Spring 2021

Integrating snp data and imputation methods into the DNA methylation analysis framework

Yuqing Su

Department: Mathematics and Statistics

INTEGRATING SNP DATA AND IMPUTATION METHODS INTO THE DNA METHYLATION ANALYSIS FRAMEWORK

by

YUQING SU

A DISSERTATION

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

MATHEMATICS WITH STATISTICS EMPHASIS

2021

Approved by

Gayla R. Olbricht, Advisor
V. A. Samaranayake
Xuerong Wen
Akim Adekpedjou
Ronald L. Frank

ABSTRACT

DNA methylation is a widely studied epigenetic modification that can influence the expression and regulation of functional genes, especially those related to aging, cancer and other diseases. The common goal of methylation studies is to find differences in methylation levels between samples collected under different conditions. Differences can be detected at the site level, but regulated methylation targets are most commonly clustered into short regions. Thus, identifying differentially methylated regions (DMRs) between different groups is of prime interest. Despite advanced technology that enables measuring methylation genome-wide, misinterpretations in the readings can arise due to the existence of single nucleotide polymorphisms (SNPs) in the target sequence. One of the main pre-processing steps in DMR detection methods involves filtering out potential SNP-related probes due to this issue. In this work, it is proposed to leverage the current trend of collecting both SNP and methylation data on the same individual, making it possible to integrate SNP data into the DNA methylation analysis framework. This will enable the originally filtered potential SNPs to be restored if a SNP is not actually present. Furthermore, when a SNP is present or other missing data issues arise, imputation methods are proposed for methylation data. First, regularized linear regression (ridge, LASSO and elastic net) imputation models are proposed, along with a variable screening technique to restrict the number of variables in the models. Functional principal component regression imputation is also proposed as an alternative approach. The proposed imputation methods are compared to existing methods and evaluated based on imputation accuracy and DMR detection ability using both real and simulated data. One of the proposed methods (elastic net with variable screening) shows effective imputation accuracy without sacrificing computation efficiency across a variety of settings, while greatly improving the number of true positive DMR detections.

ACKNOWLEDGMENTS

With many thanks to my advisor Dr. Gayla Olbricht, whose expertise was invaluable in formulating the research questions and methodology. She has been a great mentor and role model for me academically and personally for her enthusiastic, lively and kind character. I thank my committee members, Dr. V.A. Samaranayake, Dr. Akim Adekpedjou, Dr. Xuerong Wen and Dr. Ronald Frank for their insightful feedback and valuable suggestions. I am grateful for all of my instructors while studying at Missouri S&T. In addition, I thank my colleague Arnold Harder for sharing his experiences with DMR detection methods. Finally, I want to thank my family for their support during my good and hard times.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

1. INTRODUCTION

1.1. EPIGENETICS AND DNA METHYLATION

Genetics is the study of heritable changes involving modifications of the DNA sequence that exhibit variation between individuals. It includes the study of gene expression, genetic changes and multiple gene interactions. Changes to the DNA sequence are called mutations, and there are different types including deletions, insertions and translocations. Mutations can sometimes lead to the malformation of proteins, which may lead to disease. For example, sickle cell disease is caused by a single nucleotide mutation in the HBB gene that provides instructions for making one part of hemoglobin (Schnog et al., 2004). On the other hand, epigenetics is the study of heritable changes that are not associated with any alteration of the DNA sequence. Although all cells in an organism contain the same genetic information, the expression of genes can differ between cells. For example, different cell types require different genes to be active to perform their functions. Gene expression is regulated by epigenetics through different mechanisms, such as histone modifications, DNA methylation and non-coding RNA (Wei et al., 2017).

1.1.1. Mechanisms of DNA Methylation. DNA methylation (DNAm) plays an important role in gene regulation. It is one of the most studied epigenetic modifications in human cells that can affect gene expression and preserve cellular states through cell division without actually changing the DNA sequence. A nucleotide on a DNA molecule, specifically a cytosine, is methylated when a methyl group ($-\mathbf{CH}_3$) is added to the carbon-5 position of a cytosine, forming 5-methylcytosine. In mammals, DNAm is almost exclusively found in CpG dinucleotides (a compound comprised of two nucleotides, cytosine (C) and guanine (G)) (Moore et al., 2013). The "p" simply indicates that "C" and "G" are connected by a phosphodiester bond. In stem cells and in plants, methylation is also

found in the context of CHG and CHH where H is either A, T or C. A modified cytosine was first discovered in mammals by Hotchkiss (1948), who hypothesized that it was 5-methylcytosine and that it existed naturally in DNA. DNA methylation was demonstrated to be involved in gene regulation and cell differentiation in the 1980s (Holliday and Pugh, 1975; Compere and Palmiter, 1981). Further studies have revealed the important role of DNA methylation in many biological processes; including genomic imprinting (Tycko, 1997), transposable element silencing (Hollister and Gaut, 2009), stem cell differentiation (Sheaffer et al., 2014), embryonic development (Messerschmidt et al., 2014) and inflammation (Bayarsaihan, 2011), as well as cancer (Bock, 2012) and several other diseases.

1.1.2. CpG Island. A CpG island is a short part of the DNA sequence with a higher frequency of the CG dinucleotides sequence than other regions. CpG islands are often defined as a region with at least 200 base pairs (bp), a C and G percentage greater than 50%, and an observed-to-expected CpG ratio greater than 60% (Ongenaert, 2010). More stringent criteria have been proposed because this definition was unable to distinguish CpG islands from certain DNA repeat structures. Takai and Jones (2002) define a CpG island as having a minimal length of 500 bp, an observed-to-expected CpG ratio greater than 65%, and a C and G content of more than 55% are required. This largely solves the repeat problem, with the drawback that CpG islands that are smaller than 500 bp can not be predicted.

CpG islands typically occur at or near the transcription start site of genes, particularly housekeeping genes, in vertebrates. About 70% of human gene promoters have high CpG concentrations (Saxonov et al., 2006). DNA is wrapped around histone proteins forming small, packaged sections called nucleosomes. One of the common features of CpG islands is that they have less nucleosomes than other parts of DNA. This is often associated with modified histones and results in enhancing gene expression (Tazi and Bird, 1990). The CG

dinucleotides sequence is not typically methylated in the promoter region of active genes. By contrast, the CG dinucleotides sequences in the promoter region of inactive genes are usually methylated to suppress their expression (Vinson and Chatterjee, 2012).

1.1.3. DNA Methylation and Cancer. It is well recognized that DNA methylation is an important epigenetic factor influencing gene activities, including genomic imprinting, aging and carcinogenesis. Cancer cells must undergo a series of molecular-level events to have the ability to replicate without limitation, as well as to invade and metastasize (Hanahan and Weinberg, 2011). Hypomethylation describes the unmethylated state of CpG sites that are normally methylated (a decrease in methylation); whereas hypermethylation refers to the methylated state of CpG sites in a specific sequence that are normally unmethylated (an increase in methylation). In cancer, global hypomethylation is accompanied by hypermethylation of specific genes. Hypermethylation in the promoter regions of certain genes can suppress the expression of their functional proteins, including known tumor suppressor genes, leading to the silencing of those genes (Wajed et al., 2001). Epigenome-wide DNA methylation studies have shown that the methylation within functional promoter areas was associated with an increased risk of breast cancer, while the methylation of genomic regions outside the promoters was associated with a decreased risk (Severi et al., 2014). However, global hypomethylation has also been associated with oncogenesis (Das and Singal, 2004). Studies are ongoing to investigate the relationship between methylation patterns across the genome and specific types of cancer.

1.2. DNA METHYLATION TECHNOLOGIES

1.2.1. Bisulfite Sequencing. The development of technologies to measure levels of DNA methylation throughout the genome has been substantial in the past 30 years. These technological advances allow for significant improvement in understanding the role of epigenetics in medicine and biology in general. One method to detect DNA methylation at individual CpG sites is Bisulfite Sequencing combined with next generation sequencing (BS-

seq) or whole genome bisulfite sequencing (WGBS). The basic principle of this approach involves bisulfite conversion on the unmethylated cytosines. Bisulfite conversion is a process in which DNA is denatured and treated with sodium bisulfite. The unmethylated cytosines are converted to uracils, while methylated cytosines remain unchanged. Following this process, the DNA is then treated by PCR amplification where the uracils are converted to thymines (Frommer et al., 1992). Comparing the sequence of converted DNA to untreated DNA creates a methylation profile of the sample. BS-seq or WGBS is the most thorough and informative approach to measure methylation status, thus it is capable of revealing subtle methylation patterns, and it achieves the most comprehensive coverage of a genome.

However, BS-seq is a costly and time-consuming procedure because the whole genome is tested. Reduced representation bisulfite sequencing (RRBS) is an efficient alternative for analyzing the genome-wide methylation profiles on a single cytosine level (Meissner et al., 2005). RRBS examines a subset of the genome by using a restriction enzyme to extract regions with a high CG dinucleotides content. The amount of nucleotides required to sequence is only 1% of the genome. These fragments often cover key promoter regions and CpG islands. This makes RRBS more economical and efficient. Therefore RRBS is suitable for large-scale comparative methylation studies across different tissues of cell types. On the other hand, a limitation of RRBS is that it can miss some CG dinucleotides and have lower coverage of some regions.

1.2.2. Infinium Beadchips. Illumina has developed a novel bead array technology using silica microbeads. On the surface of each array, tiny silica beads are located in microwells and coated with multiple probes. Probes are a collection of DNA spots that are attached to the solid surface for hybridization with the labeled target. Different probes are attached to each bead (Steemers and Gunderson, 2005). Illumina Infinium BeadChips have provided an easy to use, time efficient and cost effective way to measure methylation levels. The technology was first introduced with the Infinium HumanMethylation27 BeadChip (HM27). Quantitative measurements of DNA methylation can be determined for 27,578

CpG dinucleotides spanning 14,495 genes using the HM27 platform. Like BS-seq, the initial step is bisulfite conversion, in which only a small amount of genomic DNA is required. Next, each sample is amplified, enzymatically fragmented, purified and then applied to the BeadChips for hybridization. There are two bead types that correspond to each CpG locus: one for the methylated and the other for the unmethylated state. Then, the array is stained with fluorescent dye and the intensities are measured (Weisenberger et al., 2008).

In 2011, an updated array called the Infinium HumanMethylation450 BeadChip (HM450) became the most widely used method for DNA methylation profiling. The HM450 array features 485,577 probes in coding and non-coding DNA regions, covering 94% of the CpG sites on the HM27 array. Coverage is targeted to gene regions with sites in the promoter region, 5' UTR, first exon, gene body, and 3' UTR of RefSeq genes. CpG islands, CpG sites outside of CpG islands, and some enhancer regions are also covered by the HM450 array, as well as differentially methylated sites identified in tumor versus normal and across several tissue types. Notably, The Cancer Genome Atlas (TCGA) (TCGA, 2021) consortium used the HM450 platform to profile more than 8500 samples from over 52 different cancer types.

There are two types of probes (Infinium I and Infinium II) on the HM450 array. Both probe types have assay chemistry technologies that are utilized to enhance the depth of coverage for methylation analysis. An illustration of the two probe types can be found in Figure 1.1. The Infinium I assay, also used in HM27, employs two bead types per CpG locus: one for the methylated and one for the unmethylated states. The Infinium II design uses one bead type, with the methylated state determined at the single base extension step. The addition of the Infinium II design enables each of up to three CpG sites to be either methylated or unmethylated on the probe with no impact on the result for the queried site. For the HM450 BeadChip, about 30% of CpG sites are measured using Infinium I probes and 70% of CpG sites are measured by the Infinium II probes. In 2016, the new Illumina Infinium Methylation EPIC array was released that can provide DNA methylation

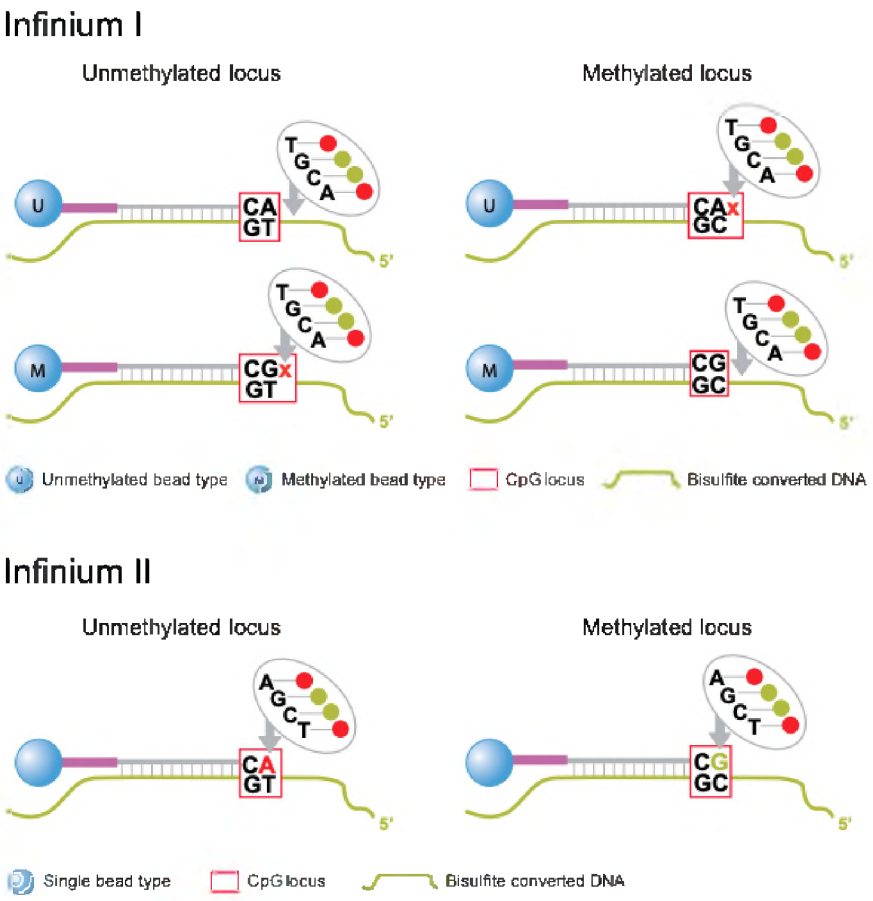


Figure 1.1. Two types of Infinium probes. The Infinium I probes (top) have two bead types: one for the methylated and one for the unmethylated states. The Infinium II probes (bottom) has one bead type with the methylated state determined at the single base extension step. Figure from Illumina (2012).

levels for a total of 863,904 CpG sites. The EPIC array includes over 90% of the HM450 probes, as well as additional probes dedicated to the enhancers revealed by the Functional Annotation of the Mammalian Genome project (FANTOM5) and the Encyclopedia of DNA Elements project (ENCODE). FANTOM5 and ENCODE are both public research projects aiming to identify functional elements in the human genome. Overall, data from the EPIC array at single loci are highly reproducible across technical and biological replicates and demonstrate high correlation with HM450 and WGBS data (Pidsley et al., 2016). In this

work, the HM450 platform is utilized since data are accessible from TCGA. However, the methods can be generalized to methylation data obtained by other technologies, such as the EPIC array, WGBS or RRBS.

1.3. SNP AND METHYLATION MICROARRAYS

A single-nucleotide polymorphism (SNP) is a single nucleotide substitution in the DNA sequence (Figure 1.2). Typically there are two possible nucleotides altering at a given position (Vignal et al., 2002). SNPs are a common type of genetic variation among other DNA sequence mutations such as deletions, insertions and translocations. A variant is classified as a SNP when more than 1% of the population does not share the same nucleotide at the specific position on the genome. In humans, the occurrence rate of SNPs is about 0.1%, meaning that there is one SNP in every 1,200 to 1,500 base pairs (Shastry, 2002). SNPs can occur anywhere in the genome, including in the coding regions of genes where they could lead to the changes in gene function and expression. SNPs can be identified through hybridization-based or enzyme-based methods. A SNP array is one detection method based on the hybridization of the fragmented DNA sequence and the immobilized allele-specific oligonucleotide probes (LaFramboise, 2009). DNA methylation microarrays also have a connection to SNP arrays. DNAm arrays interrogate DNA methylation states by sodium bisulfite conversion which transforms an epigenetic difference between a modified cytosine (including 5-methylcytosine (5-mC) and 5-hydroxymethylcytosine (5hmC)) and an unmodified cytosine to a genetic C/T SNP (Frommer et al., 1992). Therefore, the DNAm microarrays are essentially SNP arrays because the Infinium arrays obtain the methylation intensity at a particular location by checking whether there is a C/T SNP present.

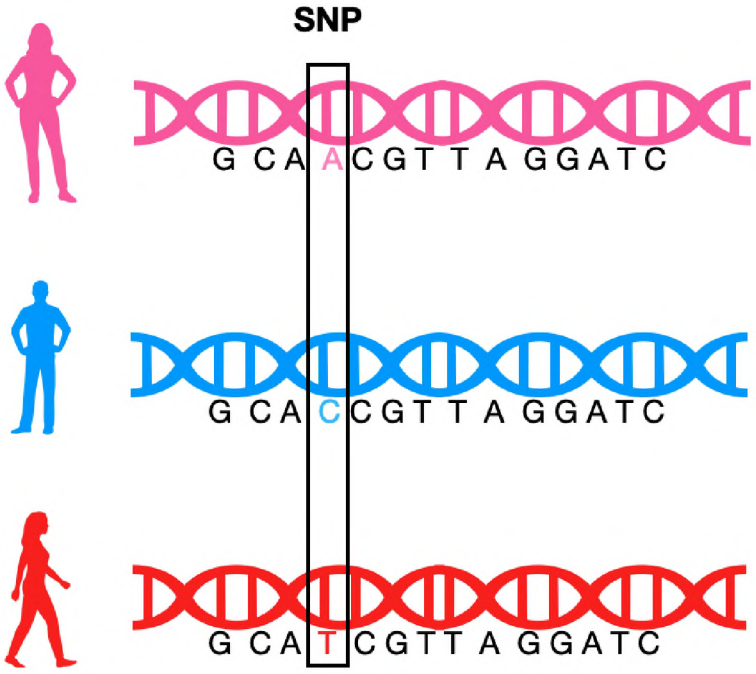


Figure 1.2. Single nucleotide polymorphism (SNP). At the fourth base, a SNP is present. The DNA sequence varies within a population at this site.

1.4. PREPROCESSING OF METHYLATION DATA

1.4.1. Probe Filtering. Some probes on the Illumina methylation arrays (including the HM450 array used in this work) are prone to giving inaccurate values that do not represent the underlying methylation state. This can arise due to a number of different reasons, that are each considered for initial filtering. One way this can happen is when a probe is present in low quantities because of amplification artifacts or mutation, resulting in a mismatched intended sequence. Such probes should be filtered out since they mostly carry background noise. To distinguish signal from noise, detection p-values are used. The background distribution is assumed normal, and the parameters are estimated using negative control

probes. The p-value is computed using a Z-test (Heiss and Just, 2019). A small detection p-value indicates that the measured intensity is very likely to be a true (significant) signal and not background noise. Typically, significance levels of 0.05 or 0.01 are used. A sample is considered a bad sample when over 10% of the probes are problematic based on the detection p-values, and should be removed from the analysis. Bead counts are also a consideration of the probe quality. Usually, probes with less than 3 beads in at least 5% of samples per probe are filtered out. A larger proportion of non-CpG-target probes (Probe ID starting with "ch") are potentially cross-hybridizing probes. Cross-hybridization is the tendency for chains of nucleic acids to bind to other chains of nucleic acids that have similar but not identical sequences. This makes the results difficult to interpret (Reilly et al., 2006). Of the 3,091 non-CpG probes on the HM450 microarray, only 39% can be mapped with a perfect match to the correct genomic location annotated by Illumina (Chen et al., 2013). Thus, all non-CpG probes are filtered out. Moreover, there are 65 built-in SNP probes (Probe ID starting with "rs") querying high-frequency SNPs in the HM450 array for the purpose of quality control, and they are typically removed in preprocessing steps.

The existence of SNPs can affect DNA methylation readouts in the Infinium arrays. SNPs can increase mismatches close to the 3' end of the probe sequence and interfere with successful extension. It can also change the CpG dinucleotide sequence and therefore the ability of cytosines to be methylated. A special case would be the presence of an actual C/T polymorphism instead of the C/T introduced by bisulfite conversion. Also, for a Type I Infinium probe, the color channel depends on the extension base. If a SNP is present in the extension base, a potential color change could happen. Specifically, the color switch can be caused by an A/G SNP but not an A/T SNP, because A and T bases are both labeled with red fluorophores, and C and G are labeled with green. Therefore, probes with any SNP of global minor allele frequency (MAF) over 1% and within 5 bp from their targets, Infinium

I probes with putative color channel switching SNPs, and Infinium II probes with SNPs of global MAF over 1% affecting the extension base are suggested to be filtered out (Zhou et al., 2016).

When aligning probe sequences to the human genome, some probes map to multiple sites. It has been shown that probes with non-unique alignments display significantly greater variance in methylation levels than uniquely mapping probes (Nordlund et al., 2013). Therefore, the multi-hit probes are usually filtered out in the preprocessing step. Typically, probes located in chromosome X and Y are removed to avoid sex related methylation biases.

1.4.2. Normalization. As mentioned in section 1.2, the Infinium HumanMethylation450 and the EPIC BeadChip use two different types of chemical assays for their probes. This probe design can potentially cause problems for data analysis if they are not handled properly. It is shown that Infinium I and II probes usually have different distributions of methylation values, and that Infinium II probes are relatively less accurate and more sensitive for detecting extreme methylation values (Dedeurwaerder et al., 2011). In order to eliminate the influence of different probe types, remove sources of technical variation between measurements, as well as cancel background noise of the data, several different normalization methods have been developed.

Quantile normalization (Bolstad et al., 2003), first used in gene expression data, uses the mean intensity of the probes with the same rank from all studied arrays to replace the intensity of a probe. This helps make the distribution of probe intensities the same for each array. A peak-based correction (PBC) method (Dedeurwaerder et al., 2011) estimates the methylation peaks for the two probe types separately, then rescales the Infinium II values according to the initial range of Infinium I. The subset-quantile within array normalization (SWAN) method (Maksimovic et al., 2012) is based on normalization methods from microarray gene expression platforms. An average quantile distribution is determined using a subset of probes defined to be biologically similar based on CpG content. The intensities of the remaining probes are then adjusted by interpolation onto the distribution of the subset

probes. The β-mixture quantile normalization (BMIQ) method (Teschendorff et al., 2013) decomposes the methylation profiles of Infinium I and Infinium II probes into two mixtures of three methylation states (unmethylated, partially methylated and fully methylated), and then quantile normalizes the three distributions of the Infinium II profile corresponding to those of the Infinium I profile.

1.5. DIFFERENTIAL METHYLATION TESTING

1.5.1. Site Level Testing. A common goal of methylation studies is to discover individual CpG sites that have significantly different methylation levels between different conditions (e.g., normal vs. disease). These differentially methylated sites can be of substantial importance for the identification of novel disease biomarkers. In recent years, many statistical methods were developed for different types of methylation data to detect differentially methylated CpG sites.

For BS-seq, data can be summarized as counts of methylated and unmethylated reads at any given site. Fisher's exact test (FET) was one of the first approaches used to detect differentially methylated sites (Lister et al., 2009). However, FET does not account for the inherent biological variation that is present across biological replicates and it assumes independence between cytosine sites. BSmooth (Hansen et al., 2012) is an alternative approach that uses a "signal-to-noise" statistic to quantify differential methylation evidence at individual CpG sites by combining top ranked differentially methylated cytosines (DMCs), which are found using a t-statistic approach with either a quantile or direct t-statistic cutoff. BSmooth is not used directly for inference of differential sites, but rather uses the site level statistics to find differentially methylated regions. The beta-binomial model is an alternative statistical model for replicated BS-seq DNA methylation measurements. The beta-binomial distribution is the binomial distribution in which the probability of success

in each of n trials is not fixed but randomly drawn from a beta distribution. It can account for both sampling and epigenetic variability. The beta-binomial model is used by methylSig (Park et al., 2014) and others for site level differential methylation detection.

Different from the count-based data obtained from BS-seq, DNA methylation arrays provide fluorescence intensities that are quantified as the relative level of methylated and unmethylated probes. Specifically, two types of data are used for downstream analyses. The β-value is an estimate of the methylation level using the ratio of intensities between methylated and unmethylated alleles. They range between 0 and 1. Ideally, a value of 0 indicates that all copies of the CpG site in the sample were completely unmethylated, and a value of 1 indicates that every copy of the site was methylated. The β-value is defined below:

$$\beta = \frac{\max(Methylated, 0)}{\max(Methylated, 0) + \max(Unmethylated, 0) + \alpha}.$$

The α in the denominator is used to stabilize the estimate when both the methylated and unmethylated intensities are low. The α value is set to 100 by default. Note that after correcting for background noise, the methylated and unmethylated intensities may have a negative reading. To avoid this, $\max(Methylated, 0)$ and $\max(Unmethylated, 0)$ are used to reset any negative values to 0.

The other commonly used methylation measure is called an M-value. It is calculated as the log_2 ratio of the intensities of methylated probes versus unmethylated probes, as defined below:

$$M = \log_2\left(\frac{\max(Methylated, 0) + \alpha}{\max(Unmethylated, 0) + \alpha}\right).$$

The α (by default equals 1) in the calculation is added in order to prevent unexpected large changes due to small intensity estimation errors. M-values can range from negative infinity to positive infinity. When the methylated and unmethylated probes have the same intensity value, the M-value is 0. Positive M-values indicate more methylation is occurring than not.

The α value in the calculations of both β-value and M-value is typically negligible due to the fact that more than 95% of CpG sites have intensities higher than 1000 (Du et al., 2010) and thus it typically does not have a large impact on the calculated methylation level. The relationship between the M- and β-values can be expressed as:

$$M = \log_2 \frac{\beta}{1-\beta}. \tag{1.1}$$

Although the β-values are useful for interpretation, there are some advantages to using the M-values for statistical analysis, such as homogeneity of variance (Du et al., 2010). Also M-values range from negative infinity to positive infinity, making it more suitable to use statistical methods that have a normality assumption. Thus, M-values are recommended by Du et al. (2010) for conducting differential methylation analysis.

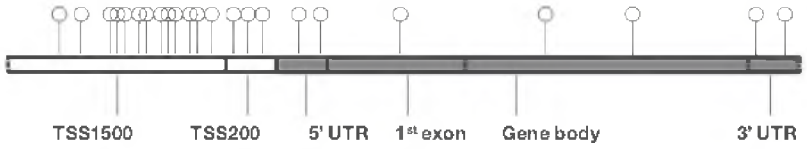
Several statistical methods have been proposed for DNA methylation microarray data to identify cytosine sites with significant differential methylation, including CpGassoc (Barfield et al., 2012), MENT (Baek et al., 2013), IMA (Wang et al., 2012), and COHCAP (Warden et al., 2013). The limma method (Smyth, 2004) is an approach first developed for detecting differential expression in gene expression microarray data, but it can also be used to test for differential methylation in DNA methylation microarray data. This method is further described since it is used in downstream region level analysis employed in this work. For DNA methylation studies, as well as other genomic studies like gene expression, typically only a small number of biological replicates are available. However, the studies are very complex, involving different aspects of biological processes and a large number of variables. It is challenging to find statistically significant and precise features between different conditions. The limma method (Smyth, 2004) tried to solve this problem by fitting a linear model to the M-value of each genomic position, then using empirical Bayes methods to estimate moderated t-test statistics. Global parameters are estimated using all the variables at once, which enables the incorporation of correlated neighboring genomic

features. The empirical Bayes approach is equivalent to shrinkage of the estimated sample variances toward a pooled estimate. It borrows information between probes in order to moderate the residual variances, and ensures that small sample inference can be conducted with reliable and stable results (Ritchie et al., 2015).

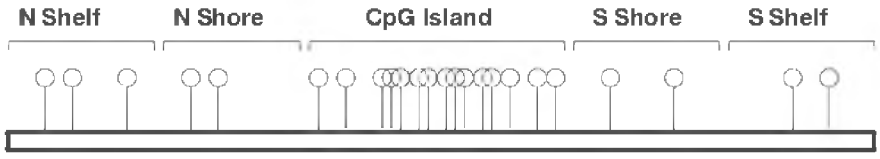
1.5.2. Region Level Testing. While there are benefits of analyzing differential methylation at the site level, there are reasons both biologically and statistically to test differential methylation at the region level. It is shown that strong correlation exists between CpG methylation levels over short distances. This correlation dissipates the further away sites are from each other, such that it is no longer detectable at sites over 1000 bp apart (Eckhardt et al., 2006). Differential methylation targets are most commonly clustered into short regions. So it is meaningful to look at the differential methylation at a region level. Also, when the difference in methylation is small and undetectable at the site level, the persistence in small methylation differences over a region will provide a higher power for detection.

1.5.2.1. Defining regions. There are two ways to define a region when performing region level differential methylation testing. The first approach is to use predefined regions. The density of probes on HM450 data varies across the genome, with higher coverage in the promoter regions of genes and CpG islands (Illumina, 2012), as shown in Figure 1.3. Some differential methylation region (DMR) detection methods, such as IMA, COHCAP and QDMR (Zhang et al., 2011), concentrate on high density areas using predefined regions, compromising only a subset of the HM450 probes. This approach may miss meaningful clusters outside the predefined ones, but it can reduce the number of tests that need to be accounted for when controlling the false discovery rate.

The second way to define regions is to use a post-hoc aggregation method based on the data. After conducting the initial analysis on each cytosine site, probes are included in a region if they have significant site level differential methylation and are within a certain



Feature Type	Genes Mapped	Percent Genes Covered	Number of Loci on Array
NM_TSS200	14895	0.79	2.56
NM_TS1500	17820	0.94	3.41
NM_5'UTR	13865	0.78	3.34
NM_1stExon	15127	0.80	1.62
NM_3'UTR	13042	0.72	1.02
NM_GeneBody	17071	0.97	8.97
NR_TSS200	1967	0.65	1.84
NR_TSS1500	2672	0.88	2.92
NR_GeneBody	2345	0.77	5.34



Feature Type	Islands Mapped	Percent Islands Covered	Average Number of Loci on Array
Island	26153	0.94	5.08
N_Shore	25770	0.93	2.74
S_Shore	25614	0.92	2.66
N_Shelf	23896	0.86	1.97
S_Shelf	23968	0.86	1.94

Figure 1.3. HumanMethylation450 BeadChip coverage of different regions. Figure from Illumina (2012).

distance of other significant sites. Bumphunter (Jaffe et al., 2012) and DMRcate (Peters et al., 2015) both use this approach to define a region. These methods are described in more details below since they will be utilized in this work.

1.5.2.2. Testing methods. Many DMR testing methods have been developed, such as Bumphunter, DMRcate and ProbeLasso (Butcher and Beck, 2015). In this study, Bumphunter and DMRcate are used since they are the most commonly used statistical methods for DMR detection in HM450 data. These methods are briefly described in this section

and further details are provided in Chapter 2. Bumphunter is a data analysis pipeline developed to identify DMRs associated with disease (Jaffe et al., 2012). A statistical model is developed to take into account batch effects, which are a potential problem in large scale and high throughput studies with many samples. Batch effects are the unnoticed correlation between subgroups of samples which may be caused by experimental environments, such as the temperature and humidity. First, a linear regression model regressing the methylation value on the group status is applied to model differential methylation between the case and the control groups at each CpG site. This model can also incorporate batch effects. The slope coefficient corresponding to the group variable is then smoothed using loess. Clusters of consecutive probes for which coefficients higher than a predetermined threshold are identified as candidate regions (bumps). Permutation tests, which permute sample labels to create a null distribution of candidate regions, are then conducted to estimate the statistical significance of the candidate DMRs.

DMRcate is a data-driven approach that can be used with WGBS data as well as HM450 array data (Peters et al., 2015). First, a linear model is fit at each CpG site using the limma method (Smyth, 2004). A Gaussian kernel with bandwidth λ is used to smooth the estimated test statistics. The Gaussian kernel is calculated with a standard deviation $\sigma = \frac{\lambda}{C}$, where C is a scaling factor for the bandwidth. Smoothed test statistics are then modeled and a p-value is calculated for each site. DMRs are defined by grouping the significant CpG sites that are at most λ nucleotides from each other.

1.6. MOTIVATION

Human genomes are complex and are regulated at multiple levels. Various types of genomic data offer different aspects of complicated biological processes. Due to recent advances in high-throughput technologies, multiple types of genomic data (e.g. gene expression, methylation, SNP) can be collected on the same individual. The Cancer Genome Atlas (TCGA) is one of the most comprehensive cancer genomics programs. TCGA hosts

a database with genomic sequence, expression, methylation, and copy number variation data on over 11,000 individuals, with samples in over 30 different types of cancer. TCGA is led by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to map genomic and epigenomic changes with the goal of accelerating new discoveries in cancer-related research and to improve the prevention and treatment of cancer (Wang et al., 2016). The International Cancer Genome Consortium (ICGC) is another genomic data consortium, which provides data on genomic, transcriptomic and epigenomic abnormalities, as well as somatic mutations in over 50 different cancer types (Hudson et al., 2010).

Integrating and combining multiple types of genomic data can provide researchers with deeper insights into complex biological processes and help scientists reveal disease mechanisms for exploration, prevention and treatment. However, it is challenging to combine these different layers of information. This study focuses on the integration of DNA methylation and SNP data to maximize the utilization of genomic information and improve biologically meaningful discoveries. When analyzing DNA methylation data, SNP probes are filtered out in a preprocessing step based on the population minor allele frequency and their location relative to the target sites. Any potential SNP is filtered out for all individuals due to the potential issues they pose for measuring methylation accurately. However, for each individual, one may or may not have the specific allele associated with the SNP. When SNP data are available, researchers can use the information to recover probes that are not affecting the quality of the methylation array. For those probes that are actually influenced by SNPs, imputation methods are proposed and investigated. Missing data are a common issue in different areas, including biology, genomics, social sciences and financial studies. Handling the missing value problem simply by deleting the missing instances can result in losing useful information. Simple solutions such as replacing the missing value with the mean may falsely lower the variability. This research will develop sophisticated imputa-

tion methods based on the specific data structure. The nature of DNA methylation data and the correlation between neighboring sites will be considered when imputing missing methylation values.

In this dissertation, imputation methods are proposed based on the characteristics of DNA methylation data and these are described in Chapter 2. Predictive models built under the regularized linear regression framework are used to impute the missing values. The shrinkage approaches for the regression models include ridge, LASSO and elastic net. Moreover, functional principal component analysis is applied to perform linear models as an alternative imputation method. The proposed methods are evaluated and compared to existing imputation methods via a simulation study and analysis of real data, described in Chapter 3. The natural structure of the DNA methylation data is retained by using the real data when conducting the simulation. Two types of DMRs are investigated to mimic the methylation patterns in human genome. The performance of the imputation methods are assessed in terms of imputation accuracy and DMR detection ability. Finally, in Chapter 4, a summary of the work and discussion of future research is provided.

2. DATA AND METHODS

2.1. DATA

2.1.1. Data. In this research, Infinium HumanMethylation450 BeadChip (HM450) DNA methylation data on breast cancer patients were obtained from The Cancer Genome Atlas (TCGA) (TCGA, 2021). Measurements of methylation levels on 485,577 CpG sites were given for the normal tissue and tumor tissue of 86 individuals. For 3 individuals, methylation data were only available for the tumor tissue. The raw data (provided in .idat files) were downloaded using the DTT UI from National Institutes of Health, the user interface (UI) design version of the Data Transfer Tool (DTT). Single-nucleotide polymorphism (SNP) data on "Pathogenic Germline Variants in 10,389 Adult Cancers" (Huang et al., 2018) were acquired from Genomic Data Commons (GDC). BCFtools (Li et al., 2009), a tool to process binary variant call format (BCF) and variant call format (VCF) files, was used to extract information on the 89 individuals for this study from the compressed VCF file of the combined variant calls.

The methylation data were processed through the Chip Analysis Methylation Pipeline (ChAMP) (Tian et al., 2017) in Bioconductor version 3.12 and R version 3.6.3. In addition to the raw .idat files, a table stating the sample names and their treatment groups is also required. The treatment group is acquired from the "Sample" code of each sample's TCGA barcode (Figure 2.1). In the code, '01' indicates the tumor sample type and '11' is the normal sample type.

Prior to analyzing the DNA methylation data, several pre-processing steps are needed. One of these steps involves filtering out probes for different reasons, as described in Chapter 1. First, the data are filtered based on the detection p-values. Detection p-values measure the likelihood that the total intensity of the probes is generated by a background

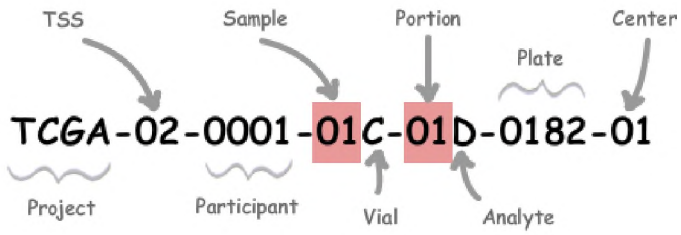


Figure 2.1. The Cancer Genome Atlas (TCGA) barcode label explanation. Figure obtained from (TCGA, 2011). 'TSS' is short for tissue source site.

distribution. If the detection p-value is below a specified significance level, it means the observed value of total intensity is unlikely to be generated by the background noise, thus the probe is considered detected (Heiss and Just, 2019). In this study, any probe with detection p-values greater than 0.01 are filtered out (Hernandez-Vargas et al., 2010). The percentage of problematic probes (based on the detection p-values) for each sample is also monitored. When the proportion is above 0.1, the sample will be removed from the analysis. Probes with less than three beads in at least 5% of samples per probe are also filtered out. All non-CpG probes and multi-hit probes are removed due to potential cross-hybridization and misinterpretation they may cause on methylation levels. In this research, the probes on the sex chromosome are kept because all the samples are females. If the samples are from both sexes, the probes located in chromosomes X and Y are suggested to be filtered out to exclude possible sex bias (Ma et al., 2013). A normalization step is also required before further analysis in order to adjust the bias caused by probe types and technical variation. In this work, the peak-based correction (PBC) normalization (Dedeurwaerder et al., 2011) is applied to the datasets before DMR detection.

Both a real dataset and a simulated dataset are used to study and evaluate proposed imputation methods with respect to the potential SNP probes. To improve the computation time without losing generality, in the real data analysis, a piece on the genome of a reasonable length is considered for the analysis. A total of 7,987 probes with genomic locations between

1 and 13,800,000 base pairs on Chromosome 1 is considered. In a typical analysis, the probes with a potential SNP are also recommended to be filtered out based on the list provided by Zhou et al. (2016). After performing the standard filtering criteria (including potential SNP probes), 6,838 probes are remaining for analysis on this segment. This set of probes is referred to as the incomplete dataset in the real data analysis.

The SNP filtering is based on the potential for an individual to have a SNP at a particular location based on population data. However, many individuals will not have a SNP but rather have the common variant in the population. In this research, SNP data are integrated into the filtering phase and each sample is checked to determine if they actually have a SNP or have the common variant at each potential location. This allows the recovery of probes that do not have an actual polymorphism. By integrating the SNP data of all the individuals into the filtering process, a dataset with the most information available is generated. This is called the complete data, which contain 7,668 probes. After this step, a list of DNA methylation probes that are actually affected by SNPs is obtained. These probes can not be recovered since they actually contain true SNPs. Imputation methods are developed to fill these positions in order to improve downstream analysis such as differential methylation region (DMR) detection.

2.1.2. Characteristics of DNA Methylation Data. Methylation is not a random process. Researchers have found that close neighboring CpG sites are likely to share the same methylation status (Sun et al., 2019). That is, the DNA methylation level of a given site is highly correlated with the methylation levels of neighboring probes. Neighboring probes are defined in terns of their physical proximity based on their genomic location on the chromosome (i.e., how far away in base pairs (bp) the sites are from each other). This phenomenon could be due to the working distance range of DNA methyltransferase (DNMT) in changing the methylation status of CpG sites (Jia et al., 2007). DNMT transfers the methyl group to DNA and could methylate two CpGs within its working distance range

in one binding event. The correlation may also be due to the influence of the nearby CpG sites in the recruitment of DNA methyltransferase or demethylase enzymes (Lövkvist et al., 2016). Demethylase enzymes remove the methyl group from methylated CpG sites.

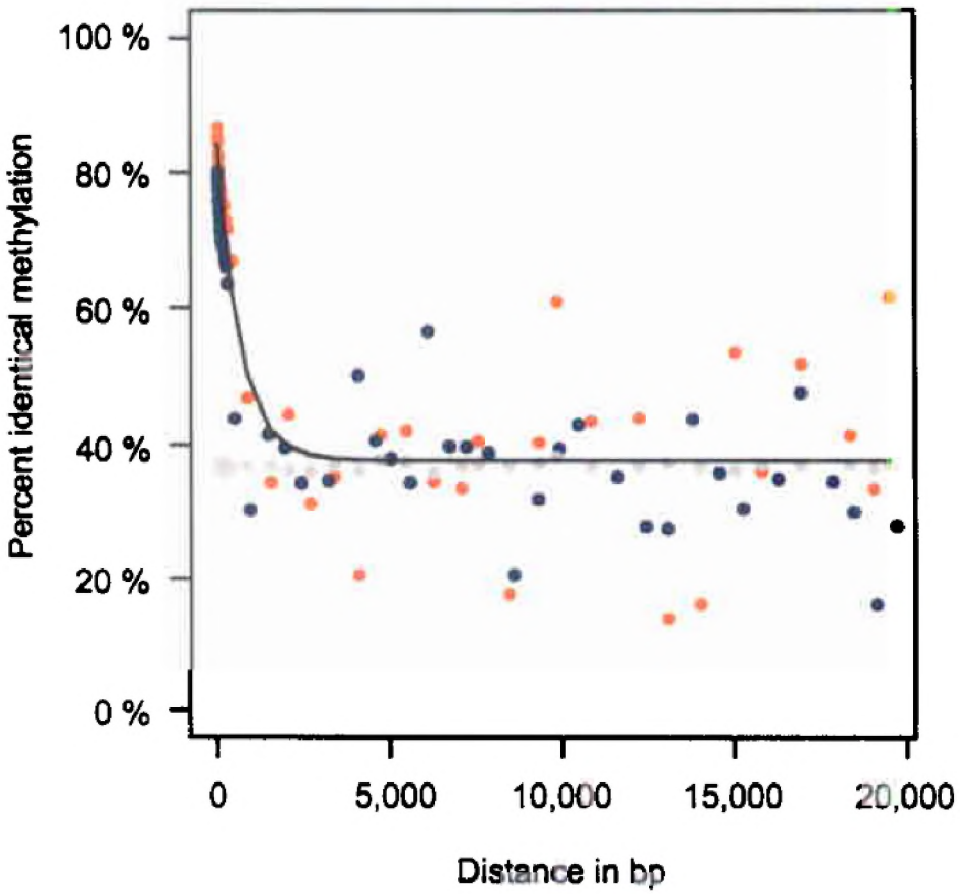


Figure 2.2. Correlation between co-methylation and spatial distance in genomic base pairs (bp). The methylation values represented by the orange dots are averaged over 25,000 individual measurements. Grey dots represent CpG methylation values based on re-sampling of random CpG positions. Blue dots represent CpG methylation values based on re-sampling of amplicons (obtained by PCR amplification). Correlation between CpG methylation and spatial distance is not detectable at distances larger than 1,000 bp. Figure from Eckhardt et al. (2006).

Eckhardt et al. (2006) mentioned that the methylation level of nearby probes have a significant correlation over short (up to 1,000 bp) distances. As shown in Figure 2.2, the correlation decreases rapidly at 1,000 bp and beyond. In the following statistical analysis

steps, it is important to consider the genomic position of each probe and the potential of neighboring correlation when performing model building, variable selection and missing value imputation.

2.2. MISSING VALUE PROBLEM

Missing data can happen for many reasons that are beyond the control of the experimenter. For example, missing data can arise due to technical errors, improper data collection, respondents refusing to answer (e.g., in surveys) or participants that drop out. For array based genomic technologies, missing values may be caused by image corruption or low resolution (Troyanskaya et al., 2001). Missing values in DNA methylation are introduced due to various reasons, such as filtering out probes with low detection p-values or low bead count, as well as removing multi-hit and non-CpG probes (Maksimovic et al., 2012). One of the main sources of missing DNA methylation data is due to filtering out SNP probes. Addressing this issue is the main focus of this work.

The missing data mechanisms are categorized into three classes by Rubin (1976). If the probability of a missing value's occurrence is the same for all cases, and there is no correlation between the missing and the observed data, the missing data are called missing completely at random (MCAR). If the probability of being missing is the same within a group of the data, then the missing data are called missing at random (MAR). MAR means that there might be systematic differences between the missing and observed data, but these differences can be explained by some observed variables. For example, in a clinical trial studying blood pressure, some records are missing. People using manual blood pressure monitors tend to not record their blood pressure reads, whereas the reads can easily be recorded and stored for automatic digital blood pressure monitor users. The missing percentage is different between people using different devices, but it can be explained by a variable separating the two groups. Moreover, the variable is not related to the value that is missing. A violation of the rule would be when people with higher blood pressure record

their results but people with normal blood pressure don't record as often. The missing rate is related to the blood pressure values. This situation is the third type of missing data mechanism, missing not at random (MNAR). Missing data are classified as MNAR when there is a relationship between the tendency of a value to be missing and its value.

The underlying missing data mechanism has an impact on the choice and performance of different imputation methods. However, for real data, the missing mechanism cannot be revealed by studying the data itself. In order to make reasonable assumptions, knowledge of the data and the data collection process is required. Due to the randomness of experimental and technological errors, MCAR/MAR is assumed for the HM450 data. For example, the missing value is higher at the positions where probes fail to capture target sequence, but the missing pattern is independent of the value itself (Lena et al., 2019a).

2.3. IMPUTATION METHODS

2.3.1. Traditional Solution for Handling Missing Values. One of the most frequently used solutions for the missing value problem is listwise deletion. In this approach, all cases with missing values will be omitted from an analysis. This default method is convenient, but it reduces the sample size radically and can waste potentially useful information present in the deleted entries. For example, an individual answering a survey many answer only part of the questions and useful information may be contained in the subset of data that is available for that subject. Also, when the missing pattern is not MCAR, listwise deletion introduces large bias to the estimated mean (Little and Rubin, 2002).

Depending on the data type, there are several other convenient approaches to address missing values. If the variable is quantitative, the missing data can be replaced by averaging the non missing observations of the variable over all samples. If the missing value is qualitative, the mode of the non-missing observations can be used. However, mean imputation is problematic because it will shrink the standard deviation of the original distribution, and disturb the relationship between variables (Van Buuren, 2018). For example, values that are

imputed by a variable's mean have almost no relationship with other variables. As a result, the correlation between variables is biased toward zero, which is not wanted in data analysis. In this research, the mean imputation method is applied on the missing methylation data to provide a basis for comparison to a commonly used and easy approach. For each CpG site, the missing values among the samples are replaced with the mean over the methylation values of the non-missing samples of the same group for that site. As a result, if a CpG site has more than one missing entry, they will have the same imputed value.

For longitudinal data, last observation carried forward (LOCF) and baseline observation carried forward (BOCF) are widely used to address missing values, especially in clinical trials. In the LOCF method, the value at the last time point prior to a subject dropping out is used as the imputed value for all later time points in the study. Assumptions of LOCF are strong, including the assumption that observations do not change when data are missing and that a single data point can be used to estimate a distribution of potential values (Molenberghs et al., 2014). BOCF instead uses the baseline value of an individual as the imputed value for any missing data in the time sequence. It is suggested that the effect of actual outcomes as well as the reason for the missing values should be investigated before choosing this approach (Liu-Seifert et al., 2010).

2.3.2. KNN Imputation. K-nearest neighbors (KNN) imputation was proposed by Troyanskaya et al. (2001) to handle missing values in gene expression microarray data. The gene expression data are arranged in a matrix with genes in the rows and samples or experiments in the columns. For a gene with a missing expression value in sample i, K other genes will be found, which have an expression value in sample i. The missing value is estimated by a weighted average of the selected K gene expression values in sample i. The weight is decided by the similarity of each gene to the gene with missing value. It is found that the method is insensitive to the exact number of K within the range of 5-20. Similarities are measured by calculating the Euclidean distance between two genes using the rest of samples other than sample i.

The idea of this method is applicable to DNA methylation imputation because it uses the summary value from similar genes. With methylation data, the missing value is estimated by a weighted average of methylation levels from K=10 CpG sites in the same sample with the missing value. The KNN method will be applied in this work as one of the comparison methods since it is a current approach in genomic literature. A drawback of the KNN method is when the missing rate is very high at a particular CpG site, the method fails when all the neighbors are missing in a particular position.

2.3.3. MethyLImp. In 2019, Lena et al. proposed a linear regression model for missing value imputation specifically for DNA methylation data. The idea aims to capture the correlation between methyation levels of CpG sites by a linear regression model. The missing values are imputed by iteratively performing linear regression on the available data. The methylation data are organized in matrix form, with each methylation probe (CpG site) being treated as a column variable and the rows correspond to each sample.

In the first step of methyLImp (Lena et al., 2019a), the first CpG site with missing values is found. The method also searches for other CpG sites which have missing values in the same samples as the first CpG site. The non-missing values for this site (or sites) are denoted as $\mathbf{Y_1}$ and the missing values as $\mathbf{Y_2}$. The imputation goal is to find the estimates of $\mathbf{Y_2}$. Next, the cases with no missing value in the submatrix with only samples in $\mathbf{Y_1}$ is denoted as $\mathbf{X_1}$. $\mathbf{X_2}$ is the matrix whose entries include the same columns (sites) as $\mathbf{X_1}$ and same rows (samples) as $\mathbf{Y_2}$. An illustration of how these matrices are defined can be found in Figure 2.3. Any column (site) with missing values not included in $\mathbf{Y_1}$ and $\mathbf{Y_2}$ are not used in that specific imputation iteration. In this example, $\mathbf{Y_1}$ is a two-column matrix instead of a vector. The missing values in site 1 and site 6 will be imputed together because samples with missing values are the same for those two sites.

	Site1	Site2	Site3	Site4	Site5	Site6	Site7	Site8	Site9	Site10	Site11
Sample 1	0.1287	0.1276	0.1223	0.1068	0.0395	0.0838	0.1038	0.0668	0.1917	NA	0.2502
Sample 2	0.0785	0.0920	0.0689	0.0633	0.0280	0.1500	0.0635	0.0441	0.1723	0.0974	0.1109
Sample 3	0.6056	0.6791	0.4810	0.6137	0.0743	0.1673	NA	0.4409	0.1163	0.5679	0.5585
Sample 4	NA	0.0952	0.0878	0.0854	0.0206	NA	0.1032	0.0794	0.1773	NA	0.2697
Sample 5	0.1842	0.2200	0.1935	0.1949	0.0399	0.1438	0.1788	0.1229	0.1855	0.2595	0.3443
Sample 6	NA	0.1301	0.1148	0.1290	0.0526	NA	0.1137	0.0593	0.1675	0.1777	0.2275

$$Y_1 = \begin{bmatrix} 0.1287 & 0.0838 \\ 0.0785 & 0.1500 \\ 0.6056 & 0.1673 \\ 0.1842 & 0.1438 \end{bmatrix}$$

$$X_1 = \begin{bmatrix} 0.1276 & 0.1223 & 0.1068 & 0.0395 & 0.0668 & 0.1917 & 0.2502 \\ 0.0920 & 0.0689 & 0.0633 & 0.0280 & 0.0441 & 0.1723 & 0.1109 \\ 0.6791 & 0.4810 & 0.6137 & 0.0743 & 0.4409 & 0.1163 & 0.5585 \\ 0.2200 & 0.1935 & 0.1949 & 0.0399 & 0.1229 & 0.1855 & 0.3443 \end{bmatrix}$$

$$X_2 = \begin{bmatrix} 0.0952 & 0.0878 & 0.0854 & 0.0206 & 0.0794 & 0.1773 & 0.2697 \\ 0.1301 & 0.1148 & 0.1290 & 0.0526 & 0.0593 & 0.1675 & 0.2275 \end{bmatrix}$$

Figure 2.3. Example of the matrix definitions from the methyLImp method.

To address the limited range of the β-values between 0 and 1, a logit function $logit(p) = \log(\frac{p}{1-p}), p \in [0, 1]$ is applied on the $\mathbf{Y_i}$'s and the model is set up as:

$$logit(\mathbf{Y}) = \mathbf{X} \cdot \alpha + \epsilon. \tag{2.1}$$

Here $\mathbf{Y}$ corresponds to the $\mathbf{Y_1}$ matrix, $\mathbf{X}$ is the $\mathbf{X_1}$ matrix, α are the regression coefficients and ϵ is the error term. The error term is assumed to be independently and identically distributed with a normally distribution. The coefficients α of the regression model are estimated by using the pseudo-inverse of $\mathbf{X}$:

$$\hat{\alpha} = \mathbf{X}^{-1} \cdot logit(\mathbf{Y}).$$

The pseudo-inverse $\mathbf{X}^{-1}$ is computed using the singular value decomposition (SVD) of $\mathbf{X}$ (Golub and Reinsch, 1970).

The SVD of an $N \times p$ matrix $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{UDV}'$$

where $\mathbf{U}$ is an $N \times p$ matrix and $\mathbf{V}$ a $p \times p$ matrix. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal to each other. $\mathbf{D}$ is a $p \times p$ diagonal matrix with entries $d_1 \geq d_2 \geq ... \geq d_p$, and the entries are called the singular values of $\mathbf{X}$.

Then, the Moore-Penrose pseudo-inverse (Penrose, 1955) has the form:

$$\mathbf{X}^{-1} = \mathbf{VD}^{-1}\mathbf{U}'$$

where $\mathbf{D}^{-1}$ is formed from $\mathbf{D}$ by taking the reciprocal of all the non-zero entries and leaving the zeros as they are.

After obtaining $\hat{\alpha}$, the estimates of α, the missing methylation levels are predicted as follows:

$$\hat{\mathbf{Y}}_2 = logit^{-1}(\mathbf{X_2} \cdot \hat{\alpha}).$$

In summary, the MethyLImp method uses $\mathbf{X_1}$ and $\mathbf{Y_1}$ to build a regression model and then predicts the missing values in $\mathbf{Y_2}$ by fitting $\mathbf{X_2}$ in the model. An R-package implementing this method called 'methyLImp' is available at GitHub (Lena et al., 2019b). Lena et al. (2019a) compared methyLImp to existing methods including mean, KNN, SVDmiss (Fuentes et al., 2006), softImpute (Mazumder et al., 2010), imputePCA (Husson and Josse, 2013) and missForest (Stekhoven and Bühlmann, 2012). MethyLImp was shown to perform equally or better than these methods and with good computational efficiency. The imputation methods proposed in this work are compared to methyLImp since it is the primary imputation method for DNA methylation data available.

2.4. REGULARIZED LINEAR REGRESSION IMPUTATION

The number of CpG sites in DNA methylation data is much larger than the sample size. For example, HM450 methylation data has over 450,000 probes but the sample size in most studies is usually limited to around 100 individuals. When using a regression model to impute missing values, this high dimensional problem is not negligible in the models. A common problem of models with a large number of variables is multicollinearity. Multicollinearity is the condition where two or more predictor variables in a statistical model are linearly related (Dormann et al., 2013). The existence of multicollinearity can result in increased variance of regression coefficients, which will lead to unstable estimation of parameter values. For least squares regression, the regression coefficients α are estimated as $\hat{\alpha} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where $\mathbf{X}$ is the design matrix and $\mathbf{Y}$ is the response vector. When the columns of the design matrix $\mathbf{X}$ are highly correlated, $\mathbf{X}'\mathbf{X}$ is almost singular, leading to the instability of $\hat{\alpha}$ with small changes in the data.

Regularized linear regression imputation is proposed in this research to deal with the issues posed by high dimensional data. The regularization approach involves adding a constraint to the loss function. A loss function is used to penalize the prediction errors when fitting the model (Hastie et al., 2009). For example, assume f is the function to predict $\mathbf{Y}$ based on the input $\mathbf{X}$. A convenient loss function is the squared error loss:

$$l = (\mathbf{Y} - f(\mathbf{X}))^2. \tag{2.2}$$

Least square estimators are obtained by minimizing the squared error loss function. Regularization methods involve adding different penalty terms to the loss function, which prevent coefficients from taking unreasonable values and help with the risk of overfitting. Three different regularization methods (ridge regression, LASSO and elastic net) are explored in this work for incorporation into the imputation process as described below.

The imputation methods in this research iteratively evaluate the complete subset of variables with missing entries. The iteration starts with the first CpG site measured on a chromosome and works sequentially through the genomic locations. Here, the variables are the methylation levels at different CpG sites. The input data is organized as each row representing a sample and each column representing a probe for a particular CpG site as shown in Figure 2.3. Several lists are generated: the list of all sample names S, all column names C, and the names of columns with missing values V. The first step is to find the variables with missing values V_1 in V. Denote the list of row names of those missing positions of V_1 as S_{NA}. $\mathbf{Y_1}$ is the vector or data matrix with columns V_1 and rows in S but not in S_{NA}. If the missing value positions are exactly the same for more than one variable, $\mathbf{Y_1}$ will be a matrix instead of a vector. $\mathbf{X_1}$ is the submatrix with the same rows as $\mathbf{Y_1}$ and the columns in C but not in V. $\mathbf{X_2}$ is the submatrix with the samples in S_{NA} and same columns as in $\mathbf{X_1}$. Finally, $\mathbf{Y_2}$ is the vector or submatrix of missing values at variables V_1 and samples in S_{NA}. After obtaining $\mathbf{Y_1}$, $\mathbf{X_1}$, and $\mathbf{X_2}$ in the first iteration, methods are applied to fit a generalized linear model with $\mathbf{Y_1}$, $\mathbf{X_1}$, and then predict $\mathbf{Y_2}$ by feeding $\mathbf{X_2}$ into the fitted model. After the missing value(s) are imputed at variable(s) V_1, the lists S, C, V and S_{NA} are updated accordingly. The algorithm will search for the next variable(s) with missing values and the complete samples, and construct new vectors or matrices $\mathbf{Y_1}$, $\mathbf{X_1}$, and $\mathbf{X_2}$. In this step, the imputed values for variable V_1 will be treated as complete entries. The iteration will stop when S and V are empty, meaning all the missing values have been imputed.

This work is inspired by methyLImp to impute missing values in DNA methylation data by utilizing a regression model. When forming the sub-matrices for imputation in the previous steps, the number of variables is large compared to the number of samples available, resulting in a high dimensional data problem. The proposed methods in this research use regularization when fitting the model instead of using the pseudo-inverse of the design matrix, which is not unique. To address the issues posed by high dimensional

data, regularization (also known as shrinkage) approaches are used to shrink estimated coefficients towards zero relative to the least squares estimates to reduce the variance and help prevent overfitting (Hastie et al., 2009). Depending on the type of shrinkage approach, some of the coefficients may be estimated to be exactly zero. In this work, the two best-known regression regularization techniques (ridge regression (Hoerl and Kennard, 1970) and the LASSO (Tibshirani, 1996)) are employed in the imputation step that involves fitting a generalized linear model. An additional approach, the elastic net (Zou and Hastie, 2005), is also explored that combines the ridge and LASSO methods.

2.4.1. Ridge Regression. Ridge regression was proposed by Hoerl and Kennard (1970). By adding a small constant value λ to the diagonal entries of $\mathbf{X'X}$, the least square estimator's stability can be improved. The ridge regression estimator is $\hat{\alpha}^{ridge} = (\mathbf{X'X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X'Y}$. Ridge regression shrinks the regression coefficients by imposing a penalty on their size:

$$\hat{\alpha}^{ridge} = \arg\min_{\alpha}\{\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 + \lambda \sum_{j=1}^{p} \alpha_j^2\}. \tag{2.3}$$

The first component in Equation 2.3, $\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2$, is typically called the residual sum of squares or the Sum of Squares Error (SSE) and it represents the squared error loss described previously. The second component of Equation 2.3, $\lambda \sum_{j=1}^{p} \alpha_j^2$, is referred to as the penalty term, which performs the shrinkage of the coefficients. $\lambda \geq 0$ is the tuning parameter that controls the amount of shrinkage. The larger the value of λ, the more shrinkage towards zero is applied to the coefficients. Equation 2.3 can equivalently be written as:

$$\hat{\alpha}^{ridge} = \arg\min_{\alpha} \sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 \tag{2.4}$$

subject to l_2 penalty

$$\sum_{j=1}^{p} \alpha_j^2 \leq t$$

with a one-to-one correspondence between λ and t. This approach works via the trade-off between bias and variance. A small bias is allowed in the coefficient estimates to reduce the variance and make the estimates more stable.

The R package 'glmnet' is used to do the regularization by the cyclic coordinate descent (CCD) method, which was developed by Friedman et al. (2010). To determine the tuning parameter λ, 100 values are generated. Two parameters are needed for the sequence of λ. The first one is the largest value for λ such that all the coefficients are zero (denoted λ_{max}). Note that $\lambda_{max} = \infty$ for ridge regression, so a value is picked corresponding to the coefficients close to zero. The second one is a pre-determined ratio of the smallest value of the generated λ sequence to λ_{max}. When the number of samples is greater than the number of probes in the model, the ratio is set to be 0.0001. In this study, the number of probes is greater than the number of samples, thus the ratio is set to be 0.01 to increase the penalty for complexity. λ_{min} is obtained by $\lambda_{min} = 0.01 \cdot \lambda_{max}$. The ten-fold validation error for each λ value is computed. The tuning parameter with the smallest cross-validation error is used to fit the model for each iteration.

2.4.2. LASSO Regression. The least absolute shrinkage and selection operator (LASSO) was proposed by Tibshirani (1996), which is an l_1 penalized least squares method for linear models. The residual sum of squares (SSE) is minimized with a constraint that the sum of the absolute values of the coefficients is less than a constant. This approach is similar to ridge regression, but the use of the l_1 penalty instead of the l_2 penalty can force certain coefficients to be zero. This is different from ridge regression which never sets the value of coefficients to be exactly zero. Ridge regression can be challenging for model interpretation, whereas LASSO yields a sparse model that results in variable selection by identifying the predictors with non-zero coefficients.

The LASSO estimate is defined by:

$$\hat{\alpha}^{LASSO} = \arg\min_{\alpha}\{\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 + \lambda \sum_{j=1}^{p} \mid \alpha_j \mid\}, \tag{2.5}$$

where λ is the parameter that controls the shrinkage. Equation 2.5 can also be written as:

$$\hat{\alpha}^{LASSO} = \arg\min_{\alpha} \sum_{i=1}^{N} (y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 \tag{2.6}$$

subject to l_1 penalty

$$\sum_{j=1}^{p} \mid \alpha_j \mid \leq t$$

with a one-to-one correspondence between λ and t. This makes the solutions nonlinear in y_i and there is no closed form expression as in ridge regression, so the minimization problem needs to be solved analytically. Efficient algorithms have been developed to compute the estimates of LASSO, such as the least angle regression algorithm (Efron et al., 2004). The same tuning procedures are adopted from ridge regression to choose the optimal value for λ with ten-fold cross validation from a sequence of 100 generated λ values. To generate the 100 values, λ_{max} is selected to be the value that makes all the coefficients zero.

To visualize differences in estimation for LASSO and ridge regression, consider the simple case when there are two variables with corresponding coefficients α_1 and α_2. LASSO has the constraint function $\mid \alpha_1 \mid + \mid \alpha_2 \mid \leq t$. This implies that LASSO coefficients have the smallest loss function for all points that lie within the square, given by $\mid \alpha_1 \mid + \mid \alpha_2 \mid \leq t$. Ridge regression has the constraint function $\alpha_1^2 + \alpha_2^2 \leq t$. Figure 2.4 shows the shape of the constraint regions for LASSO (square) and ridge regression (circle), along with the contours of the residual sum of squares. If the sum of squares hits one of the corners of the square, then the coefficient corresponding to the axis is shrunk to zero.

If some of the probes have no correlation with the true methylation levels at the specified sites, LASSO outperforms ridge regression by shrinking the coefficients of those probes to zero. One limitation of LASSO occurs when there are two or more highly correlated sites and LASSO randomly selects one of them. In methylation data, multiple

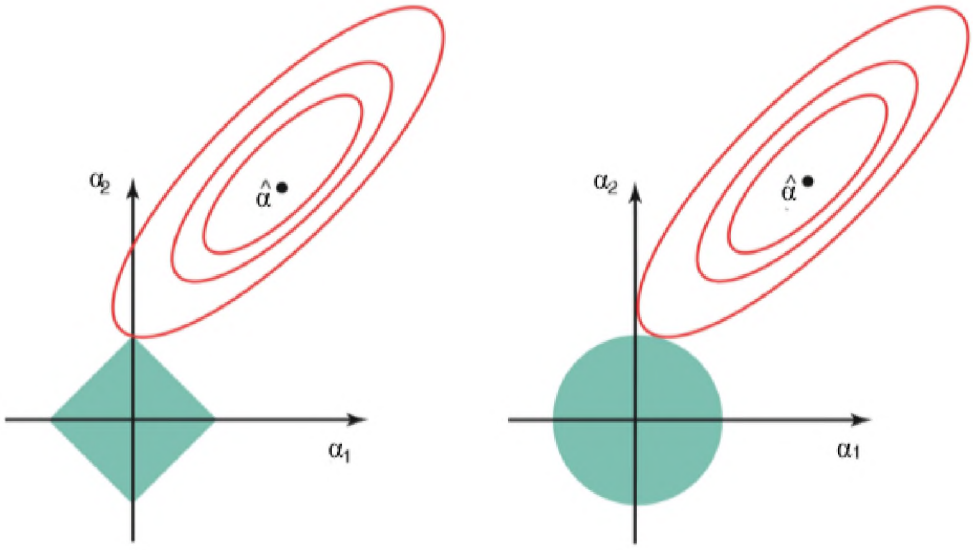


Figure 2.4. Estimation picture for LASSO (left) and ridge regression (right). The green areas are the constraint functions and the red curves are the contours for the least squares error functions. Figure modified from (Hastie et al., 2009).

CpG sites work together on a biological process and the correlation among them should be high. LASSO will only pick one site in the same group, making the model less interpretable by researchers since potentially important sites are filtered out.

2.4.3. Elastic Net Regression. A compromise between ridge and LASSO was proposed by Zou and Hastie (2005) as the elastic net penalty. The elastic net estimate is:

$$\hat{\alpha}^{enet} = \arg\min_{\alpha} \{\sum_{i=1}^{N}(y_i - \alpha_0 - \sum_{j=1}^{p} x_{ij}\alpha_j)^2 + \lambda \sum_{j=1}^{p}(\theta_1 \mid \alpha_j \mid + \theta_2\alpha_j^2)\} \tag{2.7}$$

where

$$\theta_1 + \theta_2 = 1.$$

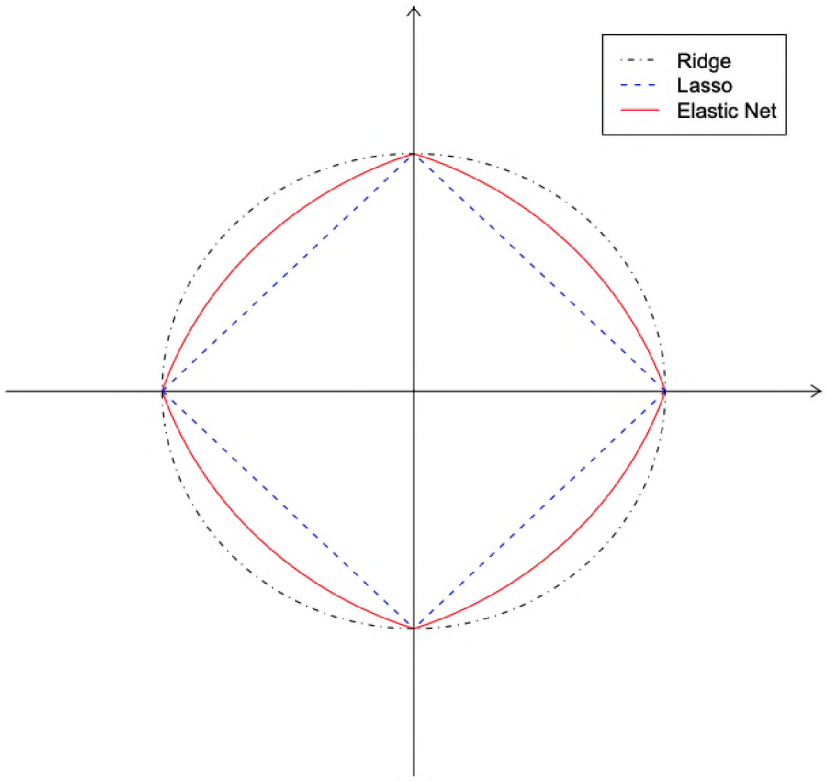


Figure 2.5. Comparison of the constraint functions for ridge, LASSO and elastic net regression. The solid red line represents elastic net regression when $\theta_1 = \theta_2 = 0.5$. Figure from Zou and Hastie (2005).

The l_1 penalty part of the elastic net generates a sparse model and the l_2 penalty part removes the limitation of LASSO that can only select a limited number of variables. Thus, important variables that work together can be included in the model together. Moreover, it stabilizes the l_1 regularization path. Compare the matrix form of the elastic net estimator:

$$\hat{\alpha}^{enet} = \arg\min_{\alpha} \alpha'(\frac{\mathbf{X'X} + \lambda_2\mathbf{I}}{1 + \lambda_2})\alpha - 2\mathbf{y'X}\alpha + \lambda_1 \mid \alpha \mid_1 \tag{2.8}$$

and the lasso estimator:

$$\hat{\alpha}^{LASSO} = \arg\min_{\alpha} \alpha'(\mathbf{X'X})\alpha - 2\mathbf{y'X}\alpha + \lambda_1 \mid \alpha \mid_1 \tag{2.9}$$

where λ_1, λ_2 are two fixed non-negative numbers. Note that $\theta_1 = \frac{\lambda_2}{\lambda_1+\lambda_2}$ and $\theta_2 = \frac{\lambda_1}{\lambda_1+\lambda_2}$. Denote $\hat{\Sigma} = \mathbf{X}'\mathbf{X}$, which is the sample version of the correlation matrix Σ. Notice the following term in Equation 2.8:

$$\frac{\mathbf{X}'\mathbf{X} + \lambda_2\mathbf{I}}{1 + \lambda_2} = (1 - \gamma)\hat{\Sigma} + \gamma\mathbf{I}$$

where $\gamma = \lambda_2/(1 + \lambda_2)$ shrinks $\hat{\Sigma}$ towards the identity matrix $\mathbf{I}$. Equations 2.8 and 2.9 show that the elastic net penalization is equivalent to replacing $\hat{\Sigma}$ with its shrunk version in the LASSO.

A mixing parameter determines the type of penalty for regularization. As shown in Figure 2.5, the elastic net penalty is a mixture of the ridge and LASSO penalties. The mixing parameter is zero for ridge regression, one for LASSO regression and can vary between 0 to 1 for elastic net regression. In this research, the elastic net regularization method with the mixing parameters 0.2, 0.5 and 0.8 are explored. The same procedures to select the tuning parameter (λ) are adopted from ridge regression. The optimal value for λ is chosen using ten-fold cross validation from a sequence of 100 generated λ values generated.

2.4.4. Summary. In this research, linear regression with regularized methods are proposed as imputation methods for missing DNA methylation data. As mentioned earlier in Section 2.4, for each CpG site with missing values, $\mathbf{Y_2}$ is imputed by generating predictions from applying $\mathbf{X_2}$ to model that was fit using $\mathbf{X_1}$ and $\mathbf{Y_1}$. An improvement on imputation performance over methyLImp in terms of imputation accuracy is expected for the proposed methods because the potential problems caused by the nature of genomic data, such as high dimensionality and multicollinearity, are addressed. However, the computational efficiency is a challenge for the proposed methods because a ten-fold cross validation is required for parameter tuning for each iteration when fitting the model with $\mathbf{Y_1}$ and a high dimensional $\mathbf{X_1}$. Section 2.5 describes the proposed solutions to handle this issue.

2.5. REGULARIZED REGRESSION WITH VARIABLE SCREENING

In the linear model set up shown in Figure 2.3, the response variable $\mathbf{Y_1}$ can be a matrix when the missing value positions are the same in more than one variable. The imputation will be performed using the same group of variables $\mathbf{X_1}$ for all the columns in $\mathbf{Y_1}$. However, in reality the variables in the submatrix $\mathbf{Y_1}$ are typically uncorrelated. Here a method to impute them individually instead of altogether is proposed. Additionally, the submatrices $\mathbf{X_1}$ and $\mathbf{X_2}$ involve thousands of variables, which make the computation cost very high. To solve this problem, an extra step is added in each iteration of the imputation process. After the formation of $\mathbf{Y_1}$, $\mathbf{X_1}$ and $\mathbf{X_2}$ (note here $\mathbf{Y_1}$ is always a vector), the dimensions of $\mathbf{X_1}$ and $\mathbf{X_2}$ are reduced to the length of $\mathbf{Y_1}$ (i.e., the number of samples) to get $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$. The selection is based on the Pearson correlation between the predictors and the response. Then, the regularization methods proposed in the previous section (ridge regression, LASSO, and elastic net regression) are applied to $\mathbf{Y_1}$, $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ to obtain imputed values for $\mathbf{Y_2}$. The steps for each imputation iteration are summarized below:

1. Matrices $\mathbf{Y_1}, \mathbf{X_1}$ and $\mathbf{X_2}$ are formed according to the description in Section 2.4 and Figure 2.3. Here, $\mathbf{Y_1}$ is a vector since each variable is imputed separately.

2. $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ are obtained by reducing the dimensions of $\mathbf{X_1}$ and $\mathbf{X_2}$ to the length of $\mathbf{Y_1}$ (i.e., the number of samples) using the Pearson correlation as a selection criteria.

3. A regularized linear regression model is fit using $\mathbf{Y_1}$ and $\widetilde{\mathbf{X}}_1$.

4. $\mathbf{Y_2}$ is predicted using $\widetilde{\mathbf{X}}_2$ and the model in the previous step.

After each iteration, the algorithm will move to the next CpG site with missing value(s) and repeat the steps above until all missing entries are imputed.

2.6. FUNCTIONAL DATA ANALYSIS IMPUTATION

The methylation level of each probe is dependent on the neighboring probes as described in Section 2.1.2. The methylation level values can be viewed as a curve over the genome. However, the previously described approaches do not incorporate this inherent ordering with neighboring dependency directly into the methodology. In this section, a functional data approach is proposed for imputation to address these issues. The basic concepts of functional data analysis (FDA) are introduced before proposing the functional principal component analysis (FPCA) imputation method.

The DNA methylation measurements for each sample can be treated as one single observation with underlying structure, rather than multiple observations of independent variables. The key assumption of functional data analysis is that there exists a function X to represent the intrinsic structure of the data and the function is smooth. This can be expressed as:

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{i,j} \tag{2.10}$$

where Y_{ij} represents the observed methylation level of individual i at the genomic location t_{ij} (i.e., CpG site), $X_i(t_{ij})$ are the smooth functional data and $\epsilon_{i,j}$ is the error that account for "roughness" in the raw data. Here, i is the individual sample ($i = 1, \ldots, n$) and j is the genomic location ($j = 1, \ldots, n_i$).

2.6.1. Basis Function. To approximate the data as a function, a basis function system is needed. A system of basis functions is a set of known functions, denoted as ϕ_k that are independent of each other. Let $k = 1, 2, 3, ..., K$ where K is the total number of basis functions. A linear combination of the basis functions constructs the desired function of the data as follows:

$$X(t) = \sum_{k=1}^{K} c_k \phi_k(t) \tag{2.11}$$

where c_k are the coefficients corresponding to the basis functions ϕ_k. Some common basis functions include the monomial system $(1, t, t^2, t^3, \ldots)$, the Fourier series system $(1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \sin(3\omega t), \cos(3\omega t), \ldots)$, and the exponential basis system $(e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, \ldots)$ (Ramsay and Silverman, 2007).

2.6.1.1. Fourier series. For periodic data, the Fourier series basis functions are commonly used, since the trigonometric sin and cos functions are periodic. Suppose the function repeats itself over a time period T, and let $\omega = 2\pi/T$. The Fourier series basis functions are defined as follows:

$$\begin{aligned}
\phi_1(t) &= 1 \\
\phi_2(t) &= \sin(\omega t) \\
\phi_3(t) &= \cos(\omega t) \\
\phi_4(t) &= \sin(2\omega t) \\
\phi_5(t) &= \cos(2\omega t) \\
\phi_6(t) &= \sin(3\omega t) \\
&\vdots \\
\phi_K(t) &= \cos(m\omega t).
\end{aligned} \tag{2.12}$$

The total number of basis functions is K where $K = 2m+1$. Fourier series have traditionally been used as basis functions in the past due to their computational efficiency. The Fourier basis is useful for extremely stable functions and ideally for data with some degree of periodicity (Ramsay and Silverman, 2007). However, this basis is not appropriate for data with discontinuities in the function or in low order derivatives of the function.

2.6.1.2. Splines. Often, non-periodic functions are approximated by spline functions. Especially for data involving a large number of observations, spline function basis systems have been developed. Splines are polynomial segments joined end-to-end, but the segments are constrained to be smooth at the joining points. The joining points are called

knots. The order m of the polynomial is the number of its highest power (degree) plus one. Being smooth at the breakpoints means the function values should be equal at these points. Further, the derivatives up to order $m-2$ are also required be the same at the breakpoints. The spline function is determined by the order m of the polynomial segments and the knot sequence τ_l where $l = 1, \ldots, L-1$. The number of parameters needed is $m + L - 1$.

The B-spline basis system is the most popular spline system. It was developed by De Boor (2001). The B-spline basis system has the following properties: (1) Each basis function is a spline function defined by m and τ, (2) any linear combination of the basis functions is a spline function, and (3) any spline function defined by m and τ can be expressed as a linear combination of these basis functions in the system. This system also has a compact support property, which states that a B-spline basis of order m is positive over no more than m intervals, and these intervals are adjacent. This property makes splines also computationally efficient.

2.6.2. Roughness Penalty. The coefficients of the B-spline functions can be determined by least squares estimation. Consider the error sum of squares (SSE):

$$SSE = \sum_{j=1}^{n} (y_j - X(t_j))^2. \tag{2.13}$$

Here the notation for the i index is removed for simplicity. To ensure the fitted curve is smooth, a simple linear smoother is obtained by finding the c_k's that minimize the following least squares criteria:

$$\sum_{j=1}^{n} [y_j - \sum_{k=1}^{K} c_k \phi_k(t_j)]^2. \tag{2.14}$$

The method is suitable under assumptions that ϵ_j's in model 2.10 are independently and identically distributed with mean zero and constant variance.

Fitting the data and the smoothness of the curve are two competing desires. The least squares approach can be modified to incorporate a roughness penalty to address this issue. Roughness penalty methods for smoothing work by optimizing a fitting criterion

by penalizing the roughness of the curvature. The curvature of a function at t can be characterized by the square of the second derivative $(D^2X(t))^2$. The roughness of a function can thus be defined as the integral of this value $\int (D^2X(t))^2 dt$. A parameter λ is used to control the roughness by minimizing the penalized squared error (PENSSE) (Ramsay and Silverman, 2007):

$$PENSSE_\lambda(X) = \sum [(y_j - X(t_j)]^2 + \lambda \int (D^2X(t))^2 dt \tag{2.15}$$

where $DX(t)$ is the slope of $X(t)$, $D^2X(t) = \frac{d^2}{dt^2}X(t)$ is the second derivative of $X(t)$ and thus represents its curvature. λ is a smoothing parameter measuring compromise between fit and smoothness. As λ increases, the roughness will be penalized more and $X(t)$ will become linear. As λ decreases, the penalty is reduced and $X(t)$ will fit the data better. The Smoothing Spline Theorem (Ramsay and Silverman, 2007) states that the function $X(t)$ that minimizes $PENSSE_\lambda(X)$ is a spline function of order 4 with a knot at each sample point t_j. Therefore, unequal spacing of the sampling points is not a problem, since smoothing splines automatically take care of high density areas in the data and areas with fewer observations.

2.6.3. Functional Principal Component Analysis Imputation. Principal component analysis (PCA) (Jolliffe, 2002) is a dimension reduction tool for multivariate data. Principal components are a new set of variables where each component is a linear combination of the original variables. The weights in the first components are chosen to maximize variance. Each subsequent component maximizes remaining variation and is orthogonal to all other components. The principal components are computed and then used for a change of basis on the data. This allows the dominant modes of variation in the data to be represented in a small subset of components. Most of the time, the first few principal components are enough to explain the majority of the variability in the data, and the remaining principal components will be discarded, resulting in dimension reduction in the data. For an $n \times p$ data matrix $\mathbf{X}$, each column is a vector of observations on one variable. A linear combination a

of the columns of matrix $\mathbf{X}$: $\sum_{j=1}^{p} a_j x_j = \mathbf{X}a$ is aimed to achieve maximum variance, where $Var(\mathbf{X}a) = a'\mathbf{S}a$ and $\mathbf{S}$ is the covariance matrix. With the restriction $a'a = 1$, maximizing $a'\mathbf{S}a$ provides the solution that a is a unit norm eigenvector of the covariance matrix $\mathbf{S}$, with corresponding eigenvalue λ's. The covariance matrix $\mathbf{S}$ is a $p \times p$ real symmetric matrix, and thus should have exactly p real eigenvalues. The eigenvectors are defined to be orthonormal, such that $a_j' a_{j^*} = 1$ when $j = j^*$ and 0 otherwise. This ensures each set of linear combination is uncorrelated. By using the top k largest eigenvalues, the data could be represented with most of the variance explained. $\mathbf{X}a_k$ are called the principal components, the eigenvectors a_k are called the principal component loadings.

PCA was extended to functional data and became widely used in functional data analysis to capture the dominant modes of variation in the smoothed curves. Functional principal component analysis (FPCA) converts infinite-dimensional functional data to a finite-dimensional vector of random scores. The underlying stochastic process can be represented by a finite sequence of uncorrelated random variables. These variables are called the functional principal component scores (FPC scores). Similar to PCA, usually only a finite subset of the sequence is used that captures most of the variation.

The following formulation illustrates how FPCA can be formulated in terms of DNA methylation data and used for imputation. Assume that the methylation levels across a chromosome have the pattern of function X, and X has an unknown smooth mean function $\mu(t)$ and a covariance function which is defined as:

$$cov(X(s)), X(t)) = G(s,t) \tag{2.16}$$

where $s, t \in T$, and T is the genomic location. $G(s,t)$ can be expanded with the orthogonal expansion:

$$G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t) \tag{2.17}$$

where λ_k is the set of eigenvalues and ϕ_k are the corresponding eigenfunctions that form an orthonormal basis set with a unit norm in l^2. The underlying pattern for the ith sample can be expressed as:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} c_{ik}\phi_k(t) \tag{2.18}$$

where ϕ_k is the kth eigenfunction, and

$$c_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt \tag{2.19}$$

is a coefficient projecting $(X_i - \mu)$ in the direction of ϕ_k.

The data $Y_i(t_{ij})$ is the jth observation of the random function $X_i(\cdot)$ at a random genomic location t_{ij}, also denoted as Y_{ij}, which can be represented as:

$$Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} c_{ik}\phi_k(t_{ij}) + \epsilon_{ij}. \tag{2.20}$$

Here ϵ_{ij} represents the measurement random errors of the ith sample at jth genomic location, and are assumed to be independent and identically distributed with mean 0 and variance σ^2. From Equations 2.19 and 2.20, it can be shown that:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{\infty} \int (X_i(t) - \mu(t))\phi_k(t)dt\phi_k(t) + \epsilon. \tag{2.21}$$

The infinite series in Equation 2.18 can be truncated by L such that the first L components explains at least $\tau_\lambda \times 100\%$ of the total variance, that is:

$$L = min\{L \geq 1 : \frac{\sum_{k=1}^{L} \hat{\lambda}_k}{\sum_{k=1}^{M} \hat{\lambda}_k} \geq \tau_\lambda\} \tag{2.22}$$

where M is the largest number of components with $\hat{\lambda}_k > 0$ and τ_λ is a user defined threshold between 0 and 1. When the observations $Y_i(t_{ij})$ are missing for some j, the missing entries can be imputed by the predicted values $\hat{X}_i(t)$.

To estimate $\hat{X}_i(t)$, the estimated values of $\hat{\mu}(t)$, $\hat{c}_{ij}$ and $\hat{\phi}_k(t)$ are needed. To find the estimates of the eigenfunctions ϕ and eigenvalues λ, the eigenequation can be expressed as:

$$G\phi = \lambda\phi. \tag{2.23}$$

The estimate of G is obtained by smoothing the empirical covariances (Yao et al., 2005). Then the eigen-decomposition procedure is applied to the covariance function estimate to get the estimated eigenvalues $\hat{\lambda}_{ij}$ and eigenfunctions $\hat{\phi}_k(t)$.

The estimate of c_{ik} cannot be calculated easily through the approximation of Equation 2.19 because if the number of repeated observations is small or if there are missing positions, the integral is not accurate. Also the true $X_i(t)$ cannot be observed. The observations are $Y_i(t_{ij}) = X_i(t_{ij}) + \epsilon_{ij}$ and bias will be introduced if X_i is replaced by Y_i. An approach first proposed by Yao et al. (2005) provides a solution to these issues of estimating the c_{ik}'s. The eigenfunction basis is estimated from the data, and functional principal component score estimates are obtained by a conditioning step. The assumption is that the functional principal component scores c_{ik} and the error term $\epsilon_{i,j}$ are jointly Gaussian. The conditional functional principal component scores are:

$$E(c_{ik} \mid \mathbf{Y}_i) = \lambda_k \boldsymbol{\phi}'_{ik} \boldsymbol{\Sigma}^{-1}_{\mathbf{Y}_i} (\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{2.24}$$

where λ_k is the kth eigenvalue, $\mathbf{Y}_i = [Y_i(t_{i1}), \ldots, Y_i(t_{in_i})]'$, $\boldsymbol{\mu}_i = [\mu_i(t_{i1}), \ldots, \mu_i(t_{in_i})]'$, $\boldsymbol{\phi}_{ik} = [\phi_k(t_{i1}), \ldots, \phi_k(t_{in_i})]'$, and $\boldsymbol{\Sigma}_{\mathbf{Y}_i}$ is the covariance matrix of $\mathbf{Y}_i$, with dimension $n_i \times n_i$. The $\boldsymbol{\Sigma}_{\mathbf{Y}_i}$ is represented as:

$$\boldsymbol{\Sigma}_{\mathbf{Y}_i} = cov(\mathbf{Y}_i, \mathbf{Y}_i) = cov(\mathbf{X}_i, \mathbf{X}_i) + \sigma^2 \mathbf{I}_{n_i}. \tag{2.25}$$

In scalar form this is:

$$(\mathbf{\Sigma}_{\mathbf{Y}_i})_{j,l} = G(t_{ij}, t_{il}) + \sigma^2 \delta_{jl} \tag{2.26}$$

with $\delta_{jl} = 1$ if $j = l$ and 0 if $j \neq l$.

The estimated scores in Equation 2.24 are obtained by:

$$\hat{c}_{i,k} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}'_{ik} \hat{\mathbf{\Sigma}}^{-1}_{\mathbf{Y}_i} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) \tag{2.27}$$

where $\hat{\boldsymbol{\mu}}_i = [\hat{\mu}_i(t_{i1}), \ldots, \hat{\mu}_i(t_{in_i})]'$ is the estimate of $\boldsymbol{\mu}_i$, $\hat{\boldsymbol{\phi}}_{ik} = [\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{in_i})]'$ is the estimate of $\boldsymbol{\phi}_{ik}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{Y}_i}$ is the estimate of $\mathbf{\Sigma}_{\mathbf{Y}_i}$. The estimated score $\hat{c}_{i,k}$ will be used as the functional covariates to perform functional linear regression with a scalar response.

2.6.4. Functional Linear Models. In traditional linear regression models, the dependent variable and the independent variables are scalars. One way to set up a functional linear model is with a scalar dependent variable y_i, but replace the independent variables by a function $x_i(t)$:

$$y_i = \alpha_0 + \int x_i(t)\alpha(t)dt + \epsilon_i. \tag{2.28}$$

One functional linear regression approach is to regress $\mathbf{Y}$ on the principal component scores as functional covariates (Ramsay and Silverman, 2007), and it will be referred to as Functional Principal Component Regression (FPCR) in this dissertation. A subset of the 100 nearest available probes to the probe with missing values is utilized in the modeling to capture a relevant set of neighboring probes. $\mathbf{Y_1}$ is an $n_1 \times 1$ vector of the samples with complete entries for the probes with missing values. $\mathbf{Y_2}$ is an $n_2 \times 1$ vector and it represents the missing entries that need to be imputed. $\mathbf{X_2}$ is the $n_2 \times 100$ matrix with data on the 100 neighboring probes with complete data and with the same rows of $\mathbf{Y_2}$. $\mathbf{X_1}$ is an $n_1 \times 100$ matrix sharing the same rows with $\mathbf{Y_1}$ and the same columns with $\mathbf{X_2}$.

The R package 'fdapace' is used to find the principal component scores via the Principal Analysis by Conditional Estimation (PACE) algorithm (Yao et al., 2005). The first step is to estimate the mean function μ as in Equation 2.18 based on the pooled data of

all individuals $\begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{bmatrix}$, with local linear smoothers (Fan and Gijbels, 1996) for function and surface estimation. A one-curve-leave-out cross-validation is used to choose the smoothing parameter. The conditional expectation method is then used to estimate the FPC scores. Finally, the response $\mathbf{Y_1}$ is regressed on the functional principal component scores to build the functional linear model. The model is then used to predict $\mathbf{Y_2}$. After each iteration, the algorithm will search for the next CpG site with missing value(s). The process of FPCR model fitting and prediction will be repeated. Imputation is completed when there is no missing value remaining in the dataset.

2.7. DMR DETECTION

The proposed imputation methods based on regularized regression and FPCR will be compared with mean imputation, KNN imputation, as well as the methyLImp method. To evaluate these methods, imputation accuracy is important since it represents how close the imputed values are to the real values. However, the goal of imputation is to obtain statistically valid results from the incomplete data. Thus, the quality of the imputation should also be evaluated with respect to this ultimate goal of DMR detection. A simulation study will be conducted to evaluate the imputation accuracy, and more importantly, the influence of imputation on DMR detection. Bumphunter and DMRcate are commonly used methods for DMR detection that are used in this work to evaluate the imputation performance on DMR detection. An overview of these methods is provided below.

2.7.1. Bumphunter. The Bumphunter method implemented in the Bioconductor package 'ChAMP' is used to find DMRs. The statistical model used by Bumphunter is:

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \gamma_k(t_j)Z_{i,k} + \sum_{l=1}^{q} a_{i,j}W_{i,l} + \epsilon_{i,j} \tag{2.29}$$

where Y_{ij} is the epigenomic measurement at the jth genomic locus for individual i, t_j denotes the location on the genome of the jth locus, $\mu(t_j)$ is the baseline level of epigenomic measurement, X_i is the condition of interest, $\beta(t_j)$ measures the association between X_i and the epigenomic measurement Y_{ij} at location t_j, Z's are potential measured confounders (e.g. sex, age, race), each column of Z represents a different confounder, $\gamma_k(t_j)$ is the effects of confounder k at locus t_j, W represents potential unmeasured confounders or batch effects (e.g. temperature, humidity), $a_{l,j}$ is the effect of the unmeasured confounder l on locus t_j, and $\epsilon_{i,j}$ is the unexplained variability.

In the Bumphunter analysis pipeine, the linear regression model 2.29 is fit by regressing the methylation value Y_{ij} on the group X_i to model differential methylation between the case and the control groups at each CpG site. The slope $\hat{\beta}(t_j)$ is then smoothed using the loess method with a smoothing window ranging from 300 to 900 base pairs to get the smoothed $\tilde{\beta}(t)$. For most genomic positions, the $\hat{\beta}$'s are zero because the methylation levels at these positions are not significantly different between groups. Each point is weighted based on the standard error obtained from the linear model. The smoother works well to reduce the effect of outliers. Clusters of consecutive probes for which all the smoothed $\tilde{\beta}(t)$ values that are greater than a predetermined threshold are identified as candidate regions (bumps) $R_n, n = 1, \ldots, N$. The maximum gap is a user determined distance. When neighboring probes are less than that distance, they will be included in one region. Next, clusters are defined using the following criteria: 1) the cluster has at least 4 probes, and 2) the probes inside one cluster are all less than or equal to 500 base pairs. The 99th percentile of the slope estimates is used as a cutoff to determine the candidate regions. This means the values of the estimate of the methylation profile above the cutoff or below the negated cutoff are treated as candidate regions.

Permutation tests, which permute sample labels to create the null distribution of candidate regions $\hat{R}_n, n = 1, \ldots, \hat{N}$, are then conducted to estimate the statistical significance of the candidate regions. The regions that are produced in the permutations are considered

null regions and can provide a null distribution for R_n . This method can solve the problem of correlated measurement errors, batch effects and so on. The number of resamples is set to be 10. Each of the 10 permutations will generate an estimated null distribution. The p-value is the percent of candidate regions obtained from the permutations that are as extreme as the observed region. False discovery rates (FDR) are calculated based on the p-values, and Q-value is defined as the minimum FDR at which the associated area may be called significant. The family-wise error rate (FWER) is also calculated, which is the proportion of permutations that had at least one region as extreme as the observed region.

2.7.2. DMRcate. The Bioconductor package 'DMRcate' is also used to find DMRs. At each CpG site, a linear model is fit using the limma (Smyth, 2004) method. The square of the t statistic $Y_i = t_i^2$ is used as the local statistic at each site i. The use of the squared t statistic allows the method to obtain the magnitude between methylation levels of two groups instead the direction of effect. Gaussian smoothing is then applied to the test statistics using a given bandwidth λ. Next, suppose there are n CpG sites on a chromosome; $x_1 < x_2 < \cdots < x_n$ representing all the locations. A Gaussian smoother is used to smooth the Y_i at locations x_i for each chromosome. The Gaussian kernel weights are $K_{ij} = \exp\left(\frac{-[x_i - x_j]^2}{2\sigma^2}\right)$, where σ is the kernel scale factor, $\sigma = \lambda/C$. The value for the bandwidth λ is set to be 500. As mentioned in Section 2.1.2, the correlation on methylation levels between sites over longer distances is not noticable. C is also user defined, and is set to be 5. Smoothed test statistics are then modeled using the method of Satterthwaite (Satterthwaite, 1946), and a p-value is calculated for each site. Significant sites are reported after Benjamini–Hochberg adjustments on p-values. Finally, DMRs are defined by grouping the significant CpG sites that are at most λ nucleotides from each other.

3. RESULTS

3.1. OVERVIEW

In this chapter, an analysis based on real data as well as a simulation study are presented to evaluate the performance of the imputation methods proposed in this work compared to the existing methods. The DNA methylation data described in Section 2.1.1 are utilized both for the real data analysis and to guide settings in the simulation study. The set-up of the simulation study is first described, followed by a discussion of how the imputation methods will be evaluated in both the real and simulated data. Results are then given for the real data followed by results for the simulation study.

Three existing methods (mean, KNN, methyLImp) are compared to the proposed methods on imputation accuracy and ability to detect true differentially methylated regions (DMRs). A total of 11 proposed methods are compared, which can be categorized into three groups. The first group includes the regularized methods: ridge regression (Ridge), LASSO, elastic net with 0.2 mixing parameter (elastic net 0.2), elastic net with 0.5 mixing parameter (elastic net 0.5), and elastic net with 0.8 mixing parameter (elastic net 0.8). The second group includes all of these regularized methods with variable screening and imputation on a site by site basis (1 by 1) rather than altogether. These methods are denoted the same as above with 1by1 at the end: Ridge 1by1, LASSO 1by1, elastic net 0.2 1by1, elastic net 0.5 1by1, and elastic net 0.8 1by1. The final alternative approach evaluated is the functional principal component regression (FPCR) method.

3.2. SIMULATION STUDY

The purpose of the simulation study is to investigate the performance of the proposed imputation methods on imputation accuracy and the ability to detect differentially methylated regions. It is important to simulate the data in a way that preserves properties of real methylation data. The HM450 dataset described in Section 2.1.1 is utilized to help create a simulated dataset with a realistic structure. The distribution of the HM450 probes is related to the length of each chromosome. As shown in Figure 3.1, Chromosome 1 has the most number of probes, Chromosome 6 has the second most number of probes and Chromosome Y has the least number of probes. For computational efficiency, simulation studies are performed on the 36,611 CpG probes located on the entire Chromosome 6. The 86 Normal samples are preprocessed as previously described, resulting in 31,362 probes after the filtering steps. This is recognized as the incomplete dataset. Among the filtered probes, 5,076 probes are filtered out because of being potential SNPs. After integrating the SNP data from "Pathogenic Germline Variants in 10,389 Adult Cancers" (Huang et al., 2018), 4,917 probes are restored since SNPs were not present in any of the samples. This provides the complete dataset with 36,279 probes. Thus, only 159 probes with true SNPs are excluded.

The next step involves identifying a set of regions in which methylation differences will be applied. To accomplish this, Adjacent Site Clustering (Sofer et al., 2013) is implemented to find region clusters on Chromosome 6 of the 86 Normal samples. The algorithm merges a set of methylation sites wedged between two highly correlated CpG sites that are located physically close to each other along a chromosome. More specifically, the criteria is to merge two CpGs with Spearman correlation greater than 0.5 and are within 200 base pairs into a cluster. This resulted in 2,478 clusters (14,801 probes total) with 4 or more probes, among which 2,088 clusters contain 10 or fewer probes. 250 clusters are randomly

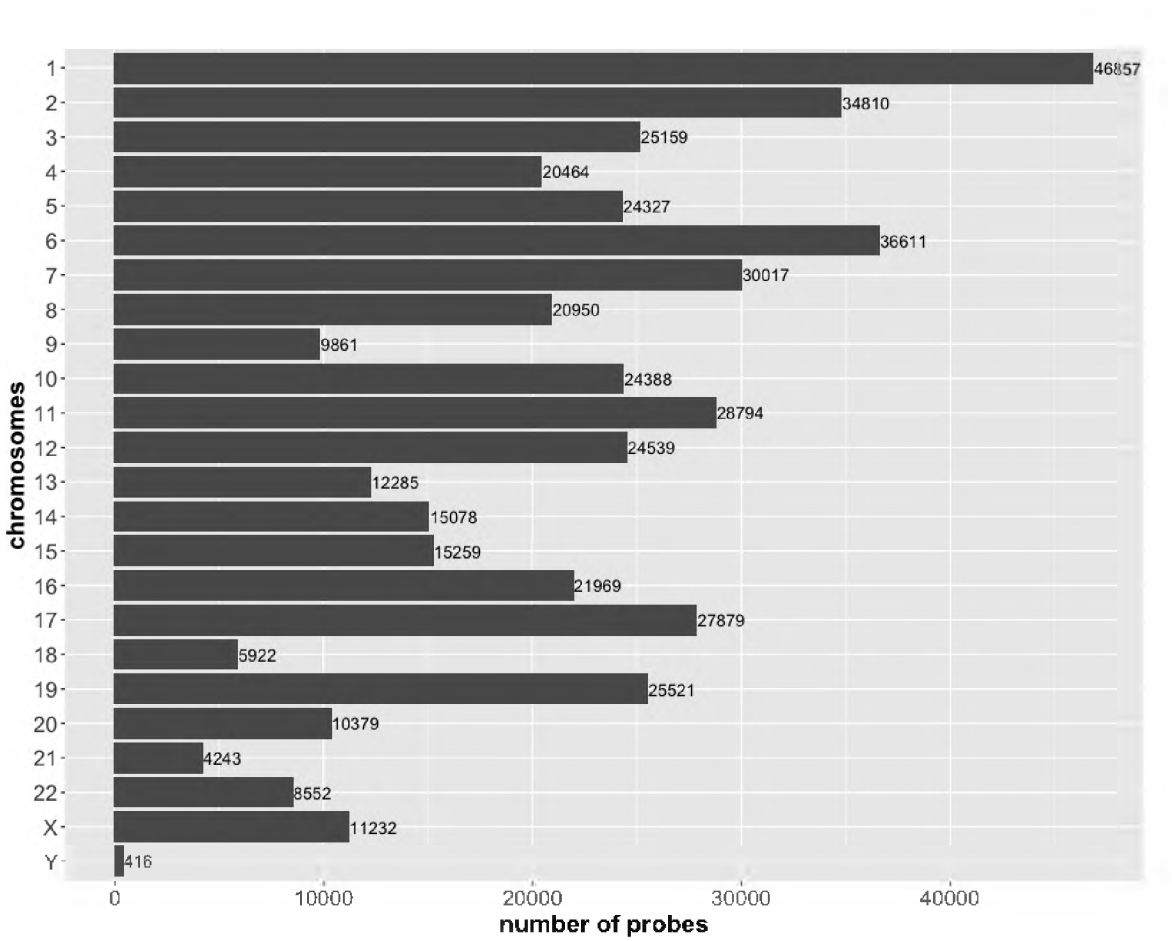


Figure 3.1. The count of HM450 probes on each chromosome.

selected from the 2,088 clusters. It is found that 96 of these clusters have SNP probes that could be restored from the 4,917 probes. A differential methylation effect is added to these 96 regions to evaluate the imputation techniques.

To ensure the nature of real data is well preserved, two key points are implemented in the simulation steps. First, real datasets from the same group (Normal group) are used as the base to add differential methylation effects. Moreover, the parameters used in the simulation process are derived from summarized results of the real data analysis between Tumor and Normal groups. The 86 Normal samples are randomly divided into two groups. Before introducing differentially methylated regions (DMRs), the two groups are compared using the DMR detection methods Bumphunter and DMRcate to make sure there is no DMR

flagged. The β-values are first converted to M-values using Equation 1.1. The M-values are not bounded between 0 and 1, thus after adding the differential methylation effects, the issue of out of limit values is avoided. An effect size of 1.5 is determined by comparing the difference between Tumor and Normal groups of the real data.

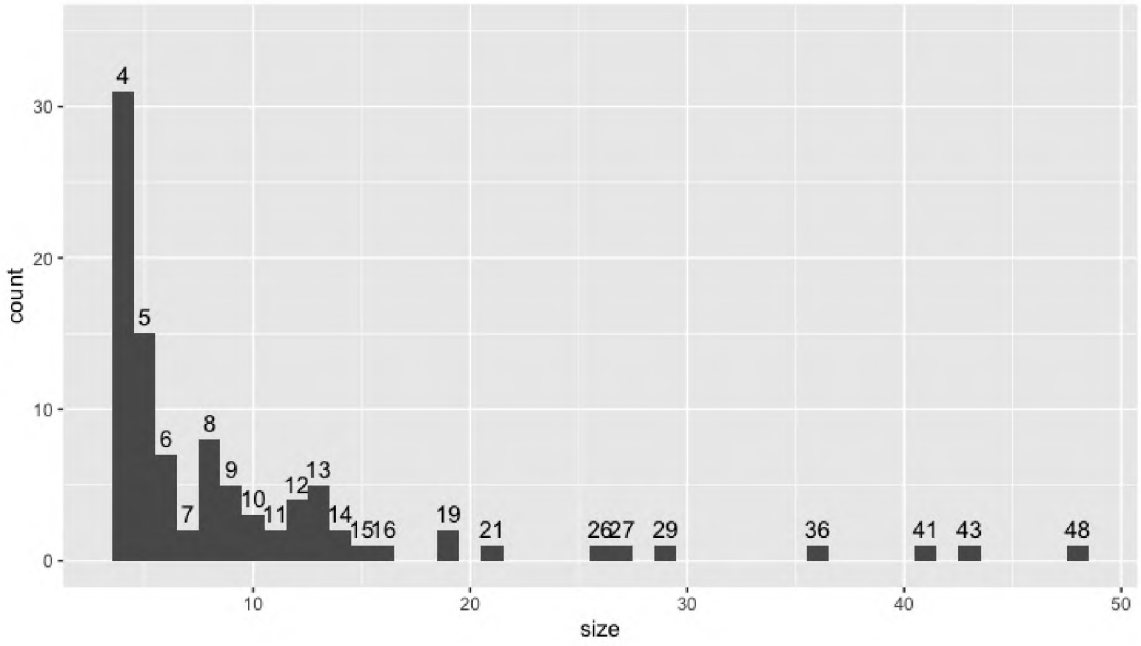


Figure 3.2. Histogram of the sizes for all 96 clusters selected to be DMRs. The number of probes in each cluster is marked on top of each bin. There are 46 clusters with 4 or 5 probes, and 8 clusters with more than 20 probes.

The details of the cluster size for all 96 clusters selected to be DMRs are shown in Figure 3.2. It is a right skewed histogram, with more small clusters than large clusters. There are 46 clusters with 4 or 5 probes, and 8 clusters with more than 20 probes. To ensure the added treatment effects do not cancel out existing differences in M-values, for each CpG probe in the methylation cluster, the group means of the two groups are first compared. Treatment effects are then added to the group of probes with higher mean M-values.

Previous studies have found that hyper- and hypomethylation can happen in the same regulatory region with one followed immediately by the other (Day et al., 2013). The methylation levels in the simulation are designed to mimic this situation in real data. For half of the clusters, a type 1 simulation is applied by adding a treatment effect of 1.5 to M-values of the probes with higher average M-values (Figure 3.3 a). In this case, values may be added to different treatment groups inside a cluster. For the other half of the clusters, a type 2 simulation is applied as follows. The group mean of the M-values for each CpG probe in each cluster is compared. For the group which has more probes with higher M-values, a treatment effect of 1.5 is added to the M-values in the same group for the entire cluster (Figure 3.3 b).

3.3. EVALUATION CRITERIA

3.3.1. Evaluation of Imputation Accuracy. In both the real and simulated datasets a subset of the probes are randomly selected to be missing at different rates. It is important to evaluate how accurate the imputed values are compared to the true values for the different imputation methods. Performance of the imputation methods on accuracy are assessed by using four different measures (Lena et al., 2019a). These measures are used to evaluate imputation accuracy in both the real data and the simulated data imputation.

The imputed or predicted values are denoted as P, and the true values are denoted as T. The Root Mean Square Error (RMSE) metric measures the square root of the average squared difference between the predicted and the true values. It is the most widely used metric for performance assessment of missing data imputation approaches and is given below:

$$RMSE(P,T) = \sqrt{\frac{\sum_{i=1}^{n}(P_i - T_i)^2}{n}}. \tag{3.1}$$

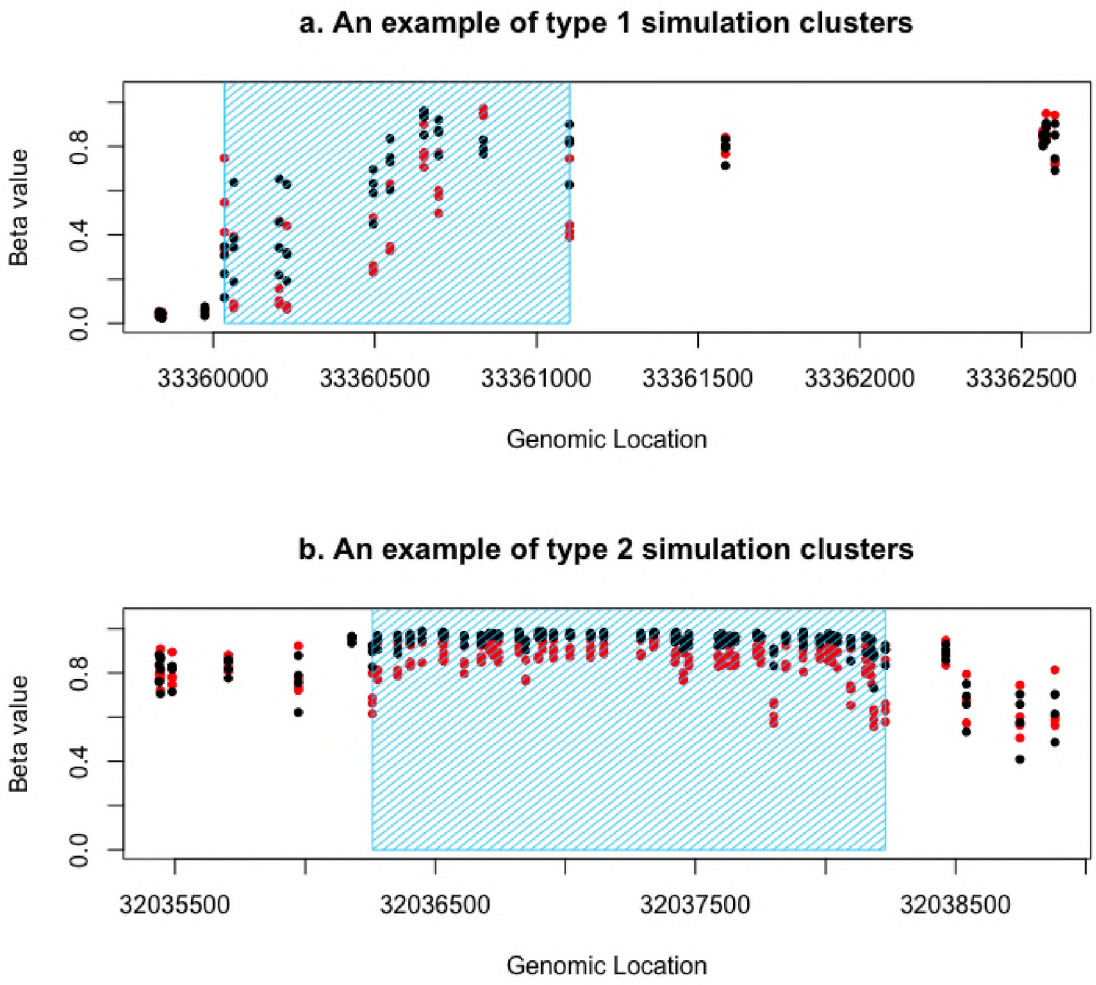


Figure 3.3. Two types of simulation clusters. β-values verses the genomic location are shown here to illustrate the effects added. Black dots represent the Normal group and red dots represent the Tumor group. The shaded areas are the differentially methylated clusters. a, the effects are added to the probes with higher average methylation levels between two groups. b, the effects are added to all the probes in a cluster.

The Mean Absolute Error (MAE) metric measures the average absolute difference between the predicted and true values. It provides the average error to expect on the imputed value. Note that by Jensen's inequality, RMSE $\geq$MAE. The MAE is given below:

$$MAE(P,T) = \frac{\sum_{i=1}^{n} \mid P_i - T_i \mid}{n}. \tag{3.2}$$

The Pearson Correlation Coefficient (PCC) metric measures the amount of linear correlation between the predicted and true values. The PCC is given below:

$$PCC(P,T) = \frac{\sum_{i=1}^{n}(P_i - \bar{P})(T_i - \bar{T}))}{\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})^2}\sqrt{\sum_{i=1}^{n}(T_i - \bar{T})^2}} \tag{3.3}$$

where $\bar{P}$ and $\bar{T}$ represent the mean value of P and T, respectively. The Mean Absolute Percentage Error (MAPE) metric expresses the accuracy as percentage of error on the true value. It gives an estimation of the error in terms of the magnitude of actual value. The MAPE is given below:

$$MAPE(P,T) = \frac{100}{n}\sum_{i=1}^{n}\frac{\mid P_i - T_i \mid}{\mid T_i \mid}. \tag{3.4}$$

Note that smaller values of RMSE, MAE, and MAPE indicate greater accuracy; whereas larger values of PCC are better.

3.3.2. Evaluation of DMR Detection. The goal of missing value imputation is to improve the ability to detect differentially methylated regions that are important and biological meaningful. The simulation results will be compared with true DMRs to determine if there are any improvements with respect to true positive, false positive, and false negative regions. Note that true DMRs are unknown in the real data, so this evaluation is only conducted for the simulated data. A true positive (TP) DMR is defined as a significant DMR declared by one of the detection methods (Bumphunter or DMRcate) that overlaps with a region in which a treatment effect was added to the methylation M-values. The overlap type is 'any', meaning any common genomic location between the compared regions will count as them as overlapping. A false positive (FP) DMR is defined as a significant DMR declared by one of the detection methods that does not overlap with any of the regions with added treatment effects. A false negative (FN) DMR is defined as a region with added treatment

Table 3.1. Details of probes filtered out by each step of the default filtering process.

Filtering Step	Probes Filtered Out	Remaining Probes
quality control probes	65	485,512
detection p-value	25,900	459,612
bead count	874	458,738
SNP	53,959	404,779
multi-hit	11	404,768
non-CpG	1,833	402,935

effect that does not overlap with any significant DMRs found by the detection methods. It is possible that some true methylated regions are broken down into smaller regions for certain DMR detection methods, or that more than one true region is recognized as one DMR.

3.4. RESULTS FOR REAL DATA ANALYSIS

3.4.1. SNP Integration. Following the filtering steps in Section 2.1.1 for detection p-values, bead counts, SNP probes, multi-hit probes and non-CpG probes, a total of 485,577 probes on the HM450 array are reduced to 402,935 probes. There are 65 built-in SNP probes in HM450 array for the purpose of quality control, and they are typically removed in preprocessing steps. The steps shown in Table 3.1 are sequential, meaning that each filtering step is based on the filtering result of previous step(s). For example, if the probes are filtered by bead count first, followed by detection p-value, the numbers in the second column would be different.

By integrating the germline SNP data, a large portion of probes are restored. As shown in Figure 3.4, 52,441 probes out of 53,959 (97.2%) are actually not SNP probes, thus it is not necessary to filter them out. With the large portion of probes being restored, the influence on DMR improvement is prominent. For the remaining 1,518 probes, imputation methods are developed and evaluated on these probes that cannot be restored.

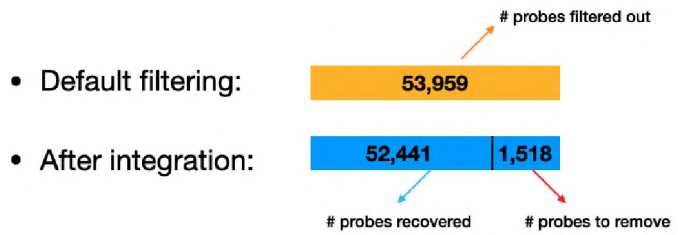


Figure 3.4. Details of probe restoration by integrating SNP data.

As briefly discussed in chapter 2, the dataset used for real data analysis is described below. Starting with a section of Chromosome 1 (7,987 probes between the genomic location 1 and 13,800,000), the filtering steps in Section 2.1.1 are conducted on the raw intensity files (idat). This step results in a dataset with 6,838 probes, and this dataset is noted as the incomplete data. After integrating the SNP data, 830 probes are restored. The dataset with 7,668 probes is called the complete dataset. Missing values are introduced in the 830 restored SNP probes at different missing rates. Imputation methods are then conducted to obtain the imputed dataset.

The imputation accuracy is compared between the true values and the imputed values. The DMR detection performance is evaluated using the complete dataset as a standard since the true DMRs are unknown for the real data. The complete dataset is the most informative since it utilizes the data from the most true probes possible for DMR detection compared to the incomplete and imputed datasets. In the next section, simulated data is used to assess the DMR detection improvements by imputation.

3.4.2. Imputation Accuracy. The performance of the imputation methods are compared by computing the RMSE, MAE, PCC and MAPE for each method. The imputation accuracy is assessed per CpG site. Three missing rates (20%, 50% and 70%) are explored. The Normal group and Tumor group are separated when conducting the

imputation. Overall, imputation methods perform better on the Normal group than Tumor group, regardless of the detection method used or the missing rate. This likely due to the heterogeneous nature of tumor samples. Additional discussion about this issue can be found in Chapter 4.

Table 3.2 shows the performance with missing rate 20%. The performance of all imputation methods work uniformly better in the Normal group than the Tumor group. Take the mean imputation method as an example. The Tumor group has a RMSE of 0.116 while the Normal group has a RMSE of 0.085. While the mean imputation has the largest RMSE in both groups, FPCR and KNN imputation perform only slightly better than mean imputation in the Tumor group. In the Normal group, the regularized linear regression imputation method using elastic net with mixing parameter 0.8 (elastic net 0.8) has the smallest RMSE, MAPE and the highest PCC. The methyLImp, elastic net 0.2, elastic net 0.8, LASSO 1 by 1, elastic net 0.2 1 by 1, elastic net 0.5 1 by 1 and elastic net 0.8 1 by 1 methods have the smallest MAE. In the Tumor group, the elastic net 0.2 1 by 1, elastic net 0.5 1 by 1 and elastic net 0.8 1 by 1 methods have the smallest RMSE, MAE and highest PCC. The elastic net 0.8 method has the smallest MAPE, followed by all three of the elastic net 1 by 1 methods.

Table 3.3 shows the performance of the imputation methods with missing rate 50%. While the mean method is still the worst, the performance of the KNN method becomes the next worst across all of the criteria in both groups. In the Normal group, all the regularized methods outperform methyLImp in terms of RMSE, PCC and MAPE. The LASSO, elastic net 0.2, elastic net 0.5, elastic net 0.8 and elastic net 0.2 1 by 1 methods have the lowest RMSE and highest PCC. The methyLImp, LASSO 1 by 1, elastic net 0.2 1 by 1, elastic net 0.5 1 by 1 and elastic net 0.8 1 by 1 methods have the lowest MAE. The elastic net 0.2 1 by 1 method has the lowest MAPE. In the Tumor group, the elastic net 0.2 1 by 1 and elastic

Table 3.2. Imputation accuracy for real data with 20% missing rate. The optimal value(s) for each criteria are in bold.

		RMSE	MAE	PCC	MAPE
Normal	mean	0.085	0.054	0.952	16.836
	KNN	0.064	0.035	0.974	11.323
	methyLImp	0.061	**0.029**	0.976	10.350
	LASSO	0.059	0.030	0.977	9.965
	Ridge	0.061	0.030	0.976	10.356
	elastic net 0.2	**0.056**	**0.029**	0.979	9.652
	elastic net 0.5	0.059	0.030	0.977	9.957
	elastic net 0.8	**0.056**	**0.029**	**0.980**	**9.625**
	FPCR	0.066	0.036	0.972	11.428
	LASSO 1by1	0.059	**0.029**	0.977	9.853
	Ridge 1by1	0.060	0.032	0.976	10.532
	elastic net 0.2 1by1	0.059	**0.029**	0.977	9.872
	elastic net 0.5 1by1	0.059	**0.029**	0.978	9.819
	elastic net 0.8 1by1	0.059	**0.029**	0.977	9.799
Tumor	mean	0.116	0.079	0.908	25.595
	KNN	0.092	0.058	0.943	17.957
	methyLImp	0.085	0.050	0.953	15.597
	LASSO	0.083	0.050	0.955	14.989
	Ridge	0.090	0.054	0.947	17.334
	elastic net 0.2	0.083	0.050	0.954	14.499
	elastic net 0.5	0.083	0.050	0.955	15.059
	elastic net 0.8	0.083	0.050	0.955	**14.214**
	FPCR	0.095	0.060	0.941	17.900
	LASSO 1by1	0.082	**0.048**	0.956	14.451
	Ridge 1by1	0.088	0.053	0.949	16.603
	elastic net 0.2 1by1	**0.081**	**0.048**	**0.957**	14.418
	elastic net 0.5 1by1	**0.081**	**0.048**	**0.957**	14.250
	elastic net 0.8 1by1	**0.081**	**0.048**	**0.957**	14.382

net 0.5 1 by 1 methods have the lowest RMSE. The elastic net 1 by 1 methods with mixing parameter 0.2, 0.5 and 0.8 have the lowest MAE and highest PCC. The elastic net 0.2 1 by 1 method has the lowest MAPE.

Table 3.4 shows the performance of all imputation methods with missing rate 70%. The KNN method performs the worst in terms of RMSE, MAE, PCC and MAPE. In the Normal group, the elastic net 0.2, elastic net 0.5, Ridge 1 by 1 and elastic net 0.2 1 by 1

Table 3.3. Imputation accuracy for real data with 50% missing rate. The optimal value(s) for each criteria are in bold.

		RMSE	MAE	PCC	MAPE
Normal	mean	0.084	0.053	0.953	16.742
	KNN	0.067	0.038	0.971	11.664
	methyLImp	0.062	**0.030**	0.975	11.469
	LASSO	**0.058**	0.031	**0.978**	10.495
	Ridge	0.059	0.031	0.977	10.725
	elastic net 0.2	**0.058**	0.031	**0.978**	10.513
	elastic net 0.5	**0.058**	0.031	**0.978**	10.485
	elastic net 0.8	**0.058**	0.031	**0.978**	10.484
	FPCR	0.064	0.036	0.973	11.078
	LASSO 1by1	0.059	**0.030**	0.977	9.855
	Ridge 1by1	0.059	0.032	0.977	10.327
	elastic net 0.2 1by1	**0.058**	**0.030**	**0.978**	**9.752**
	elastic net 0.5 1by1	0.059	**0.030**	0.977	9.789
	elastic net 0.8 1by1	0.059	**0.030**	0.977	9.821
Tumor	mean	0.118	0.080	0.906	24.784
	KNN	0.111	0.065	0.919	18.411
	methyLImp	0.090	0.054	0.946	16.255
	LASSO	0.088	0.053	0.949	15.661
	Ridge	0.095	0.058	0.940	17.477
	elastic net 0.2	0.089	0.054	0.948	16.032
	elastic net 0.5	0.088	0.053	0.949	15.771
	elastic net 0.8	0.088	0.053	0.949	15.797
	FPCR	0.097	0.061	0.937	17.907
	LASSO 1by1	0.087	0.052	0.950	15.264
	Ridge 1by1	0.091	0.055	0.945	16.592
	elastic net 0.2 1by1	**0.086**	**0.051**	**0.951**	15.182
	elastic net 0.5 1by1	**0.086**	**0.051**	**0.951**	**15.177**
	elastic net 0.8 1by1	0.087	**0.051**	**0.951**	15.219

methods have the lowest RMSE and highest PCC, The methyLImp method has the lowest MAE. The lowest MAPE is obtained by elastic net 0.2 1 by 1 method. In the Tumor group, the elastic net 0.2 1 by 1 method yields the lowest RMSE, MAE, MAPE and highest PCC.

Table 3.4. Imputation accuracy for real data with 70% missing rate. The optimal value(s) for each criteria are in bold.

		RMSE	MAE	PCC	MAPE
Normal	mean	0.085	0.054	0.952	16.657
	KNN	0.108	0.059	0.923	25.502
	methyLImp	0.063	**0.032**	0.974	10.593
	LASSO	0.063	0.034	0.974	11.086
	Ridge	0.063	0.034	0.974	11.131
	elastic net 0.2	**0.062**	0.034	**0.975**	10.925
	elastic net 0.5	**0.062**	0.034	**0.975**	10.923
	elastic net 0.8	0.063	0.034	0.974	10.920
	FPCR	0.066	0.037	0.972	11.554
	LASSO 1by1	0.064	0.033	0.973	10.725
	Ridge 1by1	**0.062**	0.034	**0.975**	11.013
	elastic net 0.2 1by1	**0.062**	0.033	**0.975**	**10.540**
	elastic net 0.5 1by1	0.063	0.033	0.974	10.578
	elastic net 0.8 1by1	0.064	0.033	0.973	10.680
Tumor	mean	0.118	0.080	0.905	24.745
	KNN	0.127	0.080	0.893	22.801
	methyLImp	0.095	0.057	0.941	17.799
	LASSO	0.097	0.059	0.938	17.774
	Ridge	0.101	0.062	0.933	18.846
	elastic net 0.2	0.096	0.059	0.939	18.043
	elastic net 0.5	0.096	0.059	0.940	17.664
	elastic net 0.8	0.096	0.059	0.939	17.642
	FPCR	0.100	0.062	0.935	18.215
	LASSO 1by1	0.096	0.057	0.939	17.040
	Ridge 1by1	0.095	0.058	0.940	17.626
	elastic net 0.2 1by1	**0.093**	**0.055**	**0.943**	**16.688**
	elastic net 0.5 1by1	0.094	0.056	0.941	16.812
	elastic net 0.8 1by1	0.095	0.056	0.941	16.917

3.4.3. DMR Detection. It is not possible to know the true regions that are differentially methylated between Tumor and Normal groups in the real data. Thus the complete dataset described in Section 2.1.1 with the most information on hand is used as a standard

for comparison. A potential false positive region is a region detected using the test data but not detected with the complete dataset. A potential false negative region is a region found by the complete dataset but not the test dataset.

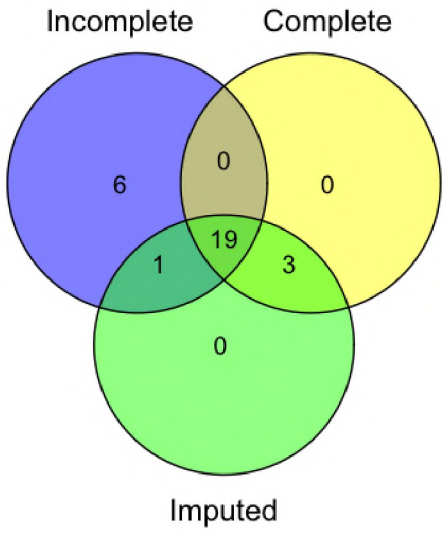


Figure 3.5. Venn diagram of DMRs detected using incomplete, complete and imputed datasets.

Based on the imputation accuracy performance, elastic net 0.2 1 by 1 method is used to impute the missing values in the dataset. Then Bumphunter method is used to detect DMRs among incomplete, complete and imputed data. As shown in Figure 3.5, 22 DMRs that overlap with the complete dataset are detected using the imputed data. This is an improvement compared to only 19 common DMRs between the incomplete and complete datasets. The number of potential false negative regions is reduced in the imputed dataset compared to the incomplete dataset. At the same time, the imputed dataset reduced the number of potential false positives to be only 1 compared to 7 using the incomplete dataset. Since the analysis on real data is conducted on a short section of Chromosome 1, these differences will accumulate when the entire genome is considered. More specifically, compared to a section with around 8000 probes, the entire genome has about 60 times more probes.

3.5. RESULTS FOR SIMULATED DATA

The simulated data are filtered by the criteria introduced in Section 2.1.1. Effects are added to the 96 selected DMR clusters in the complete dataset with 36,279 probes. Missing values are introduced to the 4,917 probes that are restored from the SNP data with the missing rates of 20%, 50% and 70%. The imputation accuracy table and DMR detection performance for the methods are given below.

3.5.1. Imputation accuracy. Table 3.5 shows the imputation accuracy for the simulated data with 20% missing rate between the different imputation methods. The elastic net 0.5, elastic net 0.8, elastic net 0.2 1 by 1 and elastic net 0.5 1 by 1 methods have the lowest RMSE. The LASSO 1 by 1, elastic net 1 by 1 with mixing parameters 0.2, 0.5 and 0.8 methods have the lowest MAE. All the elastic net methods and LASSO 1 by 1 yield the highest PCC. The elastic net 0.5 1 by 1 method has the lowest MAPE. For this missing rate, the overall performance of elastic net 0.5 1 by 1 method is the best.

Table 3.5. Imputation accuracy for simulated data with 20% missing rate. The optimal value(s) for each criteria are in bold.

	RMSE	**MAE**	**PCC**	**MAPE**
mean	0.090	0.058	0.953	19.627
KNN	0.070	0.040	0.972	13.020
methyLImp	0.069	0.036	0.973	12.490
LASSO	0.067	0.037	0.974	12.591
Ridge	0.071	0.037	0.972	12.584
elastic net 0.2	0.067	0.036	**0.975**	12.139
elastic net 0.5	**0.066**	0.036	**0.975**	12.072
elastic net 0.8	**0.066**	0.036	**0.975**	12.085
FPCR	0.074	0.043	0.969	13.874
LASSO 1by1	0.067	**0.035**	**0.975**	11.663
Ridge 1by1	0.067	0.037	0.974	12.287
elastic net 0.2 1by1	**0.066**	**0.035**	**0.975**	11.588
elastic net 0.5 1by1	**0.066**	**0.035**	**0.975**	**11.580**
elastic net 0.8 1by1	0.067	**0.035**	**0.975**	11.643

Table 3.6 shows the imputation accuracy for simulated data with 50% missing rate. In terms of RMSE, elastic net 0.5 and Ridge 1 by 1 methods perform the best. Elastic net 0.2 1 by 1 method yield the lowest MAE and MAPE. The elastic net (with mixing parameters 0.2, 0.5, 0.8), Ridge 1 by 1, and elastic net 0.2 1 by 1 methods have the highest PCC. Table 3.7 shows the imputation performance on simulated data with 70% missing rate. The Ridge 1 by 1 method has the lowest RMSE and highest PCC. The elastic net 0.2 1 by 1 method has the lowest MAE and MAPE.

Table 3.6. Imputation accuracy for simulated data with 50% missing rate. The optimal value(s) for each criteria are in bold.

	RMSE	**MAE**	**PCC**	**MAPE**
mean	0.091	0.058	0.952	19.768
KNN	0.081	0.044	0.963	14.784
methyLImp	0.073	0.039	0.970	13.337
LASSO	0.072	0.041	0.970	13.779
Ridge	0.074	0.041	0.968	13.730
elastic net 0.2	0.072	0.040	**0.971**	13.445
elastic net 0.5	**0.071**	0.040	**0.971**	13.394
elastic net 0.8	0.072	0.040	**0.971**	13.478
FPCR	0.078	0.045	0.965	15.109
LASSO 1by1	0.074	0.040	0.969	13.233
Ridge 1by1	**0.071**	0.039	**0.971**	13.313
elastic net 0.2 1by1	0.072	**0.038**	**0.971**	**12.912**
elastic net 0.5 1by1	0.072	0.039	0.970	13.028
elastic net 0.8 1by1	0.073	0.039	0.969	13.146

The average running times in seconds on a MacBook Pro with Processor 2.7 GHz Intel Core i5 and Memory 8 GB 1867 MHz DDR3 over 30 runs for a select subset of the imputation methods are recorded. One of the standard regularized imputation methods (elastic net 0.2), one of the 1 by 1 regularized imputation methods (elastic net 0.2 1 by 1) and the methyLImp method are compared. Elastic net 0.2 methods are chosen because they can represent other methods in the same method group, and their performance are stable among different settings. The elastic net 0.2 1 by 1 method is the fastest, while the elastic net 0.2 method is the slowest (Table 3.8).

Table 3.7. Imputation accuracy for simulated data with 70% missing rate. The optimal value(s) for each criteria are in bold.

	RMSE	**MAE**	**PCC**	**MAPE**
mean	0.092	0.059	0.951	20.534
KNN	0.114	0.066	0.925	31.077
methyLImp	0.078	**0.043**	0.965	15.027
LASSO	0.080	0.046	0.964	14.970
Ridge	0.080	0.046	0.964	14.991
elastic net 0.2	0.080	0.046	0.964	14.992
elastic net 0.5	0.080	0.046	0.964	14.991
elastic net 0.8	0.080	0.046	0.964	14.991
FPCR	0.083	0.048	0.961	15.488
LASSO 1by1	0.084	0.046	0.960	15.238
Ridge 1by1	**0.077**	0.044	**0.966**	14.416
elastic net 0.2 1by1	0.080	**0.043**	0.963	**14.367**
elastic net 0.5 1by1	0.081	0.044	0.962	14.630
elastic net 0.8 1by1	0.083	0.045	0.961	14.917

Table 3.8. Average running time in seconds over 30 runs

	Average time (standard deviation)
elastic net 0.2	540.62 (9.87)
elastic net 0.2 1 by 1	57.87 (1.11)
methyLImp	59.99 (3.54)

Figure 3.6 provides a visualization of the RMSE verses the missing rate to compare the standard and 1 by 1 regularized methods in both the real and simulated data. The performance between the standard regularized imputation methods and the 1 by 1 methods are similar, yet the 1 by 1 methods are much more computationally efficient. Thus only the 1 by 1 methods are further compared with other methods in Figure 3.7.

As a visualization and summary of Tables 3.2-3.7, Figure 3.7 compares the imputation accuracy of mean imputation, methyLImp, FPCR imputation and the 1 by 1 regularized methods with respect to RMSE for different missing rates in both the real and simulated data. The KNN method is not included in the figure since the RMSE is inflated dramatically

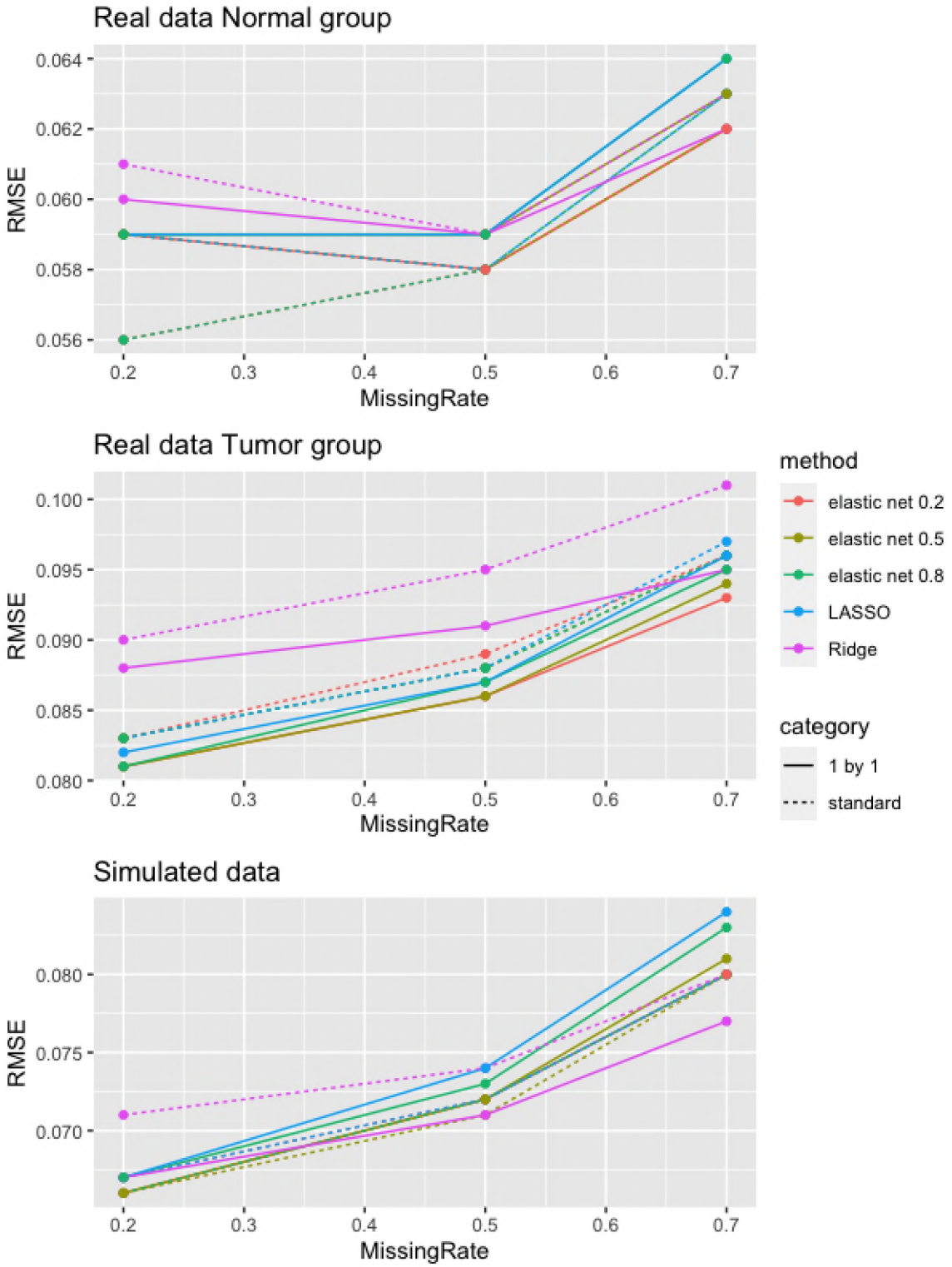


Figure 3.6. The root mean square error (RMSE) verses missing rate to compare the standard regularized imputation methods and 1 by 1 regularized imputation methods. The standard regularized imputation methods and the 1 by 1 methods have similar performance.

when the missing rate is high. The mean imputation method has the highest RMSE in all three datasets (real data Normal group, real data Tumor group and simulated data), followed by the FPCR method. The FPCR imputation has the second highest RMSE in all three datasets and for all missing rates, except for the simulated data at 70% missing rate. In the simulated data group with 70% missing rate, FPCR outperforms mean and LASSO 1 by 1 methods. The methyLImp method is shown to have higher RMSE than all the 1 by 1 regularized methods in the real data Normal group with missing rate 20% and 50%, and it only outperforms Ridge 1 by 1 and elastic net 0.8 1 by 1 methods with missing rate 70%. In this dataset, elastic net 0.2 1 by 1 method has the lowest RMSE at all missing rates. In the real data Tumor group, the elastic net methods have better performance in terms of RMSE than methyLImp. In the simulated data, all the regularized methods outperform methyLImp with 20% missing rate. Only the LASSO 1 by 1 method performed worse than methyLImp with 50% missing rate. The elastic net 0.2 1 by 1 method works the best in terms of MAE among all missing rates. When considering the overall imputation accuracy results across all datasets, the elastic net 0.2 1 by 1 method is recommended since it provides good performance and offers reasonable computational efficiency.

3.5.2. DMR detection. Using the 1 by 1 elastic net method with 0.2 mixing parameter recommended above, the DMR detection performance is assessed at the three different missing rates. Two DMR detection methods (Bumphunter and DMRcate) are applied. The regions detected as differentially methylated by both detection approaches are compared with the simulated true DMRs. Two regions are counted as overlapping if they share any common genomic locations on the chromosome. The relationship between two overlapping regions could be exactly the same (Figure 3.8 a), one region lying within the other (Figure 3.8 b), or one region partially in common with the other (Figure 3.8 c). Alternatively, one region can also overlap with multiple regions (Figure 3.8 d).

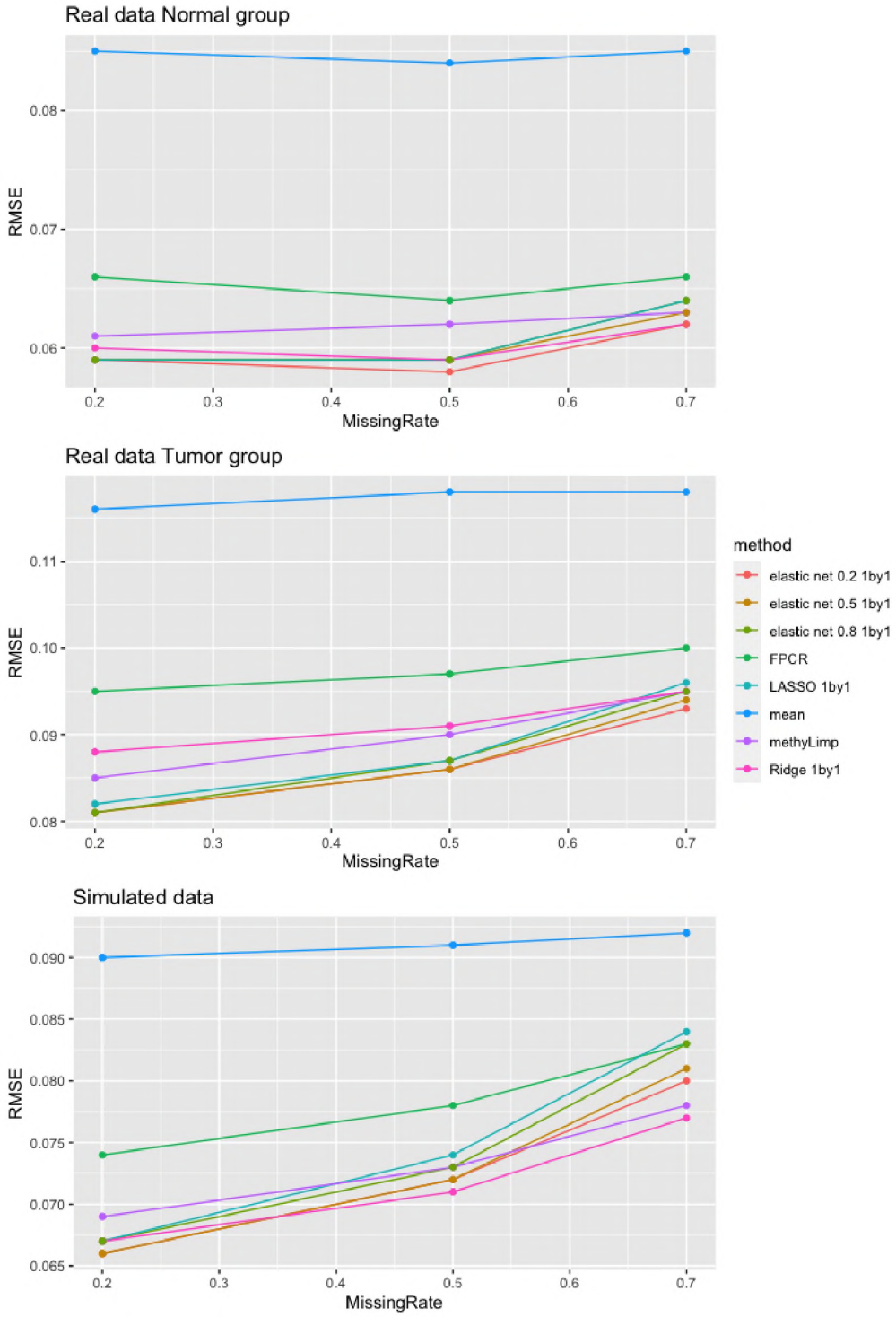


Figure 3.7. The root mean square error (RMSE) verses missing rate to compare the different imputation methods. The elastic net 0.2 1 by 1 method has good and stable performance across the three datasets and different missing rates.

Figure 3.8. Different cases of overlapping regions: a) two regions are exactly the same, b) one region is within the other region, c) two regions have partial overlap, and d) one pink region overlaps with two green regions.

There are 96 true DMRs. Figure 3.9 shows Venn Diagrams comparing the overlap in detecting the true DMRs before and after imputation using the Bumphunter method for the three different missing rates. Before imputation, using the data with missing entries, only 2 or 3 of the 96 DMRs can be found across the different missing rates. After imputation, 46 or 47 of the true DMRs can be detected. There are two numbers in the intersection of the Venn Diagrams for the 'After Imputation' results because one true DMR is broken into two regions, as shown in Figure 3.8 d. Using the Bumphunter method, the number of true positives increases by 45 (2 to 47) with 20% missing rate, and 43 (3 to 46) with 50% and 70% missing rates. The number of false positives also increases by 22 with all missing rates after the imputation method is applied. Figure 3.10 provides the results when the

DMRcate method is used for DMR detection. The imputation step increased the number of true positives by 86, 88 and 88 respectively for missing rates 20%, 50% and 70%, while also increasing the number of false positives by 58, 80, and 81, respectively.

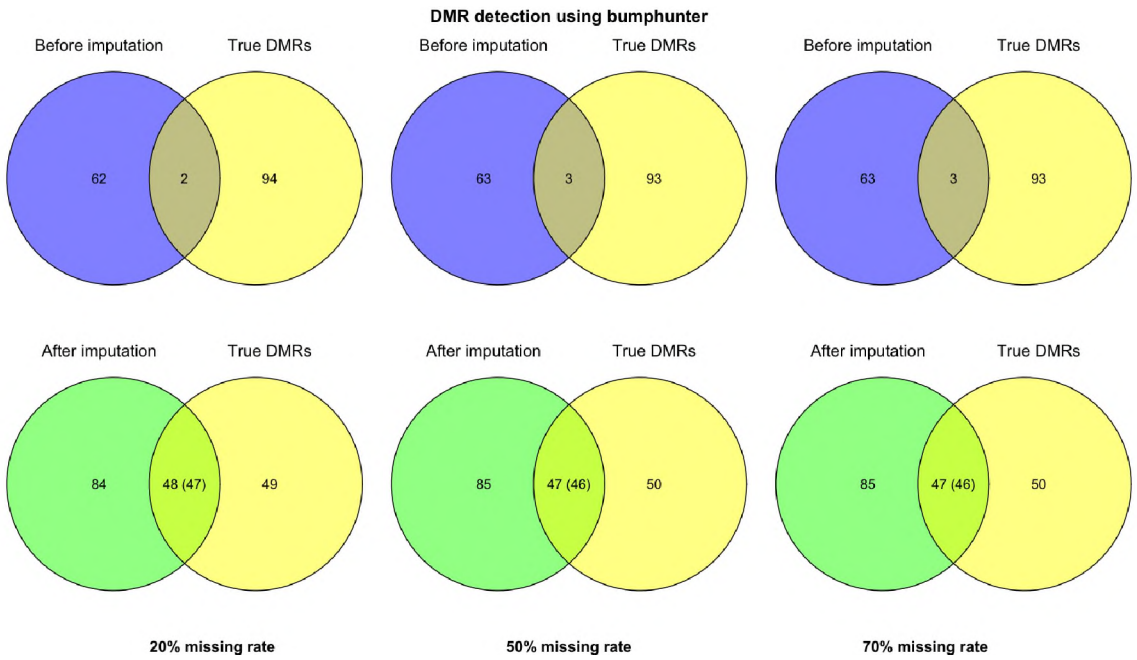


Figure 3.9. Venn diagrams to compare the DMRs found via the Bumphunter method before (top row) and after (bottom row) imputation to the true DMRs for different missing rates. When there are two numbers in the intersection, it means that one or more true regions are detected as multiple regions. Numbers in parentheses represent the number of true regions.

The proposed imputation method improved the DMR detection results despite the different DMR detection methods. To compare the DMR detection improvements among the two methods, Figure 3.11 shows the detected DMRs using the imputed dataset by Bumphunter method (blue) and DMRcate method (red) comparing to true DMRs (yellow) at different missing rates. The counts in the Venn diagrams are recorded in terms of the number of true DMRs in each part. These results show that the improvement on DMR

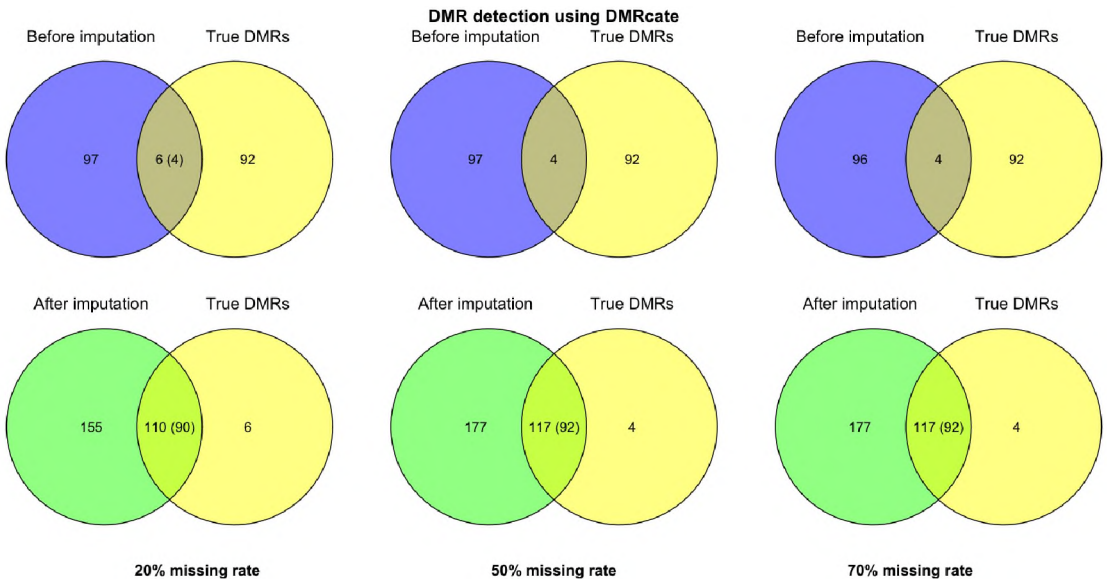


Figure 3.10. Venn diagrams to compare the DMRs found via the DMRcate method before (top row) and after (bottom row) imputation to the true DMRs for different missing rates. When there are two numbers in the overlapping part, it means that one or more true regions are detected as multiple regions. Numbers in parentheses represent the number of true regions.

detection by using the proposed imputation method is consistent for both DMR detection methods. All the true positive regions detected by Bumphunter are also detected by DMRcate method at all missing rates.

3.6. DISCUSSION OF RESULTS

When analyzing the real data, the imputation accuracy shows an apparent difference between Tumor group and Normal group. For example, using the same imputation method elastic net 1 by 1 with mixing parameter 0.2, and the 20% missing rate, the imputation accuracy in terms of RMSE is 0.059 in Normal group and 0.081 in Tumor group. The worst RMSE for the Normal group is 0.085 while the best RMSE for the Tumor group is 0.081. This may be caused by cancer heterogeneity. Previous research has shown the

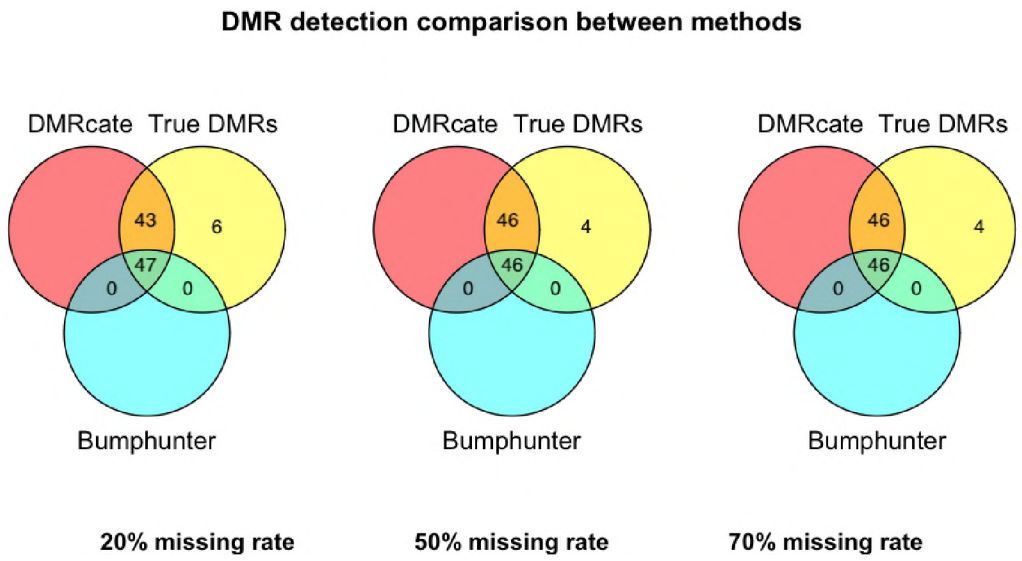


Figure 3.11. The detected DMRs using the imputed dataset by Bumphunter method (blue) and DMRcate method (red) compared to true DMRs (yellow). The counts of true DMRs are shown in the Venn diagrams.

existence of epigenetic heterogeneity among cancers (Liu et al., 2019; Teschendorff et al., 2016; Fernandez et al., 2012). In the study of Fernandez et al. (2012), DNA methylation profiles of 1505 CpG sites were examined on normal tissues and tumor tissues. It was found that little variation exists in the DNA methylation patterns of normal tissues but there was greater methylation heterogeneity among tumors. Hansen et al. (2011) suggested that the epigenetic instability of essential genomic domains in tumor cells can lead to increased methylation variability, and then contribute to cancer heterogeneity. The high variability in the methylation levels in the Tumor group can lead to the low imputation accuracy.

The KNN imputation method performed poorly at missing rates 50% and 70% as seen in Tables 3.3, 3.4, 3.6, and 3.7. The maximum percent of missing data allowed in each variable is limited for the KNN method. When the percentage is over a threshold (usually 50%), the missing value will be imputed using the overall mean of each sample. With a higher missing rate, KNN's performance is even worse than mean imputation, because the mean imputation uses the mean over all samples with complete entries of a particular methylation site, while the KNN uses the mean over variables from the same sample.

The functional principal component regression imputation methods does not perform as well as the regularized linear regression imputation. The reason is likely due to the large distance between probes. The human genome contains about 3 billion base pairs but is covered by only around 450,000 probes on the microarray. Although probes are more dense in some regions, on average neighboring probes may be too far away to maintain the correlation mentioned in Section 2.1.2.

To summarize, the regularized linear regression imputation methods proposed in this work have outperformed methyLImp in terms of RMSE, PCC and MAPE for the real data under different missing rates. For simulated data, the regularized linear regression imputation methods have better performance than methyLImp in terms of all four criteria (RMSE, MAE, PCC and MAPE) under all missing rates. The 1 by 1 regularized methods are more computationally efficient without much sacrifice in performance compared to the regularized methods. The elastic net 0.2 1 by 1 method is recommended based on its overall stable and good performance across most settings. While applying the imputation methods for DMR analysis, true positive detection are improved. Although the number of false positive detections also increased after imputation, the increase is small compared to the increase of true positive detections. Take the 20% missing rate for example, using the Bumphunter method, the number of true positives increased by 22.5 times (2 to 47) while the number of false positives increased by 0.35 times (62 to 84). Using the DMRcate method, the number of true positives increased by 21.5 times (4 to 90) while the number of false positives increased by 0.60 times (97 to 155).

4. CONCLUSION

4.1. SUMMARY

The filtering out of potential single nucleotide polymorphism (SNP) probes in the preprocessing step of DNA differential methylation studies causes an unnecessary waste of information. Incorporating SNP data into the DNA methylation analysis framework, allows a large proportion of the probes to be restored. The effects of recovering those probes are remarkable. The improvement on increasing the number of true DMRs has been demonstrated by both the real data analysis and simulation studies, which only utilize part of the genome. The effects will accumulate when the entire genome is considered.

In this research, SNP data are integrated with Infinium HumanMethylation450 BeadChip (HM450) methylation data to recover potential SNP-probes that do not actually have SNPs and apply novel imputation methods for missing data due to true SNPs or for other reasons. Missing data are categorized according to their missing mechanism as missing completely at random, missing at random or missing not at random. Missing DNA methylation data because of filtering is assumed to fall into the missing at random category. Missing rates of 20%, 50% and 70% are used to develop and test the proposed methodology.

Imputation methods are proposed in Chapter 2 for DNA methylation data. Several regularized regression methods are proposed, along with a functional data approach, and compared to three existing methods. Previous studies have shown that methylation levels are correlated with neighboring probes within short distances on the chromosome (Eckhardt et al., 2006). It has also been found that the methylation levels are highly correlated with other probes from the same sample (Zhang et al., 2015). This information can be used to aid in imputing missing methylation levels. For each probe with missing values, submatrices are extracted from the data to fit a regression model that is used to attain the imputed values.

The model is fit by using the available data at the missing probe as the response variable and data from other probes with complete information as the predictor variables. The imputation steps in this research iteratively evaluate all subsets of probes with missing entries. The input data is organized with each row representing a sample and each column representing a probe. First a predictive model is built under the regularized linear regression framework, then the missing values are imputed by prediction using complete entries of the same sample. Ridge, LASSO and elastic net regression are explored as shrinkage approaches. The tuning parameter that determines the amount of shrinkage for each model is selected by cross validation. This step makes the computational speed slow because cross validation is needed for each iteration. Therefore, variable screening before the regularization step is recommended. Also, imputing the missing values site by site is recommended since two sites may have different sets of most correlated predictors. The selection criteria for including variables in the model is the Pearson correlation between the predictors and the response variable. The number of probes used in the regression model is set to be the same as the number of samples in the model. In an alternative approach, the measurements of each sample are treated as one observation with a smooth curve representing the underlying structure based on the correlation between neighboring probes of DNA methylation data. Functional principal component analysis is performed and the component scores are used as inputs into a functional linear regression model, which is used to perform the imputations.

The proposed imputation methods are evaluated and compared to existing methods using both real and simulated data. A simulation study is conducted based on real data to keep the natural structure of the DNA methylation data. Adjacent site clustering is applied to reveal potential clusters (regions) using the normal samples of the real data. Among these clusters, a subset is randomly selected in which known effects are added to differentiate the two groups. Considering hyper- and hypomethylation patterns in the human genome (Peters et al., 2015), two types of simulated regions are applied. For 50% of the clusters,

the effect is added to the entire cluster. For the other 50% of the clusters, the effect is added at the probe resolution, meaning that only the probe with high group mean will have the effect added to the specific probe.

Performance of the proposed methods is assessed by two aspects. The first is the imputation accuracy. The regularized methods have the best overall performance with respect to imputation accuracy, followed by methyLImp, then FPCR imputation. The traditional imputation methods such as mean and KNN imputation perform worse. In terms of computational efficiency, the regularized 1 by 1 approach is more efficient with similar imputation accuracy as the regularized methods. The second way imputation performance is evaluated is by investigating the impact on DMR detection. Using simulated data with true DMRs known, imputation using the 1 by 1 approach for the elastic net with mixing parameter 0.2 increased the number of true positives and decreased the number of false negatives compared to analyzing the data without doing imputation. The number of false positive detections also increased with the imputed dataset, but this increase was minimal compared to the increase in true positive detections.

4.2. FUTURE WORK

In this research, efforts have been focused on restoring the probes that are filtered out of HM450 data for the reason of potentially having a SNP. According to Table 3.1, high detection p-values are another main reason for probe filtering. Those probes are removed because they are carrying mainly noise instead of methylation information. Imputation and simulation studies can be conducted to improve the DMR detection ability of the regions involving those probes. Also, methods and theory can be tested and modified for other types of DNA methylation data such as Illumina Infinium Methylation EPIC array and next generation sequencing data. The FPCR method has a high potential to show improvement on whole genome bisulfite sequencing (WGBS) data due to the comprehensive and dense

coverage of the genome it provides. The proposed methods in this study incorporate the correlation between genomic variables into the imputation process, so they may be generalized to other genomic data such as gene expression data.

REFERENCES

J. B. Schnog, A. J. Duits, F. A. Muskiet, H. ten Cate, R. A. Rojer, and D. P. Brandjes, "Sickle cell disease; a general overview," *Netherlands Journal of Medicine*, vol. 62, no. 10, 2004.

J. W. Wei, K. Huang, C. Yang, and C. S. Kang, "Non-coding RNAs as regulators in epigenetics (Review)," *Oncology Reports*, vol. 37, no. 1, pp. 3–9, jan 2017.

L. D. Moore, T. Le, and G. Fan, "DNA methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, jan 2013.

R. Hotchkiss, "The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography," *Journal of Biological Chemistry*, 1948.

R. Holliday and J. Pugh, "DNA modification mechanisms and gene activity during development," *Science*, vol. 187, no. 4173, pp. 226–232, jan 1975.

S. J. Compere and R. D. Palmiter, "DNA methylation controls the inducibility of the mouse metallothionein-I gene in lymphoid cells," *Cell*, vol. 25, no. 1, pp. 233–240, 1981.

B. Tycko, "DNA methylation in genomic imprinting," *Mutation Research - Reviews in Mutation Research*, vol. 386, no. 2, pp. 131–140, apr 1997.

J. D. Hollister and B. S. Gaut, "Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression," *Genome Research*, vol. 19, no. 8, pp. 1419–1428, aug 2009.

K. L. Sheaffer, R. Kim, R. Aoki, E. N. Elliott, J. Schug, L. Burger, D. Schübeler, and K. H. Kaestner, "DNA methylation is required for the control of stem cell differentiation in the small intestine," *Genes and Development*, vol. 28, no. 6, pp. 652–664, mar 2014.

D. M. Messerschmidt, B. B. Knowles, and D. Solter, "DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos," *Genes and Development*, vol. 28, no. 8, pp. 812–828, apr 2014.

D. Bayarsaihan, "Epigenetic mechanisms in inflammation," *Journal of Dental Research*, vol. 90, no. 1, pp. 9–17, jan 2011.

C. Bock, "Analysing and interpreting DNA methylation data," *Nature Reviews Genetics*, 2012.

M. Ongenaert, "Epigenetic databases and computational methodologies in the analysis of epigenetic datasets," in *Advances in Genetics*, 2010, vol. 71, pp. 259–295.

D. Takai and P. A. Jones, "Comprehensive analysis of CpG islands in human chromosomes 21 and 22," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 6, pp. 3740–3745, mar 2002.

S. Saxonov, P. Berg, and D. L. Brutlag, "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 5, pp. 1412–1417, jan 2006.

J. Tazi and A. Bird, "Alternative chromatin structure at CpG islands," *Cell*, vol. 60, no. 6, pp. 909–920, mar 1990.

C. Vinson and R. Chatterjee, "Cg methylation," *Epigenomics*, vol. 4, no. 6, pp. 655–663, 2012.

D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

S. A. Wajed, P. W. Laird, and T. R. DeMeester, "DNA methylation: An alternative pathway to cancer," *Annals of Surgery*, vol. 234, no. 1, pp. 10–20, 2001.

G. Severi, M. C. Southey, D. R. English, C. hee Jung, A. Lonie, C. McLean, H. Tsimiklis, J. L. Hopper, G. G. Giles, and L. Baglietto, "Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer," *Breast Cancer Research and Treatment*, vol. 148, no. 3, pp. 665–673, nov 2014.

P. M. Das and R. Singal, "DNA methylation and cancer," *Journal of Clinical Oncology*, vol. 22, no. 22, pp. 4632–4642, 2004.

M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 5, pp. 1827–1831, 1992.

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*, vol. 33, no. 18, pp. 5868–5877, 2005.

F. J. Steemers and K. L. Gunderson, "Illumina, inc." 2005.

D. J. Weisenberger, D. V. D. Berg, F. Pan, B. P. Berman, P. W. Laird, S. California, and U. S. C. Norris, "Comprehensive DNA Methylation Analysis on the Illumina® Infinium® Assay Platform," *Application Note: Illumina Epigenetic Analysis*, 2008.

TCGA, "The cancer genome atlas program," 2021. [Online]. Available: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Illumina, "Infinium Human Methylation 450 data sheet," *Illumina*, p. 3, 2012.

R. Pidsley, E. Zotenko, T. J. Peters, M. G. Lawrence, G. P. Risbridger, P. Molloy, S. Van Djik, B. Muhlhausler, C. Stirzaker, and S. J. Clark, "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling," *Genome Biology*, vol. 17, no. 1, p. 208, oct 2016.

A. Vignal, D. Milan, M. SanCristobal, and A. Eggen, "A review on SNP and other types of molecular markers and their use in animal genetics," *Genetics Selection Evolution*, vol. 34, no. 3, 2002.

B. S. Shastry, "SNP alleles in human disease and evolution," *Journal of Human Genetics*, vol. 47, no. 11, 2002.

T. LaFramboise, "Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances," *Nucleic Acids Research*, vol. 37, no. 13, pp. 4181–4193, 2009.

J. A. Heiss and A. C. Just, "Improved filtering of DNA methylation microarray data by detection p values and its impact on downstream analyses," *Clinical Epigenetics*, vol. 11, no. 1, jan 2019.

C. Reilly, A. Raghavan, and P. Bohjanen, "Global assessment of cross-hybridization for oligonucleotide arrays," *Journal of Biomolecular Techniques*, vol. 17, no. 2, pp. 163–172, apr 2006.

Y. A. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, feb 2013.

W. Zhou, P. W. Laird, and H. Shen, "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes," *Nucleic Acids Research*, vol. 45, no. 4, p. gkw967, oct 2016.

J. Nordlund, C. L. Bäcklin, P. Wahlberg, S. Busche, E. C. Berglund, M. L. Eloranta, T. Flaegstad, E. Forestier, B. M. Frost, A. Harila-Saari, M. Heyman, Ó. G. Jónsson, R. Larsson, J. Palle, L. Rönnblom, K. Schmiegelow, D. Sinnett, S. Söderhäll, T. Pastinen, M. G. Gustafsson, G. Lönnerholm, and A. C. Syvänen, "Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia," *Genome Biology*, vol. 14, no. 9, p. r105, sep 2013.

S. Dedeurwaerder, M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks, "Evaluation of the Infinium Methylation 450K technology," *Epigenomics*, vol. 3, no. 6, pp. 771–784, dec 2011.

B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, feb 2003.

J. Maksimovic, L. Gordon, and A. Oshlack, "SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips." *Genome biology*, vol. 13, no. 6, p. R44, jun 2012.

A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, jan 2013.

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

K. D. Hansen, B. Langmead, and R. A. Irizarry, "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions," *Genome Biology*, vol. 13, no. 10, pp. 1–10, oct 2012.

Y. Park, M. E. Figueroa, L. S. Rozek, and M. A. Sartor, "MethylSig: A whole genome DNA methylation analysis pipeline," *Bioinformatics*, vol. 30, no. 17, pp. 2414–2422, sep 2014.

P. Du, X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin, "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, vol. 11, no. 1, p. 587, nov 2010.

R. T. Barfield, V. Kilaru, A. K. Smith, and K. N. Conneely, "CpGassoc: An R function for analysis of DNA methylation microarray data," *Bioinformatics*, vol. 28, no. 9, pp. 1280–1281, may 2012.

S. J. Baek, S. Yang, T. W. Kang, S. M. Park, Y. S. Kim, and S. Y. Kim, "MENT: Methylation and expression database of normal and tumor tissues," *Gene*, vol. 518, no. 1, pp. 194–200, apr 2013.

D. Wang, L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia, and S. Liu, "IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data," *Bioinformatics*, vol. 28, no. 5, pp. 729–730, mar 2012.

C. D. Warden, H. Lee, J. D. Tompkins, X. Li, C. Wang, A. D. Riggs, H. Yu, R. Jove, and Y.-C. Yuan, "Cohcap: an integrative genomic pipeline for single-nucleotide resolution dna methylation analysis," *Nucleic acids research*, vol. 41, no. 11, pp. e117–e117, 2013.

G. K. Smyth, "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments *," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth, "Limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, jan 2015.

F. Eckhardt, J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck, "DNA methylation profiling of human chromosomes 6, 20 and 22," *Nature Genetics*, vol. 38, no. 12, pp. 1378–1385, dec 2006.

Y. Zhang, H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang, and Y. Cui, "QDMR: A quantitative method for identification of differentially methylated regions by entropy," *Nucleic Acids Research*, vol. 39, no. 9, p. e58, may 2011.

A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," *International Journal of Epidemiology*, vol. 41, no. 1, pp. 200–209, feb 2012.

T. J. Peters, M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V. Lord, S. J. Clark, and P. L. Molloy, "De novo identification of differentially methylated regions in the human genome," *Epigenetics & Chromatin 2015 8:1*, vol. 8, no. 1, p. 6, jan 2015.

L. M. Butcher and S. Beck, "Probe lasso: a novel method to rope in differentially methylated regions with 450k dna methylation data," *Methods*, vol. 72, pp. 21–28, 2015.

Z. Wang, M. A. Jensen, and J. C. Zenklusen, "A practical guide to The Cancer Genome Atlas (TCGA)," in *Methods in Molecular Biology*. Humana Press Inc., 2016, vol. 1418, pp. 111–141.

T. J. Hudson, W. Anderson, A. Aretz, A. D. Barker, C. Bell, R. R. Bernabé, M. K. Bhan, F. Calvo, I. Eerola, D. S. Gerhard *et al.*, "International network of cancer genome projects," *Nature*, vol. 464, no. 7291, pp. 993–998, apr 2010.

K.-L. Huang, R. J. Mashl, Y. Wu, D. I. Ritter, J. Wang, C. Oh, M. Paczkowska, S. Reynolds, M. A. Wyczalkowski, N. Oak, A. D. Scott *et al.*, "Pathogenic Germline Variants in 10,389 Adult Cancers," *Cell*, vol. 173, no. 2, pp. 355–370.e14, apr 2018.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, aug 2009.

Y. Tian, T. J. Morris, A. P. Webster, Z. Yang, S. Beck, A. Feber, and A. E. Teschendorff, "Champ: updated methylation analysis pipeline for illumina beadchips," *Bioinformatics*, vol. 33, no. 24, pp. 3982–3984, 2017.

TCGA, "Tcga barcode," 2011. [Online]. Available: https://docs.gdc.cancer.gov/Encyclopedia/ pages/TCGA_Barcode/

H. Hernandez-Vargas, M.-P. Lambert, F. L. Calvez-Kelm, G. Gouysse, S. McKay-Chopin, S. V. Tavtigian, J.-Y. Scoazec, and Z. Herceg, "Hepatocellular carcinoma displays distinct dna methylation signatures with potential as clinical predictors," *PLoS ONE*, vol. 5, p. e9749, 3 2010.

X. Ma, Y. W. Wang, M. Q. Zhang, and A. F. Gazdar, "Dna methylation data analysis and its application to cancer research," pp. 301–316, 6 2013.

L. Sun, S. Namboodiri, E. Chen, and S. Sun, "Preliminary Analysis of Within-Sample Co-methylation Patterns in Normal and Cancerous Breast Samples," *Cancer Informatics*, vol. 18, p. 117693511988051, jan 2019.

D. Jia, R. Z. Jurkowska, X. Zhang, A. Jeltsch, and X. Cheng, "Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation," *Nature*, vol. 449, no. 7159, pp. 248–251, sep 2007.

C. Lövkvist, I. B. Dodd, K. Sneppen, and J. O. Haerter, "DNA methylation in human epigenomes depends on local topology of CpG sites," *Nucleic Acids Research*, vol. 44, no. 11, pp. 5123–5132, jun 2016.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

D. B. Rubin, "Inference and Missing Data," *Biometrika*, vol. 63, no. 3, p. 581, dec 1976.

P. D. Lena, C. Sala, A. Prodi, and C. Nardini, "Data and text mining Missing value estimation methods for DNA methylation data," *Bioinformatics*, pp. 1–8, 2019.

R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., aug 2002.

S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018.

G. Molenberghs, G. Fitzmaurice, M. Kenward, and A. Tsiatis, *Handbook of missing data methodology*, 2014.

H. Liu-Seifert, S. Zhang, D. D'Souza, and V. Skljarevski, "A closer look at the baseline-observation-carriedforward (BOCF)," *Patient Preference and Adherence*, vol. 4, pp. 11–16, jan 2010.

G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische Mathematik*, vol. 14, pp. 403–420, 4 1970.

R. Penrose, "A generalized inverse for matrices," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, pp. 406–413, 1955.

P. D. Lena, C. Sala, A. Prodi, and C. Nardini, "R package methylimp," 2019. [Online]. Available: https://github.com/pdilena/methyLImp

M. Fuentes, P. Guttorp, and P. D. Sampson, "Using transforms to analyze space-time processes," *Monographs on Statistics and Applied Probability*, vol. 107, p. 77, 2006.

R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

F. Husson and J. Josse, "Handling missing values in multiple factor analysis," *Food quality and preference*, vol. 30, no. 2, pp. 77–85, 2013.

D. J. Stekhoven and P. Bühlmann, "Missforest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2012.

C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. Mcclean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, jan 2013.

T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

A. E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.

R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, jan 1996.

H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 67, no. 2, 2005.

J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

J. O. Ramsay and B. W. Silverman, *Applied functional data analysis: methods and case studies*. Springer, 2007.

C. De Boor, "Calculation of the smoothing spline with weighted roughness measure," *Mathematical Models and Methods in Applied Sciences*, vol. 11, no. 01, pp. 33–41, 2001.

I. Jolliffe, *Principal Component Analysis*. Springer Series in Statistics, 2002.

F. Yao, H. G. Muller, and J. L. Wang, "Functional Data Analysis for Sparse Longitudinal Data," *Annals of Statistics*, vol. 33, no. 6, 2005.

J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall, may 1996.

F. E. Satterthwaite, "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, vol. 2, no. 6, p. 110, dec 1946.

T. Sofer, E. D. Schifano, J. A. Hoppin, L. Hou, and A. A. Baccarelli, "A-clustering: A novel method for the detection of co-regulated methylation regions, and regions associated with exposure," *Bioinformatics*, vol. 29, no. 22, pp. 2884–2891, nov 2013.

K. Day, L. L. Waite, A. Thalacker-Mercer, A. West, M. M. Bamman, J. D. Brooks, R. M. Myers, and D. Absher, "Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape," *Genome Biology*, vol. 14, no. 9, p. R102, sep 2013.

Y. Liu, Y. Gu, M. Su, H. Liu, S. Zhang, and Y. Zhang, "An analysis about heterogeneity among cancers based on the DNA methylation patterns," *BMC Cancer*, vol. 19, no. 1, p. 1259, dec 2019.

A. E. Teschendorff, Y. Gao, A. Jones, M. Ruebner, M. W. Beckmann, D. L. Wachter, P. A. Fasching, and M. Widschwendter, "DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer," *Nature Communications*, vol. 7, no. 1, pp. 1–12, jan 2016.

A. F. Fernandez, Y. Assenov, J. I. Martin-Subero, B. Balint, R. Siebert, H. Taniguchi, H. Yamamoto, M. Hidalgo, A. C. Tan, O. Galm, I. Ferrer, M. Sanchez-Cespedes, A. Villanueva, J. Carmona, J. V. Sanchez-Mut, M. Berdasco, V. Moreno, G. Capella, D. Monk, E. Ballestar, S. Ropero, R. Martinez, M. Sanchez-Carbayo, F. Prosper, X. Agirre, M. F. Fraga, O. Graña, L. Perez-Jurado, J. Mora, S. Puig, J. Prat, L. Badimon, A. A. Puca, S. J. Meltzer, T. Lengauer, J. Bridgewater, C. Bock, and M. Esteller, "A DNA methylation fingerprint of 1628 human samples," *Genome Research*, vol. 22, no. 2, pp. 407–419, feb 2012.

K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg, "Increased methylation variation in epigenetic domains across cancer types," *Nature Genetics*, vol. 43, no. 8, pp. 768–775, aug 2011.

W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt, "Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements," *Genome biology*, vol. 16, no. 1, pp. 1–20, 2015.

VITA

In July 2009, Yuqing Su graduated from Shandong Agricultural University, China with a Bachelor of Engineering degree in Bioengineering. She earned her Master of Science degree in Microbiology from Shandong University, China in June 2013. She entered Missouri University of Science and Technology in August 2015 and received her Ph.D in Mathematics with Statistics Emphasis in May 2021.