of this study, FPCA can be performed to find the eigenfunctions and corresponding principal components as follows [22]. Let $X(t) = [X_1(t), X_2(t), \ldots X_N(t)]^T$ be a vector-valued function, with $X_i(t)$ denoting the methylation profile function for the ith sample among N replicates in the predefined region. A set of orthonormal basis functions are selected using either a Fourier or B-spline basis. Note that the Fourier basis is typically used for periodic data, while B-spline basis is used for non-periodic data. Both will be explored since it is unclear which will work best for methylation data. The chosen basis has P functions $\Delta(t) = [\delta_1(t), \delta_2(t), .., \delta_P(t)]$, where it is assumed that the methylation functions $X_1(t), \ldots, X_N(t)$ in the predefined region and eigenfunctions $\{\phi_1, \phi_2, ..\}$ can be expressed as a linear combination of $\delta_1(t), \delta_2(t), \ldots, \delta_P(t)$. Now the methylation profile function can be expressed as, $X(t) = C\Delta(t)$, where the ijth element in the matrix C is $C_{ij} = \int X_i(t)\,\delta_j(t)dt$, with $i = 1, \ldots, N$ and $j = 1, \ldots, P$. Similarly, $\phi(t)$ can be expressed

as $\phi(t) = \Delta^T(t)\beta$, where $\beta = [\beta_1, \dots, \beta_P]^T$ with $\beta_j = \int \phi(t)\delta_j(t)dt$. To find the eigenfunctions ϕ, or equivalently, to determine β, use Equation (1), which has the following equivalent expression [22]:

$$E\left[\begin{pmatrix}\xi_1\\ \vdots\\ \xi_P\end{pmatrix}(\xi_1 \dots \xi_P)\right]\begin{pmatrix}\beta_1\\ \vdots\\ \beta_P\end{pmatrix} = \lambda\begin{pmatrix}\beta_1\\ \vdots\\ \beta_P\end{pmatrix} \quad (2)$$

Replace $E(\xi_i\xi_j)$ with its empirical estimate from sample methylation region functions $X_1(t), \dots, X_N(t)$ to obtain an empirical version of Equation (2):

$$\frac{1}{N} C^T C\beta = \lambda\beta. \quad (3)$$

The eigenfunctions can be found by solving for the above multivariate eigenvalue (λ) and multivariate eigenvector (β). The number of eigenfunctions can be chosen based on the percentage of variance explained. In this study, 90% was used, but this can be modified to allow different function approximation accuracies [22].

2.2. TEST STATISTIC

The pooled empirical methylation profile $x_i(t)$ of cases and $y_i(t)$ of controls was used to estimate the orthonormal principal component function $\phi_j(t)$, $j = 1, \dots, k$ (eigenfunctions), employing the basis expansion method [20]. Let the corresponding principal components associated with $\phi_j(t)$ be ξ_{ij} and η_{ij}, for $x_i(t)$ and $y_i(t)$, respectively. The test statistic was defined using $\xi'_{ij}s$ and $\eta'_{ij}s$ to evaluate the difference in average principal component scores between the case and control samples. Vectors of averages of the functional principal component scores in cases and controls are denoted by $\bar{\xi} = [\bar{\xi}_1, \dots, \bar{\xi}_k]^T$ and $\bar{\eta} = [\bar{\eta}_1, \dots, \bar{\eta}_k]^T$, where $\bar{\xi}_j = \frac{1}{n_A}\sum_{i=1}^{n_A}\xi_{ij}$, and $\bar{\eta}_j = \frac{1}{n_B}\sum_{i=1}^{n_B}\eta_{ij}$, $j = 1, \dots, k$. The pooled covariance matrix *S* is defined as follows:

$S = \frac{1}{n_A+n_B-2}\left(\sum_{i=1}^{n_A}(\xi_i - \bar{\xi})(\xi_i - \bar{\xi})^T + \sum_{i=1}^{n_B}(\eta_i - \bar{\eta})(\eta_i - \bar{\eta})^T\right)$, where $\xi_i = [\xi_{i1}, \dots, \xi_{ik}]^T$, $\eta_i = [\eta_{i1}, \dots, \eta_{ik}]^T$. Let $\Lambda = \left(\frac{1}{n_A} + \frac{1}{n_B}\right) S$. Then, the test statistic is defined as $T^2 = (\bar{\xi} - \bar{\eta})^T \Lambda^{-1} (\bar{\xi} - \bar{\eta})$. Note that this is a form of a Hotelling's T^2 statistic. Under the null hypothesis of no differential methylation between the case and control group in a specific region, T^2 asymptotically follows a central $\chi^2_{(k)}$ distribution, where k is the number of functional principal components. To accurately estimate the p-value, it is best to use a large number of replicates in each group [22]. Note that since the number of regions is determined prior to testing, the false discovery rate can be controlled across the entire set of region level tests.

3. SIMULATION STUDY

3.1. DATA SOURCE

To evaluate the performance of the FPCA method a simulation study based on real RRBS data was performed. Methylation data of bisulfite-sequenced DNA was obtained from 4 patients with acute promyelocytic leukemia (APL) and 12 APL control samples. This data set was obtained under accession number GSE42119 (National Center for Biotechnology Information) [24]. The RRBS data was preprocessed using Bismark version 0.5 (a reference genome alignment tool) that maps bisulfite treated sequencing reads to a genome of interest and performs methylation calls in a single step [25].

3.2. SIMULATION PLAN

To mimic methylation profile changes accurately, a simulation was constructed from the RRBS data set described above following the same approach as in M3D [19]. The regions (CpG clusters) were defined as follows: (1) CpG sites that covered at least 75% of samples were defined as frequently covered CpG sites and (2) a maximum distance of 100 base pairs to the nearest neighbor within a region was accepted. Using these criteria, only regions with at least 20 frequently covered CpG sites were used in the analysis [16]. The simulation study focused on the first 1,000 regions on chromosome 1. Out of the 12 APL control samples in the RRBS data, 4 patients were randomly selected to use in the simulation study as controls. Four more replicates were simulated 100 times to be the testing group (i.e., cases). Of these, 250 of the CpG clusters (predefined regions) were randomly selected to apply differential methylation changes. The replicates that acted as the testing group (cases) were simulated by first adding or subtracting random Poisson ($\lambda = 1$) noise to the total number of reads at each cytosine. Uniform $[-0.1 \text{ to } 0.1]$ random noise was added to cytosine methylation levels. The methylation level L_i, defined as the ratio of methylated reads to total reads mapped to a particular cytosine, was adjusted within the 250 selected, predefined regions [19]. The degree of methylation level change was controlled by the parameter $\alpha \in [0,1]$; new methylation levels were simulated by $L_i^{new} = (1-\alpha)L_i^{old} + \alpha$ when $L_i^{old} \leq 0.5$ for hypermethylation (methylation higher in case than control) and $L_i^{new} = (1-\alpha)L_i^{old}$ when $L_i^{old} > 0.5$ for hypomethylation (methylation lower in case than control) [19].

FPCA and M3D were applied to all 100 simulated data sets under various settings. For FPCA, both Fourier and B-spline basis were investigated. In general, 15-37 knots, with

polynomial order 4, seemed to be a reasonable model for the FPCA-B-spline. To investigate the performance of the methods under different degrees of differential methylation the alpha parameter was varied as $\alpha = \{0.4, 0.6, 0.8, 1\}$. To examine the robustness of the methods for various experimental design features, two different sequencing depths (5 and 20 reads) were simulated and three different replicate numbers per group (3, 8, 12) were simulated. Methods were compared by calculating the average type I and type II error rates across 100 data sets as well as the true positive rate (TPR). The false discovery rate (FDR) was controlled at 0.05 for all analyses [26].

4. RESULTS

4.1. SIMULATION RESULTS

The results using FPCA were compared with the results using M3D. Table 2.1 summarizes the results obtained for different values of the methylation change strength parameter α and different basis expansions, based on an average sequencing depth of 20 reads. The average and standard deviation for the correct number of DMRs is given along with type I and type II errors for each method. Of the 250 truly differentially methylated regions (DMRs), FPCA under the Fourier expansion approach identified 229.85 on average, with 3.93 falsely called DMRs when α =100%. The FPCA under the B-spline expansion approach identified 229.03 true DMRs on average, with 4.08 falsely called DMRs and M3D identified 224.51 true DMRs on average, with no falsely called DMRs at $\alpha = 100\%$.

FPCA–Fourier correctly identified 229.02 DMRs on average, while FPCA-Bspline correctly identified 227.03 at a methylation level difference of 80%, with 3.07 and 3.41

falsely called DMRs on average respectively. M3D called 222.94 true DMRs on average, with no falsely called DMRs. At $\alpha = 60\%$, the FPCA-Fourier and FPCA-B-spline correctly identified 219.82 and 219.05 DMRs on average, respectively, with 2.97 and 3.03 falsely called DMRs on average; whereas M3D called 202.95 correct DMRs on average,with no falsely called DMRs. At $\alpha = 40\%$ the FPCA-Fourier and FPCA-Bspline correctly identified 212.5 and 207.88 DMRs on average, respectively, with 2.47 and 2.46 falsely called DMRs on average; whereas M3D correctly called 190.07 DMRs on average, with no falsely called DMRs.

Table 2.1. Results for Average and Standard Deviation (S.D.) of 100 Simulations Based on FPCA-Fourier, FPCA-BSpline and M3D on Average Sequencing Depth (20 Reads), with Various Levels of Strength of Methylation Change (α)

Alpha	100%			80%			60%			40%		
Methods	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**
Correct	229.85	229.03	224.51	229.02	227.03	222.94	219.82	219.05	202.95	212.5	207.88	190.07
S.D.	0.796	0.784	0.502	0.840	0.809	0.502	0.783	0.845	0.757	0.833	0.794	0.781
# Type-1	3.93	4.08	0	3.07	3.41	0	2.97	3.03	0	2.47	2.46	0
S.D.	0.794	0.977	0	0.877	1.090	0	0.892	0.934	0	0.501	0.900	0
# Type-2	20.15	20.97	25.49	20.98	22.97	27.06	30.18	30.95	47.05	37.5	42.12	59.93
S.D.	0.796	0.780	0.502	0.840	0.809	0502	0.783	0.845	0.757	0.833	0.794	0.781

In conclusion, all methods had a low average type I error rate with the maximum being 0.0054 in FPCA-Bspline when $\alpha = 100\%$. It should be noted that M3D did not produce any type I errors, making it the most conservative of the methods but at the sacrifice of higher type II errors (i.e., lower power). M3D had higher type II errors across

all values of α than both FPCA methods. Results were similar for FPCA-Fourier and FPCA-Bspline, with FPCA-Fourier having slightly lower type II errors and thus giving it a slight advantage. Across all methods, there were fewer type II errors as α increased from 40% to 100%, which is expected since it is easier to detect more extreme differences. However, it is notable that for small α values, α = 0.40, 0.60, there are more extreme differences between M3D and the FPCA methods. This indicates FPCA can improve DMR detection in more difficult situations when "the signal" is low.

In contrast, Table 2.2 displays the results based on an average sequencing depth of 5 reads. At methylation strength 100%, FPCA-Fourier and FPCA-B-spline called 223.88 and 222.5 DMRs on average, respectively, out of the 250 true DMRs, with 5.97 and 6.2 false DMRs. M3D called only 200.04 true DMRs on average, with no false DMRs. The number of truly identified DMRs decreased using FPCA-Fourier, FPCA-Bspline and M3D, when decreasing the strength of methylation change from α = 80% to 40%, as was also observed in Table 2.2.

Table 2.2 Results for Average and Standard Deviation (S.D.) of 100 Simulations Based on FPCA-Fourier, FPCA-B-Spline, and M3D on Average Sequencing Depth (5 Reads), with Various Levels of Strength of Methylation Change (α)

Alpha	100%			80%			60%			40%		
Methods	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**	**FPCA-Fourier**	**FPCA-Bspline**	**M3D**
# Correct	223.88	222.5	200.04	219.15	219.08	197.13	202.06	211.09	178.00	197.93	200.02	170.05
S.D.	0.819	0.885	0.815	0.832	0.812	0.824	0.826	0.829	0.804	0.843	0.791	0.808
# Type-1	5.97	6.2	0	6.05	6.15	0	3.94	3.71	0	4.54	5.07	0
S.D.	1.041	1.470	0	1.426	1.431	0	0.887	1.112	0	0.8946	1.029	0
# Type-2	26.12	27.5	49.96	30.85	30.92	52.87	47.94	38.91	72.00	52.07	49.98	79.95
S.D.	0.819	0.885	0.8155	0.821	0.812	0.824	0.826	0.829	0.804	0.843	0.791	0.808

Overall, similar trends were observed at 5 reads as for 20 reads, except that FPCA-Bspline had slightly lower type II errors than FPCA-Fourier at $\alpha = 0.4, 0.60$ when the sequencing depth was 5 reads (this was reversed for 20 reads). Also, all the methods had higher type II errors at 5 reads than 20 reads while still maintaining a low type I error rate on average. It should be noted that when the sequence depth is 5 reads there are more drastic differences between the FPCA and M3D methods even for the largest α = 100% (i.e., across all α).

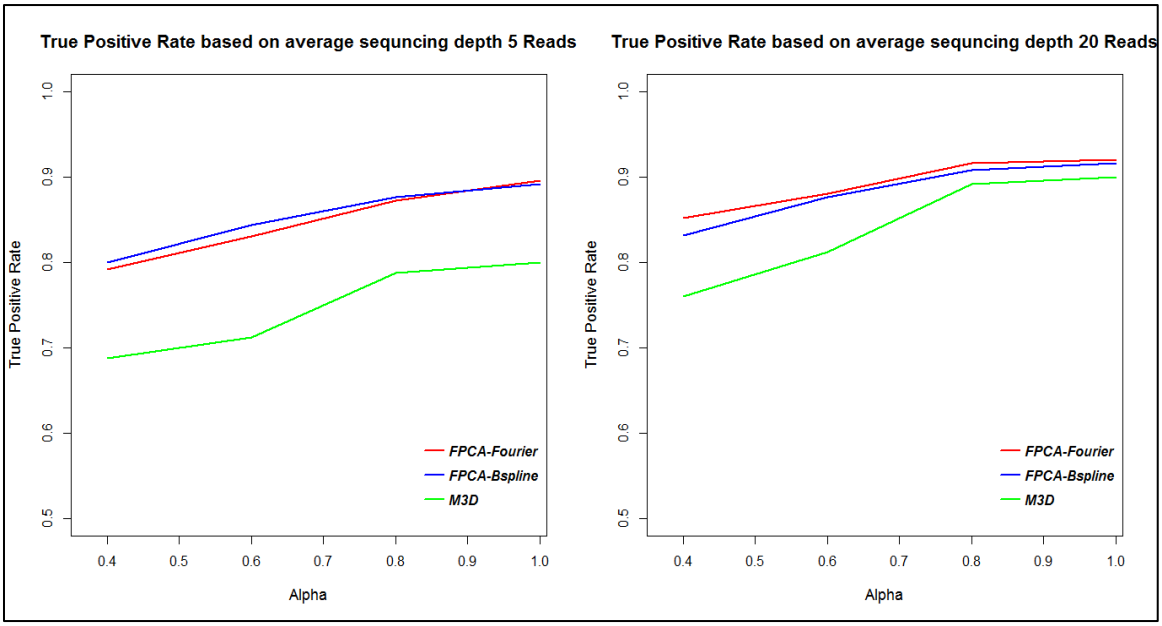

Figure 2.1. True Positive Rates Based on the Average over 100 Simulations on Average Sequencing Depths of 5 (left graph) and 20 (right graph) Reads verses α Level for Controlling the Degree of Differential Methylation for Each of Three Methods: FPCA (Fourier Expansion Approach) – Blue, FPCA (B-Spline Expansion Approach) – Red, M3D – Green.

Figure 2.1 shows the average true positive rates (TPRs) over the 100 simulated data sets for varying degrees of differential methylation (α values) for each of the three methods (FPCA-Fourier, FPCA-Bspline, and M3D) and two coverage depth (5 and 20

reads). The FPCA-Fourier method had the highest average TPR with an average sequencing depth of 20 reads across all α values, with FPCA-Bspline yielding similar but slightly lower TPRs. However, in an average sequencing depth of 5 reads, FPCA-Fourier had the highest average TPR only when α = 80%, 100%. For the lower levels of differential methylation strength $\alpha = 0.4, 0.60$, the FPCA-B-spline had the highest TPR. Overall, FPCA-Fourier and FPCA-B-spline substantially outperformed M3D with regard to TPR in both average sequencing depths (5 and 20 reads), across all levels of differential methylation strength. The coverage is also important to investigate, since low coverage can lead to less stable methylation estimates and prevent statistical significance while high coverage costs more to obtain. Figure 2.1 shows that the sequencing depth of 20 has the highest average TPR compared to average sequencing depth of 5. However, this difference is more drastic for M3D than it is for the FPCA methods. The FPCA methods maintain an average TPR between 79% and 90% for a depth of 5 reads; whereas the M3D TPR ranges from 68% to 80%.

4.2. ROBUSTNESS IN REPLICATIONS

To examine the robustness of the FPCA method to changes in replication number, simulated data sets were created for differing numbers of replicates per group, using the same approach as described as in section 3.2. Control samples from real RRBS data set were used as the control groups for 3, 8 and 12 replicates per group. This was possible since the data set contained 12 control samples. A set of 3, 8, or 12 replicates were simulated as previously described to act as the cases groups. As before, the same 250 regions were simulated to be true DMRs using $\alpha = 80\%$ and a coverage of 20 reads. The

FPCA-Fourier basis function method was used to identify DMRs with 3, 8 and 12 replicates per group since this method performed best for 20 reads and these results were compared. The false discovery rate (FDR) was controlled at 5%. The FPCA-Fourier method identified 179, 193, and 216 true DMRs out of the total of 250, with 3, 8, and 12 replicates per group, respectively.

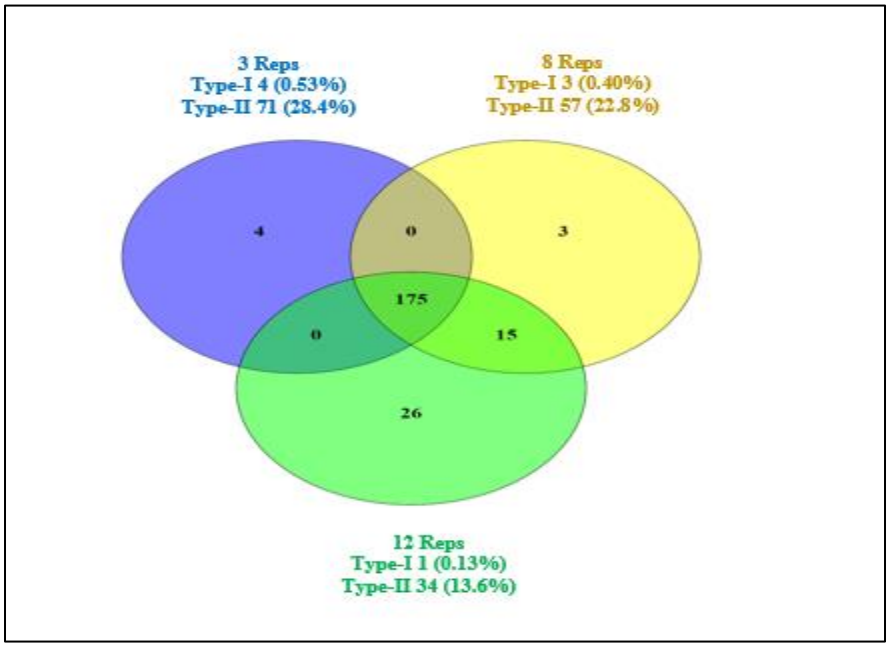

Figure 2.2. Venn Diagram of True DMRs Detected with FPCA-Fourier, for 3, 8, and 12 Replicates Per Group. The Number and Percentage of Type I and Type II Errors is also Given for Each Replicate Number.

As shown in Figure 2.2, the overlap between the three sets of true DMRs identified accounts for 70% of the total. As was expected, the testing lost power with lower replication, with 12 replicates per group identifying the most unique true DMRs and having the lowest number of type II errors, and the highest number of type II errors occurred for three replicates per group. More similarity was observed between the simulations with eight and 12 replicates as they shared 15 true DMRs uniquely, whereas the simulation with

three replicates had no unique overlap with eight or 12 replicates. Overall, the type II error rates ranged from 13.6% in the 12 replicate cases to 28.4% in the three replicate cases. Type I error was low for all three cases with the lowest being 0.13% for 12 replicates and the highest being 0.53% for three replicates. This shows that while more replicates are better, the FPCA- Fourier method exhibits a reasonable amount of robustness to smaller replicate number per group.

4.3. DMRS DETECTED IN REAL DATA

An analysis was completed using the real RRBS data described in section 3.1 with four samples from bone marrow patients with acute promyelocytic leukemia (APL) and four control samples (APL in remission). All CpG sites (with at least 20 reads) across all samples were used, including all regions with start and stop locations defined as described in section 3.2. Since this data set provided a coverage of at least 20 reads, the FPCA-Fourier method was applied since it performed the best under that setting and these results were compared to M3D.

Out of 14,000 CpG regions selected for testing, FPCA-Fourier identified 3897 DMRs and M3D identified 2603 DMRs total, with 1488 DMRs in common. Figure 2.3 confirms that the FPCA-Fourier method identified a clear group of changed profiles between the two conditions in the real data sets. The false discovery rate (FDR) was controlled at 0.05 for all analyses [26].

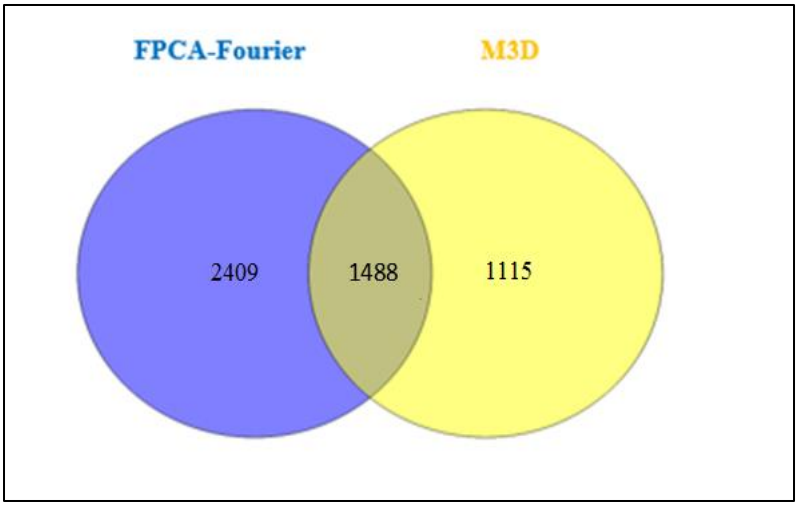

Figure 2.3. Venn Diagram Comparing the Number of Significant Differentially Methylated Regions (DMRs) Identified by the FPCA-Fourier and M3D Methods in the Real APL RRBS Data Set.

5. CONCLUSION

This research demonstrates that information from reduced representation bisulfite sequencing (RRBS) datasets can be analyzed using higher-order mathematics, specifically a functional data analysis approach. Here, a dimension reduction approach is presented, based on the Karhunen-Loève transform, to create a hypothesis test for differential methylated regions (DMRs) using functional principal components based on the spatial features of methylation profiles. This allows the investigation of dominant modes of variation in the methylation profile over a region using eigenfunctions of the covariance function. The FPCA in this study employs a few principal components that increase the power and reduce degrees of freedom in testing to make the underlying biological signals stable. An FPCA based on Fourier and B-spline functions was developed that successfully detects information from shapes of the methylation curves that cannot be identified by traditional multivariate statistics and tests for differentially methylated regions between

case and control groups. An empirical comparison, using a simulation based on real data, showed a substantial increase in the true positive rate for FPCA in comparison with the M3D approach [19], as well as considerable robustness with respect to coverage depth and replication. In general, the simulation results were similar for FPCA-Fourier and FPCA-Bspline, with FPCA- Fourier having slightly lower type II errors across most of the simulation settings thus giving it a slight advantage.

The good performance of the FPCA method is attributable to a number of factors. First, the method takes spatial correlation into account in analyzing the methylation profile. Second, the FPCA translates high-dimensional DNA methylation data into a few principal components, which greatly reduces the degrees of freedom in testing, while preserving most of the underlying biological signals. In contrast M3D, does not perform well in high dimensional DNA methylation data within a region.

The methodology proposed and illustrated here builds on the interpretation of next-generation sequencing data. The FPCA method can be applied in cancer research as well as in the pursuit of therapies to combat or prevent lupus, muscular dystrophy, and other diseases. In fact, because hypermethylation occurs early in colon cancer, detection of hypermethylation could be an important indicator of potential health problems, which might be detected using the FPCA method. In addition, future studies are needed to investigate the use of other functional data analysis techniques, such as functional linear regression or functional canonical correlation analysis, as well as incorporating smoothing penalties into the analysis. Although the FPCA framework was investigated using RRBS data, it should scale up well for utilization in whole genome bisulfite sequencing studies, but this should be investigated more fully. Finally, although the FPCA method exhibited

robustness in detecting DMRs under 1ow coverage and replications in two groups, it is of interest to extend the method to work for experiments that require testing for differences between more than two groups or that have covariate information.

REFERENCES

1. Phillips, T. The role of methylation in gene expression. Nature Ed. 2008;1(1):116.

2. Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, et al. 5′ CpG island methylation is associated with transcriptional silencing of the tumor suppressor p16/CDKN2/MTS1 in human cancers. Nature Med. 1995;1(7):686–692.

3. Lim DHK, Maher ER. DNA methylation: a form of epigenetic control of gene expression. TOG. 2010;12(1):37–42.

4. Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol. 2014;6:a019133.

5. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23 (3):555–567.

6. Krueger F, Kreck B, Franke A, Andrews SR. DNA methylome analysis using short bisulfite sequencing data. Nature Methods. 2011;9(2):145–151.

7. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6(4):468–481.

8. Urich MA, Nery, JR, Lister R, Schmitz, RJ, Ecker, JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature Protoc. 2015;10(3):475–483.

9. Hebestreit K, Dugas M, Klein H-U. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013;29(13):1647–1653.

10. Meissner A, Gnirke A, Bell, GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18):5868–5877.

11. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen, KD, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–1369. doi:10.1093/Bioinformatics/btu049.

12. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13:R83.

13. Akalin A, Garrett-Bakelman FE, Kormaksson M, Busuttil J, Zhang L, Khrebtukova I, et al. Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. PLoS Genet. 2012;8:e1002781.

14. Benoukraf T, Wongphayak S, Hadi LH, Wu M, Soong R. GBSA: a comprehensive software for analyzing whole genome bisulfite sequencing data. Nucleic Acids Res. 2013;41:e55.

15. Hansen, KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83.

16. Hebestreit K, Dugas M, Klein H. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013;29(13):1647–1653.

17. Park Y, Figueroa ME, Rozek LS, Sartor MA. MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics. 2014;30(17):2414–2422. doi: 10.1093/bioinformatics/btu339.

18. Baumann D, Doerge R. MAGI: Methylation analysis using genome information. Epigenetics. 2014;9(5):698–703.

19. Mayo T, Schweikert G, Sanguinetti G. M3D: a kernel-based test for spatial correlated changes in methylation profiles. Bioinformatics. 2015;31(6):809–816. doi: 10.1093/bioinformatics/btu749.

20. Gretton A, Borgwardt KM, Rasch M, Scholkopf B, Smola AJ. A kernel method for the two-sample-problem. Adv Neural Inf Process Syst. 2007;19:513.

21. VanderKraats ND, Hiken JF, Decker KF, Edwards JR. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013;41(14):6816–6827.

22. Xiong H, Brown JB, Boley N, Bickel PJ, Huang Het. DE-FPCA: testing gene differential expression and exon usage through functional principal component analysis. In: Datta S, Nettleton D, editors. Statistical analysis of next generation sequencing data. Switzerland: Springer International Publishing; 2014. p. 129–143.

23. Ramsay JO, Silverman BW. Functional data analysis. New York: Springer; 2005.

24. Schoofs T, Rohde C, Hebestreit K, Klein H-U, Göllner S, Schulze I, et al. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. Blood®. 2013;121(1):178–187.

25. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics. 2011;27(11):1571–1572.

26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995:289–300.

II. SMOOTHED FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS FOR DETECTING DIFFERENTIALLY METHYLATED REGIONS

M. Milad and Dr. Gayla R. Olbricht*

Department of Mathematics and Statistics, Missouri University of Science and Technology

*Corresponding Author. OlbrichtG@mst.edu

ABSTRACT

DNA methylation is a key, heritable, epigenetic modification that can alter gene expression without a DNA sequence change. Most instances of DNA methylation in mammals take place when a methyl group attaches to a cytosine when followed by a guanine (CpG dinucleotides) on the DNA sequence. DNA methylation can be measured throughout the genome at individual cytosine sites by combining bisulfite sequencing with next-generation sequencing (NGS). Although the measurements are taken at the site level, researchers are often interested in testing for methylation differences over genomic regions. Although DNA methylation has been well researched, little statistical research has been conducted to develop methods that will discover epigenomic structural variations using NGS data to identify predefined differentially methylated regions (DMRs). This study addresses this critical gap in the literature, by creating a new strategy that evaluates predefined methylation regions (DMRs) using smoothed functional principal component analysis (SFPCA). This study compares the performance of SFPCA to FPCA without smoothing and to an existing method, M3D, using real and simulated data.

1. INTRODUCTION

DNA methylation has been researched in depth because it is an important, heritable, epigenetic modification that can alter gene expression without changing the DNA sequence. In mammals, DNA methylation is almost always the result of a methyl (CH_3) group attaching to a cytosine when followed by a guanine (CpG dinucleotides) on the DNA sequence. Methylation can modify the way genes function after a methyl group has been added to the DNA. DNA methylation is one of the best characterized epigenetic modifications and its connection to human health has been heavily studied but is not yet fully understood [1]. In mammals, it is involved in various biological processes including the silencing of transposable elements, regulation of gene expression, genomic imprinting, and X-chromosome inactivation [1]. The methylation of CpG locations in promoter regions is often associated with gene silencing; however, recent research suggests that the correlation between CpG methylation with gene bodies and gene expression is more complex [2].

The most thorough method for measuring DNA methylation is bisulfite sequencing combined with next-generation sequencing (NGS), which has the advantage of quantifying single-base cytosine methylation levels across the entire genome [25]. In bisulfite sequencing, DNA is treated with sodium bisulfite, which converts unmethylated cytosine residues to uracil, but which does not affect the 5` methylcytosine residues. After PCR amplification, the uracils are converted to thymines, thus enabling a distinguishing between methylated and unmethylated cytosines. Bisulfite converted DNA fragments are then sequenced via NGS and aligned to a reference genome. The percentage of methylation at each cytosine position is found by calculating the ratio (C/C+T), that is the number of methylated reads (C) divided by the total of all methylated and unmethylated reads (C+T). Several

techniques can be used for high-throughput bisulfite sequencing, including reduced bisulfite sequencing (RRBS) [3], whole-genome shotgun bisulfite sequencing methods (BS-seq, methylC-seq) [4, 5], and target capture bisulfite sequencing[7]. Although the whole genome methods such as BS-seq and MethylC-seq provide the most complete information, cost is still a limitation for many studies. Alternatively, RRBS allows for the use of restriction enzymes, such as MspI or TaqI, to cleave at CCGG loci so as to choose an informative set of short reads to sequence [8]. This process provides more accurate and specific results within specific subsets of the genome, with greater coverage of CpG-dense regions, and is less expensive than sequencing all cytosines genome wide.

A major problem in computational epigenomics is that epigenetic signals are poorly understood. However, new statistical methods and software tools that identify differentially methylated sites and regions (DMRs) [8-10] have been developed recently to aid in understanding these complex data. Although many initial methods focused on testing individual cytosine sites, there are biological and statistical benefits of testing regions instead of sites. While the site level data may be noisy, the overall regions tend to be more informative and there are fewer of them to test, easing the burden of the multiple testing problem. Additional advantages of using DMRs are that although differences at any given site may be small, variations across a region can be detected more easily due to the high correlation neighboring sites [11] and DMRs potentially have a greater ability to predict phenotypes.

Region level testing methods can be categorized into those that operate on predefined regions verses those that define the region after site level testing has taken place, and thus cannot be defined in advance. Both methods need to control the false discovery

rate (FDR) at the region-level, but when the number of regions is not determined prior to testing this task is non-trivial, making it more difficult to control for multiple test for such methods. For example, when the regions are undefined before testing, it is impossible to extrapolate the region-level FDR control from the site-level tests in the region. As such, the best approach is to use predefined regions, defined based on annotation regions (e.g., CpG islands, CpG shores, introns, and exons), or defined based on a non-annotation criteria (Figure 3.1) [12]. Often, non-annotation regions are defined based on locating regions with a certain minimum CpG density within a specific genomic window.

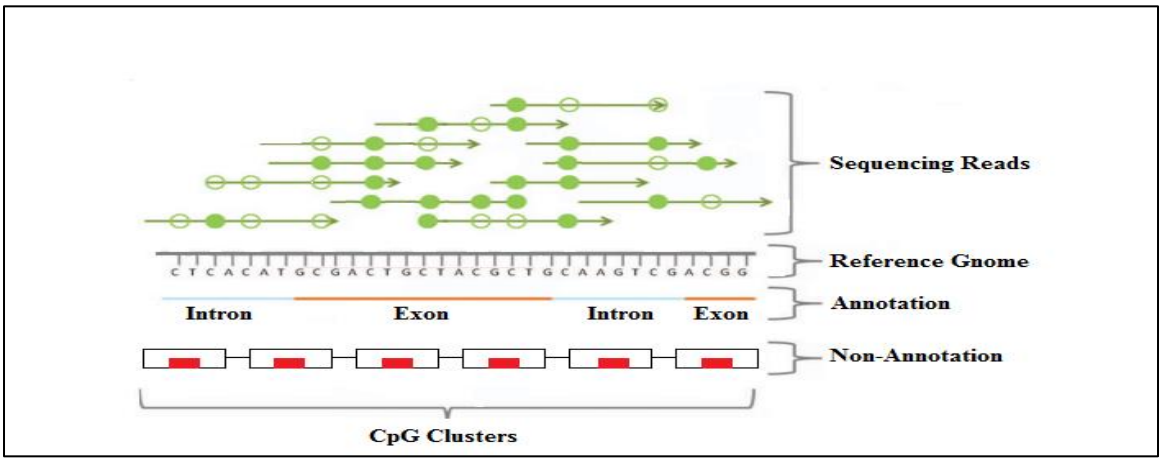

Figure 3.1. Illustration of Predefined Regions Based on Annotation and Non-Annotation Profiles. Image modified from Baumann and Doerge [12].

Both types of predefined regions have pros and cons. Annotation based regions may have a direct biological meaning but if the CpG sites are sparse or separated by more than 1000 bp, the known correlation in methylation levels between neighboring sites may be diminished for different parts of the region. When regions are defined based on CpG density within a certain neighborhood this problem is alleviated, at the sacrifice of

potentially less biologically meaningful regions. This can be somewhat overcome by determining which annotation units overlaps with the defined region.

Many approaches can be employed to define DMRs. Many region-level methods first test each cytosine site to search for DMRs, then follow the site level results with post-hoc aggregation. These are the methods mentioned previously where the regions being tested are not known in advance, and a method is needed to control the type I error rate while also letting the data determine where to look. BSmooth [8], which is a widely cited method, employs a smoothing process of methylation levels across the genome for each sample, which improves the accuracy of the methylation level estimate for any single CpG site. To discover DMRs, BSmooth combines individual cytosines with ranked (significant) differentially methylated cytosines (DMCs), which are found using t-statistics or a linear model, with a quantile or direct t-statistic cutoff [8]. Most of the newer approaches (e.g., BiSeq [13] and methylSig[14]) use local smoothing, with a beta binomial model of methylation levels at individual cytosine sites. Both BiSeq and methylSig aggregate the results of tests at discrete loci when computing a measure of significance for estimating DMRs.

Among methods that use predefined regions, MethylKit uses annotation to provide a statistical test that pools the sequencing reads across an annotated unit (e.g., gene) by group. When using multiple samples, a logistic regression with a binary predictor corresponding to the condition is applied, which can be expressed as a binomial-based test [10]. Thus methylKit still relies on post-hoc aggregation of site-level tests but the regions where aggregation takes place are known in advance due to annotation information. In contrast, the methylation analysis using genome information (MAGI) tests directly for DMRs across annotation units rather than computing measures of significance for each

region based on an examination of individual cytosine sites. This difference in methods results from the assumption, under MAGI [12], that the regions are homogeneous in terms of methylation and require no adjustments for spatial correlations between cytosine sites. Methylation levels at each cytosine site are labeled with a binary representation showing whether or not they exceed a specified decision boundary. A Fisher's Exact Test (FET) for unreplicated experiments or a logistic regression when replicates are available is performed over each region, which counts the number of cytosine sites that have changed states [12].

An alternative method, M3D [15], relies on the Maximum Mean Methylation Discrepancy (MMD) method to assess changes in the shapes of methylation profiles within the local predefined regions being tested. Regions are defined in M3D based on CpG density rather than annotation. M3D applies a machine learning technique (MMD) [23] to test the homogeneity in underlying methylation-generating distributions. The method uses a radial basis function (RBF) kernel function to construct the MMD between data sets in each region being tested and this number is modified based on changes in coverage profiles. The M3D statistics are compared to a null distribution of observed M3D statistics between replicate pairs [15]. It has been suggested that the shape of the methylation profile is a crucial factor in predicting gene expression, supporting the notion of a functional role for the methylation pattern [23]. This is one of the advantages of M3D's idea of looking at the differences in shape change over a predefined region in methylation profile. In a review of the literature, it appears that only the M3D method utilizes the shape of the methylation profile over the region (Figure 3.2).

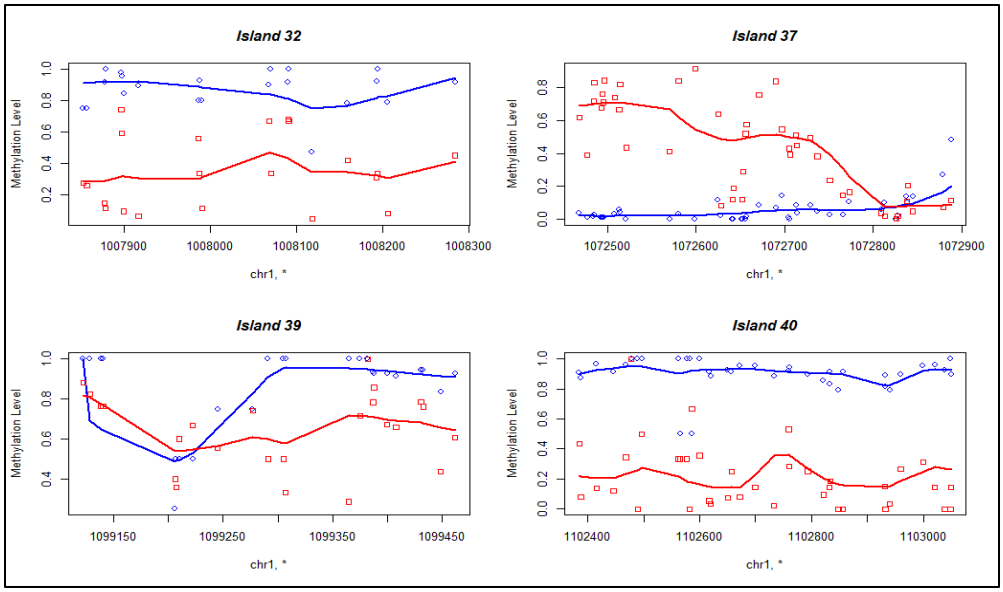

Figure 3.2. Methylation Profiles of Predefined Regions Identified by the M3D Method in a Comparison of Leukemia and Human Embryonic Stem Cells (ESC). Figure from Mayo et al. [15].

Although M3D offers certain advantages, there may be more information that can be captured about the shape of the methylation profile over a region than is used in M3D. In the previous paper (Section 2), a functional principal component analysis (FPCA) approach was proposed to capture dominant modes of variation in the methylation level across a region. FPCA was shown to greatly improve power to identify DMRs over M3D, indicating the benefit of considering additional aspects of the curve shape beyond those used in M3D. However, the observed methylation profiles are typically not smooth, which lead to substantial variability in the estimated functional principal component curves. When the epigenetic methylation function changes rapidly within the genomic region, the basis expansion in the FPCA may not provide a good estimate of the genetic variation, thus potentially decreasing the power of FPCA. In this study, this limitation was overcome by developing a smoothed FPCA (SFPCA) for testing DMRs by combining a goodness-of-fit measure with a roughness penalty on the functional principal component weight functions

to maintain the advantages of basis expansion. SFPCA explicitly accounts for adjusting spatial correlations between cytosine sites. Rather than only testing at individual cytosine sites, this method can find changes in methylation profiles across predefined regions based on CpG density. This study compares the performance of SFPCA to FPCA without smoothing and to the existing method (M3D) for shape changes in predefined regions, using real and simulated data.

2. METHODS

2.1 SMOOTHED FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

In this section, a smoothed functional principal component analysis (SFPCA) is developed for testing for region-level differential methylation. SFPCA has been shown to be beneficial in other areas of genomics, specifically in association studies where it can be used for testing across the entire allelic spectrum (rare and common) of genetic variation [18]. One goal of this paper was to employ SFPCA for a group test across a genomic region. In this approach, smoothed functional principal component scores take information across all variants in the genomic region into account, hence including all single variant variation while constructing a region level test [18]. This SFPCA method could be highly beneficial for DNA methylation data, yet no functional data analysis techniques have currently been developed for region level differential methylation detection. Here, an SFPCA method is developed for summarizing the methylation profile in region.

DNA methylation levels are often strongly spatially correlated between neighboring CpG sites [28]. The SFPCA developed below serves as dimension reduction approach of the finite methylation profile using the Karhunen-Lo*é*ve transform to show the

variability existing between RRBS datasets under differing conditions with respect to a defined region of interest [18]. This allows the investigation of dominant modes of variation in the data using the eigenfunctions of the methylation profile covariance function. Let t be a genomic position of a cytosine site within a genomic region, and let T be the length of the genomic region under consideration. If the CpG density of the genomic region is high, the region $[0, \mathrm{T}]$ can be rescaled to $[0, 1]$[18], where t is a continuous variable over the interval $[0,1]$. Assume that RRBS data are collected on samples in two conditions (e.g., cases and control) with n_A case samples and n_B control samples.

Let $X(t)$ be the centered, epigenetic methylation function to describe each region. In this study, the epigenetic methylation function represents methylation levels of CpG sites described over a predefined region. When using functional principal component analysis (FPCA) [16], the variation in the epigenetic methylation function can be expressed with a linear combination of the functional values:

$$f = \int_0^1 \xi(t)X(t)dt \qquad (1)$$

where $\xi(t)$ is a weight function. The functional principal components can be found by finding the weight function $\xi(t)$ that maximizes the variance of f [16]:

$$var(f) = \int_0^1 \int_0^1 \xi(s)R(s,t)\,\xi(t)dsdt \qquad (2)$$

where $R(s,t)$ describes the covariance function of each epigenetic methylation function for each predefined region. The methylation profiles are not normally smooth, which causes there to be considerable variability when estimating the functional principal component curves. A roughness penalty is used in combination with the functional principal component weight functions to aid with smoothness of the functional principal

component [18]. In this study, the roughness penalty on the functional principal component weight functions utilizes the integrated, squared second derivative.

The smoothed functional principal components can be found by solving the following integral equation [18]:

$$\int_0^1 R(s,t)\,\xi(s)ds = \rho[\xi(t) + \lambda \parallel D^2\xi(t) \parallel^2] \quad (3)$$

Where λ is a smoothing parameter that balances the function roughness and the fit. To reduce the SFPCA to an unsmoothed FPCA, set $\lambda = 0$.

2.2. COMPUTATION FOR SFPCA

The principal component function is an eigenfunction that is an integral function and is difficult to solve in closed form. To solve for the eigenfunction in Eq. (3), first convert the continuous eigenanalysis to an appropriate discrete eigenanalysis [16]. To obtain this conversion, Fourier basis function methods can be used [18].

Let $\delta_j(t)$ be a series of Fourier basis functions. For each j, define $\omega_{2j-1} = \omega_{2j} = 2\pi j$. Then, expand the epigenetic methylation function as a linear combination of the basis function δ_j:

$$X(t) = \sum_{j=1}^{T} C_j\delta_j(t) \quad (4)$$

Let $X(t) = [X_1(t), \dots\dots\dots, X_N(t)]^T$ be a vector-valued function, with $X_i(t)$ be a centered, square-integrable function, in this case describing the methylation level of the t^{th} CpG site in the predefined region for ith sample for N replicates. Then, select an orthonormal Fourier basis with T functions $\delta(t) = [\delta_1(t), \delta_2(t), \dots, \delta_T(t)]^T$. The joint expansion of the N methylation profiles can then be expressed as follows [18]:

$$X(t) = C\delta(t) \quad (5)$$

where C is a coefficient matrix and the covariance function of the methylation profiles can be represented as

$$R(s,t) = \frac{1}{N}\delta^T(s)C^T C\delta(t). \quad (6)$$

Also, the eigenfunction can be written as

$$\boldsymbol{\xi(t)} = \sum_{j=1}^{T} \boldsymbol{b_j}\,\boldsymbol{\delta_j(t)} \text{ and } \boldsymbol{D^4\xi(t)} = \sum_{j=1}^{T} \boldsymbol{\omega_j^4 b_j}\,\boldsymbol{\delta_j(t)} \text{ or}$$

$$\boldsymbol{\xi(t)} = \boldsymbol{\delta(t)^T b} \text{ and } \boldsymbol{D^4\xi(t)} = \boldsymbol{\delta(t)^T v_0 b} \quad (7)$$

where $\boldsymbol{b} = [\boldsymbol{b_1}, \ldots, \boldsymbol{b_T}]^T$ and $\boldsymbol{v_0} = \boldsymbol{diag}(\boldsymbol{\omega_1^4}, \ldots, \boldsymbol{\omega_T^4})$.

The right term can be expanded in Eq. 3 as

$$\boldsymbol{\xi(t)} + \boldsymbol{\lambda} \parallel \boldsymbol{D^2\xi(t)} \parallel^2 = \boldsymbol{\delta(t)^T v^{-2} b}. \quad (8)$$

where $\boldsymbol{v} = \boldsymbol{diag}((\boldsymbol{1} + \boldsymbol{\lambda\omega_1^4})^{-\frac{1}{2}}, \ldots, (\boldsymbol{1} + \boldsymbol{\lambda\omega_T^4})^{-\frac{1}{2}})$.

Substituting Equations 6 and 7 for $\boldsymbol{R(s,t)}$ and $\boldsymbol{\xi(t)}$ into the eigenequation (Eq. 3), results in the following [18]:

$$\frac{\boldsymbol{1}}{\boldsymbol{N}}\,\boldsymbol{C^T C\,b} = \boldsymbol{\rho}\,\boldsymbol{v^{-2} b}. \quad (9)$$

which can be written as

$$\left[\boldsymbol{v}\left(\frac{\boldsymbol{1}}{\boldsymbol{N}}\boldsymbol{C^T C}\right)\boldsymbol{v}\right][\boldsymbol{v^{-1}b}] = \boldsymbol{\rho}\,[\boldsymbol{v^{-1}b}] \text{ , or } \boldsymbol{v}\left(\frac{\boldsymbol{1}}{\boldsymbol{N}}\boldsymbol{C^T C}\right)\boldsymbol{vu} = \boldsymbol{\rho u} \quad (10)$$

where $\boldsymbol{u} = \boldsymbol{v^{-1}b}$. Therefore, $\boldsymbol{b} = \boldsymbol{vu}$ and $\boldsymbol{\xi(t)} = \boldsymbol{\delta(t)^T b}$ is a solution to eigenequation (Eq. 3) [18].

2.3. TEST STATISTIC

To estimate the set of orthonormal principal component functions $\xi_j(t), j = 1,2, \ldots k$ (eigenfunctions), let the pooled methylation profiles $x_i(t)$ denote the methylation level for the CpG site at genomic position t for ith case sample. Similarly define $y_i(t)$ for

the ith control samples. By the Karhunen-Lo$\acute{e}$ve decomposition [16], the smoothed functional principal component score can be obtained by $\beta_{ij} =< x_i(t), \xi_j(t) >$ and $\eta_{ij} =< y_i(t), \xi_j(t) >$, where $j = 1,2, \dots k$. Let the average vectors of the functional principal component scores in the cases and controls be $\bar{\beta} = [\bar{\beta}_1, \dots., \bar{\beta}_k]$ and $\bar{\eta} = [\bar{\eta}_1, \dots., \bar{\eta}_k]$. The pooled covariance matrix can be defined as $S = \frac{1}{n_A+n_B-2}\left[\left(\sum_{i=1}^{n_A}(\beta_i - \bar{\beta})(\beta_i - \bar{\beta})^T + \sum_{i=1}^{n_B}(\eta_i - \bar{\eta})(\eta_i - \bar{\eta})^T\right)\right]$, where $\beta_i = [\beta_{i1}, \dots, \beta_{ik}]^T$,$\eta_i = [\eta_{i1}, \dots, \eta_{ik}]^T$. Let $\Lambda = \left(\frac{1}{n_A} + \frac{1}{n_B}\right) S$. Then, the Hotelling T^2 statistic test is defined as $T^2 = (\bar{\beta} - \bar{\eta})^T \Lambda^{-1} (\bar{\beta} - \bar{\eta})$. Under the null hypothesis of no differential methylation in the region between the case and control group, T^2 asymptotically follows a central $\chi^2_{(k)}$ distribution, where k equals the number of functional principal components. For the most accurate estimate of the p-value, a large number of replicates in each treatment group should be used [17]. The false discovery rate can be controlled across all of these region level tests.

3. SIMULATION STUDY

3.1. DATA SET

To evaluate the performance of the SFPCA method, a simulation study based on real RRBS data was performed. Methylation data of bisulfite-sequenced DNA was obtained from 4 patients with acute promyelocytic leukemia (APL) and 12 APL control samples [20]. This data set was obtained under accession number GSE42119 (National Center for Biotechnology Information) [21]. The RRBS data were preprocessed using Bismark version 0.5; a reference genome alignment tool that maps bisulfite treated sequencing reads to a genome of interest and performs methylation calls in a single step [25].

3.2. SIMULATION PLAN

Using the simulation approach in M3D [1], and employing actual RRBS data, a simulation was created to accurately imitate methylation profile changes. The regions (CpG clusters) were defined as follows: (1) CpG sites that covered at least 75% of samples were defined as frequently covered CpG sites and (2) a maximum distance of 100 base pairs to the nearest neighbor within a region was accepted. Only regions with at least 20 frequently covered CpG sites were used in the analysis[18]. This investigation looked at the first 1,000 regions on chromosome 1. Four biological replicates based on the controls in the APL RRBS data set described above were randomly chosen out of the 12 as the control group. Four replicates were simulated 100 times to be the case group. Differential methylation changes for the case group were applied to 250 randomly chosen CpG clusters (predefined regions). To create the case group, data for the replicates were simulated by first adding or subtracting random Poisson ($\lambda = 1$) noise to the total number of reads at each cytosine. Uniform [-0.1 to 0.1] random noise was added to cytosine methylation levels. The methylation level L_i, which was defined as the ratio of methylated reads to the total reads mapped to an individual cytosine site, was adjusted for all cytosine sites within the 250 selected, predefined regions [15]. The parameter $\alpha \in [0,1]$ was used to control the degree of methylation level change. To simulate methylation level changes in the 250 regions, the following equations were used. If $L_i^{old} \leq 0.5$:

then $L_i^{new} = (1-\alpha)L_i^{old} + \alpha$ for hypermethylation (higher methylation in cases) (11)

else $L_i^{new} = (1-\alpha)L_i^{old}$ for hypomethylation (lower methylation in cases). (12)

The false discovery rate (FDR) was controlled at 0.05 for all analyses [21].

To investigate the performance of SFPCA, a large scale simulation was performed under various settings. The average type I and type II error rates as well as the average true positive rate was calculated across the 100 simulated data sets. Performance of SFPCA was compared to the existing method (M3D) for predefined regions, as well as to an FPCA without the smoothing technique. The Fourier basis expansion was used for both FPCA and SFPCA using 15-35 basis functions. Different degrees of differential methylation were considered by varying the alpha parameter for $\alpha = \{0.4, 0.6, 0.8, 1\}$. To examine the robustness of the methods for various experimental design features, two different sequencing depths (5 and 20 reads) were simulated and three replicate numbers per group (3, 8, and 12) were simulated.

4. RESULTS

4.1. SIMULATION RESULTS

To assess the performance of SFPCA, results from the simulation study were compared with results from M3D and Functional Principal Component Analysis (FPCA) without smoothing. The average type I error, type II error and correct number of true DMRs identified over 100 simulated data sets under varying differential methylation strength parameters α (based on an average sequencing depth of 20 reads) under the Fourier basis expansion for both FPCA and SFPCA are illustrated in Table 3.1.

With a total of 250 true DMRs at $\alpha = 100\%$, SFPCA found 239.87 true DMRs on average, with 2.20 falsely called DMRs; FPCA found 229.85 true DMRs on average, with 3.93 false positives. Using M3D located 224.51 true DMRs on average, and found no false positives. At a differential methylation strength of 80%, SFPCA correctly called 237.03 on

average, with 2.11 false positives, while FPCA correctly called 229.02 on average, with 3.41 false positives. M3D correctly identified 222.94 DMRs on average, and found no false positives. At a methylation strength of 60%, SFPCA correctly identified 226.97 DMRs on average, with 2.48 false positives, whereas FPCA correctly identified 219.05 DMRs on average, with 2.97 false positives on average. At methylation 60%, M3D correctly identified 202.95 DMRs on average, and found no false positives. At a differential methylation strength of 40%, SFPCA found 215.00 true DMRs on average, with 2.02 false positives, while FPCA found 212.5 true DMRs on average, with 2.47 false positives. In contrast, M3D correctly identified only 190.07 DMRs on average, but had no false positives.

Table 3.1. Results for Average and Standard Deviation (S.D.) of 100 Simulations Based on SFPCA, FPCA, and M3D on Average Sequencing Depth (20 Reads), with Various Levels of Strength of Methylation Change (α)

Alpha	100%			80%			60%			40%		
Methods	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**
# Correct	239.87	229.85	224.51	237.03	229.02	222.94	226.97	219.05	202.95	215.00	212.5	190.07
S.D.	0.774	0.796	0.502	0.797	0.809	0.502	0.822	0.783	0.757	0.804	0.833	0.781
# Type-1	2.20	3.93	0	2.11	3.41	0	2.48	2.97	0	2.02	2.47	0
S.D.	0.752	0.794	0	0.827	1.090	0	0.702	0.892	0	0.816	0.501	0
# Type-2	10.13	20.15	25.49	12.97	20.98	27.06	23.03	30.95	47.05	35.00	37.5	59.93
S.D.	0.774	0.794	0.502	0.797	0.804	0.502	0.822	0.783	0.757	0.804	0.833	0.781

Table 3.2 illustrates the sensitivity of the methods to fewer reads (average sequencing of 5 reads) based on systematically altering the strength of the methylation level difference α. The SFPCA method performed well for all of the α values at an average

sequence depth of 5 and 20 reads compared to other methods. In summary, all methods had low average type I error rate with the highest occurring in FPCA at a coverage of 5 reads (Table 3.2) with average type I error rate of 0.008. M3D was the most conservative of the methods as it did not produce any type I errors, but all methods controlled the type I error rate well below 0.05. Across all settings, M3D had the highest type II error rate on average. SFPCA and FPCA had similar type II error rates on average, but SFPCA had a slight advantage as it always yielded lower type II errors across all settings.

Table 3.2. Results for Average and Standard Deviation (S.D.) of 100 Simulations Based on SFPCA, FPCA, and M3D on Average Sequencing Depth (5 Reads), with Various Levels of Strength of Methylation Change (α)

Alpha	**100%**			**80%**			**60%**			**40%**		
Methods	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**	**SFPCA**	**FPCA**	**M3D**
# Correct	**225.07**	**223.88**	**200.04**	**221.14**	**219.15**	**197.13**	**211.88**	**202.06**	**178.00**	**202.9**	**197.93**	**170.05**
S.D.	**0.831**	**0.819**	**0.815**	**0.791**	**0.832**	**0.824**	**0.782**	**0.826**	**0.804**	**0.834**	**0.843**	**0.808**
# Type-1	**4.95**	**5.97**	**0**	**4.95**	**6.05**	**0**	**3.56**	**3.94**	**0**	**4.99**	**4.54**	**0**
S.D.	**1.320**	**1.041**	**0**	**1.439**	**1.426**	**0**	**1.139**	**0.887**	**0**	**1.431**	**0.8946**	**0**
# Type-2	**24.93**	**26.12**	**49.96**	**28.86**	**30.85**	**52.87**	**38.12**	**47.94**	**72.00**	**47.1**	**52.07**	**79.95**
S.D.	**0.831**	**0.819**	**0.8155**	**0.791**	**0.821**	**0.824**	**0.782**	**0.826**	**0.804**	**0.834**	**0.843**	**0.808**

Figure 3.3 shows the average true positive rates (TPRs) for varying degrees of differential methylation (α values) for each of the three methods (SFPCA, FPCA, and M3D) and two coverage depths (5 and 20). The SFPCA method had the highest average TPR at an average sequencing depth of 5 and 20 reads across all levels of α. Overall, SFPCA and FPCA substantially outperformed M3D with respect to TPR at both average

sequencing depths (5 and 20 reads), and all levels of differential methylation strength. SFPCA and FPCA performed similarly, with SFPCA always having a slight advantage that is more magnified at larger alpha values in 20 reads. All methods have larger TPR for 20 reads compared to 5 reads. However, the TPR for SFPCA and FPCA are always above 80% under both coverage levels. It is also true that TPR increases as the strength of methylation difference (α) increases, but both SFPCA and FPCA consistently maintain high average TPR greater than 80% indicating their ability to perform well even when the "signal" is smaller.

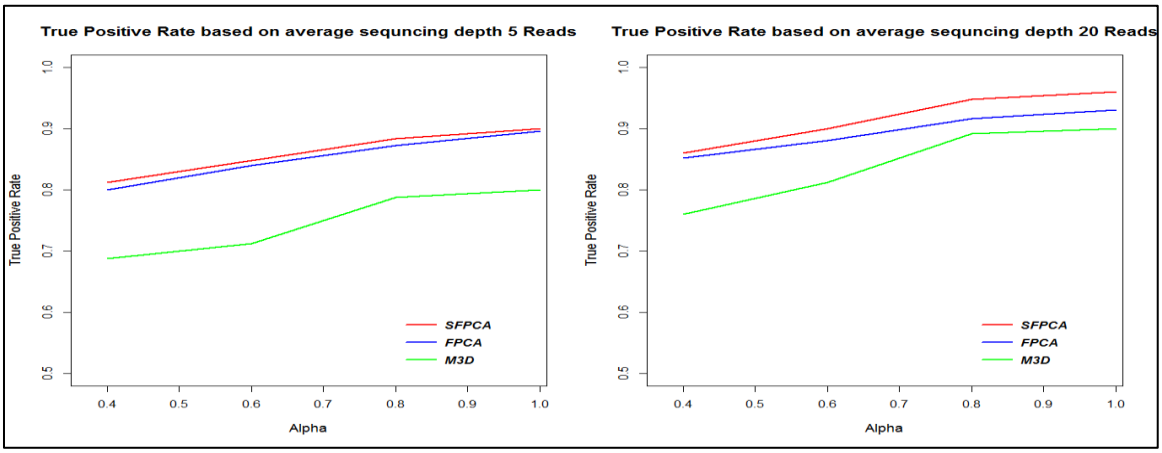

Figure 3.3. True Positive Rates Based on the Average over 100 Simulations on Average Sequencing Depths of 5 (left graph) and 20 (right graph) Reads verses α Level for Controlling the Degree of Differential Methylation Region for Each of Three Methods: Smoothed Functional Principal Component Analysis (SFPCA-Red), Functional Principal Component Analysis (FPCA-Blue) and M3D-Green.

4.2. ROBUSTNESS IN REPLICATIONS

To examine the robustness of the SFPCA method to changes in replication number, simulated data sets were created for differing numbers of replicates per group, using the same approach as described as in section.3.2. Control samples from the real RRBS data set were used as the control groups for 3, 8 and 12 replicates per group. This was possible

since the data set contained 12 control samples. A set of 3, 8, or 12 replicates were simulated as previously described to act as the cases groups. As before, the same 250 regions were simulated to be true DMRs using $\alpha = 80\%$ and coverage of 20 reads. The SFPCA method was used to identify DMRs with 3, 8 and 12 replicates per group and these results were compared. The false discovery rate (FDR) was controlled at 5%. The SFPCA method identified 195, 195, and 228 true DMRs out of the total of 250, with 3, 8, and 12 replicates per group, respectively.

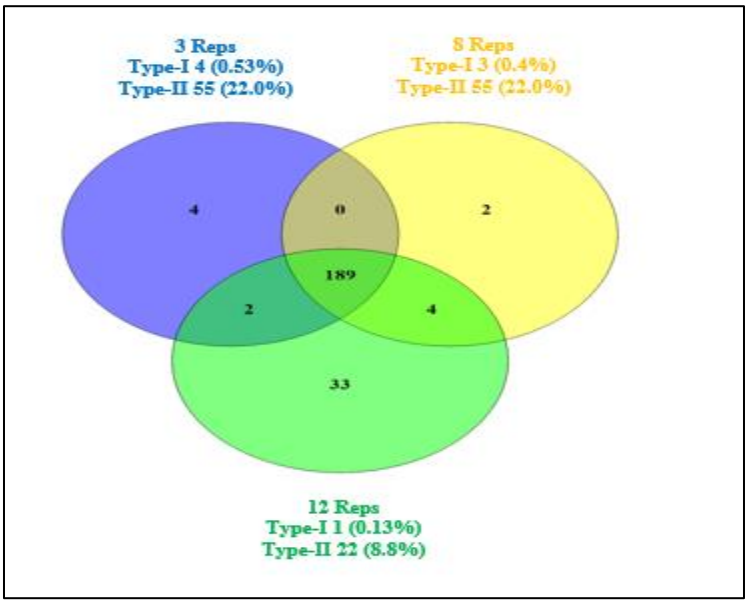

Figure 3.4. Venn Diagram of True DMRs Detected with SFPCA for 3, 8, and 12 Replicates per Group.

As shown in Figure 3.4, the overlap between the three sets of true DMRs identified accounts for 75% of the total. As was expected, the testing lost power with lower replication, with 12 replicates per group identifying the most unique true DMRs and having the lowest number type II error, and the highest number of type II errors occurred for 3 and 8 replicates per group. Overall the type II error rates ranged from 8.8% in the 12 replicate

cases to 22% for 3 and 8 replicates per group. Type I error was low for all three cases with the lowest being 0.13% for 12 replicates and the highest being 0.53% for 3 replicates. This shows that while more replicates are better, the SFPCA method exhibits a reasonable amount of robustness to smaller replicate numbers per group.

4.3. APPLICATION TO REAL DATA

An analysis was completed using the real RRBS data described in section 3.1 with four samples from patients with acute promyelocytic leukemia (APL) and four control samples (APL in remission). All CpG sites (with at least 20 reads) in both samples were used, including all region start and stop locations defined as in the simulation section.

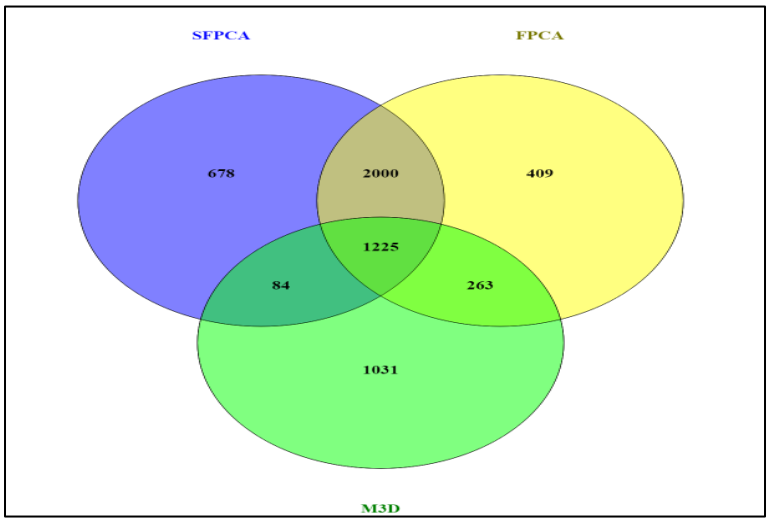

Figure 3.5. Venn Diagram Comparing the Number of Significant Differentially Methylated Regions (DMRs) Identified by the SFPCA, FPCA, and M3D Methods in the APL RRBS data set.

The false discovery rate was controlled at 5% for all analyses. Out of 14,000 CpG regions selected for testing, SFPCA and FPCA identified 3,987 and 3,897 DMRs, respectively, whereas M3D identified 2603 DMRs (Figure 3.5). Note that 1225 DMRs

were identified by all three methods, but there was much more overlap in the SFPCA and FPCA results than with both methods and M3D. These results align with simulation results the showed similarity between SFPCA and FPCA with M3D having large type II errors and identifying fewer DMR overall.

5. CONCLUSION

This study reveals that reduced representation bisulfite sequencing (RRBS) datasets can be analyzed using higher-order mathematics, using a functional data analysis approach. Region level differential methylation tests can be formed by using functional principal components that capture spatial features of methylation profiles. In this work, a smoothed functional principal component analysis (SFPCA) based on Fourier basis functions was developed to accurately identify differentially methylated regions (DMRs) between two conditions (e.g cases and control) with RRBS data. Using a simulation study based on real data, the SFPCA exhibited higher average true positive rates (TPR) when compared with FPCA without smoothing and M3D. Since low coverage can prevent statistical significance and high coverage can be costly to perform, simulations investigated how coverage depth along with different replicate numbers affected performance of the methods. SFPCA was substantially robust in relation to both coverage depth and replication maintaining high (> 78%) TPR across all settings. Overall, the SFPCA based on the Fourier basis expansion method surpassed performance of both the FPCA and M3D approaches in the simulation based on real data, as it accurately discovered more true differentially methylated regions, while maintaining a low type I error rate. Although SFPCA and FPCA exhibited similar

results, SFPCA always slightly surpassed FPCA and both approaches surpassed M3D in their ability to accurately detect true DMRs.

Even though M3D is sensitive to spatially correlated changes in methylation, it still does not allow the investigation of the dominant modes of variation in RRBS data. However, one of the best advantages of FPCA methods is to investigate the dominant modes of variation in RRBS data using the eigenfunctions of the methylation profile covariance function. The addition of a roughness penalty on the functional principal component weights to improve the smoothness appeared to be beneficial when comparing FPCA with SFPCA. The SFPCA method is superior to other currently used methods because the SFPCA scores (1) takes into account higher order properties of curve trajectory shapes when analyzing the methylation profiles and (2) accounts for correlations across all cytosine sites in the region. Further, the SFPCA statistic provides a region level comparison of the average SFPCA scores between cases and control groups by reducing information across multiple sites to a single region level test. The SFPCA approach builds on the interpretation of next-generation sequencing data by translating high-dimensional DNA methylation data into a few key factors. This greatly reduces the degrees of freedom in testing, yet it preserves the majority of the underlying biological signals.

Building on this research, the use of other functional data analysis techniques, (e.g., functional linear regression or functional canonical correlation analysis) can now be investigated. Extending the method to more complex experimental designs with more than two groups or covariates would be advantageous. Furthermore, although the effectiveness of SFPCA was investigated using RRBS, it should also work for whole genome studies but this needs to be more fully explored. Finally, although the predefined regions were defined

based on CpG density, it is also possible to apply the SFPCA method to regions based on functional annotation (e.g., CpG islands, CpG shores, and UTRs). Future studies would determine how the difference in CpG density in annotation regions affect the method performance.

REFERENCES

1. Li E, Zhang Y. DNA methylation in mammals. Cold Spring Harb Perspect Biol. 2014;6:a019133.

2. Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013;23(3):555–567.

3. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454:766–770.

4. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature. 2008;452:215–219.

5. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell. 2008;133(3):523–536.

6. Ball MP, Li JB, Gao Y, Lee J-H, LeProust EM, Park I-H, et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol. 2009;27:361–368.

7. Benoukraf T., Wongphayak S, Hadi LH, Wu M, Soong R. GBSA: a comprehensive software for analyzing whole genome bisulfite sequencing data. Nucleic Acids Res. 2013;41:e55.

8. Vaillant I, Paszkowski J. Role of histone and DNA methylation in gene regulation. Curr Opin Plant Biol. 2007;10:528–533.

9. Zilberman D, Gehring M., Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nature Genet. 2007;39:61–69.

10. Feinberg AP, Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature. 1983;301:89–92.

11. Madrigal P, Krajewski P. Uncovering correlated variability in epigenomic datasets using the Karhunen-Loeve transform. BioData Min. 2015; 8(1):20.

12. Baumann D, Doerge R. MAGI: Methylation analysis using genome information. Epigenetics. 2014;9(5):698–703.

13. Herman JG, Baylin SB. Gene silencing in cancer in association with promoter hypermethylation. N Engl J Med. 2003;349:2042–2054.

14. Feinberg AP, Tycko B. The history of cancer epigenetics. Nat Rev Cancer. 2004;4:143–153.

15. Mayo T, Schweikert G, Sanguinetti G. M3D: a kernel-based test for spatial correlated changes in methylation profiles. Bioinformatics. 2015;31(6):809–816. doi: 10.1093/bioinformatics/btu749.

16. Ramsay JO, Silverman BW. Functional data analysis. New York: Springer; 2005.

17. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. Genome Res. 2011;21(7):1099–1108.

18. Luo L, Zhu Y, Xiong M. Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. Eur J Hum Genet. 2013;21(2):217–224.

19. Hebestreit K, Dugas M, Klein H. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013;29(13):1647–1653.

20. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics. 2011;27(11):1571–1572.

21. Schoofs T, Rohde C, Hebestreit K, Klein H-U, Göllner S, Schulze I, et al. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. Blood®. 2013;121(1):178–187.

22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995; 289–300.

23. Gretton A, Borgwardt KM, Rasch M, Scholkopf B, Smola AJ. A kernel method for the two-sample-problem. Adv Neural Inf Process Syst. 2007;19:513.

24. VanderKraats ND, Hiken JF, Decker KF, Edwards JR. Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. Nucleic Acids Res. 2013;41(14):6816–6827.

25. Fouse SD, Nagarajan RP, Costello JF. Genome-Scale DNA methylation analysis. Epigenomics. 2010;2(1):105–117.

26. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen, KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049.

27. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics. 2011;27(11):1571–1572.

28. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, et al. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature Genet. 2005;37(8):853-862.

4. SUMMARY AND FUTURE WORK

The main purpose of this dissertation is to provide a novel statistical framework for identifying differentially methylated regions that contribute to biologically meaningful interpretation of reduced representation bisulfite sequencing (RRBS) data. Specific methods from functional data analysis (FDA) can be beneficial since they utilize correlation between neighboring cytosines and capture dominant modes of variation in methylation trajectories over a region. Testing differentially methylated regions through functional principal component analysis (FPCA) is described in Section 2. This research developed an FPCA method based on Fourier and B-spline basis functions that successfully tested for differentially methylated regions (DMRs) between the case and control groups in the RRBS data. An empirical comparison of FPCA to the only other similar type of region level test that explored curve shape differences, M3D[1], was made via a simulation based on real data. FPCA showed a significant increase in true positive rates in comparison with M3D, as well as considerable robustness with respect to coverage depth and replications. The FPCA based on the Fourier and B-spline methods both outperformed M3D as they both accurately detected more DMRs across all simulation settings. FPCA-Fourier and FPCA-Bspline perform similarly overall, except that FPCA-Fourier had slightly lower type II errors than FPCA-Bspline when the sequencing depth was 20 reads and this was reversed for lower degrees of differential methylation for 5 reads. Both methods maintained a type I error rate below 0.05.

Since the methylation profiles are typically not smooth across a region, this leads to substantial variability in the estimated functional principal component curves. Thus,

further improvements to the FPCA method were developed in Section 3 via a smoothed functional principal component analysis (SFPCA) for detecting differentially methylated regions. This smoothed FPCA identifies DMRs by combining a goodness-of-fit measure with a roughness penalty to maintain the advantages of basis expansion. The SFPCA is used to compare differences within a region in the average SFPCA scores between the variation of cases and controls. In this study, the SFPCA scores take into account all information across all CpG sites in a predefined genomic region based on CpG density. In comparison to the currently available M3D method, the SFPCA technique had significantly higher true positive rates (TPR) and was robust in relation to coverage depth and replications, using a simulation study based on real data. The SFPCA method also showed slight improvements in the TPR when compared to FPCA without smoothing, indicating that this additional model component was beneficial in capturing an important aspect of the DNA methylation profile over a region level.

In future research, the SFPCA and FPCA framework could be expanded to incorporate a test for more complex experimental designs that involve more than two groups or that include covariates such as age, sex or medical related information. Furthermore, this technique could be tested where the functional annotation information is used to define the regions (e.g., CpG islands, CpG shores, UTRs, introns, and exons). An investigation into how the CpG density within annotation regions differs and will affect the testing performance will be beneficial to understanding the difference in the two options. Also the SFPCA and FPCA could be expanded to plants data where the DNA methylated occurred in three sequence contexts: CG, CHG and CHH (where H=A, T or C).

REFERENCES

1. A. J. F. Griffiths, *An Introduction to Genetic Analysis*, Macmillan, New York, 2005.

2. S. R. Eddy, "Non-coding RNA Genes and the Modern RNA World," *Nature Reviews Genetics* 2, no. 12 (2001) 919–929.

3. H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore and J. Darnell, "Section 4.1, Structure of Nucleic Acids," *Molecular Cell Biology*, 4th edition. W. H. Freeman, New York, 2000. https://www.ncbi.nlm.nih.gov/books/NBK21475/.

4. A. M. Deaton and A. Bird, "CpG Islands and the Regulation of Transcription," *Genes & Development* 25, no.10 (2011) 1010-1022.

5. https://geneed.nlm.nih.gov/topic_subtopic.php?tid=15. Gene-Ed: Genetics, Education, Discovery, "DNA, Genes, Chromosomes," Feb 28, 2017.

6. J. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., "The Sequence of the Human Genome," *Science* 291, no. 5507 (2001) 1304–1351.

7. P. C. Ng and E. F. Kirkness, "Whole Genome Sequencing," in *Genetic Variation*. Humana Press, pp. 215–226, 2010.

8. C. H. Waddington, "The Epigenotype," *Endeavor* 1 (1942) 18–20.

9. V. E. A. Russo, R. A. Martienssen and A. D. Riggs, *Epigenetic Mechanisms of Gene Regulation*, Cold Spring Harbor Laboratory Press, Plainview, NY, 1996.

10. P. A. Jones and S. B. Baylin, "The Epigenomics of Cancer," *Cell* 128 (2007) 683–692.

11. R. Jaenisch and A. Bird, "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals," *Nature Genetics* 33 (2003) 245–254.

12. I. Vaillant and J. Paszkowski, "Role of Histone and DNA Methylation in Gene Regulation," *Current Opinion in Plant Biology* 10 (2007) 528–533.

13. D. Zilberman, M. Gehring, R. K., Tran, T. Ballinger and S. Henikoff, "Genome-wide Analysis of *Arabidopsis thaliana* DNA Methylation Uncovers an Interdependence Between Methylation and Transcription," *Nature Genetics* (2007) 61–69.

14. A. P. Feinberg and B. Vogelstein, "Hypomethylation Distinguishes Genes of Some Human Cancers from Their Normal Counterparts," *Nature* 301 (1983) 89–92.

15. A. P. Feinberg and B. Tycko, "The History of Cancer Epigenetics," *Nature Reviews Cancer* 4 (2004) 143–153.

16. K. D. Robertson, "DNA Methylation and Human Disease," *Nature Reviews Genetics* 6 (2005) 597–610.

17. D. S. Shames, J. D. Minna and A. F. Gazdar, "DNA Methylation in Health, Disease, and Cancer," *Current Molecular Medicine* 7 (2007) 85–102.

18. [J.-P. J. Issa and H. M. Kantarjian, "Targeting DNA Methylation." *Clinical Cancer Research* 15, no. 12 (2009) 3938-3946.

19. J. Qiu, "Epigenetics: Unfinished Symphony," *Nature 441* (2006) 143 –145. doi:10.1038/441143a.

20. S. Brinkers, H. R. Dietrich, F. H. de Groote, I. T. Young and B. Rieger, "The Persistence Length of Double Stranded DNA Determined Using Dark Field Tethered Particle Motion," *The Journal of Chemical Physics* 130, no. 21 (2009) 06B607.

21. M. D. Topal and J. R. Fresco, "Complementary Base Pairing and the Origin of Substitution Mutations," *Nature* 263, no. 5575 (1976) 285–289.

22. A. Bird, "DNA Methylation Patterns and Epigenetic Memory," *Genes & Development* 16 (2002) 6–21.

23. M. M. Suzuki and A. Bird, "DNA Methylation Landscapes: Provocative Insights from Epigenetics," *Nature Reviews Genetics* 9 (2008) 465–476.

24. E. Li and A. Bird, "DNA Methylation in Mammals," in *Epigenetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 341–356, 2007.

25. J. A. Law and S. E. Jacobsen, "Establishing, Maintaining and Modifying DNA Methylation Patterns in Plants and Animals," *Nature Reviews Genetics* 11, no. 3 (2010) 204–220.

26. J. B. Margot, A. E. Ehrenhofer-Murray and H. Leonhardt, "Interactions Within the Mammalian DNA Methyltransferase Family," *BMC Molecular Biology* 4, no. 1 (2003).

27. K. R. Van Eijk, S. de Jong, M. P. M. Boks, T. Langeveld, F. Colas, J. H. Veldink, et al., "Genetic Analysis of DNA Methylation and Gene Expression Levels in Whole Blood of Healthy Human Subjects," *BMC Genomics* 13 (2012) 636. Doi: 10.1186/1471-2164-13-636.

28. M. Gehring and S. Henikoff, "DNA Methylation Dynamics in Plant Genomes." Biochimica et Biophysica Acta 1769 (2007) 276–286.

29. R. K. Slotkin and R. Martienssen, “Transposable Elements and the Epigenetic Regulation of the Genome,” *Nature Reviews Genetics* 8 (2007) 272–285.

30. J. K. Kim, M. Samaranayake and S. Pradhan, “Epigenetic Mechanisms in Mammals,” *Cellular and Molecular Life Sciences* 66 (2009) 596–612.

31. E. J. Finnegan, “DNA Methylation: A Dynamic Regulator of Genome Organization and Gene Expression in Plants,” in *Plant Developmental Biology - Biotechnological Perspectives: Volume 2*, Springer, Berlin, Germany, pp. 295–323, 2010.

32. E. C. Berglund, A. Kiialainen and A. C. Syvänen, “Next-generation Sequencing Technologies and Applications for Human Genetic History and Forensics,” *Investigating Genetics* 2 (2011) 23.

33. K. V. Voelkerding, S. Dames and J. D. Durtschi, “Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy: A Paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology,” *The Journal of Molecular Diagnostics* 12, no. 5 (2010) 539–551. doi:10.2353/jmoldx.2010.100043.

34. E. D. Gunnarsdóttir, M. Li,, M. Bauchet, K. Finstermeier and M. Stoneking, “High-throughput Sequencing of Complete Muman mtDNA Genomes from the Philippines,” *Genome Research* 21, no. 1 (2011) 1–11.

35. R. R. Lister, C. O’Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, et al., “Highly Integrated Single-base Resolution Maps of the Epigenome in *Arabidopsis*,” *Cell* 133 (2008) 523–536.

36. A. Meissner, T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, et al., “Genome-scale DNA Methylation Maps of Pluripotent and Differentiated Cells,” *Nature* 454 (2008) 766–770.

37. D. Serre, B. H. Lee and A. H. Ting, “MBD-isolated Genome Sequencing Provides a High-throughput and Comprehensive Survey of DNA Methylation in the Human Genome.” *Nucleic Acids Research* 38 (2010) 391–399.

38. T. A. Down, V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, et al., “A Bayesian Deconvolution Strategy for Immunoprecipitation Based DNA Aethylome Analysis,” *Nature Biotechnology* 26 (2008) 779–785.

39. M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collins, F. Watt, G. W. Grigg, et al., “A Genomic Sequencing Protocol That Yields a Positive Display of 5-methylcytosine Residues in Individual DNA Strands,” *Proceedings of the National Academy of Sciences* 89, no. 5 (1992) 1827–1831.

40. D. R. Masser, D. R. Stanford and W. M. Freeman, "Targeted DNA Methylation Analysis by Next-generation Sequencing," *Journal of Visualized Experiments* 96 (2015).

41. A. Meissner, A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander and R. Jaenisch, "Reduced Representation Bisulfite Sequencing for Comparative High-resolution DNA Methylation Analysis," *Nucleic Acids Research* 33, no. 18 (2005) 5868–5877.

42. H. Gu, C. Bock, T. J. Mikkelsen, N. Jäger, Z. D. Smith, E. Tomazou, et al., "Genome-scale DNA Methylation Mapping of Clinical Samples at Single-nucleotide Resolution," *Nature Methods* 7, no. 2 (2010) 133–136.

43. H. Gu, Z. D. Smith, C. Bock, P. Boyle, A. Gnirke, and A. Meissner, "Preparation of Reduced Representation Bisulfite Sequencing Libraries for Genome-scale DNA Methylation Profiling," *Nature Protocols* 6, no. 4 (2011) 468–481.

44. G. R. Olbricht, "Statistical Methods for Next-generation Sequencing (NGS) DNA Methylation Data," Joint Statistical Meetings, San Diego, CA, 2012.

45. A. Raine, E. Manlig, P. Wahlberg, A.-C. Syvänen and J. Nordlund, "SPlinted Ligation Adapter Tagging (SPLAT), a Novel Library Preparation Method for Whole Genome Bisulphite Sequencing," *Nucleic Acids Research* 45, no. 6 (2017) e36. doi: 10.1093/nar/gkw1110.

46. F. Krueger, B. Kreck, A. Franke and S. R. Andrews, "DNA Methylome Analysis Using Short Bisulfite Sequencing Data," *Nature Methods* 9, no. 2 (2012) 145–151.

47. R. A. Harris, T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, et al., "Comparison of Sequencing-based Methods to Profile DNA Methylation and Identification of Monoallelic Epigenetic Modifications," *Nature Biotechnology* 28, no. 10 (2010) 1097–1105.

48. K. Hebestreit, M. Dugas and H.-U. Klein, "Detection of Significantly Differentially Methylated Regions in Targeted Bisulfite Sequencing Data," *Bioinformatics* 29, no. 13 (2013) 1647–1653.

49. I. C. G. Weaver, N. Cervoni, F. A. Champagne, A. C. D'Alessio, S. Sharma, S. Dymov, et al., "Epigenetic Programming by Maternal Behavior," *Nature Neuroscience* 7, no. 8 (2004) 847–854.

50. R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, et al., "Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences," *Nature* 462 (2009) 315–322.

51. A. Akalin, M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, et al., "methylKit: A Comprehensive R Package for the Analysis of Genome-wide DNA Methylation Profiles," *Genome Biology* 13, no. 10 (2012) R87.

52. Y. Park, M. E. Figueroa, L. S. Rozek and M. A. Sartor, "MethylSig: A Whole Genome DNA Methylation Analysis Pipeline," *Bioinformatics* 30, no. 17 (2014) 2414–2422.

53. A. T. L. Lun and G. K. Smyth, "De Novo Detection of Differentially Bound Regions for ChIP-seq Data Using Peaks and Windows: Controlling Error Rates Correctly," *Nucleic Acids Research* 42, no. 11 (2014) e95.

54. G. Wu, N. Yi, D. Absher and D. Zhi, "Statistical Quantification of Methylation Levels by Next-generation Sequencing," *PLoS One* 6, no. 6 (2011) e21034.

55. T. R. Mayo, G. Schweikert and G. Sanguinetti, "M3D: A Kernel-based Test for Spatially Correlated Changes in Methylation Profiles," *Bioinformatics* 31, no. 6 (2015) 809–816.

56. D. D. Baumann and R. W. Doerge, "Magi: Methylation Analysis Using Genome Information," *Epigenetics* 9, no. 5 (2014) 698–703.

57. J. O. Ramsay and B. W. Silverman, *Functional Data Analysis,* Springer, New York, 2005.

VITA

Mohamed Salem F. Milad was born in East Libya, Benghazi, Libya. In 2002, he received a B. S. in Statistics from Garyounis University, Libya. In 2004, Milad received an M.S. in Mathematics from the University of Science, Malaysia (USM). In May 2005, he joined Al Jabal Al Gharbi University, Libya, as a lecturer in the Data Analysis Department. In 2013, Milad received an M.S. in Applied Mathematics from Missouri University of Science and Technology, and in July 2017 he received his PhD in Mathematics from Missouri University of Science and Technology.