Fall 2017

Comparing region level testing methods for differential DNA methylation analysis

Arnold Albert Harder

Department:

COMPARING REGION LEVEL TESTING METHODS FOR DIFFERENTIAL DNA METHYLATION ANALYSIS

by

ARNOLD ALBERT HARDER

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

APPLIED MATHEMATICS WITH STATISTICS EMPHASIS

2017

Approved by

Dr. Gayla Olbricht, Advisor
Dr. VA Samaranayake
Dr. Xuerong Wen

ABSTRACT

Finding possible connections and solutions to help fight progression of diseases is a major area of research. Genomics is a primary path of research in disease research. Through the DNA sequence, possible connections to diseases have been found. However, most methods for fixing issues within a DNA sequence are still out of reach. One potential path is to investigate epigenetic modifications, such as DNA methylation. DNA methylation occurs when a methyl group attached to cytosines on the DNA sequence. Statistical methods can be used to identify sites or regions of significant differences in methylation levels between groups (e. g. disease vs. healthy). If these particular sites or regions can be linked to diseases they could aid in better understanding the disease pathway and could potentially be used to diagnose or treat the disease. With recent advancement in technology, tools to measure and analyze methylation levels have become easier and are more accessible. Thus, more and more researchers are investigating methylation to help identify possible connections to diseases.

The statistical tools to analyze and find significantly different methylated sites or regions have advanced. Since methylation levels have been shown to have correlation among neighboring sites, meaningful differences in methylation levels commonly occur over regions rather than individual sites. There are several statistical tools to test for these regions, including three that are the focus of this thesis: Bumphunter, Probe Lasso, and DMRcate. Each of these methods identifies regions that have significantly different methylation levels in different ways. In this thesis, a thorough examination of each of these methods is presented and all three methods are used to identify differentially methylated regions (DMRs) between HIV positive women with and without squamons cell carcinoma cervical cancer. The methods are compared to help better understand their impact of identifying DMRs.

ACKNOWLEDGMENTS

I would like to thank Dr. Gayla Olbricht and Dr. Mohamed Milad for help with understanding and learning this material. I am also grateful to Fred Hutchinson Cancer Research Center for providing the data set that was used within this thesis. In addition, I want to thank Dr. VA Samaranayake and Dr. Xuerong Wen for participating on my Masters' Thesis committee. Lastly, I would like to thank Luyang Wang for help with writing R code for this thesis.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1. DNA METHYLATION

Finding the causes of human diseases is complex and has been studied for a very long time. In the quest for potential answers, the field of genomics has begun to play an important role in recent years [1]. Genomics is the field that involves analyzing and studying genomes. Genomes are the complete set of the all the genetic material of an organism, which is primarily orginized into chromosomes. These chromosomes are made up of strands of DNA that contain genes, which code for functional units called proteins. Within the DNA, there is an arrangement of four different types of nucleotides. These nucleotides are thymine, cytosine, adenine, and guanine [1]. The arrangement of these four nucleotides contain the information required to inform the cells of the human body to perform certain functions. This transfer of information occurs when the coding units called genes under go a process called transcription to messenger RNA and then translation to proteins, which are functional units of the cell. Different genes are transcribed or "expressed" in different tissues at different times depending on what functions are needed. However, occasionally there might be an arrangement of the DNA sequence or an aberrant expression of genes that are detrimental to the human body and can lead to formation of diseases such as cancers. As a result, there is a desire to determine the locations of DNA sections that are associated with diseases. If this knowledge could be found, then it could be used to better understand the disease pathway and offer potential treatments. For diseases that are associated with DNA sequence aberrations, it is often difficult to change the underlying genetic code. If the issue is with gene expression, then mechanisms that can alter this process need to be examined and explored.

One type of epigenetic factor that might be a linked to diseases outside of the formation of DNA, is DNA methylation. DNA methylation is a chemical addition to DNA that occurs when a methyl (CH3) group is attached to the DNA sequence. DNA methylation can affect how genes function and are expressed [2]. The common location for the methyl groups to attach to DNA are at cytosine sites within cytosine-guanine dinucleotides (CpG) [2]. This modification can be inherited when a cell undergoes division. However, this chemical modification does not become a part of the DNA structure. An example of DNA methylation can be seen in Figure 1.1. The original DNA sequence is not altered by methylation, however methylation can determine if the neighboring location (gene) of the DNA will be expressed. Methylation can be changed by environmental factors [2]. It is for this reason that a set of twins may have identical DNA strands, but one of the twins might develop a disease while the other does not. Due to the recent increase in technology advancement, both testing for DNA methylation and treatments have become less expensive. This has led to a major upswing in the amount of researchers investigating DNA methylation as a possible avenue to treat major diseases.

DNA methylation can be linked with many different phenotypes, however it is predominantly studied to determine how it is related to disease. A phenotype in this thesis is a common characteristic that each individual in a identified group might possess. An example of this could be a group of women who all show similar stages of the same type of cancer. Many cancers have distinguishing methylation patterns that might be a contributing factor or an indicator of the identified disease. Researchers would like to be able to determine if the pattern is a cause or contributing factor for certain diseases or if there are other reasons and the pattern is a result of the underlying causes. If the disease is caused by methylation, then possible cures or treatments might be used that focus on the aberrant methylation patterns [2]. Such a treatment could target the methyl groups that have erroneously attached, or where the methylation was mistakenly lost. Thus, if the particular locations with aberrant methylation patterns contributing to cancer could be identified, then

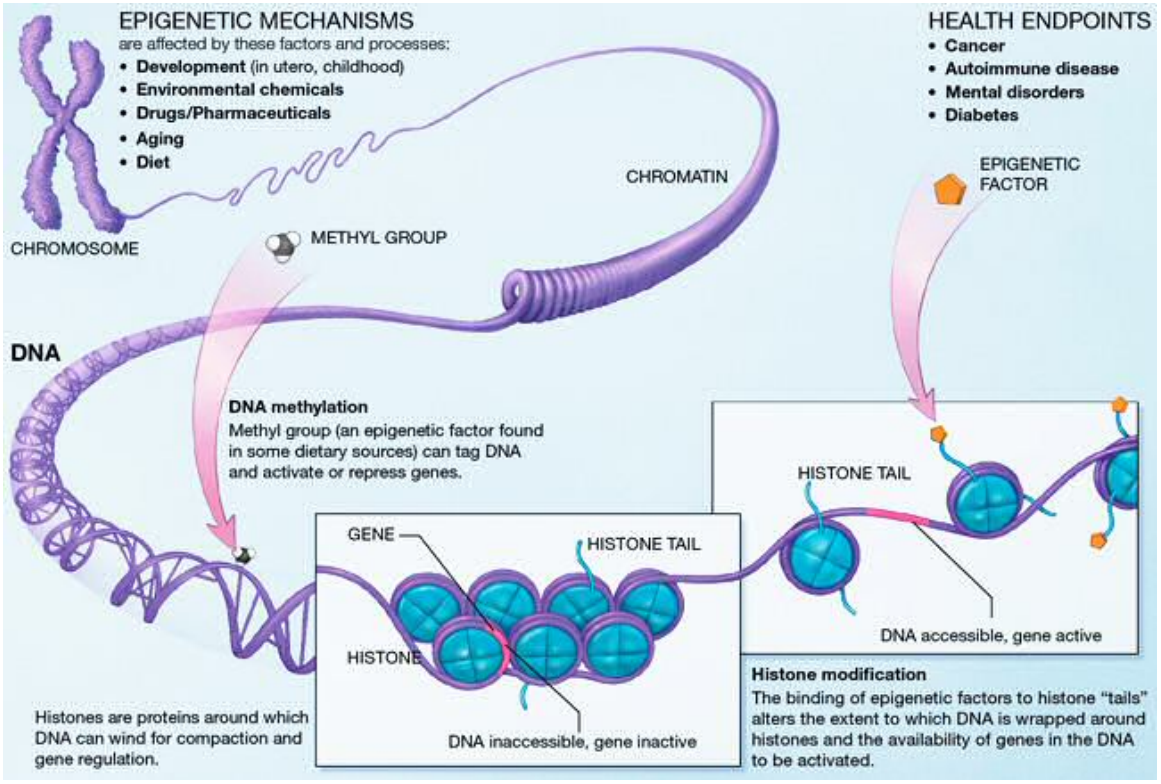

Figure 1.1. Illistration of DNA Methylation. Found at *https://commons.wikimedia.org/wiki/File:Epigenetic-mechanisms.jpg* on October 2017.

methyl groups could potentially be removed or attached to counter a cancer's progress. There are two examples of potential drugs that have been created to counter cancers in such a manner, Vidaza and Decitabine [3]. These two drugs are used against cancerous tumours by the removal of methyl groups that are shutting-off genes that contribute to the tumour suppression [3]. As the study of methylation progresses, potential future drugs might be able to treat malignant tumours, but prevent them from ever forming.

1.2. TECHNOLOGY USED TO MEASURE DNA METHYLATION

To be able to study and determine possible connections of DNA methylation with diseases, DNA samples from different individuals must be taken and DNA methylation measured. One problem with finding such connections is that the locations of important

CpG sites for testing are largely unknown at the start of a study. Thus the sites where aberrant DNA methylation patterns occur could be anywhere within the genome. To make sure that the important methyl groups are not missed, the entire genome must be tested for methylation differences between healthy and diseased individuals. Different technologies exist for measuring DNA methylation at a genome-wide level.

To test a whole genome for differential methylation, whole-genome bisulfite sequencing (WGBS) can be used to gather data about methylation levels at all CpG sites in humans [2]. This process helps identify methylated sites by using the compound sodium bisulfite which only affects unmethylated sites. The sodium bisulfite targets the unmethylated cytosine sites and changes them into uracils [2]. A uracil is a nucleotide that is used in place of thymine in a strand of RNA [1]. The methylated sites however are unaffected and remain the same. The bisulfite converted DNA is then sequenced via next generation sequencing (NGS) techniques. This gives the ability to identify the possible methylated sites by examining around 28 million CpG sites for any potential unchanged sites [2]. WGBS is the most preferred method to gathering data about methylation in a single subject's genome since all potential sites of methylation would hopefully be identified. Unfortunately, WGBS has a high cost per subject. Most studies prefer a large sample size of subjects; however with a limited budget, the sample size when using WGBS is limited. This might mean that the sample size is smaller than needed to attain reasonable statistical power to find true connections to the disease. Another reason WGBS is not widely used is due to the fact that data collection and analysis methods take a high amount of training and technology that may not be accessible to most researchers. Therefore, based on technology, training, and financial cost factors, the WGBS approach is not the most feasible for many researchers/studies.

A more practical tool used to identify DNA methylation levels is a type of microarray called the Illumina Infinium Bead chips. This type of microarray is a glass chip that has probes that are short sets of nucleotides that correspond to CpG sites that are attached to the chip to help measure methylation levels [1]. This method also uses sodium bisulfite;

however, unlike WGBS, beadchips only target a subset of CpG sites and measure methylation levels using predetermined probes on the microarray. Sodium bisulfite is used on DNA samples in a similar method as WGBS. The sodium bisulfite targets unmethylated sites and changes them to uracils while ignoring methylated sites. Then the samples then are whole-genome amplified and broken into smaller sections with enzymatic fragmentation [4]. These amplified fragments of the sample are then applied to the beadchip. These fragments will interact with one of two types of probes: Infinium 1 (Type 1) and Infinium 2 (Type 2) [5].

The Type 1 or Infinium 1 probes contain what are known as two "beads." A bead is made up of oligonucleotides, which are short synthetic DNA strands that contain sequences complementary to those that surround the specific CpG sites of interest [6]. These are important for being able to determine the methylation level at CpG sites. Type 1 probes possess two separate beads that allows for detecting methylation levels with one bead and unmethylation with the other bead at a determined site. This is accomplished by creating a different oligonucleotide for each bead, one that will pair with the bisulfite converted sequence when the site has been converted to a uracil (unmethylated) and one that will pair when the site remains a cytosine (methylated) [4]. Once this pairing (or hybridization) occurs, the information is reported as signals of light, the intesity of which provides a measure of methylation and unmethylation. The light intensity is found after the beadchip is fluorescently stained and intensity of stains are measured to determine levels of methylation and unmethylation.

The other type of probe is the type 2 or Infinium 2. With this probe all the steps are the same as type 1 except that when the sample is introduced to the probe, there is only a single bead. This single bead has the capability to measure both methylation and unmethylation levels at the site level. Type 2 probes create two signals for methylation and unmethylation by allowing the correct nucleotide from the bisulfite converted DNA to hybridize and be tagged with the specific dye color [7, 8]. The probe is also stained and

intensities found in a similar way as type 1 probes. The intensity of green dye channel represents the level of methylation and the intensity of the red dye channel represents the level of unmethylation.

With the levels of intensities at each site, a numerical value called a β value can be calculated to determine the percentage of methylation at a site and is found by the following: $\beta = (Max(M, 0))/(Max(M, 0) + Max(U, 0) + 100)$ [9] Through the beads, a Max(M,0), which is the intensity of methylation at the site, can be found. The values for Max(U,0), which is the intensity of unmethylation, can also be found [9]. The constant of 100 is used to avoid problematic β values in which both methylation and unmethylation levels have low measurements [10]. The advantage of the type 2 probe is that more of these probes can be put onto the array than type 1 probes [6]. This allows for more of the gene sequence to be tested. The advantage of the type 1 probe is that this probe is able to measure extremes of methylation and unmethylation intensities better than type 2 probes [7]. Together these two types of probes can be used throughout the beadchip to test predetermined site locations. Since the beadchips are not able to cover the entire genome, the results are less comprehensive than using WGBS. Since the probes are predetermined and only a subset of the CpG sites, the beadchips are able to be produced easier than WGBS, thereby reducing the cost of beadcips compared to WGBS[2]. Therefore, there is a trade-off between accuracy and cost/use for WGBS and beadchips. However, since most researchers/studies have a limited budget, the beadchips are often the preferred method. The beadchips also provide an easier way to conduct testing on larger sample sizes, which can aid in improving statistical power for testing.

Currently, there are three major types of beadchips. The first version was the Human Methylation 27K Bead Chip (HM27). This version had 27,578 probes for testing CpG sites and had the ability to test up to 12 samples per beadchip. The next version of this microarray technology was the HM450 Bead Chip (450K). It also had the capability of testing 12 samples per beadchip, however each beadchip contains 485,577 probes for testing

corresponding CpG sites. The latest version as of 2016 is the EPIC Bead Chip (EPIC) which has over 850,000 probes that are used to identify methylation levels at these sites [2]. As the technology for microarrays has advanced, the number of probes has increased enabling the researcher measure methylation at more potential CpG sites that may be connected to diseases. As the number of probes on the beadchips increases, then the coverage across the genome gets similar to WGBS. Currently, the most widely used microarray is the 450K, however this will likely be replaced by EPIC in the near future.

The order of probes and the corresponding CpG sites that are tested follow a certain pattern. It has been noted that certain segments of the genome contain a higher concentration of CpG sites than expected by chance. These known high concentrations of CpG sites are known as "CpG islands." CpG islands often occur near transcription factor start sites of genes and thus the methylation status of them could affect expression of the nearby gene. The areas neighboring the CpG islands where the concentration of CpGs starts to decrease (~ 2kb from CpG islands) are known as "CpG shores." Lastly the area of sequences neighboring the CpG shores (~ 4kb from CpG shores) where CpG sites are sparse can be known as "CpG shelves [2]." The remaining areas of the sequence that are sparse in the number of CpG sites are known as "CpG oceans." The names and definitions of these regions differ slightly depending on the researcher. Since methylation occurs at CpG sites, the higher concentration of CpG sites is of interest for testing. Thus the probes in the beadchips focus on testing CpG islands and shores. For the 450K beadchips, the distribution of probes to each region is the following: 32 % of probes for CpG islands, 23 % for CpG shores, 10 % for CpG shelfs, and 36 % for CpG oceans [11]. As the technology improves, more probes will be able to test the shelves and oceans.

1.3. STATISTICAL METHODS FOR ANALYZING DNA METHYLATION DATA

When analyzing the data produced from DNA methylation microarrays, either single sites or entire regions can be investigated. DNA methylation rarely occurs in isolation with a lone methyl group being added to a cytosine site but rather often occurs in clusters throughout the genome. Testing for differential methylation occurs when one compares two groups of samples, such as treatment and control, and tries to determine if there is a significant difference in the average levels of methylation at a certain location or set of locations between the two groups. This testing can be conducted at either the individual site level or the region level for the DNA. Methylation and thus meaningful differences in methylation levels are more likely to occur over a region of the genome structure[2].

Methods have been developed to help identify these regions that are referred as differently methylated regions (DMRs). These regions can be further investigated to determine if they are affecting the expression of nearby genes. When testing the regions, DNA methylation is easier to identify when the regions range in size from several hundred base pairs to several megabases. In the past, there was evidence to support that there is correlation between adjacent sites' methylation levels; meaning that if a site is methylated then the neighboring sites are more likely to also be methylated [3]. Since regions are more common than single site methylation, it is of interest to identify regions that have a significant amount difference between average DNA methylation levels of the two examined phenotypes, such as a disease group and a healthy group. Testing for DMRs enables the possibility of finding links between methylation and the disease in question, which may lead to potential diagnostic tools or treatments. The measurement of DNA methylation levels is a continuous measurement. Any possible association between DNA methylation levels at specific genomic locations and disease status is the goal.

A DMR consists of a set of CpG sites that have been grouped together based on having a significant difference in their average methylation levels between groups. Figure 1.2 is an example of the methylation levels (β values) for a DMR. Notice there is a difference

between the two distributions of site level methylation of the two groups of SCC and Negative. DMRs can have a significant amount of variation in methylation levels between sites in the region. This means that DNA methylation measurements can have measurement error. Thus, when trying to identify DMRs, the analysis must take this into account. DNA methylation is typically concentrated in regions throughout the genome. Thus, locations of methylation will not be as dispersed throughout the genome. The benefit of this is that, when analyzing the genome, less of it has to be examined when looking for methylation than examining the entire DNA sequence.

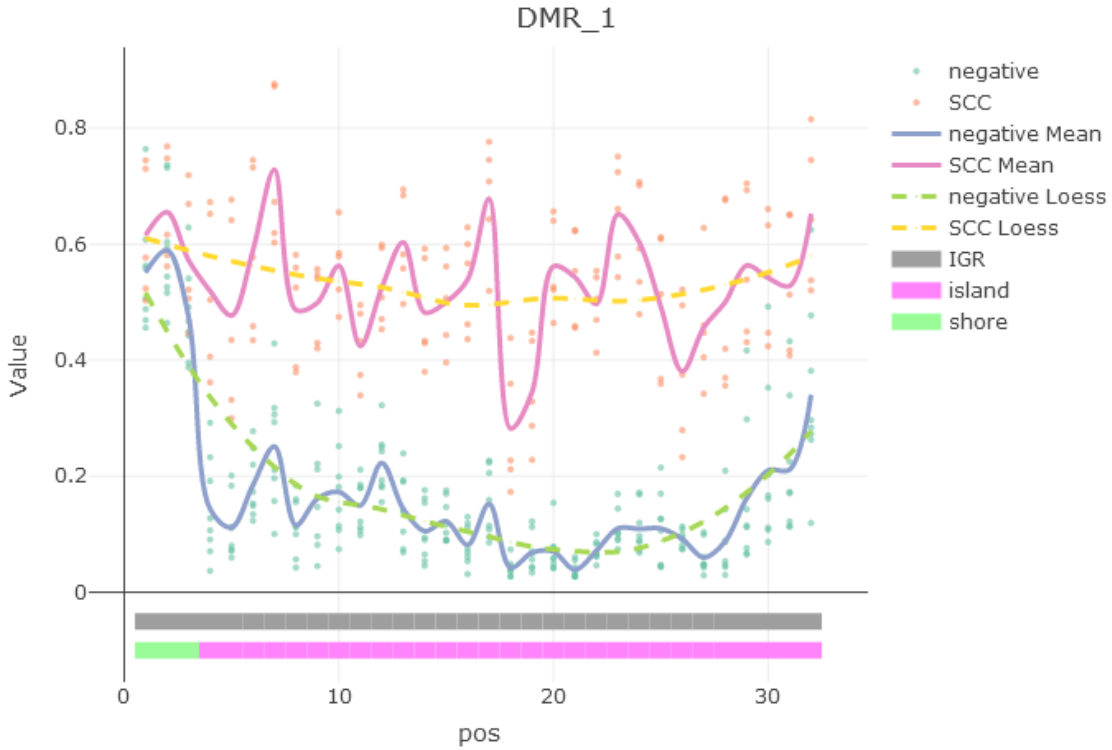

Figure 1.2. Example of methylation levels for a DMR. Plot of Methylation level (β value) versus genomic position. The two groups of interest (negative and SCC) are shown in different colors.

1.3.1. Pre-processing. Other possible factors might have an effect on the formation of a disease or DNA methylation, thus taking away the significance of a possible connection between the disease and differentially methylated sites or regions. Therefore, these effects or variables also need to be considered. Specifically, certain aspects of the data collection

process can affect the data quality at some sites or in some samples and preforming filtering steps can help ensure low quality data are removed. One issue can be attributed to the probes that are used in measuring the data. This is usually called site level filtering. Probes could be defective and not measuring the data correctly, so the final results could be skewed. There are six different criteria that are used to determine if a probe or any data produced from the probe need to be disregarded. The first involves filtering out probes that possess a detection p-value that is less than 0.01. This is a good sign that the probe might be defective and not correctly working. The proper action is to remove the probes that significantly failed at detecting methylation levels. The next step is to remove any probe that in at least five percent of the samples for each probe has less than 3 beads. If five percent or more of the samples collectively have less than three beads at a particular probe, than the probe will be removed. Third, remove any probes that are associated with non-CpG sites. Since DMRs occur at CpG sites, probes from anywhere else will not have any pertinent information about DMRs. Fourth, filter out probes that are associated with known single nucleotide polymorphisms(SNPs). The reason is that SNPs can make the probes have difficulty in identifying methylation at CpG sites and can cause change in color channels from the probes [12]. The probes to be filtered out can be identified by using the database of General Recommended Probes that was compiled by Zhou [12]. Fifth, probes that are categorized as multi-hit should be removed. Multi-hit will be where the probe is methylated in all samples [13]. This will defined by the Nordlund's Genome Biology article[13]. Lastly, disregard any probes located on the X or Y chromosome. The X and Y chromosome determines the sex of the sample. If there is a difference in the sex of the samples then the male and female beta values for the remaining chromosomes will be processed separately [5].

Another potential data quality issue is that particular samples could be outliers, thus skewing the data. So after site level filtering is done, sample level filtering also needs to be performed. A common approach to perform this filtering is to look at the beta distributions for all of the samples and determine whether any of the distributions are significantly

different than the remaining distributions. The data are then normalized to counter any bias that might arise through the difference between type 1 and type 2 probes on the microarray. The common method used to normalize the data is Beta Mixture Quantile Dilation (BMIQ) [14]. The last significant cause of error is through a batch effect. This means samples for a study have sub-groups that have correlation that is not attributed to the variables that are being examined with the study. This is usually detected by the presence of extreme outliers that are not representative of the population that the samples are supposed to represent. An important example of this is when samples for the same study were collected at different processing dates. Thus, a possible sub-group might form that has a correlation that could be attributed to a different person collecting the samples or the samples being taken from a different environment [3]. It is important to try to reduce possible batch effects with a proper experimental design. However, if batch effect can not be removed then a correction on the data needs to be done. For this thesis, there were no known batch effects, thus no correction on the data was necessary.

1.3.2. Methods for DMR Testing. Many methods for identifying significant DMRs in DNA microarray data use different types of smoothing methods as one of the steps in their testing framework. The method of smoothing involves finding a curve that highlights the main patterns in a set of data that does not have constraints and is not heavily influenced by individual data points. There can be different methods to obtain such a smoothed curve. One method, called local regression or Loess (or Lowess) involves gathering points that will be used in the curve fitting by examining several neighborhoods within the data. The data will be divided into sections of points that are located closely in terms of their input values and be grouped into neighborhoods. An example for this thesis would be grouping methylation levels together based on their CpG site locations being close together. All methods that use neighborhoods use a bandwidth or smoothing parameter that the researcher can set to determine the size of each neighborhood that will be used with the method. Within these neighborhoods, a polynomial regression fit (typically linear or quadratic) is obtained and

then the fitted values from the regression are gathered for the overall Loess smoothing. In the polynomial fitting, the weighted least squares approach is used so that data points within each neighborhood that are further away from the center are given less weight than those near the center, as determined by a specified weight function. The tricube weight function is commonly used. To combat the issue of weighting some values less, a new neighborhood is constructed in order for the regression fits to be centered around the previously less weighted values. The resulting fitted values are once again collected and the values are given weights. The regression fitting is repeated for multiple neighborhoods, potentially up to the total number of data points. To make the procedure more robust to outliers, the regression fit in each neighborhood can be repeated a determined number of times. In each fitting, the weights can be updated according to the size of residuals in the previous fitting so that points with large residuals receive less weight [15].

Another option for smoothing is the method of running medians. Smoothing is done by obtaining a collection of median output values calculated from subsets of the data. The medians are determined from a subset of output values that are grouped based on similarity of their input values. For example, median values would be found from the methylation values that come from CpG sites that are close to each other. These medians of output values are found by sequentially defining subsets of the data based on their adjacent input values throughout the remaining data (i.e., moving windows). The resulting medians are then used as the smoothed values [15].

The last major smoothing type used within this thesis is the Gaussian Kernel smoothing method. Kernel smoothing methods provide weighted averages of the output values (e. g. methylation levels) located in a neighborhood based on similar input values (e. g. CpG site location). The bandwidth λ defines the width of the neighborhood and the weighted averages are calculated using a moving window. The major key for this method is the use of weights which are defined by a "kernel" function. Weights help determine the importance

that needs to be given to certain surrounding CpG sites when determining the smoothed value. The methods used in this thesis utilize Gaussian Kernel Weights, which are defined as the following:

$$K_{ij} = exp((-[x_i - x_j]^2)/2\sigma^2)$$ [16]

This is applied to the x_i and x_j CpG sites where σ is the kernel scale factor, which should be proportional to the bandwidth λ. Thus σ is set to be $\sigma = \lambda/C$, where λ and C are pre-chosen [16]. This is then used at each potential x_i CpG site to help produce the following three equations:

$$S_{KY}(i) = \sum_{j=1}^{n} K_{ij} * Y_j$$

$$S_K(i) = \sum_{j=1}^{n} K_{ij}$$

$$S_{KK}(i) = \sum_{j=1}^{n} (K_{ij})^2$$ [16]

These summations are calculated with the kernels for all the x_i CpG sites on a chromosome for methylation levels or test statistics. The $S_{KY}(i)$ summation represents the smoothed model, whereas the other two summations are used in other aspects of the DMR testing method [16].

In addition to smoothing, another factor important to DMR testing is imposing some type of control over the amount of false positives (Type 1 errors) that could occur across the multiple tests conducted. At the site level, hundreds of thousands of hypothesis tests could be conducted, one for each CpG site. Although region level testing reduces the number of tests, there are still multiple tests and many methods rely on site level significance. In many genomic studies, it is common to control the false discovery rate (FDR). The FDR is defined as:

$$FDR = E(V/R \mid R > 0)Pr(R > 0)$$ [9]

Let "R" be the number of differentially methylated sites detected (i.e. discoveries). This number "R" can be broken down into the number of correctly identified differentially methylated sites ("S"), and the number of falsely chosen differentially methylated sites ("V") [9]. Thus, V/R represents the proportion of false discoveries which is conditioned

on the number of discoveries being greater than zero. Through this, the testing error rate will be controlled. Specifically, Benjamini and Hochberg introduced a method to control the FDR at level α [17]. An adjusted p-value can be calculated from this method and tests with an adjusted p-value less than α will reject the null hypothesis.

To test for DMRs, three main methods are considered in this thesis: Bumphunter, Probe Lasso, and DMRcate. These are described in detail below. Bumphunter is a statistical method that uses smoothing at the site level to help detect potential DMRs. Bump hunter uses linear regression modeling and smoothing techniques on the data to help determine differentially methylated regions that are considered significant [3]. The first step is to fit the following statistical model:

$$Y_{ij} = \mu(t_j) + \beta(t_j)X_i + \sum_{k=1}^{p} \gamma_k(t_j)Z_{i,j} + \sum_{l=1}^{q} a_{l,j}W_{i,l} + \varepsilon_{i,j} \tag{1.1}$$

[3]

With DNA methylation data, the j^{th} CpG site from the genome for individual i, will have a methylation level of Y_{ij}. This model will be fit from linear modeling the data through the use of a framework similar to limma, which is commonly used in other areas of genomics. The Y_{ij} will be a transformed β value from the beadchip. The transformation will be:

$M = log_2(\beta/(1-\beta))$ [9]

This transformation is done so that the Y_{ij} values are not bounded between zero to one like β values. At the j^{th} CpG site of the genome, t_j represents the CpG site will be location within the genome. The average DNA methylation at the CpG sites of the control group (baseline) is represented by $\mu(t_j)$. The X_i term denotes the phenotype group (e.g. disease or healthy) within which individual i falls. $\beta(t_j)$ represents a possible association between X_i and Y_{ij} at location t_j. Any p number of confounders that might exist will be shown with Z's. The effects that confounder k has at t_j is represented with $\gamma_k(t_j)$. Any confounders or batch effects that are not able to be measured will be shown with W. The amount of effect of the l^{th} unmeasured confounder at t_j will be represented with $a_{l,j}$. For data used

within this thesis, there will be no Z's or W's so these terms can be removed from the model. Any unexplained variability or error will be denoted with the error term $\varepsilon_{i,j}$, with variances depending on the location $\sigma^2(t_j)$. The primary interest is to identify locations in the genome, shown with t_j, where $\beta(t_j)$ does not equal zero (i.e. "bumps"), thus indicating that there is a possible connection between a phenotype such as disease status and the location of methylation. These sites will be grouped into regions of interest that might be DMRs. These are portrayed with $R_n, n = 1, \cdots, N$ where $\beta(t_j) \neq 0$ for all $t \in R_n$ [3].

The goal of Bumphunter is to identify these regions or "bumps" across the genome and this is accomplished by a set of four steps. First, $\beta(t_j)$ must be estimated for each t_j. These values of $\beta(t_j)$ are then used to help create a smoothing function, $\beta(t)$. The smoothing within this thesis will be done with the Loess method. This smoothing function will be used to help determine any anomalies or bumps in the function to help predict possible R_n or the possible regions where there might be significant differentially methylation occurring. R_n will be made up of smoothed β values that are either more than or less than a chosen threshold value K. The area of each region is then calculated as a metric to indicate the strength of differential methylation in the region. Permutation techniques are then employed to assign statistical uncertainty to each of the previous regions that were identified with the smooth function [3]. To do this, the phenotype variable X_i is permuted and the bumphunter process of linear modeling, obtaining smoothed βs, and finding areas for regions of interest is repeated a large number of times (e.g. $B = 1000$). The areas of the regions that are produced in the B permutations are considered null areas and can provide a null distribution for the region areas. This distribution can then be used to find a p-value for each potential region. Multiple p-values are going to be calculated, one for each region. This causes a problem when multiple hypothesis are conducted. To combat this, false discovery rate (FDR) will be controlled at level α using the Benjamini and Hochberg approach [17]. The

resulting bumps deemed significant will be considered significant DMRs. Note that an alternative bootstrap method may be used, especially when the model contains batch effects to improve computational efficiency.

Probe Lasso is a method that can offset potential bias which might occur from the unequal concentrations of CpG sites represented by probes in different types of genomic regions (e.g. CpG islands, open sea). It does this by controlling a window size for which the possible DMRs will be defined against the local density of CpG sites with the probes present on the array. A "flexible window" or "lasso" is "thrown" around identified CpG sites with a probe that exhibit significant difference in methylation. The sites that are deemed to be significant are determined by site level testing for significant differential methylation levels based on the reported β values. A linear model similar to that of Bumphunter as in equation (1.1) without the batch or confounding effects is fit using a method called limma [18]. The limma approach differs from Bumphunter in the way the error variance is estimated. Limma uses an empirical Bayes approach to shrink individual site level variance to a common pooled estimate. Site level significance is established by testing whether $\beta(t_j) = 0$. The lassos have a center located at the probe and a set mean radius on either side of the probe. The lasso will be able to expand larger in regions that are more sparse with potential sites (e.g. the open sea). In high concentration of CpG sites, the lassos will be shrunk (e.g. CpG islands). The average will be decided by the researcher. Any sites that fall within this lasso that show significance in being differentially methylated are then considered part of a potential DMR, with the hope of meeting the minimum number of sites which are within the lasso's boundary to be considered a DMR. A site within the lasso is deemed significant using the same limma approach that was used to determine significance for the sites at the centers of the lassos. If there are overlaps of lassos, then the lassos and the probes within the adjacent lassos are combined together to make a potentially larger DMR. The amount of separation between lassos to be considered combined together is a parameter that will be set by the researcher. The minimum number of sites within each

lasso that needed to be defined as a DMR and radius of the lassos are two parameters for the Probe Lasso test that are important in determining DMRs [19]. The p-value for each DMR is found by using Stouffer's method to weight site level p-values and calculate a combined p-value for the region [20]. Stouffer's method is used because neighboring sites have been shown to be correlated with each other in terms of having similar methylation levels. Therefore, Stouffer's method uses weights to account for the distance between sites, which is based on a correlation matrix of the normalized beta values from the data within each DMR. With this matrix, sites that are uncorrelated with other sites are weighted more heavily, while sites that show correlation in the methylation levels are down weighted. After this, the weighted p-values of sites can be combined to obtain a p-value of the DMR itself. Due to the multiple p-values being found at the site level, a multiple hypothesis type 1 error rate can occur when finding the DMR p-value [19]. The false discovery rate for testing multiple DMRs is controlled at level α using the Benjamini and Hochberg approach [17].

The final method considered for DMR testing in this thesis is DMRcate. DMRcate finds test statistics for identifying differential methylation at each of the CpG sites similar to Bumphunter and Probe Lasso. This is done through the use of the limma framework to conduct linear modeling to obtain a test statistic, t, for testing the difference in methylation at the site level. This test statistic is based on a test for $\beta(t_j) = 0$ as in equation (1.1) described in Bumphunter using M values and a variance shrinkage estimation procedure. However, in this method, the direction of the difference of methylation between the two compared groups does not matter. At a site, a cancerous group's methylation level being higher than the noncancerous group is not important. The only important information for this test is the magnitude between the methylation levels in the two groups. Thus the test statistic is squared, t^2, to obtain magnitude not direction. Gaussian Smoothing, as discussed earlier, is then applied with the new test statistics. The matrix that is used to smooth the model is determined with a known bandwidth parameter of λ [16]. This parameter is also later used to determine which sites should be clustered to form significant DMRs.

The new values for the test statistics obtained from smoothing are then modeled using the Satterthwaite method [21]. Satterthwaite models the smoothed statistics $S_{KY}(i)$ by a scaled χ^2 distribution where it's parameters are determined with $S_K(i)$ and $S_{KK}(i)$ that were obtained from the Gaussian Smoothing. The P-values are then found from this model by utilizing a χ^2 distribution and the p-values produced are then adjusted with the Benjamini and Hochberg correction to control the FDR. Then sites or probes that are found to be significant with their adjusted p-values are grouped into possible DMRs. The grouping of significant sites is controlled with the bandwidth value λ that was used with the Gaussian Smoothing earlier [16]. This allows this method to use the minimum adjusted p-value in the regions to be a representative p-value of the entire region. A significant parameter for DMRcate is the chosen false discovery rate (FDR). This determines which sites will be deemed to be significant and potentially be clustered into DMRs.

1.4. SUMMARY

When determining if there is a possible connection between DNA methylation and developing diseases, the 450K microarray is commonly used to collect information for large scale studies with many samples. Typically at least two different phenotypes will be investigated to determine their association with DNA methylation patterns. An example could be a group of individuals that have tested positive and negative in possessing a type of cancer. When analyzing the resulting data from the microarrays, filtering and normalizing the data to counter any error that might result from the samples or probes is first performed. Once this is done, either site or region level testing is done to determine locations where there is a significant difference in methylation levels between the two or more groups represented in the samples.

Concentrations of identified CpG sites that exhibit significant differential methylation can be clustered and tested for being differentially methylated regions (DMRs). To determine if there are any DMRs and where they might be, numerous testing methods can be

used on the gathered data using statistical techniques. In this thesis, the Bumphunter, Probe Lasso, and DMRcate methods are used and compared in their ability to identify significant DMRs in the data.

2. EVALUATING REGION-LEVEL TESTS FOR 450K METHYLATION DATA

2.1. DATA SET

The DNA methylation data that will be examined in this thesis was obtained from 450K microarrays by Fred Hutchinson Cancer Research Center in 2013. Data were obtained from a group of either Kenyan or Senegalese women who are HIV positive and have various stages of pre-cancerous cervical lesions. There were 5 types of groups among the women that were tested: three different types of cervical pre-cancerous lesions of CIS, CIN2, CIN3, a group with Squamous Cell Carcinoma (SCC), and a group of women without any pre-cancerous cervical lesions (Negative). It is of interest to determine whether there are any significant differences in methylation levels among the different cancer stages. Within this thesis, the SCC and negative group will be used and examined. The negative group is deemed the 'control' and the SCC group is the 'test' group to help determine if there are possible links between methylation levels in certain regions of the genome and the existence of the pre-cancerous lesions. There are 5 women with SCC cervical cancer and 8 women with no pre-cancerous lesions.

2.2. ANALYSIS METHODS WITH R/BIOCONDUCTOR SOFTWARE

To process and analyze the data obtained from DNA methylation microarrays, the statistical program R is often used since many researchers develop packages for method implementation and annotation packages are available that can be easily integrated into results. In R, a suite of packages for analyzing genomic data is available in a software called Bioconductor. Packages within Bioconductor can be used to test for differential DNA methylation in either regions or single sites in diverse ways. The primary package used for processing the DNA methylation microarray data is "minfi". With this package, methylation

levels at individual sites can be calculated and potentially significant methylation level differences at CpG sites can be identified. Also, this package is the primary package that enables the site and sample level filtration that was discussed in chapter 1. In conjunction with using "minfi" for site level methylation analysis, numerous packages can be used to help identify possible DMRs such as the following: "Bumphunter", "Probe Lasso", and "DMRcate." There are other methods and packages that can help identify possible DMRs however, the previous three will be the ones examined within this thesis [3].

Recently, a package known as "The Chip Analysis Methylation Pipeline" (ChAMP) was developed that combined several processing and analysis methods into a single package. ChAMP is a pipeline that lets the user pre-process data from 450K or EPIC microarrays, normalize the data, conduct a Batch effect correction if possible, detect possible differential methylated positions (DMP), and detect possible differentially methylated regions (DMRs) [22]. This pipeline as of July 25, 2017, lets the user conduct DMR testing with Bumphunter, DMRcate, and Probe Lasso [3, 16, 19].

With each of these three methods for region level testing, various parameters for the tests can be changed and adjusted through the common ‘champ.DMR()’ function in the ChAMP package. Several parameters are shared between all three tests including the minimum number of probes that are required to from a region and the minimal p-value to determine if a DMR is defined as significant. However, some parameters are specific to the individual methods. Some important parameters solely for the Bumphunter test include maximum length of DMR and the integer value for the number of permutations used in the null distribution for the test. For DMRcate, important parameters include: the maximum distance between SNP and CpG probes to be taken out of data, the lambda value for the bandwidth used in the Gaussian kernel for smoothing, and the scaling factor used for the bandwidth. Probe Lasso parameters include the minimum radius of each lasso used, the

minimum distance between neighboring DMRs, minimum size of DMRs, and minimum significant value for probes to be counted with the DMR. Each of these parameters can be changed and fine-tuned to better detect possible DMRs [22].

The ChAMP package also has a default pipeline that can automatically call distinct functions involved in analyzing the data, including preprocessing, normalization, finding DMPs, clustering, block finding, and finding DMRs. A general overview of the various aspects of analyzing 450K data is given below.

The data are first loaded and interpreted from .idat files that are obtained from the 450K microarray to determine the methylation levels at all CpG sites by using the 'champ.load()' command [5, 12, 23]. In addition to determining methylation levels, this function also filters out any sites with probes that have been deemed to fail, probes that with less than 3 beads in less than five percent of the samples, any probes that do not contain any CpG sites, any probes that can be related to SNPs, any multi-hit probes, and any probes with data from Chromosome X or Y. This is done through the use of the package of "minfi." To determine if there are any quality control issues with the samples the function of 'champ.qc()' can be utilized. This will help create several figures that can be used to determine if any of the subjects should be removed. Experimental error, such as from differing probe types, needs to be corrected by normalization of the data using the 'champ.norm()' function [14]. This is done through an option of four types normalization methods, "BMIQ," "SWAN," "FunctionalNormalize," and "PBC" [8, 14, 23, 24]. The default setting and the method used in this thesis was the BMIQ function. The advantage that BMIQ has over the other three methods is that it is able to run in parallel. This means that if the computer can support more cores being run parallel to each other, then the faster BMIQ will normalize the data [14].

ChAMP also has the ability through the use of the "minfi" package to analyze site level differential methylation. This is done with the ChAMP package function of 'champ.DMP()' [25, 26]. Individual CpG sites are examined for differential methylation

rather than testing for DMRs. Some additional options include identifying and counteracting batch effects with the 'champ.runCombat()' function and identifying blocks with the ChAMP function of 'champ.Block()'. However, for this thesis the only ChAMP functions used were the 'champ.load,' 'champ.qc,' 'champ.norm,' and 'champ.dmr.' Through the use of these functions, the Bumphunter, Probe Lasso, and DMRcate tests can be utilized. Although the other functions were not used, ChAMP does give a flexible overall package that lets the user analyze both 450K and EPIC microarrays in multiple ways. It also lets the user produce plots of the various results, many of which are connected to the genomic annotation to aid in interpreting results [22].

2.3. EVALUATING DIFFERENT PARAMETERS

Within the ChAMP package there are several values that can be changed prior to evaluating DMRs within the 'champ.DMR()' function. For simplicity, these values that are input to the function are called parameters. Depending on whether the test is Bumphunter, Probe Lasso, or DMRcate, the parameters can vary. Firstly, there are 7 parameters that exist across all three tests. They are beta, pheno, arraytype, method, minProbes, adjPvalDmr, and cores. Beta is a matrix of normalized beta values representing the methylation level for each CpG site. Pheno represents the phenotype or factor to be analyzed. In the data used in this thesis, the phenotype is the cervical cancer status(Control/Negative or Cancerous/SCC). Arraytype lets the user determine what type of technology was used to collect the data. The two options are 450K or EPIC with the 450K being used here. Method enables the choice of Bumphunter, Probe Lasso, or DMRcate. MinProbes decides the minimium amount of probes needed for determining what will defined as clusters for potential DMRs. For example, if the minProbes is set to 7, then there will not be any DMRs that would have 6 or less probes in their range. AdjPvalDmr lets the user choose a p-value threshold to determine if a DMR is found to be significant. The cores setting helps decrease processing time by running the function in parallel[22].

The parameters that are solely used in Bumphunter are maxGap, cutoff, pickCutoff, smooth, smoothFunction, useWeights, permutations, B, and nullMethod. MaxGap is the largest possible length between sites to be considered part of the same DMR. Cutoff determines what regions or bounds for the genome should be tested for DMRs. PickCutoff is the value used with the Bumphunter algorithm as a cutoff value with its permutation process. Smooth determines if the SmoothFunction from ChAMP will be used to smooth the data. There are two options for smoothing types in the SmoothFunction: loessByCluster or runmedByCluster. The loessByCluster denotes the loess smoothing method, while runmedByCluster denotes running median smoothing method [27]. The parameter of UseWeights affects the loessByCluster by being able to determine the weight functions that it will use. If runmedBycluster was used with SmoothFunction, then UseWeights is ignored. The method of creating the null distribution has a choice between bootstrap and permutation and is chosen with the parameter nullMethod. Permutations allows the user to create a null distribution by using a chosen matrix that will help randomize the data. However, this parameter is only used if Bootstrap is not used to randomize the data. When creating the null distribution using the bootstrap, the parameter that determines the number of times that resampling occurs with the data is the B[3, 22].

The parameters for only Probe Lasso are meanLassoRadius, minDmrSep, minDmrsize, adjPvalProbe, Rplot, PDFplot, and resultsDir. MeanLassoRadius determines that each probe will have a lasso that is a chosen radius. MinDmrSep allows the user to decide for different DMRs, what the minimum distance will be that separates them from each other. MinDmrSize decides for all possible DMRs what the minimum size they must be for the DMRs to be considered significant. This means that the length of the DMR must be at least some chosen value of minDmrSize. AdjPvalProbe decides for each probe, the p-value threshold used to consider the probe to be part of a potential DMR. PDFplot gives the option of creating a plot of the DMRs that were created with the lassos. The ResultsDir parameter indicates where to store the final results on the user's computer [19, 22].

Finally, the unique parameters for DMRcate are rmSNPCH, fdr, dist, mafcut, lambda, and C. RmSNPCH gives the option of filtering the Beta (or M values) that are reported for the probes by distance from SNPs. Fdr determines the value for the false discovery rate. Dist allows the user to determine the largest possible distance between SNPs and CpGs for filtering. Mafcut determines if probes can be removed from the data by determining the smallest value for the allele frequency of each probe. Lambda decides the Gaussian kernel bandwidth to be used in conjunction with the smooth-function estimation that is used in DMRcate to help determine DMRs. For the Gaassian kernel, the parameter that determines what factor the bandwidth should be scaled at is C [16, 22].

2.4. SUMMARY

In this thesis, DMRs will first be identified using default parameters of each method. Results will be compared by investigating the number of DMRs identified by each method and summarizing various aspects of the identified DMRs. Similarity of DMRs identified between the methods will be examined by calculating the average percentage of overlap in genomic area covered by DMRs commonly identified by the methods. Further analysis will be conducted to better understand how certain parameters affect the results and attempt to find settings where all methods identified the same number of DMRs. Although it is uunknown where the true DMRs are located, the goal of this thesis is to provide an illustration of how the three methods can be used to find DMRs and provide a comparison of the results in real data. Investigating the different parameter settings and comparing the three methods provides insights into similarities and differences between the methods and their results.

3. RESULTS

3.1. LOADING AND FILTERING

When the .idat files for the data are first loaded into ChAMP, there are a subset of 485,512 probes that have information about methylation and unmethylation for each sample. The filtration process then removes certain probes based on criteria described previously. In the first step, any probes that have a p-value of detection of less than 0.01 were removed. For these data, this resulted in 2642 probes being removed. Next, any probes with beadcounts that are less than 3 in at least 5 percent of the samples were removed. This amounted to 4537 probes being filtered out. Then only probes that have association with CpG sites were to be used with any testing for DMRs. The number of probes taken away for this filtration was 2999. The fourth filtration involved removal of any probes that have SNP association, resulting in 49,692 probes being removed from the data. The fifth step eliminated multi-hit probes. Thus, 7059 probes were taken out of the data. Lastly, 10,106 probes were removed due to being located on either the X or Y chromosomes. The final number of site level probes that used in the DMR testing was 408,477.

For sample level filtration, a raw density plot was produced and can be seen in Figure 3.1. There were no noticeable samples that appeared as outliers. Thus, all 13 samples were used for the DMR testing.

3.2. PARAMETER SETTINGS FOR IDENTIFYING DMRS

All three methods (Bumphunter, Probe Lasso, and DMRcate) were first implemented to identify DMRs under their default parameter settings and these results are compared. Then an effort was made to better understand how the parameter settings affect the results and compare the methods when a similar number of DMRs are identified for all methods.

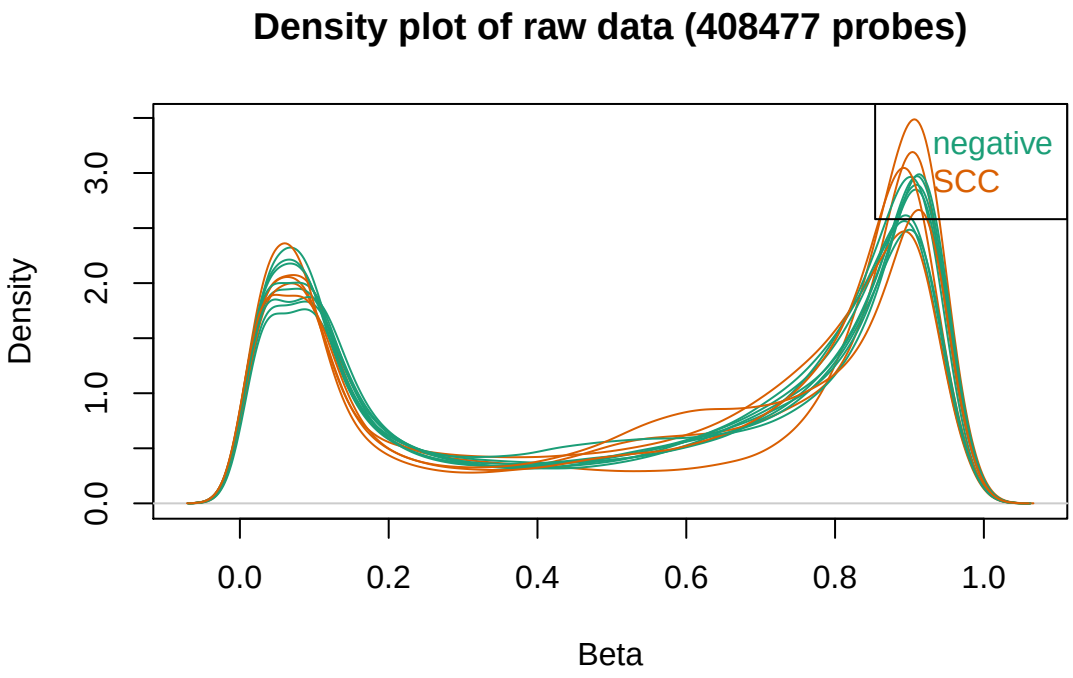

Figure 3.1. Raw Density Plot of Beta values for all 13 samples. Samples in different phenotype groups (Negative vs SCC) are labeled in different colors.

To accomplish this, parameters were altered for two of the methods to achieve the same numbers of DMRs as identified in the default values of the third method. This procedure was performed using the defaults for each of the three methods. Table 3.1 summarizes the analyses that were conducted and introduces notation that will be used throughout the remainder of this chapter. Tables 3.2, 3.3, and 3.4 provide the values of the parameters used in Bumphunter, Probe Lasso, and DMRcate respectively for each analyses. With all three methods, an attempt was made to keep all common parameters (minprobes, adjPvalDMR) that the three methods share the same. This was done so that a better comparison between methods could be made. The only exception to this was the parameter that determines the number of cores when running DMRcate. In order for DMRcate to be able to run, the

number of cores must be set to 1 instead of the default of 3. If kept at 3, then the test was not able to run. However, this should only affect the computational efficiency but not the results.

To change the number of DMRs identified by a specific method, parameters unique to the method type were changed. In Bumphunter and DMRcate only a single parameter was changed to find similar numbers of DMRs as defaults of the other methods. In Bumphunter, the maxGap was changed as highlighted in Table 3.2. As the maxGap parameter increases, the number of DMRs identified increases. In DMRcate, the FDR was changed as is highlighted in Table 3.4. From the results, with an increase to FDR cutoff that declares individual sites significant, there was a corresponding increase in the number of DMRs identified. Probe Lasso on the other hand, required changing two parameters, meanLassoRadius and adjPvalProbe as highlighted in Table 3.3. The need to change two parameters was due to the need to get similar number of DMRs as the other two methods and this was accomplished with the combination of both parameters rather than a single parameter change. Increasing both the mean Lasso Radius and adj Pval probe seemed to increase the number of DMRs identified.

There were certain pieces of information about the resulting DMRs with each method that are important to note. The important pieces of information are the number of DMRs identified with each method, which Chromosome each DMR is located on, and the width of each DMR. All three of these were reported in the results of each test. Another important piece of information was the number of CpGs in each DMR. However, this was only reported with DMRcate. Thus, to help find these values for both Bumphunter and Probe Lasso, an annotation table with the location of all CpGs known as "Illumina450ProbeVariants.db" was used in conjunction with the results of the DMRs from both Bumphunter and Probe Lasso [28]. With each DMR identified, a function was made to determine if any CpG sites from the annotation table were located within the bounds of the DMR. The number of CpGs in the DMRs might not be exactly correct since the annotation table includes all CpG prior to

Table 3.1. Notation for each of the nine analyses conducted

Notation	Description
BMP	Bumphunter with default parameters
PL	Probe Lasso with default parameters
DC	DMRcate with default parameters
BMP-PL	Bumphunter with changed parameters to get a similar number of DMRs as Probe Lasso default
BMP-DC	Bumphunter with changed parameters to get a similar number of DMRs as DMRcate default
PL-BMP	Probe Lasso with changed parameters to get a similar number of DMRs as Bumphunter default
PL-DC	Probe Lasso with changed parameters to get a similar number of DMRs as DMRcate default
DC-BMP	DMRcate with changed parameters to get a similar number of DMRs as Bumphunter default
DC-PL	DMRcate with changed parameters to get a similar number of DMRs as Probe Lasso default

Table 3.2. Bumphunter Parameters

Method	BMP	BMP-PL	BMP-DC
minProbes	7	7	7
adjPvalDmr	0.05	0.05	0.05
cores	3	3	3
maxGap	300	19	570
cutoff	NULL	NULL	NULL
pickCutoff	TRUE	TRUE	TRUE
smooth	TRUE	TRUE	TRUE
smoothFunction	loessByCluster	loessByCluster	loessByCluster
useWeights	FALSE	FALSE	FALSE
permutations	NULL	NULL	NULL
B	250	250	250
nullMethod	bootstrap	bootstrap	bootstrap
Number of DMRs	277	1	494

Table 3.3. Probe Lasso Parameters

Method	PL-BMP	PL	PL-DC
minProbes	7	7	7
adjPvalDmr	0.05	0.05	0.05
cores	3	3	3
meanLassoRadius	1180	375	1770
minDmrSep	1000	1000	1000
minDmrSize	50	50	50
adjPvalProbe	0.3	0.05	0.3
Number of DMRs	277	1	490

Table 3.4. DMRcate Default Parameters

Method	DC-BMP	DC-PL	DC
minProbes	7	7	7
adjPvalDmr	0.05	0.05	0.05
cores	1	1	1
rmSNPCH	T	T	T
fdr	0.0329	0.001	0.05
dist	2	2	2
mafcut	0.05	0.05	0.05
lambda	1000	1000	1000
C	2	2	2
Number of DMRs	276	1	492

filtering. Thus, extra CpGs could be included. However, this should give an idea about what the numbers of CpGs in each DMR might be. The final metric important for comparing methods is how much the DMRs from each test overlap with DMRs identified with other methods. That is, the percentage of overlap in the genomic area covered by DMRs identified by two methods is calculated. To be able to find this, the package known as "IRanges" was used [29]."

3.3. COMPARING BMP, PL, DC

The first comparison was made when Bumphunter, Probe Lasso, and DMRcate were kept at their default parameters, except for the number of cores changed to 1 for DC. As seen in Tables 3.5, 3.6, and 3.7, BMP was able to identify 277 DMRs, PL was able to find 1, and DC found 492. As seen in Table 3.5, both BMP and DC identify DMRs that have similar average and standard deviation of widths of DMRs, while PL's width is smaller. A histogram of the widths of DMRs is given in Figure 3.2. It can be seen that DC found a higher frequency of longer DMRs that BMP, even though the averages are similar. From Table 3.6 and Figure 3.3, BMP and DC had similar numbers of the mode and standard deviation of the Chromosome number, while once again PL was different from the other two. It is interesting to note that in Table 3.7 and Figure 3.4, PL appears to have more than twice as many CpGs in its shorter DMR than either of the other two tests. Also from Table 3.8, 233 of the DMRs exhibited overlap between BMP and DC. Of these, a high average of overlap exists at 94.8 % when comparing the DMRs found between BMP and DC. Figure 3.5 reveals that the majority of these 233 exhibited over 90% overlap. Also notable is that the single overlap between the DMR from PL was identified with both BMP or DC.

Table 3.5. Width of DMRs with BMP, PL, and DC

Method	Number of DMRs	Average Width of DMRs	Standard Deviation of Width of DMRs
BMP	277	1175.018	634.300
PL	1	654	0
DC	492	1363.980	769.245

Table 3.6. Chromosome of DMRs with BMP, PL, and DC

Method	Number of DMRs	Mode of Chromosomes of DMRs	Standard Deviation of Chromosome of DMRs
BMP	277	1.000	5.832
PL	1	6	0
DC	492	1.000	6.020

Table 3.7. Number of CpG sites in DMRs with BPM, PL, and DC

Method	Number of DMRs	Mean of Number of CpGs	Standard Deviation of Number of CpGs
BMP	277	11.671	4.300
PL	1	31	0
DC	492	11.376	5.465

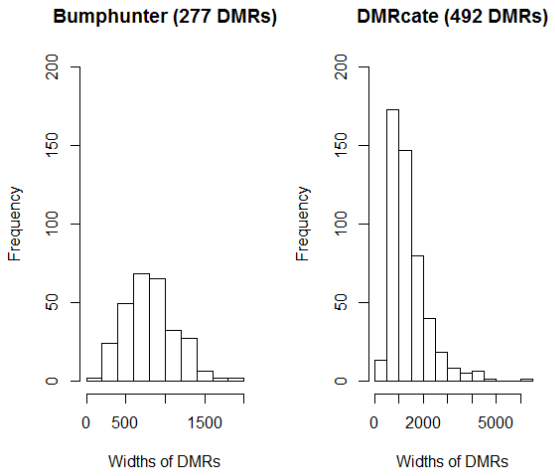

Figure 3.2. Width of the DMRs found with BMP and DC

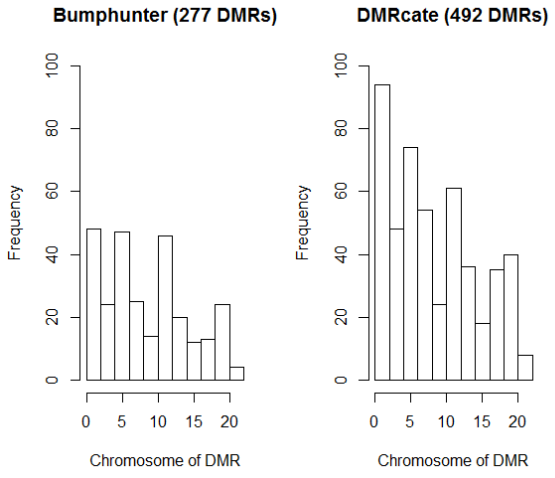

Figure 3.3. Chromosomes of the DMRs found with BMP and DC

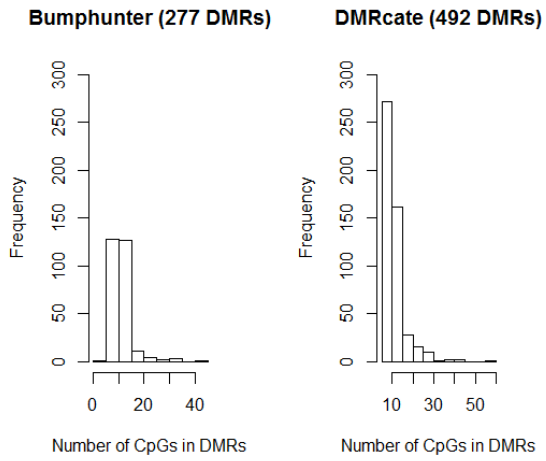

Figure 3.4. Number of CpGs for each of the DMRs found with BMP and DC

Table 3.8. Percentage overlaps between BMP, PL, and DC

Comparison	Number of Overlaps	Mean of Percentage of Overlap	Standard Deviation of Percentage Overlap
BMP vs PL	1	1	0
PL vs DC	1	1	0
BMP vs DC	233	0.949	0.128

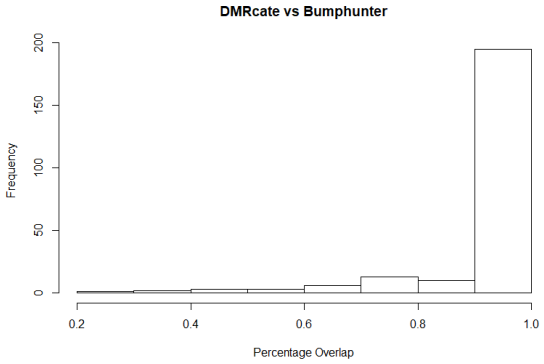

Figure 3.5. Percentage of overlap between BMP and DC

3.4. COMPARING BMP, PL-BMP, DC-BMP

The next comparison of interest is to investigate results when DMRcate and Probe Lasso have altered parameters to achieve the same number of DMRs (277) as Bumphunter default settings. When parameters were changed for both DC-BMP and PL-BMP, then 277 and 276 DMRs, respectively, were found. When the FDR parameter for DC-BMP was reduced in value, the number of significant sites to be used in DMRs was also decreased due to a more stringent inclusion threshold. This resulted in a reduction in DMRs. When PL-BMP parameters of MeanLassoRadius was increased in value, this created a larger area around sites deemed significant to have lassos. This resulted in more CpG sites within each lasso, and the sizes of DMRs growing. When AdjPvalProbe was increased, then the significance threshold used to determine which sites to fall within each lasso increased and thus made less stringent. This resulted in more sites being able to be collected with each lasso. With all three methods identifying similar numbers of DMRs, a comparison of the DMRs identified between the methods can be conducted. The width statistics given in Table 3.9 and Figure 3.6 reveal that the widths of DMRs between BMP and DC-BMP are similar, whereas PL-BMP is drastically different in both average and standard deviation. From Table 3.10 and Figure 3.7, all three methods have different mode values for locations of DMRs. The number of CpGs in the DMRs from BMP and DC-BMP are similar, as seen in Table 3.11 and Figure 3.8, but PL-BMP has both a higher average and variation in CpG density. Lastly, when examining the overlaps between methods in Table 3.12 and Figure 3.9, the highest total number of overlaps between two methods was found between BMP and DC-BMP (176). While the comparisons of BMP with PL-BMP, and PL-BMP with DC-BMP showed comparable numbers of overlaps (101 and 110, respectively), the PL-BMP and DC-BMP methods appear to have a much higher average of the percentage overlap (69.98% compared to 33.81%).

Table 3.9. Width of DMRs with BMP, PL-BMP, and DC-BMP

Method	Number of DMRs	Average Width of DMRs	Standard Deviation of Width of DMRs
BMP	277	1175.018	634.300
PL-BMP	277	9820.430	8135.753
DC-BMP	276	1163.442	615.362

Table 3.10. Chromosome of DMRs with BMP, PL-BMP, and DC-BMP

Method	Number of DMRs	Mode of Chromosomes of DMRs	Standard Deviation of Chromosome of DMRs
BMP	277	1.000	5.832
PL-BMP	277	6.000	5.609
DC-BMP	276	11.000	6.019

Table 3.11. Number of CpG sites in DMRs with BMP, PL-BMP, and DC-BMP

Method	Number of DMRs	Mean of Number of CpGs	Standard Deviation of Number of CpGs
BMP	277	11.671	4.300
PL-BMP	277	26.361	30.882
DC-BMP	276	11.384	5.334

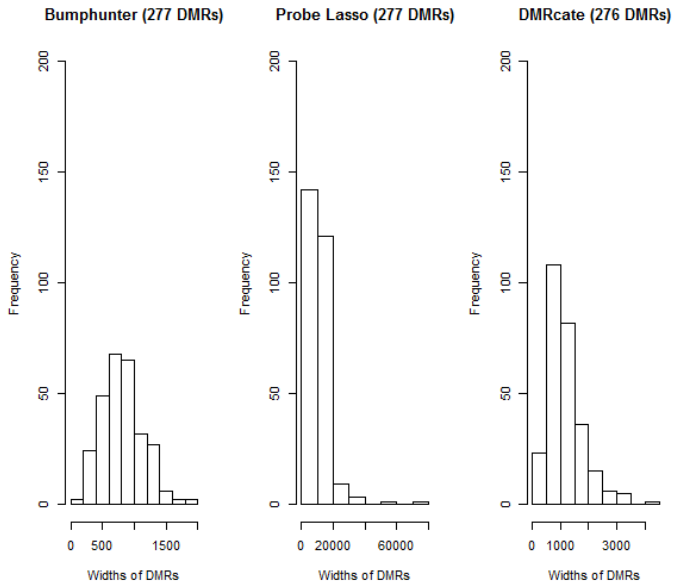

Figure 3.6. Width of the DMRs found with BMP, PL-BMP, and DC-BMP

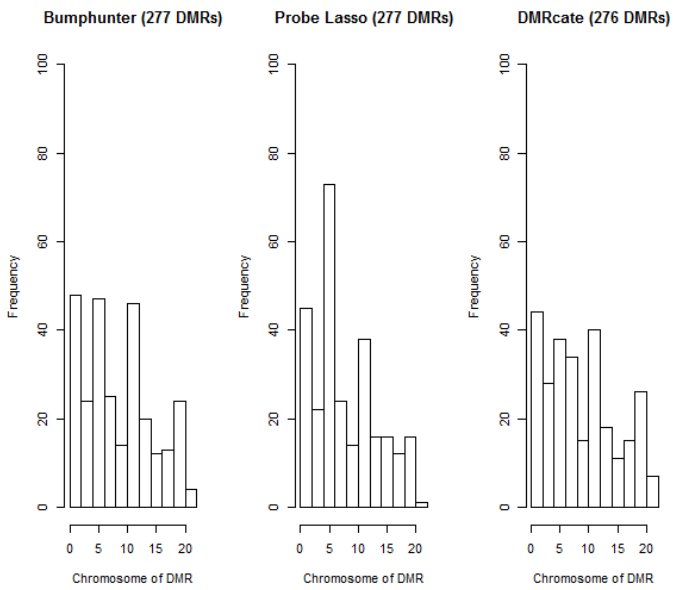

Figure 3.7. Chromosomes of the DMRs found with BMP, PL-BMP, and DC-BMP

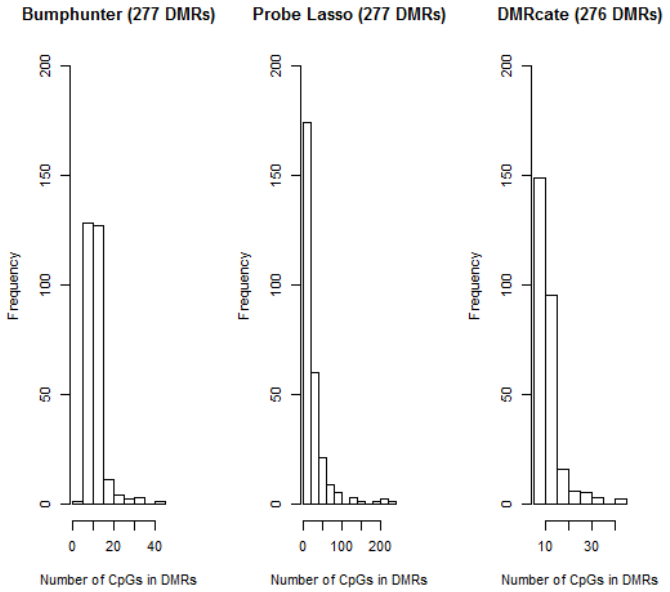

Figure 3.8. Number of CpGs in the DMRs found with BMP, PL-BMP, and DC-BMP

Table 3.12. Percentage overlaps between BMP, PL-BMP, and DC-BMP

Comparison	Number of Overlaps	Mean of Percentage of Overlap	Standard Deviation of Percentage Overlap
BMP vs PL-BMP	101	0.338	0.334
PL-BMP vs DC-BMP	110	0.700	0.311
BMP vs DC-BMP	176	0.689	0.282

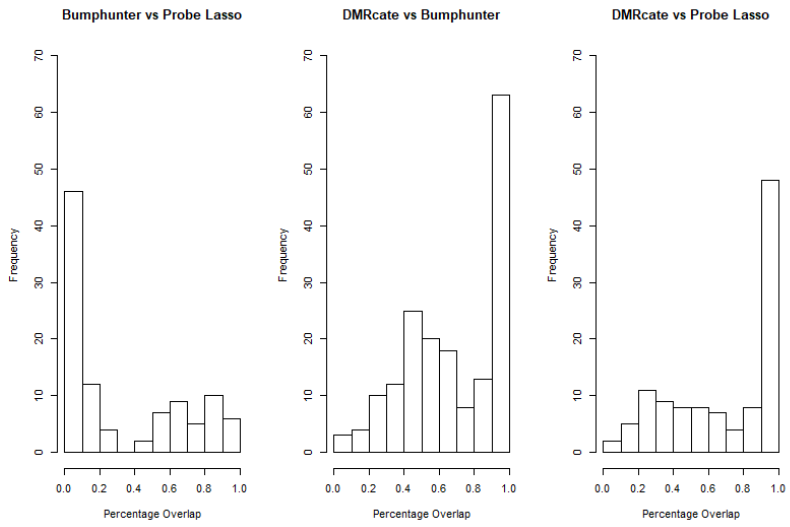

Figure 3.9. Percentage of overlap between BMP, PL-BMP, and DC-BMP

3.5. COMPARING BMP-PL, PL, DC-PL

In this comparison, the parameters of BMP-PL and DC-PL were changed so that each test produced only one DMR as significant in order to compare to the PL default results. When the MaxGap parameter of BMP-PL decreased, the distance between sites to be considered part of the same DMR became smaller. This resulted in less sites being grouped together and decreasing the number of DMRs found. The parameters of DC-PL were changed to reduce the number of DMRs and in the same manner as discussed earlier with DC-BMP. The results for PL, BMP-PL, and DC-PL were compared for the single DMR found with the three methods. When looking at the widths of the DMRs found with these methods, Table 3.13 shows that PL and DC-PL appear to have similar width while BMP-PL is a significantly shorter DMR. As seen in Table 3.14, all three methods identified a DMR that was located on Chromosome 6. Once again, PL and DC-PL have similar number of CpGs in their DMRs as seen in Table 3.15. There was only a single overlap of DMRs between any of the tests. From Table 3.16, the percentage of the single overlap between PL and DC-PL was 82.68%.

Table 3.13. Width of DMRs with PL, BMP-PL, and DC-PL

Method	Number of DMRs	Average Width of DMRs	Standard Deviation of Width of DMRs
BMP-PL	1	33	0
PL	1	654	0
DC-PL	1	791	0

Table 3.14. Chromosome of DMRs with PL, BMP-PL, and DC-PL

Method	Number of DMRs	Mode of Chromosomes of DMRs	Standard Deviation of Chromosome of DMRs
BMP-PL	1	6	0
PL	1	6	0
DC-PL	1	6	0

Table 3.15. Number of CpG sites in DMRs with PL, BMP-PL, and DC-PL

Method	Number of DMRs	Mean of Number of CpGs	Standard Deviation of Number of CpGs
BMP-PL	1	9	0
PL	1	31	0
DC-PL	1	28	0

Table 3.16. Percentage overlaps between PL, BMP-PL, and DC-PL

Comparison	Number of Overlaps	Mean of Percentage of Overlap	Standard Deviation of Percentage Overlap
BMP-PL vs PL	0	0	0
PL vs DC-PL	1	0.827	0
BMP-PL vs DC-PL	0	0	0

3.6. COMPARING BMP-DC, PL-DC, DC

When the parameters for DC were left at default except for the number of cores being 1, the two tests of BMP-DC and PL-DC had their parameters changed to also reflect a similar number of identified DMRs. When BMP-DC had the parameter MaxGap increased in value, this made the distance between sites to be considered part of the same DMR to be greater. This resulted in more CpG sites to be clustered together and a increase in the number of DMRs found from this method0. When the parameters for PL-DC were changed to increase the number of DMRs, it created a similar effect as discussed earlier for PL-BMP. From Table 3.17 and Figure 3.10, DC was able to identify 492 DMRs with its default parameters, while BMP-DC and PL-DC identified 494 and 490 respectively. Both average and standard deviation for the widths were drastically different when comparing either of BMP-DC or DC to PL-DC. PL-DC on average identified DMRs that were more then ten times as large as either of the other methods. The chromosome that the DMRs were found on are similar for all three tests. The Table 3.18 and Figure 3.11 shows that all three tests

have similar mode value for the DMRs with Chromosome 1. When the number of CpGs per DMR was examined, PL-DC stood out, as seen with Table 3.19 and Figure 3.12, with on average having more than twice the amount of CpGs then either BMP-DC or DC. Lastly, when percentage of overlaps between tests was examined in Table 3.20 and Figure 3.13, the most number of overlaps was between BMP-DC and DC. This pair also exhibited the highest average of percentage overlap and the lowest standard deviation than the other two comparisons.

Table 3.17. Width of DMRs with DC, BMP-DC, and PL-DC

Method	Number of DMRs	Average Width of DMRs	Standard Deviation of Width of DMRs
BMP-DC	494	1245.393	736.602
PL-DC	490	16626.730	13471.940
DC	492	1363.980	769.245

Table 3.18. Chromosome of DMRs with DC, BMP-DC, and PL-DC

Method	Number of DMRs	Mode of Chromosomes of DMRs	Standard Deviation of Chromosome of DMRs
BMP-DC	494	1.000	5.852
PL-DC	490	1.000	5.609
DC	492	1.000	6.020

Table 3.19. Number of CpG sites in DMRs with DC, BMP-DC, and PL-DC

Method	Number of DMRs	Mean of Number of CpGs	Standard Deviation of Number of CpGs
BMP-DC	494	12.130	4.976
PL-DC	490	29.308	43.442
DC	492	11.376	5.465

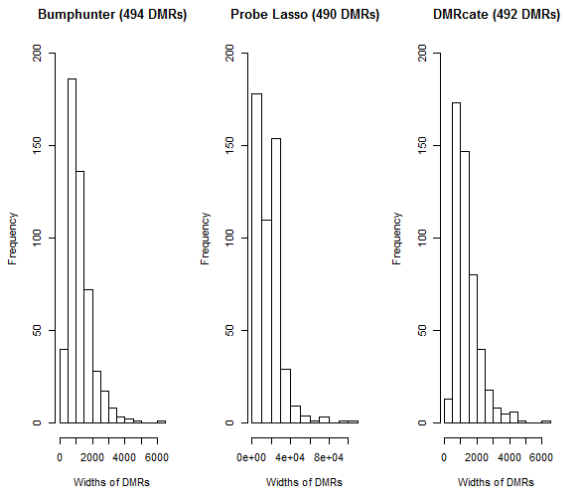

Figure 3.10. Width of the DMRs found with DC, BMP-DC, and PL-DC

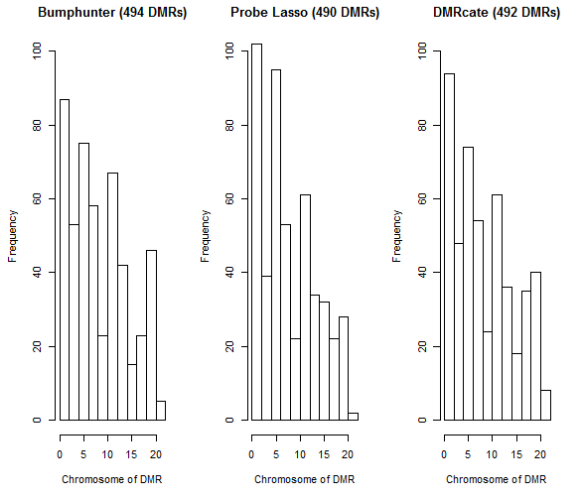

Figure 3.11. Chromosomes of the DMRs found with DC, BMP-DC, and PL-DC

Table 3.20. Percentage overlaps between DC, BMP-DC, and PL-DC

Comparison	Number of Overlaps	Mean of Percentage of Overlap	Standard Deviation of Percentage Overlap
BMP-DC vs PL-DC	241	0.279	0.320
PL-DC vs DC	256	0.284	0.314
BMP-DC vs DC	355	0.892	0.171

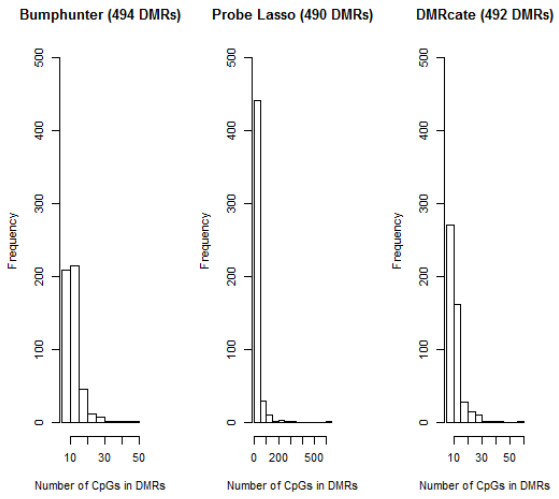

Figure 3.12. Number of CpGs in the DMRs found with DC, BMP-DC, and PL-DC

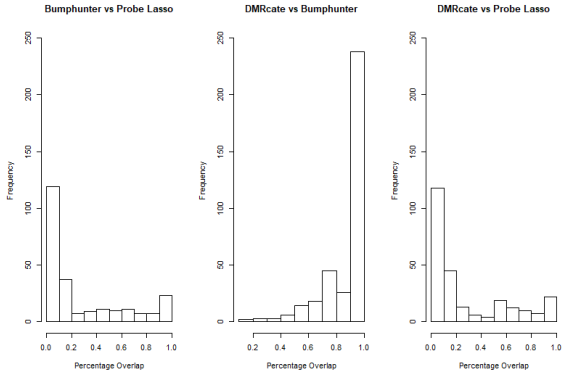

Figure 3.13. Percentage of overlap between DC, BMP-DC, and PL-DC

4. CONCLUSIONS

4.1. SUMMARY

DNA methylation data from a 2013 cervical cancer study was analyzed with the use of the R package,"ChAMP," to find potentially significant DMRs within the genome. With the use of three methods for identifying DMRs ("Bumphunter," "Probe Lasso," and "DMRcate"), DMRs were found when comparing a group of eight HIV positive women with Squamous Cell Carcinoma (SCC) cervical cancer to a group of five HIV positive women with no cervical cancer. The overall conclusion that can be drawn from the final results is that the three methods produced different results. In addition to the differences in DMRs, the outputs that each method produces was also different. While all three reported the Start, End, and Chromosome on which each DMR was located, DMRcate also reported both the gene that each DMR was a part of and how many CpG sites each DMR contained. Probe Lasso reported the genes associated with the DMRs, but did not give the number of CpG sites. Bumphunter reported neither of these pieces of information. Both the number of CpGs and gene association are useful to know when examining what type of DMRs are being reported and possible relationships the DMRs might have with the expression of the genes. If either of these pieces of information are missing then it can be done manually instead of through a known function or package. This method is prone to being both time intensive and mistake prone. Improvements to both Bumphunter's and Probe Lasso's packages are desirable in the future to counter these limitations.

When examining the analyzed results, throughout all nine analyses conducted involving Bumphunter, Probe Lasso, and DMRcate, both DMRcate and Bumphunter produce somewhat similar results as seen with the amount of percentage overlaps between DMRs identified with the two methods. They also showed similar means and standard deviations

with the DMRs' widths, Chromosome location, and number of CpGs. The only radical difference is the number of DMRs found in both methods with default parameters. The 492 DMRs found with DMRcate was more than 1.5 times larger than the 277 DMRs found with Bumphunter. The default parameters for all three methods produce vastly different number of identified DMRs with Probe Lasso finding only one DMR. To get a comparable number of DMRs among the three methods, the parameters of each method had to be changed. When Bumphunter and DMRcate are able to find a similar number of DMRs, the number of DMRs in common to both methods and percentage overlap of those DMRs are relatively high. When comparing all three methods, Probe Lasso was usually an outlier. The widths and number of CpG sites for each DMR was significantly larger for Probe Lasso than either of the other two tests. There could be many contributing factors to this outcome. One cause could be the drastic increase in the significance threshold for probes to be considered part of the DMRs created with Probe Lasso. By increasing this value, the method was least stringent and accepted probes into DMRs that would normally not be accepted with either Bumphunter or DMRcate. Another factor in the inability of Probe Lasso to identify multiple DMRs under the default setting, could be the type of microarray used. Since Probe Lasso is based on meeting a minimum number of significant CpG sites to be considered part of a DMR, then increasing the total number of probes could potentially allow more DMRs to be identified. This would mean that the use of EPIC might encourage a better identification of DMRs than 450K with Probe Lasso.

When looking at the results, an important piece of information that needs to be considered with DMRs is the gene with which the DMR is associated. If the DMR is part of a gene or its transcription factor start site, this could affect how the gene will be expressed and the outcome produced when this happens. Thus, both Probe Lasso and DMRcate are advantageous in that any genes associated with each DMR are reported in their results. Bumphunter does not report this information and, if desired, this information must be obtained manually, which was completed for this thesis. Genes for the top five

best DMRs for all nine analyses were found. Bumphunter's DMRs are ranked based on the pvaluearea, Probe Lasso are ranked with dmrP, and DMRcate does not report the p-values used for obtaining the DMRS in the results for the researcher. However, DMRcate does rank the DMRs from the lowest p-value to highest and provides the ranking for the researcher.

Table 4.1 provides the gene associations for any DMR found in the top five in any of the nine analyses. The table contains columns that include the gene name, which method found this gene, and any possible association between these genes to diseases and body functions. The associations was obtained through the Database of RefSeqGene [30]. This database is a collection from various researchers that have found connections between genes and diseases. With this database specific genes can be looked up and the diseases can be found. For example the SALL1 gene had associated DMRs with the methods of BMP-DC, DC, DC-BMP, and PL-DC. From RefSeqGene, SALL1 is linked with protein coding and has known connections to Townes-Brocks syndrome and Bronchio-oto-renal syndrome whenever the gene contains defects [30]. With this information, a researcher will have a better idea of how the DMRs located in genes might be linked to diseases.

4.2. FUTURE/DISCUSSION

Even though comparisons between the three methods could be made from the results, the answer of which method with what parameters would most effective at correctly identifying the DMRs cannot be determined since it is unknown where the true DMRs are located. Thus, the amount of Type 1 and Type 2 errors for the nine different analyses could not be calculated and there was no way to find these pieces of information without knowing the true DMRs. To be able to gather this information, a simulation study would be necessary. There are no existing public codes or packages that would let a user to do this. Thus, a new and original code would have to be made. So in the future, to help answer what method would be best, a code to create a simulation study that examined different possible

scenarios would have to be made. Then, different parameter settings could be more fully investigated to learn which methods and parameters perform well in different scenarios. This could help inform researchers on the most suitable method to use for their data.

Another issue that was found was the drastically different results that Probe Lasso found when compared to either of the two other methods. For example, the default parameters for this method were only able to find one DMR from the data. A possible solution to this might be the type of microarray that is used to gather the data. If EPIC was used then there would be nearly double the number of probes on the array. Since Probe Lasso is significantly dependent on identifying clusters and high concentrations of probes that are showing differences in methylation, an increase to probes used in testing samples might improve in the ability to find DMRs. If possible, a comparison could be made on the DMRs found with Probe Lasso when the samples were taken with 450K and EPIC microarrays. Also, as discussed earlier, when the parameters for Probe Lasso were changed to identify more DMRs, the DMRs found were outliers when compared to either Bumphunter or DMRcate. One possible avenue to correct this with Probe Lasso would be to explore how changing other parameters might effect the DMRs found. Changing other parameters might produce more comparable DMRs with Probe Lasso to either Bumphunter or DMRcate.

Another important issue was the difference in the number of DMRs found with Bumphunter and DMRcate when both tests used their default parameters. Even though there were considerable overlaps between the DMRs from Bumphunter and DMRcate, DMRcate still identified more DMRs. Further investigation needs to be done to determine if this difference exists with larger number of samples. If so, then knowing a possible way to determine what parameters to change in order for a researcher to be able to get similar results from both tests may be informative. Overall, further investigation is needed to better understand the difference in performance of the methods under different parameter settings.

Table 4.1. Gene and known associations for top five DMRs identified in any of the nine analyses.

Gene Name	Methods	Association
EDNRB	BMP, BMP-DC, DC-BMP	Hirschsprung disease type 2
OR2I1P	BMP, BMP-DC, DC, DC-PL, DC-BMP, PL	No known disease, Olfactory receptors
EYA4	BMP, BMP-PL, BMP-DC, PL-BMP	dilated cardiomyopathy 1J
ELMO1	BMP	glioma cell invasion, diabetic nephropathy
SALL1	BMP-DC, DC, DC-BMP, PL-DC	Townes-Brocks syndrome, bronchio-oto-renal syndrome
ADCYAP1	BMP-DC, DC	multiple mature peptides
GFRA1	DC, DC-BMP	Hirschsprung disease
PAX6	DC, DC-BMP, PL-DC	aniridia, Peter's anomaly
RP11	DC, PL-DC, PL-BMP	No information through database
ZIC4	PL-DC	Dandy-Walker malformation, X-linked visceral heterotaxy, holoprosencephaly type 5
ZIC1	PL-DC	medulloblastoma
TRIM40	PL-DC, PL-BMP	regulates inflammation and carcinogenesis in the gastrointestinal tract
SNORD116	PL-DC, PL-BMP	small nucleolar RNA, C/D box
RIC3	PL-BMP	encodes proteins to be resistant to inhibitors
TBX18	PL-BMP	Effects embryonic development

However, by comparing the methods using the approach in this thesis, many conclusions about the similarity and differences between the methods when applied to a real data set could be made.

APPENDIX

Table A1. Top 5 DMRs determined by p.valArea for BMP

	seqnames	start	end	width	p.valueArea
DMR_3	chr13	78492568	78494067	1499	5.01E-05
DMR_1	chr6	29521013	29521803	790	5.01E-05
DMR_4	chr6	133561756	133562774	1018	0.000267032
DMR_10	chr6	28602543	28603437	894	0.000500684
DMR_2	chr7	37488162	37488936	774	0.000550753

Table A2. Top result for BMP-PL

	seqnames	start	end	width	p.valueArea
DMR_1	chr6	133562461	133562494	33	0.041666667

Table A3. Top 5 results determined by p.valArea for BMP-DC

	seqnames	start	end	width	p.valueArea
DMR_14	chr16	51183988	51190201	6213	4.31E-05
DMR_1	chr6	29520698	29521803	1105	4.31E-05
DMR_8	chr13	78492568	78494064	1496	5.17E-05
DMR_27	chr6	133561614	133564578	2964	5.17E-05
DMR_2	chr18	904523	909154	4631	0.000112123

Table A4. Top 5 results determined by dmrP for PL-BMP

	seqnames	start	end	width	dmrP	dmrpRank	ensemblID	geneSymbol
DMR_251	chr6	85471060	85485442	14383	3.02E-55	1	ENSG00000112837	TBX18
DMR_210	chr6	30094504	30096205	1702	1.83E-44	2	ENSG00000204614	TRIM40
DMR_252	chr6	133561581	133562614	1034	1.00E-36	3	ENSG00000112319	EYA4
DMR_104	chr15	25302950	25335070	32121	3.48E-29	4	ENSG00000207014; ENSG00000207464; ENSG00000207191; ENSG00000207442; ENSG00000207133; ENSG00000207093; ENSG00000206727; ENSG00000200661; ENSG00000206609; ENSG00000206193; ENSG00000206621; ENSG00000207174; ENSG00000207263; ENSG00000206656; ENSG00000206688; ENSG00000207460; ENSG00000207236; ENSG00000207159	SNORD116-3; NA;SNORD116-5; SNORD116-6; SNORD116-7; SNORD116-8; SNORD116-9; SNORD116-10; SNORD116-11; SNORD116-14; NA;SNORD116-15; SNORD116-16; SNORD116-17; SNORD116-18; SNORD116-19
DMR_56	chr11	8190225	8195246	5022	3.00E-24	5	ENSG00000166405; ENSG00000246820	RIC3; RP11-379P15.1

Table A5. Top result determined with dmrP for PL

	seqnames	start	end	width	dmrP	dmrpRank	ensemblID	geneSymbol
DMR_1	chr6	29521136	29521789	654	3.32E-138	1	ENSG00000237988	OR2I1P

Table A6. Top 5 results determined by dmrP for PL-DC

	seqnames	start	end	width	dmrP	dmrpRank	ensemblID	geneSymbol
DMR_214	chr16	51184983	51211725	26743	2.62E-52	1	ENSG00000103449	SALL1
DMR_310	chr3	147075348	147128550	53203	1.82E-44	2	ENSG00000243620; ENSG00000174963; ENSG00000241202; ENSG00000152977	RP11-649A16.1; ZIC4;ZIC4-AS1; ZIC1
DMR_392	chr6	30094265	30096485	2221	2.09E-44	3	ENSG00000204614	TRIM40
DMR_190	chr15	25289307	25344324	55018	1.82E-37	4	ENSG00000207063; ENSG00000207001; ENSG00000207014; ENSG00000207464; ENSG00000207191; ENSG00000207442; ENSG00000207133; ENSG00000207093; ENSG00000206727; ENSG00000200661; ENSG00000206609; ENSG00000206193; ENSG00000206621; ENSG00000207174; ENSG00000207263; ENSG00000206656; ENSG00000206688; ENSG00000207460; ENSG00000207236; ENSG00000207159; ENSG00000207375; ENSG00000207279	SNORD116-1; SNORD116-2; SNORD116-3; NA;SNORD116-5; SNORD116-6; SNORD116-7; SNORD116-8; SNORD116-9; SNORD116-10; SNORD116-11; SNORD116-14; NA;SNORD116-15; SNORD116-16; SNORD116-17; SNORD116-18; SNORD116-19; SNORD116-23; SNORD116-24
DMR_110	chr11	31825004	31828505	3502	1.64E-35	5	ENSG00000007372	PAX6

Table A7. Top 5 results ranked for DC-BMP

	seqnames	start	end	width	no.cpgs	Stouffer	overlapping.promoters
DMR_1	chr6	29520527	29521803	1277	30	4.30E-38	OR2I1P-201; OR2I1P-001
DMR_2	chr16	51183988	51188129	4142	29	1.73E-26	SALL1-001; SALL1-201; SALL1-202; SALL1-005; SALL1-002; AC009166.5-001; SALL1-003
DMR_3	chr10	118030848	118034031	3184	22	1.02E-20	GFRA1-002; GFRA1-003; GFRA1-001; GFRA1-201; GFRA1-004
DMR_4	chr11	31824973	31828040	3068	19	2.98E-16	PAX6-016; PAX6-006; PAX6-014; PAX6-015; PAX6-007; PAX6-025; PAX6-029; PAX6-024; PAX6-026
DMR_5	chr13	78492216	78494462	2247	42	9.85E-16	EDNRB-003; EDNRB-001; EDNRB-201; EDNRB-002; RNF219-AS1-013

Table A8. Top result ranked for DC-PL

	seqnames	start	end	width	no.cpgs	Stouffer	overlapping.promoters
DMR_1	chr6	29521013	29521803	791	28	3.74E-39	OR2I1P-201; OR2I1P-001

Table A9. Top 5 results ranked for DC

	seqnames	start	end	width	no.cpgs	Stouffer	overlapping.promoters
DMR_1	chr6	29520527	29521803	1277	30	2.48E-38	OR2I1P-201; OR2I1P-001
DMR_2	chr16	51183988	51190201	6214	38	7.58E-29	SALL1-001; SALL1-201; SALL1-202; SALL1-005; SALL1-002; AC009166.5-001; SALL1-003
DMR_3	chr10	118030848	118034357	3510	23	9.52E-20	GFRA1-002; GFRA1-003; GFRA1-001; GFRA1-201; GFRA1-004
DMR_4	chr18	904523	909154	4632	26	5.48E-19	ADCYAP1-005; ADCYAP1-002; RP11-672L10.2-004; ADCYAP1-003; RP11-672L10.2-003; ADCYAP1-004; RP11-672L10.2-002; RP11-672L10.3-001; RP11-672L10.2-001
DMR_5	chr11	31824327	31828715	4389	21	9.90E-17	PAX6-016; PAX6-006; PAX6-014; PAX6-015; PAX6-007; PAX6-025; PAX6-027; PAX6-029; PAX6-024; PAX6-026

REFERENCES

[1] Anthony JF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.

[2] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation of the illumina methylationepic beadchip microarray for whole-genome dna methylation profiling. *Genome biology*, 17(1):208, 2016.

[3] Andrew E Jaffe, Peter Murakami, Hwajin Lee, Jeffrey T Leek, M Daniele Fallin, Andrew P Feinberg, and Rafael A Irizarry. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, 41(1):200–209, 2012.

[4] DJ Weisenberger, D Van Den Berg, F Pan, BP Berman, and PW Laird. Comprehensive dna methylation analysis on the illumina infinium assay platform. *Illumina, San Diego*, 2008.

[5] Martin J Aryee, Andrew E Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P Feinberg, Kasper D Hansen, and Rafael A Irizarry. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*, 30(10):1363–1369, 2014.

[6] Tiffany J Morris and Stephan Beck. Analysis pipelines and packages for infinium humanmethylation450 beadchip (450k) data. *Methods*, 72:3–8, 2015.

[7] illumina. Humanmethylation450 beadchip achieves breadth of coverage using two infinium chemistries. Nucleic Acids Res, 2012. [Online]. Available: http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote hm450 data analysis optimization.pdf. [Accessed October 2017].

[8] Sarah Dedeurwaerder, Matthieu Defrance, Emilie Calonne, Hélene Denis, Christos Sotiriou, and François Fuks. Evaluation of the infinium methylation 450k technology. *Epigenomics*, 3(6):771–784, 2011.

[9] Dongmei Li, Zidian Xie, Marc Le Pape, and Timothy Dye. An evaluation of statistical methods for dna methylation microarray data analysis. *BMC bioinformatics*, 16(1): 217, 2015.

[10] Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

[11] Juan Sandoval, Holger Heyn, Sebastian Moran, Jordi Serra-Musach, Miguel A Pujana, Marina Bibikova, and Manel Esteller. Validation of a dna methylation microarray for 450,000 cpg sites in the human genome. *Epigenetics*, 6(6):692–702, 2011.

[12] Wanding Zhou, Peter W Laird, and Hui Shen. Comprehensive characterization, annotation and innovative use of infinium dna methylation beadchip probes. *Nucleic acids research*, 45(4):e22–e22, 2017.

[13] Jessica Nordlund, Christofer L Bäcklin, Per Wahlberg, Stephan Busche, Eva C Berglund, Maija-Leena Eloranta, Trond Flaegstad, Erik Forestier, Britt-Marie Frost, Arja Harila-Saari, et al. Genome-wide signatures of differential dna methylation in pediatric acute lymphoblastic leukemia. *Genome biology*, 14(9):r105, 2013.

[14] Andrew E Teschendorff, Francesco Marabita, Matthias Lechner, Thomas Bartlett, Jesper Tegner, David Gomez-Cabrero, and Stephan Beck. A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450 k dna methylation data. *Bioinformatics*, 29(2):189–196, 2012.

[15] Michael H Kutner, Chris Nachtsheim, and John Neter. *Applied linear regression models*. McGraw-Hill/Irwin, 2004.

[16] Timothy J Peters, Michael J Buckley, Aaron L Statham, Ruth Pidsley, Katherine Samaras, Reginald V Lord, Susan J Clark, and Peter L Molloy. De novo identification of differentially methylated regions in the human genome. *Epigenetics & chromatin*, 8(1):6, 2015.

[17] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[18] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

[19] Lee M Butcher and Stephan Beck. Probe lasso: a novel method to rope in differentially methylated regions with 450k dna methylation data. *Methods*, 72:21–28, 2015.

[20] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1. 1949.

[21] Franklin E Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

[22] Yuan Tian, Tiffany Morris, Lee Stirling, Andrew Feber, Andrew Teschendorff, Ankur Chakravarthy, Maintainer Yuan Tian, and Block GUI. Package âĂŸchampâĂŹ. 2017.

[23] Jean-Philippe Fortin, Timothy J Triche Jr, and Kasper D Hansen. Preprocessing, normalization and integration of the illumina humanmethylationepic array with minfi. *Bioinformatics*, 33(4):558–560, 2016.

[24] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. Swan: Subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome biology*, 13(6):R44, 2012.

[25] James M Wettenhall and Gordon K Smyth. limmagui: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18):3705–3706, 2004.

[26] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25, 2004.

[27] Rafael A Irizarry, Martin Aryee, Hector Corrada Bravo, Kasper D Hansen, Harris A Jaffee, Maintainer Rafael A Irizarry, Suggests RUnit, Epigenetics biocViews DNAMethylation, and MultipleComparisons Infrastructure. Package âĂŸbumphunterâĂŹ. 2013.

[28] L Butcher and T Morris. Illumina450probevariants. db: annotation package combining variant data from 1000 genomes project for illumina humanmethylation450 bead chip probes. *R package version*, 1(0), 2013.

[29] Michael Lawrence, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[30] Brister JR Ciufo S Haddad D McVeigh R Rajput B Robbertse B Smith-White B Ako-Adjei D Astashyn A Badretdin A Bao Y Blinkova O Brover V Chetvernin V Choi J Cox E Ermolaeva O Farrell CM Goldfarb T Gupta T Haft D Hatcher E Hlavina W Joardar VS Kodali VK Li W Maglott D Masterson P McGarvey KM Murphy MR O'Neill K Pujar S Rangwala SH Rausch D Riddick LD Schoch C Shkeda A Storz SS Sun H Thibaud-Nissen F Tolstoy I Tully RE Vatsan AR Wallin C Webb D Wu W Landrum MJ Kimchi A Tatusova T DiCuccio M Kitts P Murphy TD Pruitt KD O'Leary NA, Wright MW. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res, January 2016. 44(D1):D733-45.

VITA

In December 2012, Arnold Harder graduated from University of Alaska Anchorage with a B.S. in Mathematics and a minor in Statistics. He received a Master of Science in Applied Mathematics with a Statistics Emphasis from Missouri University of Science and Technology in December 2017.