Differential Methylation Analysis of Colon Tissues

by

Xiaoyu Yan

A Thesis Presented to the
FACULTY OF THE USC KECK SCHOOL OF MEDICINE
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
BIOSTATISTICS

August 2020

Table of Contents

Abstract

DNA methylation is a common phenomenon that occurs at CpG dinucleotides in eukaryotes, often playing a critical role in gene expression regulation. There is much evidence that shows an association between DNA methylation and disease, especially cancer. Microarrays are a widely used technique to measure DNA methylation in a high-throughput manner. To analyze methylation data from such large generated datasets, a wide variety of statistical approaches and methods have appeared and have been adopted in both preprocessing/QC and final analysis. One aspect of such analyses is that of differential methylation, which plays an important role in understanding the information contained in the human methylome in epigenetic studies.

In this thesis, we analyzed data resulting from samples from both colon tumors and from normal colon tissue. Specifically, we conduct a comparison looking at variation between normal and tumor tissue. As part of this analysis, we describe the statistical methods that are applied in both preprocessing and analyzing the data. Our data arose from the Illumina Infinium MethylationEPIC array, and we conducted our analysis using R packages such as SeSAMe and ChAMP. Our focus here is on detecting differential methylation status at both the site and region level. After detecting differential methylated positions and regions, we also show the significant CpG-related genes and enriched pathways.

Introduction

Completed in April 2003, the human genome project has mapped and sequenced the entire human genome including 28.5 hundred billion nucleotides, covering 99% of the autosomal genome, and led to the Big Data era of analysis of the genetic code (International Human Genome Sequencing Consortium, 2004). This thesis focuses on one aspect of such data, methylation and its role in cancer. Past research has shown that hypomethylation is seen in genes related to some human cancers (Feinberg and Vogelstein, 1983).

With the development of two major molecular biological techniques, sequencing and microarrays, the amount of information we have regarding genomics, epigenomics, and transcriptomics has exponentially expanded. For example, an Illumina InfiniumÍ Global Screening Array used by 23andMe can identify 600,000 single nucleotide polymorphisms (SNPs). SNPs are markers that are sometimes correlated with observable phenotypes or that sometimes contribute to disease. Such mass information about structural and functional characteristics of the genome and transcriptome under various conditions is beneficial to our understanding of the causes and progression of diseases.

Epigenetic studies have revealed that epigenetic change may play an important role in regulating gene expression; perhaps a more important role than that played by nucleotide variation in the genome (Berger et al., 2009). DNA methylation and histone modification, small RNA interference, and nucleosome and chromatin remodeling are typical forms of epigenetic changes.

This thesis focuses on DNA methylation. DNA methylation is a common epigenetic mechanism that modulates gene transcription and plays a role in genomic instability. It, therefore, plays an important role in human disease (Robertson, 2005). In the DNA methylation process, a methyl group is transferred by

DNA methyltransferase to the cytosine nucleotide that is adjacent to a guanine nucleotide in the DNA sequence along the 5ù to 3ù direction (see Figure 1). Such a cytosine-phosphate-guanine site is referred to as a CpG site, and these sites are scattered throughout the genome. Regions containing a high-frequency (>55%) of CpG sites are called CpG islands (CGIs), which appear in 40% of gene-promoters, and CpGs in such islands usually are unmethylated (see Figure 2). While some studies restrict their attention to specific features, such as CpG islands, it has been shown that including other nearby regions, such as shores (~2Kb from islands, > 75% of tissue specific differentially methylated regions can be found) and shelves (~4Kb from islands), can provide insight for studies of colorectal cancer (Visone et al., 2019). For that reason, in this paper, we will consider all CpGs, regardless of location.

In mammals, 70%-80% of CpG sites are methylated in different domains of the genome (Jones, 2012). Heterochromatin is normally hypermethylated, and if hypomethylated will promote genetic instability, leading to cancerization (Sharma, Kelly and Jones, 2010). On the contrary, if hypermethylation aberrantly occurs in a promoter, transcription factors may fail to bind, or methyl-CpG-binding proteins (MBPs) may introduce co-repressor molecules to repress downstream gene activity (Klose and Bird, 2006). The silencing of tumor suppressor genes and pro-differentiation factor genes also is indicative of cancer formation (Jones and Baylin, 2002). DNA methylation sometimes occurs within the gene body, and evidence in mouse has shown that this can preclude spurious transcription inception to avoid the production of aberrant proteins and ensure transcription fidelity, providing an explanation of the hypomethylation characteristic of gene bodies in tumors (Neri et al., 2017).

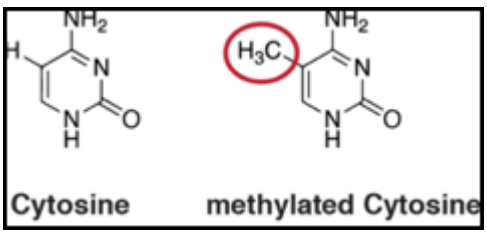

Figure 1: Chemical make-up of methylation of CpG sites. (wikipedia.org)

There is some evidence that epigenetic variation can be transgenerationally transmitted. Thus, events influencing the original generation might not only directly affect the first-generation but also act on the next generation (Heard and Martienssen, 2014; Klosin et al., 2017; Horsthemke, 2018). One hypothesis for the mechanism of transgenerational inheritance is that it is based on small RNA regulation (Houri-Zeevi and Rechavi, 2017). The epigenetic modification potentially occurs in any cell cycles in any type of cell. For example, during T-cell differentiation, although the progeny does not have DNA sequence specificity, they develop to perform different functions by responding to the epigenetic processes of transcription factor changes in the nuclear context (Wilson, Rowell and Sekimata, 2009). Such advancement in our understanding of the mechanisms of epigenetic inheritance provides the in-depth insights that may lead to widespread use of epigenetic therapy (Egger et al., 2004).

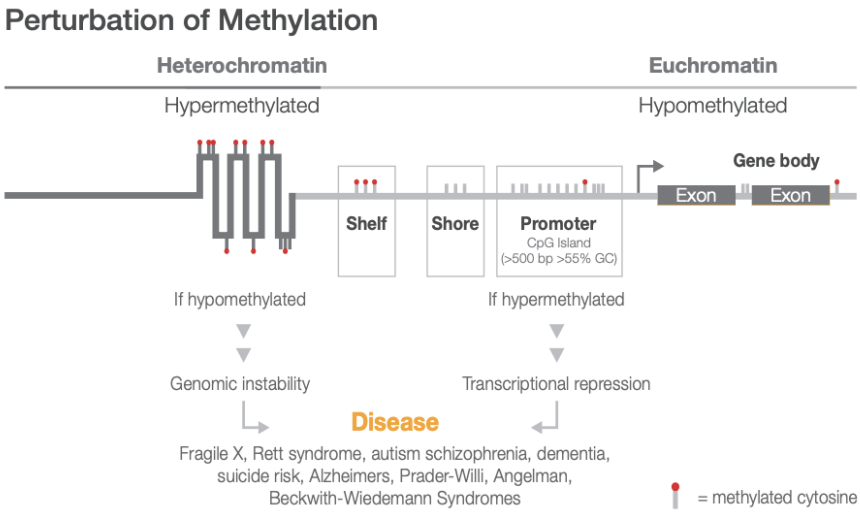

Figure 2: Scenarios in which methylation occurs (illumina.com)

Methylation has tissue specificity and, as discussed above, heritability (Silva and White, 1988). Much of the heritability evidence has been found from zebrafish, in which the hypermethylated DNA methylome of sperm was seen to direct the methylation status for the embryo and thereby potentially reprogram the maternal genome to demethylate or to experience de novo methylation (Jiang et al., 2013).

In contrast to the inheritance discussed above, DNA methylation is also known to be highly correlated with age. For embryonic stem cells and induced pluripotent stem cells, the DNA methylation level is approximately zero and the whole-genome is active with a hypomethylated methylome. As growth and development occur, methylation (and methylation variation) gradually increases. Aging is also a major factor associated with specific region hypermethylation in tumor cells (Ahuja et al., 1998). In general, there is a trend for methylation to increase in low methylation level regions (i.e. Islands). Conversely, since most CpG sites are actually outside of CpG islands, methylation tends to be lost with age elsewhere (Jones, Goodman and Kobor, 2015). Age-related methylation phenomena, such as epigenetic drift, have been found in Alzheimerùs disease (Wang, Oelze and Schumacher, 2008) and an epigenetic clock has been discovered in HIV infection (Horvath and Levine, 2015). Thus, DNA methylation can be an indicator to measure tissue age and to distinguish different parts of the genome (Horvath, 2013; Horvath and Raj, 2018).

Overall, methylation is a dynamic and complicated form of variation, with many interrelated processes occurring. As such, there is much interest in discovering the methylation state of tissue and understanding how that might relate to features such as disease status, particularly for cancer. Various methods have been developed to detect methylation changes or assess variation of target DNA methylation regions and each has its advantages in terms of stability, throughput capacity, resolution and cost (Kurdyukov and Bullock, 2016). We summarize those methods in Tables 1a&b below.

In this thesis, we focus on data from colon tumor (and normal) samples. A growing body of evidence has been produced to show that tumorigenesis is a multistep, multiphase process involving numerous genes and that cancer stem cells are probably the origin of all malignant cells in primary tumors (Jordan, Guzman and Noble, 2006). Genetic changes, environmental differences and reversible changes in cell characteristics also influence the phenotypic and functional heterogeneity of cancer cells within the same tumor (Meacham and Morrison, 2013). Many cancer epigenetic studies have shed light on the transformation from a normal cell to a cancer stem cell, and note that this is driven by methylation (Barker et al., 2009; Zeuner et al.,

2014; Gehart and Clevers, 2019). In order to decode the high volume of information now being measured from the genome, and more specifically here the methylome, statistical analysis methods have been developed for the analysis of methylation patterns. The variation of methylation data patterns between tumors and controls, and even between tumors, results from the process of epigenetic drift and can be used to measure the kinship of the cells, infer tumor onset and trace the progression of a colorectal tumor (Siegmund, Marjoram and Shibata, 2008; Siegmund et al., 2009; Hong et al., 2010). Differential methylation regions (DMRs), our focus here, are defined to be regions in which the degree of methylation varies between two groups of interest. They often contain SNPs associated with diseases. Methods to uncover the most informative DMRs to classify cell phenotype have been developed (Ziller et al., 2013). Our present study aims to discover whether there are regions of differential methylation between colorectal tumor tissues and normal tissues (specifically, colon crypts containing stem cells).

Table 1a: Current methods for high-throughput DNA methylation analysis (Suzuki and Bird, 2008)

Pretreatment method	General basis	Resolution	Advantages	Disadvantages
Bisulphite conversion	Sodium bisulphite converts unmethylated cytosine to uracil, whereas methylated cytosines are protected from conversion	High: single base resolution	Applicable to any samples	Complete conversion is essential
Methylation-sensitive restriction enzyme methods				
RLGS; HELP assay	DNA is differentially fragmented with a methylation-sensitive restriction enzyme. Following size fractionation, this method enriches methylated DNA	Moderate	Relatively simple	Analysis limited to methylation at restriction sites
McrBC digestion	DNA digestion with a methylation-specific restriction enzyme, McrBC. Following size fractionation, this method enriches unmethylated DNA	Moderate	Effective in degrading most methylated DNA	–
Affinity purification methods				
Methylated DNA immunoprecipitation (MedIP)	Immunoprecipitate DNA containing methylated cytosines using a monoclonal antibody	Moderate	The antibody is commercially available. Precipitates methylated cytosines in all contexts	High m5C density required
MBD affinity purification (MAP)	Immunoprecipitate DNA containing methylated CpG using an MBD column	Moderate	Only methylated CpGs are recovered	High m5CpG density required
CXXC affinity purification (CAP)*	Immunoprecipitate DNA containing unmethylated CpG using a CXXC-domain column	Moderate	A direct method to extract unmethylated DNA	High CpG density required

*X could be any residue. HELP, HpaII tiny fragment enrichment by ligation-mediated PCR; m5C, 5-methyl cytosine; m5CGI, CGI containing m5C; MBD, methyl-binding domain; RLGS, restriction landmark genome scanning.

Table 1b: Current methods for high-throughput DNA methylation analysis (Suzuki and Bird, 2008)

Readout method	Sample pretreatment method	General basis	Resolution	Other features	Uses
DNA microarrays					
Oligonucleotide arrays	Bisulphite conversion, methylation-sensitive restriction enzyme or affinity purification methods	Short (25-mer) or long (60-mer) oligonucleotide array	Moderate	–	Tiling genomic arrays, promoter arrays and custom arrays
SNP arrays		SNP selective probe array	Moderate	–	Detection of allele-specific DNA methylation
BeadArray (Illumina)	Bisulphite conversion	Ratio of the methylated and unmethylated PCR products is determined at single CpG sites	High: single-base resolution, quantitative	A large set of primers needs to be designed	Detection of methylation polymorphisms (96 samples assayed in parallel)
Sequencing					
Standard sequencing	Bisulphite conversion	Sanger sequencing	High: single-base resolution, quantitative	–	Expensive and labour intensive for genome-wide analysis
Direct large-scale sequencing	Bisulphite conversion, methylation-sensitive restriction enzyme or affinity purification methods	Short-read sequencing (Solexa sequencing: 40 million reads of 25–35 bases; 454 sequencing: 400,000 reads of >100 bases)	High: single-base resolution, quantitative	High-quality reference sequence is required	Fast and relatively inexpensive. Genotype information can be obtained simultaneously

Data

The data in this thesis were produced by Illumina's Methylation EPIC BeadArray. Microarrays are a variation detection tool utilizing arrays of DNA probes to produce methylation data. A chip, or Bead Array, can measure thousands of samples, identifying known point mutations, structural mutations or gene expression and methylation states. With genome information being extensively applied in precision medicine, agriculture and many other areas, different types of chips have been developed to meet those demands. The Illumina platform has widely adopted the Infinium technology to provide good quality high-throughput data. Our study used the Illumina Infinium MethylationEPIC BeadChip, which covers over 850,000 CpG sites. The high-resolution technology used in this chip is based on quantitative testing of single nucleotide polymorphisms across the whole genome (for more details see Bock, 2012).

The dataset consists of 63 samples, including colon tumor glands and colon crypts from individuals of different ages. Our focus in this particular study is to use a subset of that data to compare colon tumor glands to normal colon crypts.

Our total sample sizes are shown in Table 2 below:

Table 2: Sample types and sizes

	Colon Tumor Gland	Colon Crypt
Total	40 (from 11 individuals)	23 (from 5 individuals)

After preprocessing and checking missing values rates, 4 colon tumor gland samples are excluded for having a high proportion of missing values. 3 individuals provided both colon tumor gland samples as well as colon crypt samples, while all 11 individuals who provided colon tumor gland also provided bulk samples.

Analysis Pipeline - Methods and Results

There are several pipelines focusing on processing and analyzing Illumina Infinium DNA methylation data. Perhaps the most widely-used are SeSAMe (Zhou et al., 2018), ChAMP (Morris et al., 2014), and missMethyl (Phipson, Maksimovic and Oshlack, 2016). SeSAMe is known for having stricter preprocessing standards, including comprehensive probe quality masking, bleed-through correction in background subtraction, nonlinear dye bias correction, nondetection calling and control for bisulfite conversion based on C/T-extension probes. It also provides sex or ethnicity inference based on specific probes to help with sample quality control. In addition, it includes a variety of options regarding the detection of differential methylation and segmented copy number. For that reason, we exploit the SeSAMe pipeline in this thesis.

A variety of preprocessing steps performed before conducting the final analysis. In this section, we walk through each of those steps, following the best practices in the SeSAMe User Guide (Zhou, Laird and Shen, 2017).

First, we read the IDAT files and checked for missing values - these correspond to probes that have low mapping quality scores, caused by things such as cross-hybridization with other loci in the genome. These loci are then masked from further analysis. This results in a loss of ~12% of probes on average. (This number is typical for such experiments.)

Second, a further set of underperforming probes are excluded by testing for low intensity signals across all samples. We filtered out those probes using the p-value from a test for out-of-band array hybridization (Zhou et al., 2018). We removed all probes for which the p-value was 0.05 or less. This resulted in us removing a further ~5% of probes on average (so ~17% of probes removed in total).

Third, background subtraction is performed by leveraging normal exponential deconvolution, controlling out-of-band probes to reduce high levels of background noise and increase the sensitivity of measurement (Triche et al., 2013).

Fourth, some of the probes lost in steps one and two are rescued using nonlinear scaling to correct the signal intensity difference between the red and green channels. Probes come in two types: Type-I and Type-II (Pidsley et al., 2016). This processing step shifts the Beta-values of Type-I probes using nonlinear quantile interpolation to correct the residual dye bias that is seen for those probes. This results in the two channels having similar performance and in recovering around 1.5% of probes (So now we have lost a total of ~15.5% probes on average).

Finally, using this remaining set of probes, Beta-values, which one can think of as being the proportion of methylated cells at that CpG, are obtained. Each of our samples was processed through this pipeline with SeSAMe to obtain Beta-values at each non-removed CpG site for each sample. The Beta-value can be retrieved by gene, region, or probe name. SeSAMe also provides quality control functions to check the mean signal intensity, and hybridization efficiency (bisulfite conversion using GCT scores (Zhou et al., 2017)). These are also reported as M-values, which are a logit transformation of Beta-values. For example, Beta-values of 0.2, 0.5 and 0.8 correspond to M-values of -2, 0 and 2, respectively. The Beta-value has a more intuitive biological interpretation, but the M-value is more statistically valid for the differential analysis of methylation levels. Therefore, the M-value method is used for differential methylation analysis while the Beta-value statistics are often reported for results interpretation (Du et al., 2010).

We show the distribution of CpGs sites that were removed as ömissing÷ during the first and second steps of the QC process above. We show results for tumor gland samples, but the normal samples performed similarly. In Figure 3, each column represents one sample and each row represents a CpG probe. Missing Beta-values are shown in red, while levels of grey indicate the magnitude of Beta-values. We see that the overall proportion of poorly-performing probes varies greatly across samples.

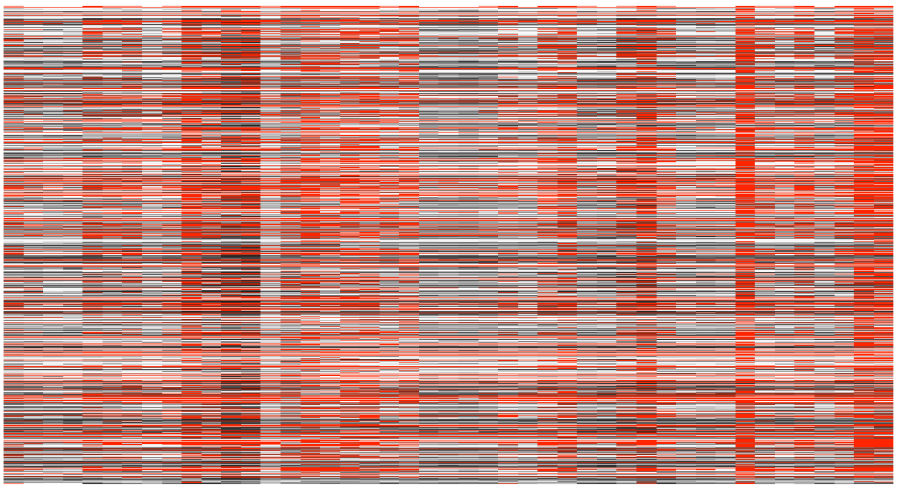

Figure 3: Beta-value matrix of all tumor gland samples (red: missing value; white: ~0; black: ~1).

Finally, in Figure 4a and 4b, we show the overall distribution of Beta-values before (top) and after (bottom) quality control for all 63 samples. We see that after normalization the distribution of samples within each group (normal or tumor) shows much greater consistency.

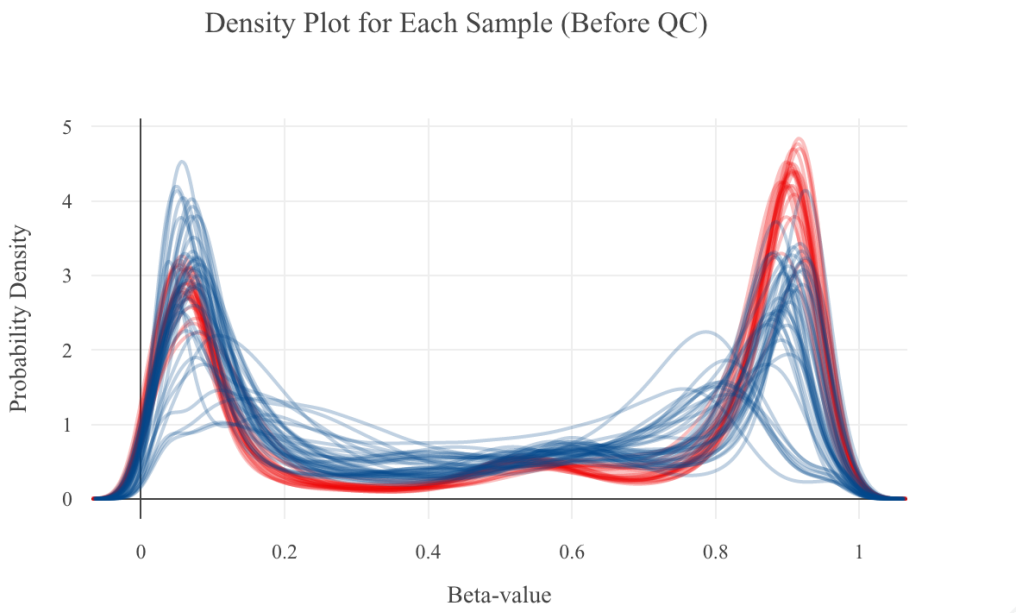

Figure 4a: Data distribution and clustering before quality control (865,918 probes) (blue: colon crypt; red: colon tumor).

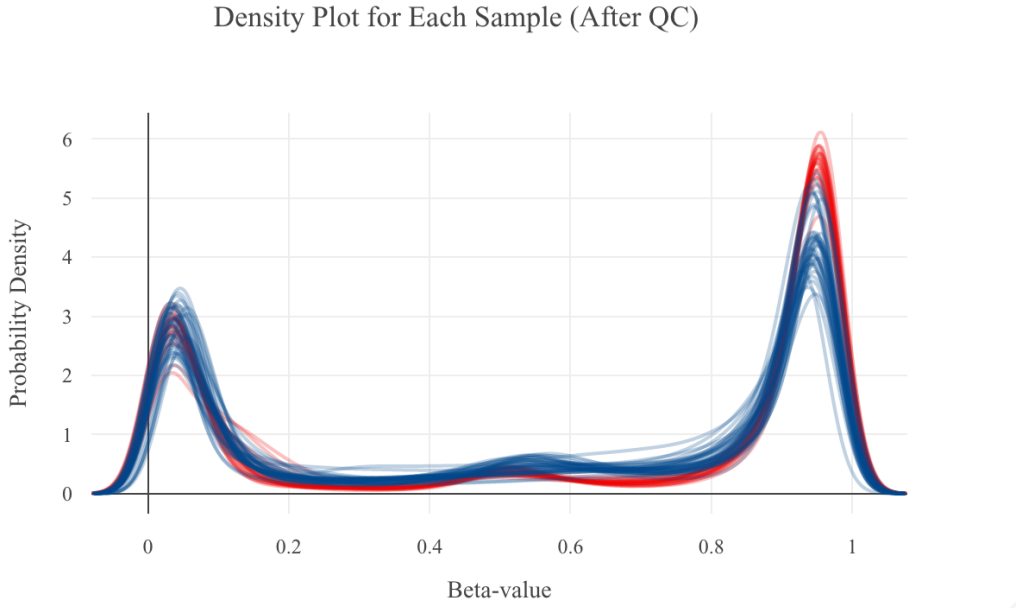

Figure 4b: Data distribution and clustering after quality control (340,103 probes) (blue: colon crypt; red: colon tumor).

Considering the potential for unwanted batch effects and other unexpected variations, we used the Remove Unwanted Variation (RUV) algorithm to adjust for such systematic effects (Molania et al., 2019). This algorithm specifies a set of negative control variables and uses it to detect and remove unwanted variation. Specifically, we use the version RUV-4 to correct for this in our data (Jacob, Gagnon-Bartsch and Speed, 2016). We then take these adjusted M-values forward for further analysis.

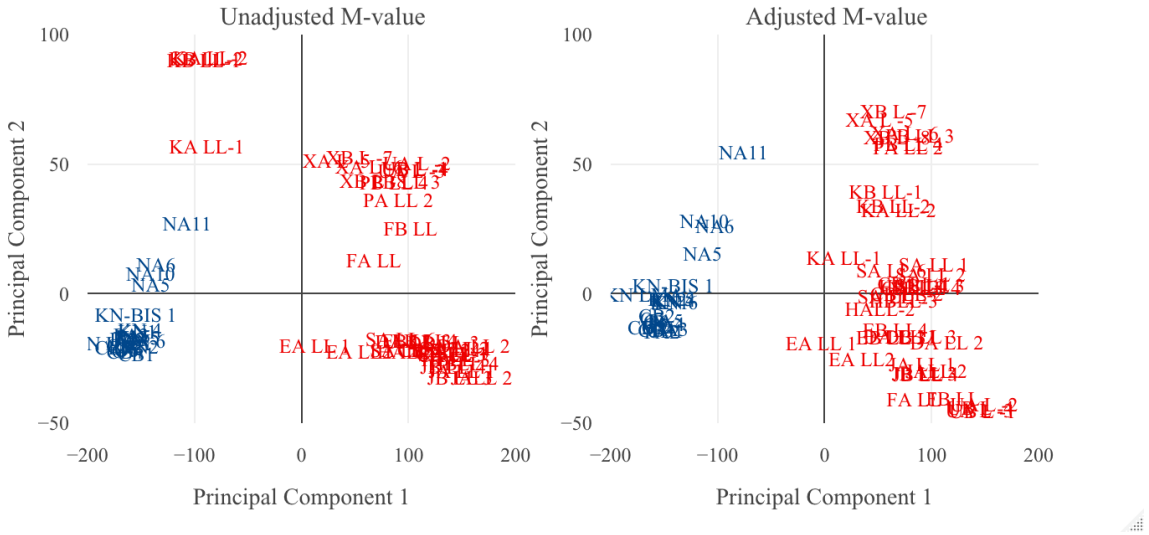

Figure 5: Data distribution before and after adjusting for unwanted variation (blue: colon crypt; red: colon tumor).

Variance for each CpG site is estimated among all samples. We extracted the 1000 most variable positions and performed principal component analysis (Gower, 1966) to show the distances between each sample. Results are shown in Figure 5. While the effects are not dramatic, we see that after adjusting for unwanted variation, samples of the two colors, each representing a given tissue type, are separated more distinctly.

Differential methylation analysis

We then followed the ChAMP pipeline to detect differentially methylated positions and regions using the limma package (Ritchie et al., 2015). Then we used the DMRcate package to obtain differentially methylated gene module results (Peters et al., 2015). The differential methylation position testing method

of limma employs linear models and an empirical Bayes framework based on Gaussian model theory to calculate the p-value, and uses the Benjamini-Hochberg procedure to control the false discovery rate (Smyth, 2004).

Differentially methylated positions (DMPs)

The first step in this analysis was to build a design matrix with tumor glands and colon crypts, and then a linear model with tissue type and individual ID as the two main factors. We then set the contrast to detect differences between tumor gland and colon crypt. Differences between tissue types are calculated within each individual, and then these differences are compared across individuals to determine whether there is an overall significant difference in the mean methylation level for each CpG site. The difference in mean methylation is labeled as logFC which is analogous to the log fold-change in gene expression analysis. Empirical Bayes methods are used to combine information across all the CpGs to obtain more precise estimates of gene-wise variability (Smyth, 2004). After testing for differentially methylated loci with our 40 tumor gland samples and 23 colon crypt samples, 105,811 significant differential methylated CpG sites are identified (adjusted p-value < .05, absolute log FC > 1).

The results of these process are displayed below. We show the features of the final set of significant differential methylated probes (i.e., CpG sites) that were included in our subsequent analysis of differential methylation. We begin by showing the overall distribution of the significant probes, and of these probes, the distribution of the hypermethylated probes and hypomethylated probes over genome in Figure 6. The first part of this figure shows how methylation sites are sited in terms of being within, near, or outside of, CpG islands (CGI). We see that about 22% of total probes are located in island areas, 55% in opensea area, 7% in shelf area (immediately adjacent to an island) and 15% in shore area (between øshelfùand øopenseaù). About 58% of these sites being hypermethylated are located in island, 16% in opensea, 3% in shelf and 23

in shore area. While about 5% hypomethylated sites are located in island area, 73% in opensea, 9% in shelf and 14% in shore. Generally speaking, in these samples, the island area contains most hypermethylated probes and the opensea area contains the most hypomethylated probes. Similarly, in the second part of this figure, we show the CpG locations in terms of gene structures. Specifically, the proportions of probes in the 1st exon, 5ùuntranslated region (UTR), exon boundary, intergenic region (IGR) and near transcription start sites (TSS) are shown below. The third part of the figure shows the proportion in terms of both of these categorizations simultaneously.

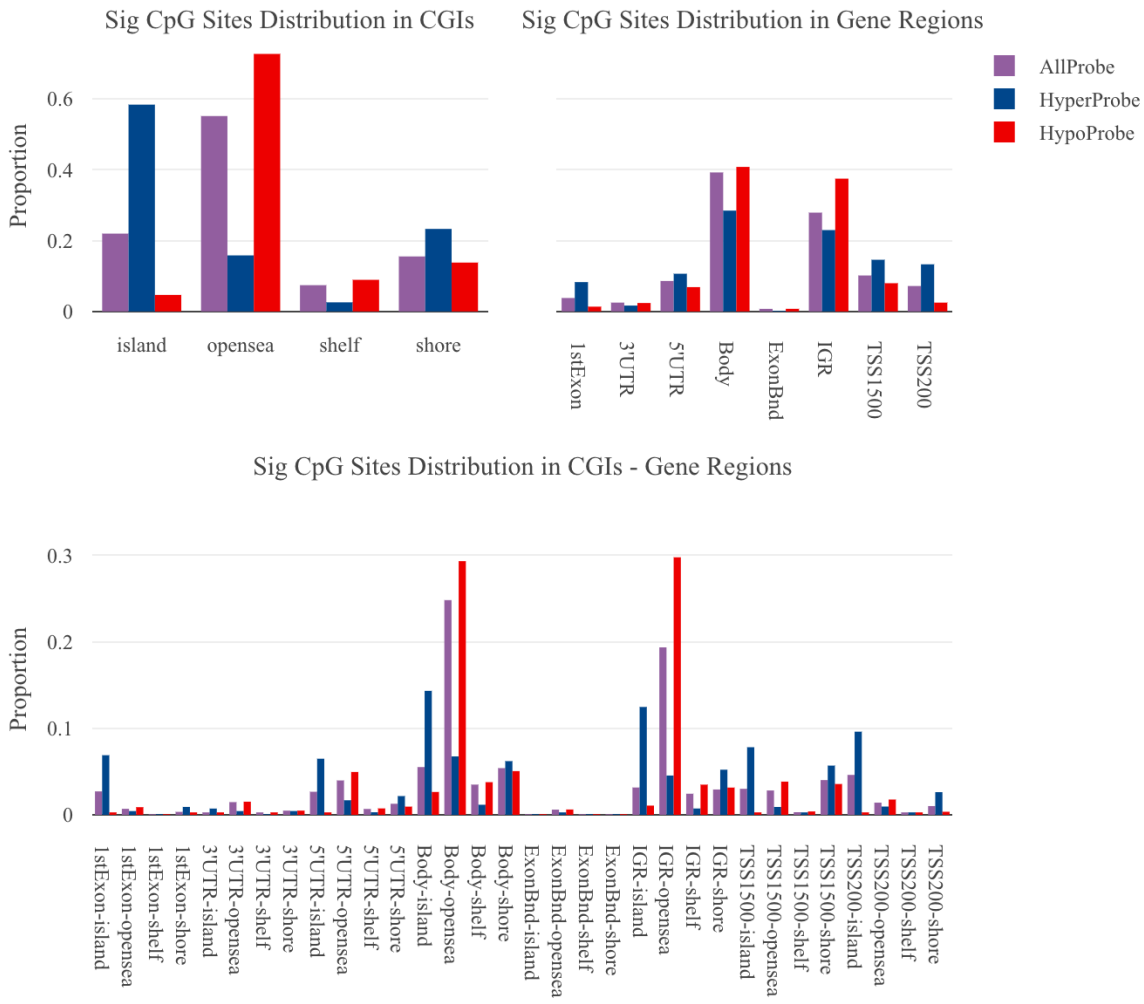

Figure 6: Overview of significant DMPsùlocations.

As an example, we also show the top 4 differentially methylated CpGs in Figure 7. The cluster CG IDs are cg12753986, cg14672084, cg26284735, and cg02037307.

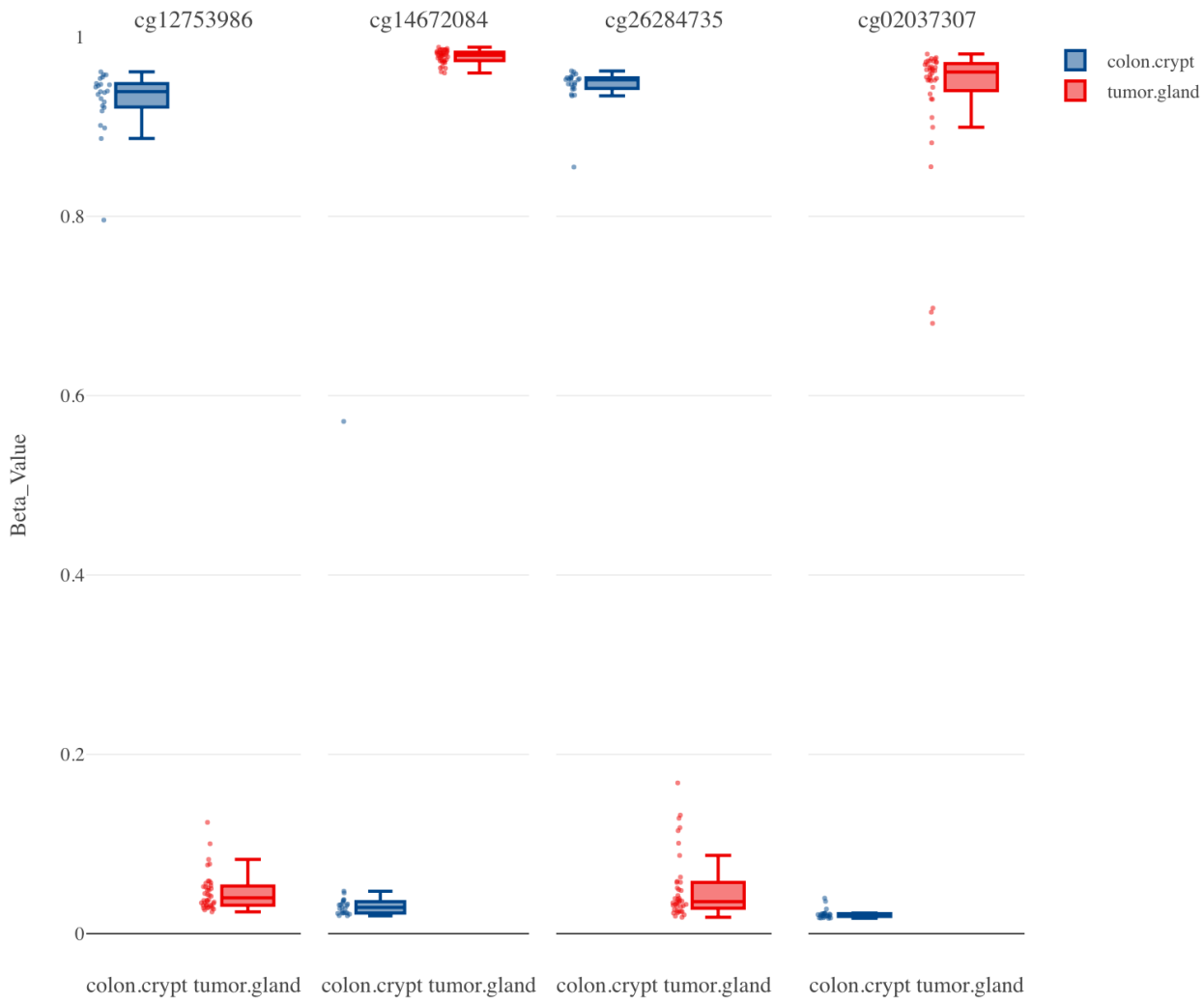

Figure 7: Top 4 significantly differentially methylated CpGs.

To improve interpretation, functional annotation from the Ensembl genome database project was used to indicate the nearest gene to each probe (Hubbard et al., 2002). We indicate 16,158 unique genes associated with the differentially methylated positions in tumor gland and colon crypts.

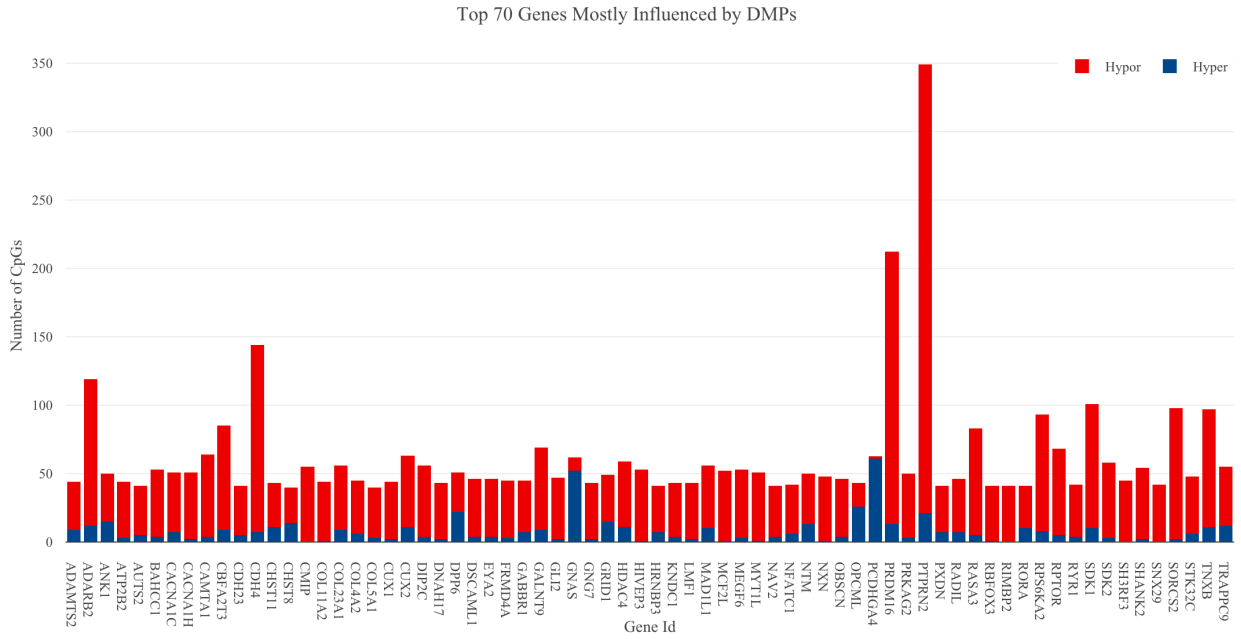

Figure 8: Top 70 significantly CpG-related genes.

In Figure 8 we show the top 70 genes enriched for significantly differential methylated CpGs and the number of hypermethylated CpGs or hypomethylated CpGs near each gene. Most of them are hypomethylated when comparing colon tumors to colon crypt.

As an example of the behavior within a particular gene, in Figure 9 we show the results for ZNF331 (Zinc-finger protein 331). The figure shows the number of measured CpGs and their position and mean methylation level for each group. The mean methylation level of tumor samples across this CpG island of 20 CpGs are significantly lower than that of colon crypts (adjusted p-value <.05, absolute logFC > 1).

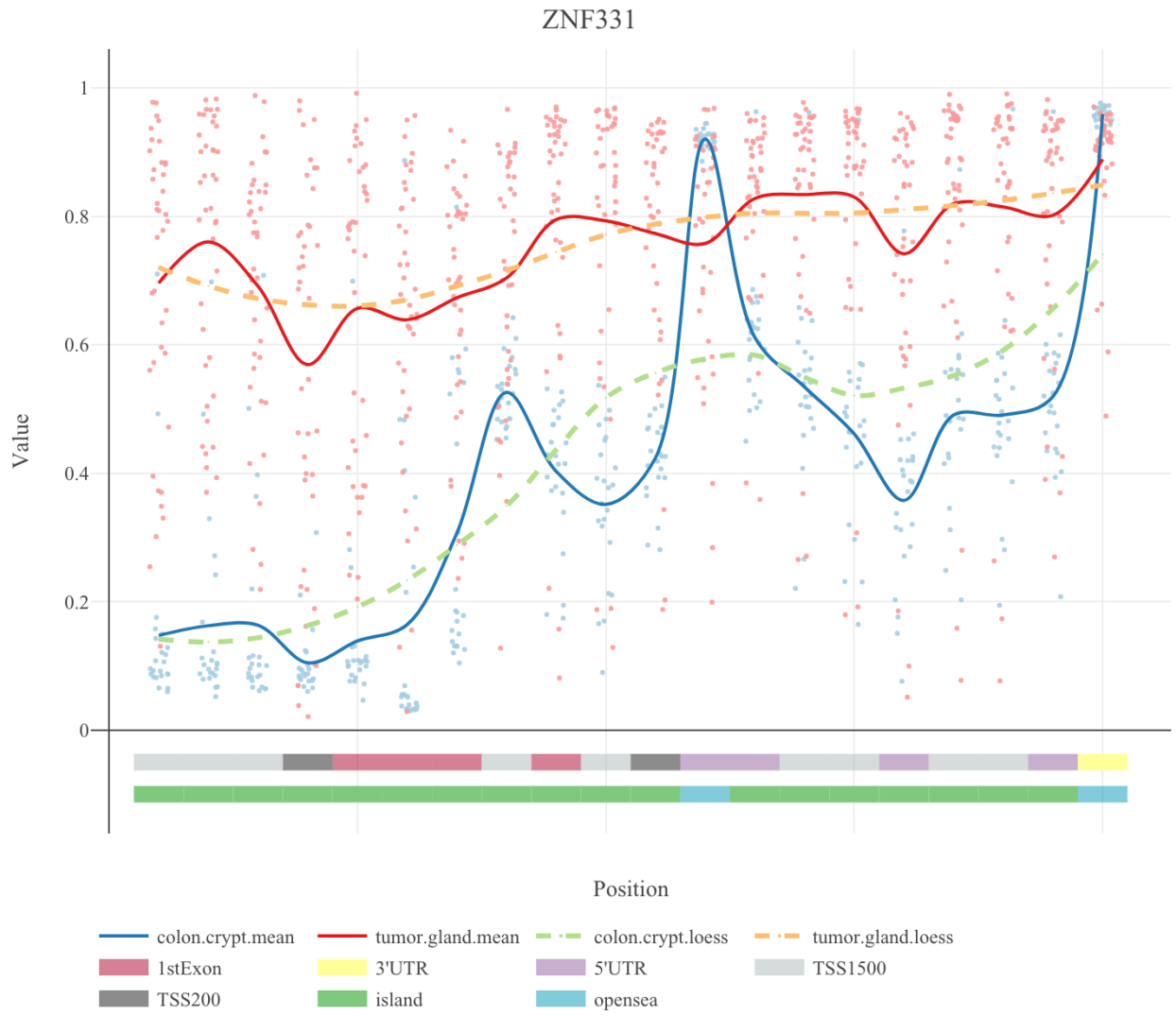

Figure 9: methylation level of the significant DMPs and related gene ZNF331

According to previous studies, ZNF331 is a tumor suppressor and was reported to be a potential biomarker for colorectal cancer detection with sensitivity (71%) and specificity (98%) (Yu et al., 2013; Vedeld et al., 2018). The result above is consistent with that knowledge. In particular, the CpGs before the 1st exon have much higher methylation levels in tumor samples than that in colon crypts.

Differentially methylated regions (DMRs)

Differential methylation status is detected in a large set of widely-dispersed CpGs. Therefore, we aimed to detect the regions in which most differential methylation occurred. The DMRcate package can extract the most differential methylated regions from BeadChip Array samples through kernel smoothing, and its performance is superior to Bumphunter and Probe Lasso (Peters et al., 2015). After using this package to test for differential methylation using FDR-correction, 11,259 significant regions were found (adjusted p-value <.05, absolute logFC > 1). The distance from the probe to the next consecutive probe in any given region was set to be less than 1000 nucleotides. As an example, one of the differentially methylated regions on Chromosome 1, at about 50.8825 mb, is shown in Figure 10 below.

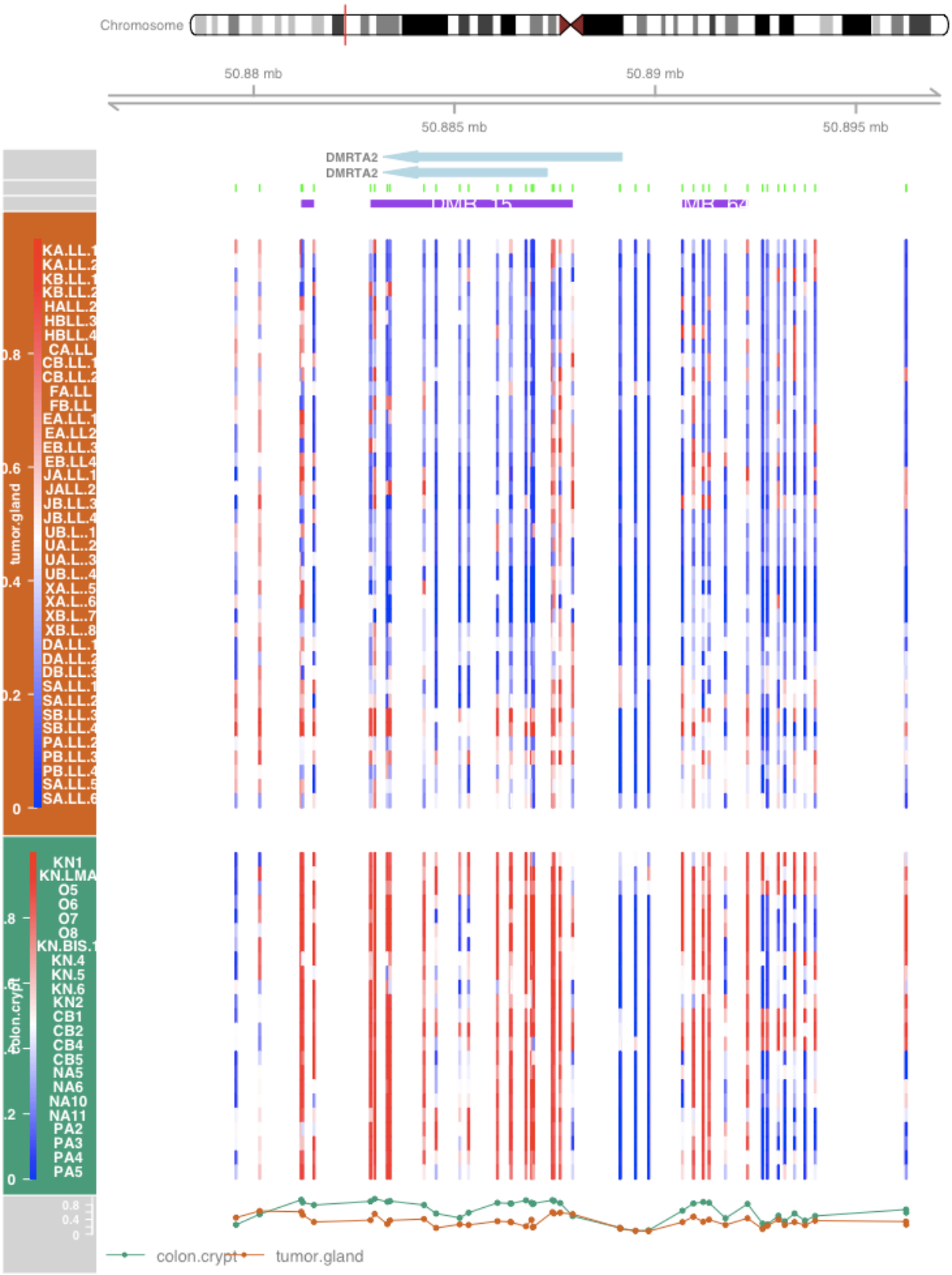

Figure 10: A DMR in genomic context

The figure shows the location of the DMR in the genome. It also shows the position of the nearby gene DMRTA2 (Doublesex- and mab-3-related transcription factor A2), and the positions of the CpG probes (marked green). The methylation levels of the individual samples are shown as a heat map (methylation level scale: 0-1; Blue=0; Red=1) and the mean methylation levels for the colon crypt and tumor gland groups are presented at the bottom. DMRTA2 is a transcription factor related to various cancers according to the Human Protein Atlas (http://www.proteinatlas.org) (Uhlen et al., 2017). The mean methylation level of tumor glands here is significantly lower than that of colon crypt, as assessed through DMRcate analysis.

Gene ontology analysis

After performing a differential methylation analysis, we obtained a long list of significant CpG sites to interpret (adjusted p-value < .05, absolute logFC > 1). In order to better interpret this, as a final step we performed a gene-set enrichment analysis. Thus, the genetic pathways in which differentially methylated CpGs were over-represented were identified. We identified gene ontology molecular functions and metabolic pathways significantly enriched for probes, with adjusted meta p-value less than 0.05, using the missMethyl package (Phipson et al., 2016). This helped us to gain an understanding of the biological processes that are involved. The results are shown in Table 3.

Table 3: Top 20 differential enriched pathways

	ONTOLOGY	TERM	N	DE
GO:0005509	MF	calcium ion binding	631	528
GO:0098978	CC	glutamatergic synapse	346	312
GO:0005887	CC	integral component of plasma membrane	890	716
GO:0030054	CC	cell junction	531	455
GO:0005886	CC	plasma membrane	4271	3052
GO:0009986	CC	cell surface	530	439
GO:0007155	BP	cell adhesion	375	319
GO:0007165	BP	signal transduction	827	661
GO:0005576	CC	extracellular region	1865	1319
GO:0062023	CC	collagen-containing extracellular matrix	355	294
GO:0007268	BP	chemical synaptic transmission	146	133
GO:0005201	MF	extracellular matrix structural constituent	105	99
GO:0007411	BP	axon guidance	164	150
GO:0045211	CC	postsynaptic membrane	174	155
GO:0030198	BP	extracellular matrix organization	189	168
GO:0016324	CC	apical plasma membrane	294	248
GO:0005178	MF	integrin binding	111	102
GO:0000122	BP	negative regulation of transcription by RNA polymerase II	707	562
GO:0070588	BP	calcium ion transmembrane transport	109	101
GO:0009897	CC	external side of plasma membrane	173	150

(BP = Biological Process; CC = Cellular Component; MF = Molecular Function; N = Number of genes in the GO term; DE = number of genes that are differentially methylated) (adjusted p <.001, FDR < 1%)

From the enrichment test, we see that the most differential methylation regulated gene expression and influenced the activity of cells. The top enriched pathways are related to cell junction and adhesion and membrane components.

Discussion

In this study we identified differentially methylated positions and regions between tumor gland and normal colon crypt samples. Our results are generally consistent with previous studies. Ultimately, we detected a significant number of genetic pathways that are enriched for differentially methylated genes when comparing tumor and normal samples. Our study demonstrates the use of several popular pipelines and methods for the analysis of methylation data.

One limitation of our study is sample size. The analyses we conducted were computationally intensive and non-trivial to run on a laptop. Furthermore, the size of the dataset was relatively small. As such, our results should be regarded with care, and ultimately, we would wish to support them with further analysis. For epigenome-wide association studies, a large sample size is necessary in order to obtain good statistical power to detect a significant effect. For example, 98 twin pairs were required to reach 80% EWAS power, and 112 pairs of cases and controls were needed to detect a 10% mean methylation difference between affected and unaffected subjects at a genome-wide significance threshold of $p = 1 \times 10^{-6}$ (Tsai and Bell, 2015). However, in this thesis, we only used 40 samples of tumors and 23 samples of colon crypt. Another limitation was due to computational capacity and efficiency. The computations in this thesis were processed by a 2.3 GHz dual-core Intel Core i5 processor using a personal computer which has 16 GB 2133 MHz memory. Under this circumstance, the processes used often hung or collapsed. In the future, we will apply high-performance computing clusters to address this problem.

Rather than imputing missing values for probes, in this study we simply removed such probes. While this choice is pragmatic, and was made for computational reasons, it assumes that the missing values are missing completely at random, which may introduce bias if not true. In the future, we might apply imputation methods, such as the K-Nearest Neighbors algorithm, to impute the missing values (Di Lena et al., 2019).

Normalization for methylation array data also needs to be considered. For consistency, we used the method from SeSAme, because the method is superior in reducing artifactual detection and is computationally friendly. Other methods exist and ultimately it would be good to explore how they might affect the final outcome of studies such as ours.

Regarding differential methylation testing, the key strength of using limma's linear modeling approach is the accommodation of arbitrary experimental complexity. Simple designs such as ours can then be handled relatively easily. Furthermore, if we wish to consider sex, age, and race, this approach can also deal with that. However, the samples in the present study did not have information regarding those covariates.

References

Berger, S.L., Kouzarides, T., Shiekhattar, R. and Shilatifard, A. (2009) An operational definition of epigenetics. *Genes & Development*, **23**, 781ô783.

Bock, C. (2012) Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, **13**, 705ô719.

Di Lena, P., Sala, C., Prodi, A. and Nardini, C. (2019) Missing value estimation methods for DNA methylation data. *Bioinformatics*, **35**, 3786ô3793.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L., et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, **11**, 587.

Feinberg, A.P. and Vogelstein, B. (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, **301**, 89ô92.

Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325ô338.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., et al. (2002) The Ensembl genome database project. *Nucleic Acids Research*, **30**, 38ô41.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931ô945.

Jacob, L., Gagnon-Bartsch, J.A. and Speed, T.P. (2016) Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. *Biostatistics (Oxford, England)*, **17**, 16ô28.

Jiang, L., Zhang, J., Wang, J.-J., Wang, L., Zhang, L., Li, G., et al. (2013) Sperm, but Not Oocyte, DNA Methylome Is Inherited by Zebrafish Early Embryos. *Cell*, **153**, 773ô784.

Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature*

Reviews Genetics, **13**, 484ô492.

Jones, P.A. and Baylin, S.B. (2002) The fundamental role of epigenetic events in cancer. *Nature Reviews Genetics*, **3**, 415ô428.

Jones, M.J., Goodman, S.J. and Kobor, M.S. (2015) DNA methylation and healthy human aging. *Aging Cell*, **14**, 924ô932.

Klose, R.J. and Bird, A.P. (2006) Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*, **31**, 89ô97.

Molania, R., Gagnon-Bartsch, J.A., Dobrovic, A. and Speed, T.P. (2019) A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Research*, **47**, 6073ô6083.

Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., et al. (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics (Oxford, England)*, **30**, 428ô430.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., et al. (2017) Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, **543**, 72ô77.

Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., V Lord, R., et al. (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin*, **8**, 6.

Phipson, B., Maksimovic, J. and Oshlack, A. (2016) missMethyl: an R package for analyzing data from Illuminaùs HumanMethylation450 platform. *Bioinformatics*, **32**, 286ô288.

Pidsley, R., Zotenko, E., Peters, T.J., Lawrence, M.G., Risbridger, G.P., Molloy, P., et al. (2016) Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, **17**.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**, e47ôe47.

Robertson, K.D. (2005) DNA methylation and human disease. *Nature Reviews Genetics*, **6**, 597ô610.

Sharma, S., Kelly, T.K. and Jones, P.A. (2010) Epigenetics in cancer. *Carcinogenesis*, **31**, 27ô36.

Silva, A.J. and White, R. (1988) Inheritance of allelic blueprints for methylation patterns. *Cell*, **54**, 145ô 152.

Smyth, G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, 1ô25.

Suzuki, M.M. and Bird, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, **9**, 465ô476.

Tsai, P.-C. and Bell, J.T. (2015) Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology*, **44**, 1429ô1441.

Uhlen, M., Zhang, C., Lee, S., Sj¿stedt, E., Fagerberg, L., Bidkhori, G., et al. (2017) A pathology atlas of the human cancer transcriptome. *Science*, **357**.

Vedeld, H.M., Nesbakken, A., Lothe, R.A. and Lind, G.E. (2018) Re-assessing ZNF331 as a DNA methylation biomarker for colorectal cancer. *Clinical Epigenetics*, **10**, 70.

Yu, J., Liang, Q.Y., Wang, J., Cheng, Y., Wang, S., Poon, T.C.W., et al. (2013) Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer. *Oncogene*, **32**, 307ô317.

Zhou, W., Triche, T.J., Jr, Laird, P.W. and Shen, H. (2018) SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Research*, **46**, e123ôe123.