Saarland University
Center for Bioinformatics
Max Planck Institute for Informatics
Master's Program in Bioinformatics

Master's Thesis

Identification and Analysis of Methylation Call Differences between Bisulfite Microarray and Bisulfite Sequencing Data with Statistical Learning Techniques

submitted by

Matthias Döring

on June 12, 2014

Supervisor

Dr. Nico Pfeifer

Reviewers

Dr. Nico Pfeifer
Prof. Dr. Thomas Lengauer, Ph.D.

Statement

I hereby confirm that this thesis is my own work and that I have documented all sources used.
I declare that the submitted digital and printed versions of this thesis correspond to each other. I permit Saarland University to duplicate and publish this thesis.

Saarbrücken, June 12, 2014

Matthias Döring

Abstract

Methylation of DNA is an epigenetic modification known to play a prime role in gene silencing and is an important topic in epigenetic research. Although accurate methods for measuring DNA methylation exist, technology-dependent errors give rise to inconsistencies between method results. We studied DNA methylation measurements from the Infinium HumanMethylation450 microarray ($\beta_{450\text{K}}$) and whole genome bisulfite sequencing (β_{WGBS}) to evaluate whether there are locus-specific measurement differences, $\Delta\beta = \beta_{450K} - \beta_{\text{WGBS}}$, and whether this effect is predictable using statistical learning techniques. The ability to determine inconsistent methylation measurements based on bisulfite-converted sequencing data alone would be valuable for epigenome-wide association studies, in which such positions could be excluded.

We built support vector regression models based on Illumina bisulfite-sequencing data of HepaRGd7R2 to predict $\Delta\beta$. A measure for read similarity was obtained via numerical and string kernels as well as set kernels. We introduced the notion of hybrid string kernels to afford a similarity measure for both, numeric and string input simultaneously. Feature importance was analyzed using kernel-target alignment and positional oligomer importance matrices.

Using a read-based set kernel, we found that the predicted values of $\Delta\beta$ correlated significantly with the observed outcomes ($r \cong 0.73$, p-value $< 2.2 \cdot 10^{-16}$). This model made use of CpG positions, which implicitly code for the sample's methylation. To obtain a model independent of β_{WGBS}, we excluded the CpG positions and still found a significant correlation ($r \cong 0.37$, p-value $< 2.2 \cdot 10^{-16}$). Features beside the reads played only a minuscule role in the emergence of inconsistent methylation measurements. To our knowledge, this is the first time someone was able to show that differences between β_{450K} and β_{WGBS} are predictable from the sequence, hinting at the over- or underestimation of methylation status for specific loci using either technique.

Acknowledgments

First of all, I would like to thank Prof. Dr. Thomas Lengauer, Ph.D. for giving me the opportunity of writing my Master's thesis in his group at the Max Planck Institute for Informatics. My deep thanks go to Prof. Dr. Jörn Walter for providing the high-quality biological data used in this project. Considering supervision, I am greatly indebted to Dr. Nico Pfeifer for his expertise and kindness. In addition, I would like to give my sincere thanks to Dr. Karl Nordström and Pavlo Lutsik from Jörn Walter's lab for their helpful advice and enduring interest in the project.

I am obliged to Dr. Glenn Lawyer for his advice on my thesis and for the great conversations over a cup of coffee. My appreciation goes to Fabian Müller and Peter Ebert for their aid in computational epigenetic matters. For providing technical support and resources, as well as server maintenance, I want to thank Joachim Büch and Georg Friedrich. My gratitude goes to my fellow students Nora Speicher, Anna Feldmann, and Valentina Galata for the lectorate of my thesis and all others involved, in particular Adrin Jalali and Sarvesh Nikumbh. Finally, I would like to thank my parents for their continuing support and everything they have done for me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Role of Epigenetics

Epigenetics is defined as *[t]he study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence* [66]. Focusing first on the impact of gene function modulation through epigenetics, consider the different cell types present in our body. All somatic cells in an organism share the same deoxyribonucleic acid (DNA). However, cells are highly differentiated and required to fulfill specific functions. While all cells share the same genotype, their phenotypes differ drastically from each other.

An example of contrasting phenotypes in cells is myocytes and neurons. Myocytes are responsible for muscular contraction. They are characterized by myofibrils, which are long protein cords composed of the myofilaments actin, myosin, and titin. Muscle shortening is caused by a pulling action initiated by myosin that leads to a compression of the actin filaments. Neurons, in contrast, are responsible for processing and transporting information from one cell to another via electrical and chemical signals. Signals are transducted through axons. The axon of one neuron attaches to another neuron at its dendrites where the incoming signal is propagated in terms of electrochemical signals that are transmitted across synapses, the junctions between two neurons.

Looking at the functional aspects of the genome, it is not enough to only consider the genome by itself. One needs to study the epigenome, that is, the complete set of (heritable) chemical changes to DNA and histones, in order to get a better understanding of active genes. Therefore, epigenetics has recently found its way into areas of research that were previously solely focused on genetics. Cancer research is a prime example. Before the emergence of epigenetics, the main cause of cancer was thought to be the aggregation of damage to DNA, e.g., through mutations or chromosomal abnormalities. With the advent of epigenetics, however, certain patterns of

epigenetic modifications were identified indicating cancerous cells and specific cancer types [5]. This gives a possible molecular explanation for the meaningful environmental component in the emergence of cancer [see 46].

Epigenetic changes can be meiotically inherited. In principle, epigenetic information can be passed from generation to generation in the same fashion as genetic information. What distinguishes the epigenome from the genome is its greater susceptibility to environmental influence. This attribute allows short-term adaptions to the environment in accordance with the inheritance of acquired characteristics proposed by Lamarck [65]. With this, epigenetics revolutionized evolutionary theory, which was previously based mainly on Darwin's theory. According to Darwin, evolution occurs on a time-scale of thousands of years and is only based on random changes in genetic makeup leading to an increase in fitness. Albeit environmental influence plays a key role in epigenetics, the epigenome is probably even more influenced by stochastic epigenetic events, e.g., during mitosis when epigenetic marks are transmitted from mother to daughter chromatids [59].

Multiple studies demonstrated the heritable character of epigenetic modifications and their relation to phenotype and environment, e.g., reduction in spermatogenic capacity and decreased rate of fertility in rats after environmental insult [2] or the effect of paternal grandfathers' nutrition on mortality rates in grandsons [58]. A groundbreaking study conducted by Weaver et al. [82] was able to show a relation between maternal care, epigenetic reprogramming, and stress response in rat pups. Their research exposed that epigenetic states can be established through behavioral programming and that it may be possible to reverse the changes associated with these states. In a recent study, transgenerational epigenetic effects were observed in mice [17]. The parental generation was subjected to an olfactory experience that was associated with a fear-inducing event. Subsequent generations without conditioning still exhibited behavioral sensitivity to the odor that was used to condition the parental generation and exhibited hypomethylation in the corresponding olfactory gene. With this finding they revealed the heritable nature of stress-induced behavior.

The measurement of epigenetic modifications, such as methylation of DNA, is necessary to study epigenetic effects. Although, state-of-the-art technologies allow cheaper, faster, and more accurate measurements than previously, technology-associated biases and error profiles exist. It is therefore crucial to perform low-level processing of data to correct for these errors. Nevertheless, it is not possible to attain fully accurate measurements of methylation. In the following, we present the goal of this thesis, which is concerned with handling such errors.

1.2 Purpose and Problem Statement

The purpose of this thesis is to work towards assigning confidence values to individual cytosine-phosphate-guanine (CpG) dinucleotides, in order to represent the accuracy of their methylation measurement. Currently, there is no available tool that is able to indicate whether one can trust a measured methylation value. However, for bisulfite-sequencing of DNA there are many sources of errors. Among them are platform-dependent errors (e.g., lagging strands in Illumina technology), technology-based errors (e.g., DNA polymerase errors), errors during bisulfite conversion, but also errors during post-processing (e.g., mapping and calling methylation). Although, many approaches tackling these problems exist, it is impossible to procure completely error-free measurements. Therefore, our working hypothesis is that there has to exist some signal in the data allowing for the determination of methylation call fidelity. The ability to provide a measure of confidence for methylation calls would be a great asset for association studies, as this would allow researchers to weight individual positions appropriately.

For our approach, we use methylation measurements of a single hepatic cell sample provided by two technologies, whole-genome bisulfite sequencing (WGBS) on the Illumina platform and the Infinium HumanMethylation450 BeadChip. In a supervised learning scenario, we utilize WGBS data as input features. The outcomes are constructed by forming the differences in beta-values between the two methods for each measured CpG. Our goal is the prediction of differences in methylation between the two methods using only WGBS data in order to identify positions with inconsistent measurements. Positions with high absolute methylation differences indicate that at least one of the methods was inaccurate. The usage of Infinium microarray data as a gold-standard is motivated by its good agreement with WGBS data (greater than $r = 0.94$). An important component of this work is the identification and interpretation of those features indicative of methylation call differences and therefore represent suitable features for predictive models.

1.3 Related Work

To our knowledge, there currently exists no tool that determines confidence values for CpG positions with regard to methylation calls. However, there are approaches dealing with the identification of sequence motifs in next-generation sequencing reads indicative of base miscalls. These motifs could also influence the accuracy of methylation measurements, which is why they are of concern. In the following, we delineate an approach by Allhoff et al. [1] that was concerned with finding error-inducing sequence patterns.

1.3.1 Discovering Motifs that Induce Sequencing Errors

Allhoff et al. [1] were concerned with the discovery of motifs associated with errors in next-generation sequencing of DNA in order to increase fidelity of single nucleotide polymorphism (SNP) calls. They used a statistical approach that relies on knowledge about how context-sensitive errors (CSEs) emerge. Since a CSE is caused only by sequence motifs preceding but not following it, one can perform a strand-specific analysis. When a position does not agree with the reference only in reads of one directionality, that is, either + or -, then these mismatches are subject to strand bias and one can infer the existence of an error-causing motif.

Fisher's exact test provides a robust statistical framework to assign a p-value to the hypothesis that read direction and number of mismatches at the current position are independent of each other. To perform the test, the authors constructed a 2×2 contingency table, where rows represent read direction and columns indicate matches and mismatches (Table 1.1). With Fisher's test one can identify whether the distributions of the rows agree with each other (no strand bias) or not (strand bias).

Table 1.1: Contingency table for Fisher's exact test. Shown are the number of strand-specific position matches and mismatches. Context-sensitive errors (CSEs) are associated with a strand bias with regard to mismatches, i.e., for a CSE, one would expect the row distributions to differ from each other.
Source: Allhoff et al. [1]

Strand	**Match**	**Mismatch**	**Total**
Forward (+)	a	b	f
Backward (-)	c	d	k
Total	m	s	n

Given a contingency table $\boldsymbol{a}$, its probability given the null hypothesis H_0 (read direction and number of mismatches are independent) can be computed as

$$\Pr_{H_0}(\boldsymbol{a}) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}}$$

The p-value of $\boldsymbol{a}$ is given by

$$p\text{-value}(\boldsymbol{a}) = \sum_{\boldsymbol{a}'} \Pr_{H_0}(\boldsymbol{a}').$$

It determines the probability of finding a table at least as extreme as the current one. This is done by summing up the probabilities of tables $\boldsymbol{a}'$ for which $\Pr_{H_0}(\boldsymbol{a}') \leq \Pr_{H_0}(\boldsymbol{a})$ holds. For a sufficiently low p-value, the null

hypothesis is rejected, i.e., it is likely that the observed position is subject to strand bias.

To efficiently compute the contingency tables, Allhoff et al. [1] considered only positions following a set of certain motifs by setting a motif length q and a maximum number of allowed wildcard characters n. CSE-associated motifs were detected by adding all q-gram contingency tables matching that motif. Then, using Fisher's exact test, they computed p-values for each motif and significance was determined with Bonferroni thresholding to deal with multiple hypothesis testing. This approach was able to confirm previously discovered motifs, but also identified other error-indicative motifs for several data sets and sequencing technologies.

There are several points that demarcate the presented approach from the work done in this thesis. First, while Allhoff et al. [1] were concerned with DNA, we deal with bisulfite-converted DNA. Second, our main goal is not the explicit identification of sequence motifs, but rather the identification and usage of features, e.g., sequence reads, to form a predictive model that is able to identify positions with accurate methylation measurements.

1.4 Thesis Structure

Having talked about the implications of epigenetics on gene function, inheritance, and the corresponding research in this chapter, we will focus on the molecular mechanisms at the core of epigenetics and introduce methods for measuring DNA methylation in Chapter 2. After that, Chapter 3 provides information about the materials used in this thesis, that is, data from Illumina WGBS and the Infinium HumanMethylation450 BeadChip. Before we deal with the particulars of how we tackled the presented problem in Chapter 5, we give an overview of the methodological background in Chapter 4, where we, among others, cover data processing, the supervised learning scenario, and support vector machines (SVMs), with a focus on kernel functions. Then, we present our findings in Chapter 6 and discuss them in Chapter 7. Finally, in Chapter 8, we draw conclusions about the project and present an outlook regarding future work in the field.

Chapter 2

Biological Background

2.1 Epigenetic Modifications

An epigenetic modification is any chemical alteration of DNA or its structural organization that leads to mitotically and/or meiotically heritable changes in gene function. The two major epigenetic mechanisms are DNA methylation and histone modification, i.e., changes in the structural units that DNA is associated with. To understand the modifications of DNA and histones, we will first give a short overview of chromosomal structure. After that, we will consider the molecular basis for DNA methylation and its impact, in particular in CpG context. Finally, we discuss the different types of histone alterations, as well as their ramifications. Since this thesis deals solely with DNA methylation data, we will just present histone modifications to receive a better understanding of the complex framework of epigenetic changes.

2.1.1 Chromosomal Structure

The DNA of an organism is organized as chromosomes located in the cell nucleus, see Fig. 2.1. Chromosomes are structural units composed of three elements: DNA, protein, and ribonucleic acid (RNA). Of these elements only DNA is responsible for storing the organism's genetic information. Protein and RNA serve other purposes. As an analogy, let us consider a warehouse. While chromosomes represent individual shelves in which wares (genes) are stored, RNA and proteins are the workers in that warehouse. RNA has a managing role as it decides which wares are shipped out and proteins work in the packaging department. In biological terms, DNA-associated RNA's primary role is a regulatory one: It is responsible for increasing or decreasing the expression of specific genes. The histone proteins that are bound to DNA are crucial for DNA condensation, that is, compaction of DNA. Condensation of DNA results in chromatin, whose level of condensation depends on the current cell cycle state defining the level of transcriptional activity (Fig. 2.2). Stages in the cell cycle exhibiting high levels of transcriptional ac-

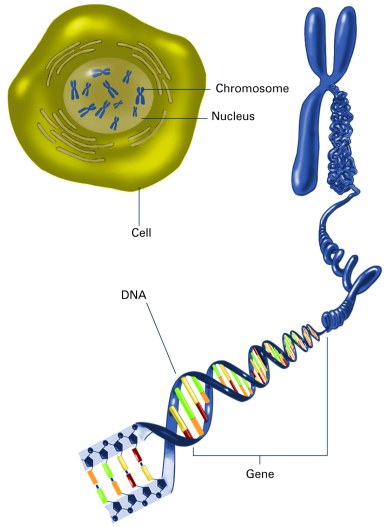


Figure 2.1: Organization of DNA in the form of chromosomes.
Source: National Institute of General Medical Sciences [56]

tivity, such as the interphase, are characterized by less condensed chromatin; stages with little transcriptional activity exhibit more compact chromatin, e.g., in the metaphase during mitosis. In the following we are only concerned with the structure of potentially active chromosomes, for which two types of chromatin structures are relevant, namely the beads-on-a-string structure and the 30 nm fiber (Fig. 2.2). The loosely-packed beads-on-a-string structure is considered *euchromatin*, which is associated with active transcription; the densely-packed 30 nm fiber is associated with gene silencing and termed *heterochromatin*. Both structures are shaped by the presence of histones, specific proteins that enable DNA to wrap itself around them. These units of histones and DNA form the *beads* in the beads-on-a-string structure and are called nucleosomes. Further wrapping of histones leads to tight arrays of nucleosomes giving rise to the 30 nm fiber. Its structure is based on solenoids, helical windings of at least five nucleotides. Since nucleosomes are the basic effectors of chromatin states, we will now discuss their structure in more detail.

Nucleosome Structure

Nucleosomes consist of a segment of about 147 base pairs of DNA wrapped around a histone octamer. This octamer is comprised of two copies of each of the four core histones H2A, H2B, H3, and H4. Each histone possesses an N-terminal tail, an unstructured sequence of amino acids that protrudes from

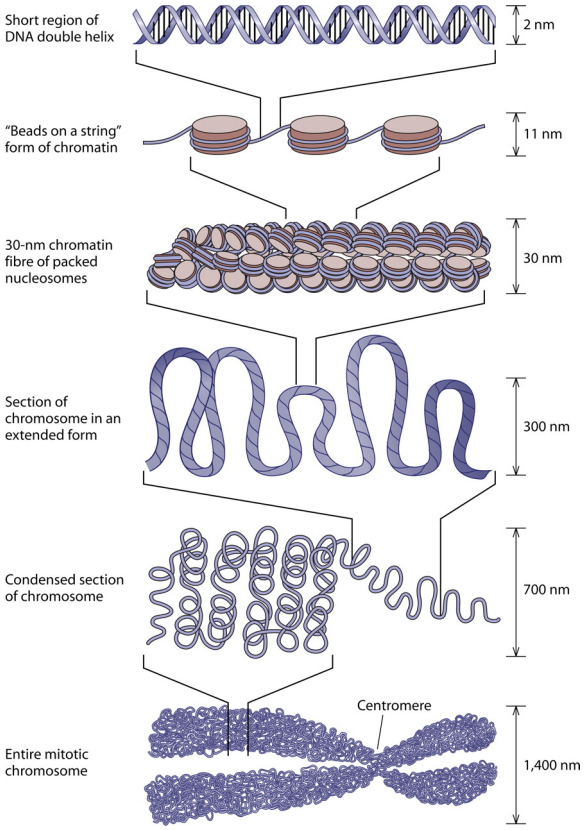


Figure 2.2: Chromatin structure. The beads-on-a-string form of chromatin is associated with active transcription (euchromatin), while the more condensed form, the 30 nm fiber, is associated with transcriptional repression (heterochromatin).
Source: Jansen and Verstrepen [32]

the core globular histone core. There are more than 120 direct interactions between protein and DNA and several hundred water-mediated ones, making nucleosomes tight complexes.

2.1.2 DNA Methylation

Methylation of DNA can occur either at cytosine or adenine bases and is associated with gene repression. It plays a crucial role in: cell differentiation from stem cells to specialized cells, embryonic development, masking endogenous retroviral genes, X chromosome inactivation in females, and the development of cancer [60, 7, 41]. In the following, we are concerned solely with 5-methylation of cytosine (Fig. 2.3), in particular preceding a guanine. This is because 5-methylation of cytosine is the only form of DNA methylation present in mammals and mainly occurs in CpG dinucleotide context.

The percentage of methylated CpGs in mammals ranges from about 60% to

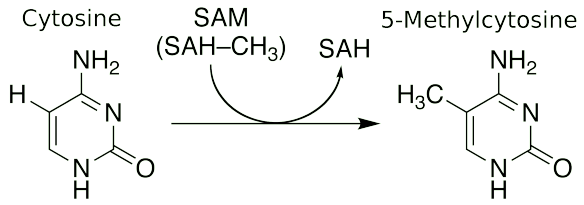


Figure 2.3: Methylation of cytosine. DNA methyltransferases (DNMTs) are the enzymes that facilitate methylation of DNA. They require S-adenosyl methionine (SAM), a cosubstrate that acts as a methyl-donor.
Source: Adapted from Wikipedia [84]

90% [21]. CpG dinucleotides typically form clusters, so-called CpG islands (CPIs). They are often part of gene regulatory regions; in mammals, about 40% of genes contain CPIs in their promoter or exonic regions and they are typically unmethylated [24]. According to Gardiner-Garden and Frommer [26], CPIs are defined by three characteristics: a length of at least 200 bp, a CpG percentage of at least 50%, and an observed-to-expected CpG ratio, $\frac{\#CpG}{\#C \cdot \#G} \cdot N$, greater than 60%. Hypermethylation of these regions results in gene silencing and has, for example, been observed in principal tumor types, such as colon, lung and prostate tumors [5]. Breast cancer is associated with CPI methylation of tumor suppressing microRNA (Fig. 2.4). There are two ways in which DNA methylation can affect gene expression. In the first, methylation of DNA sterically hinders binding of transcriptional proteins with DNA. In the second, methylation of DNA leads to the recruitment of specific proteins, so-called methylated-CpG-binding domain proteins (MBDs). These proteins then form complexes with other proteins, e.g., histone deacetylases (HDACs) or chromatin remodeling complexes, which cause the formation of heterochromatin. Fig. 2.5 shows the interplay between methylation of DNA and histone modifications.

2.1.3 Histone Modifications

Histone modifications are covalent, post-translational attachments of functional groups to histones. They are not considered an epigenetic modification by some researchers, because there is little evidence that histone modifications are actually heritable [44]. Since our working definition of epigenetics does not necessitate heritability, we still consider them an epigenetic modification. There exist various possible modifications of histones, among them acetylation, ubiquitination, and methylation (Fig. 2.6). Histone modifications can occur either at histone tails or their globular core domain. Historically, due to technological limitations, research was focused on modifications affecting the histone tails, although multiple positions in the globular do-

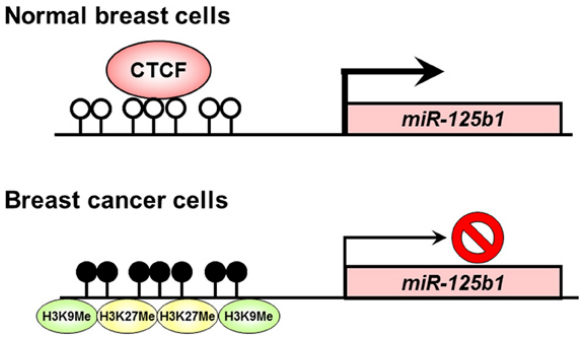


Figure 2.4: CpG island (CPI) silencing in cancer. In non-cancer cells, the presence of CCCTC-binding factor (CTCF) prevents silencing at miR-125b1 via epigenetic mechanisms such as histone modification or DNA methylation and induces an open chromatin structure. However, cancerous breast cells, which exhibit loss of CTCF are associated with CPI methylation and histone modifications associated with a repressive transcriptional state. Open circles in the CPI indicate unmethylated cytosine, while filled circled indicate methylation. The depicted gene, miR-125b1, codes for a micro RNA that is possibly involved in tumor suppression.
Source: Saito et al. [67]

main are now also known for their potential to be modified [52]. Histone modifications act in two ways. First, some tail modifications, for example acetylation and phosphorylation, can influence chromatin structure through electrostatic interactions by changing the charge of the tail. Modifications of the histone core involve similar structural modifications of the nucleosome. The second, primary mechanism of action caused by tail modifications, seems to be binding affinity alteration of non-histone proteins interacting with chromatin. Two examples that illustrate how histone modifications impact chromatin structure and how they modulate binding of regulative proteins are acetylation and recruitment of heterochromatin protein 1 (HP1). Acetylation (Fig. 2.7) of histones leads to a change in nucleosome structure resulting from a reduction in the positive charge of histones. Since the phosphate backbone of DNA is negatively charged, the extent of electrostatic interaction between histones and DNA is diminished. Hence, after acetylation, the highly condensed heterochromatin adopts a more relaxed structure and becomes easily accessible for transcription. An example for the second mechanism of action is the recruitment of HP1 via H3K9 methylation [70]. If lysine (K) at the ninth position of histone H3 is methylated, HP1 is able to bind. It can then form a complex with a DNA methyltransferase (DNMT)1, which effects gene silencing via DNA methylation, see Fig. 2.8. The concept that a combination of various modifications in one or multiple histone tails of a nucleosome is responsible for the recruitment of chromatin-binding proteins was termed

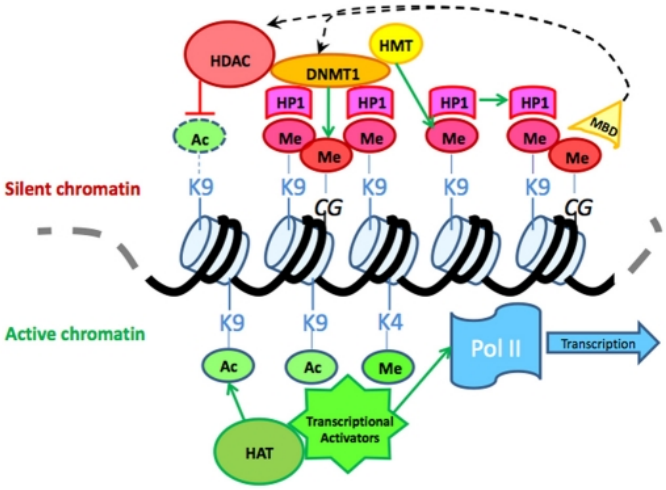


Figure 2.5: Interplay of epigenetic modifications. methylated-CpG-binding domain proteins (MBDs), which specifically bind to methylated DNA, interact with histone modification enzymes such as histone deacetylases (HDACs) and histone methyltransferases (HMTs), thereby reinforcing the repressive chromatin state. Methylation of histones can lead to DNA and further histone methylation via binding of heterochromatin protein 1 (HP1). These repressive modifications are shown in shades of red in the upper part of the figure. The lower part of the figure illustrates acetylation of histones, which is facilitated by histone acetyl transferases, allowing for the transcription of DNA via DNA polymerase II (shades of green).
Source: Conerly and Grady [10]

the histone code [33]. In contrast to the genetic code, the histone code has not yet been elucidated completely due to its complexity. For the genetic code, we know that triplets made up of the four nucleic bases adenine (A), cytosine (C), guanine (G), and thymine (T) are responsible for coding one of twenty amino acids, giving rise to 4^3 possible combinations. The challenge of the histone code is manifold: There may exist a complex interplay between modifications on each of the eight tails present in a nucleosome, effects of modifications are position-dependent, and there is a large number of possible modifications. In the next section, we deal with determination of DNA methylation, the other major mode of epigenetic modification.

2.2 Measuring Genome-Wide DNA Methylation

Since this thesis focuses on DNA methylation rather than histone modifications, we only deal with methods measuring methylation. First of all, a more formal definition of what one understands under the term *methylation level*

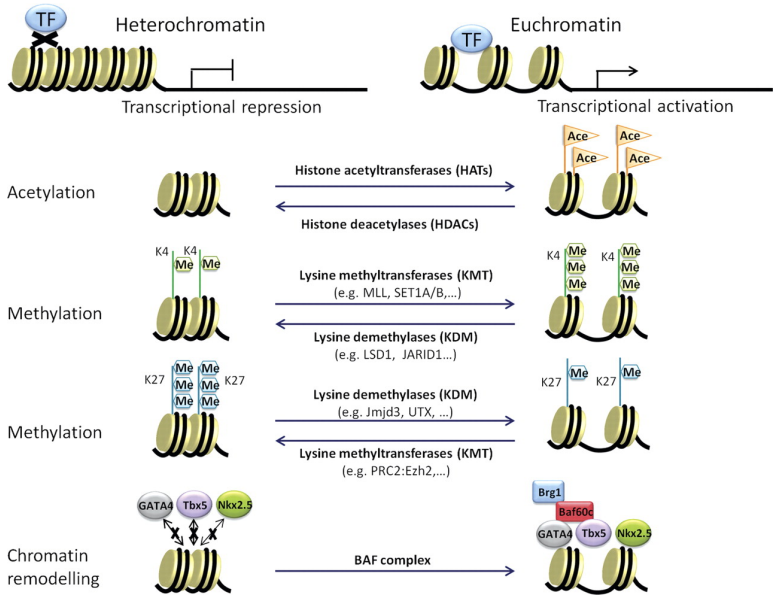


Figure 2.6: Regulation of chromatin structure through histone modifications and chromatin remodeling complexes.
Source: Ohtani and Dimmeler [57]

is given. After that, we introduce the bisulfite reaction, which constitutes the foundation for the two technologies used in this thesis, WGBS and the Infinium HumanMethylation450 BeadChip.

2.2.1 Representing Levels of Methylation

The Beta-Value

The level of methylation of a position in the genome is described by its beta-value. Methods used to determine methylation levels give two types of signals: one signal for methylated positions (M) and another for unmethylated ones (U). The beta-value describes the ratio of methylated versus non-methylated signal [79] and is given by

$$\beta = \frac{M}{U + M}.$$

It ranges from 0 to 1 and has an intuitive biological interpretation: A value of 0 indicates no methylation and a value of 1 full methylation. Hemimethylation, i.e., a locus that exhibits methylation in 50% of the cases, is represented by a beta-value of 0.5. This phenomenon can be interpreted as differential methylation in the two copies of a chromosome, that is, intermediate allelic methylation or strand-specific methylation. Fig. 2.9 exemplifies these

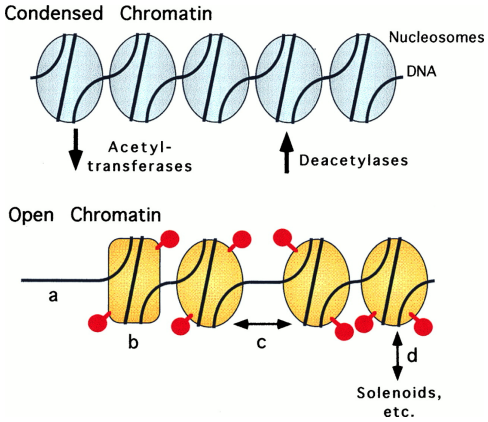


Figure 2.7: Acetylation of histones. Addition of acetyl groups to histones is facilitated by histone acetyl transferases (HATs), while their removal is effected by histone deacetylases (HDACs). The more relaxed, acetylated gold-colored chromatin (euchromatin) is associated with transcriptional activity, while the condensed, deacetylated variant shown in blue (heterochromatin) is associated with repression. Red spheres represent acetyl groups linked to histones. An open chromatin structure can be the result of (a) changes in nucleosome occupancy, (b) changes in nucleosome conformation, (c) changes in interactions between nucleosomes, or (d) changes in higher-order chromatin packaging and structure.
Source: Privalsky [62]

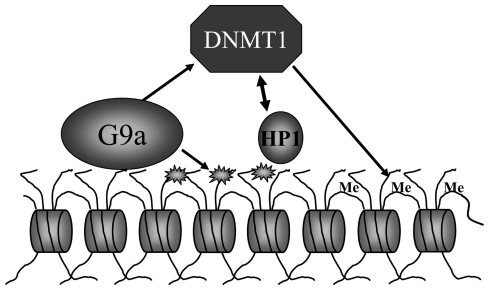


Figure 2.8: H3K9 methylation effects DNA methylation via cooperation of heterochromatin protein 1 (HP1) and DNMT1. G9A dimethylates H3K9 and in this way creates a binding site for HP1. Through an interaction of HP1 with DNMT1, DNMT1's methyltransferase activity is stimulated, leading to DNA methylation. DNMT1 stabilizes HP1 binding.
Source: Smallwood et al. [70]

three methylation states. Due to its simple biological interpretation, the beta-value is the established method for reporting methylation.

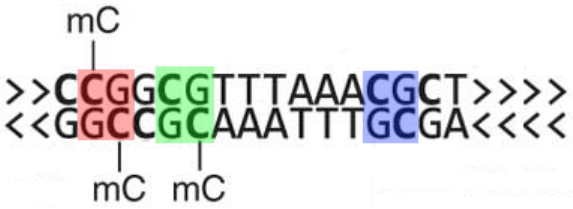


Figure 2.9: Interpretation of the beta-value. Full methylation ($\beta = 1$), hemimethylation ($\beta = 0.5$), and no methylation ($\beta = 0$) are indicated by the red, green, and blue CpG positions, respectively.
Source: Adapted from Krueger et al. [39]

2.2.2 Method Overview

Before we delve into methods for methylation profiling, we will first introduce the chemical reaction that forms the basis of these methods.

The Bisulfite Reaction

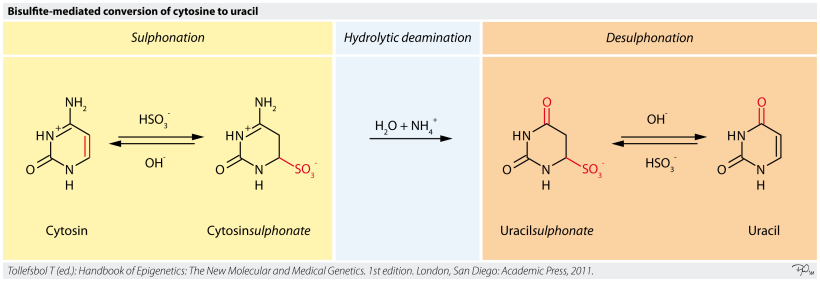


Figure 2.10: The bisulfite reaction is comprised of three steps: sulphonation, hydrolytic deamination, and desulphonation. It results in the transformation of unmethylated cytosine to uracil. Methylated cytosine, on the other hand, is not affected by bisulfite treatment.
Source: Tollefsbol [77]

The bisulfite reaction (Fig. 2.10) allows the determination of methylation levels in the sequence through the conversion of unmethylated cytosine to uracil, while 5-methylcytosine remains unaffected. The use of polymerase chain reaction (PCR) for amplification of sample DNA results in the replacement of uracil by thymine. After the alignment of sample DNA to a reference genome, methylation can be measured by considering nucleotide exchanges induced by bisulfite conversion. The process of calling methylation based on bisulfite-converted DNA is illustrated in Fig. 2.11. Technically, there exist two protocols for bisulfite treatment. The protocol shown in Fig. 2.11 is considered non-directional. In non-directional protocols, bisulfite-conversion

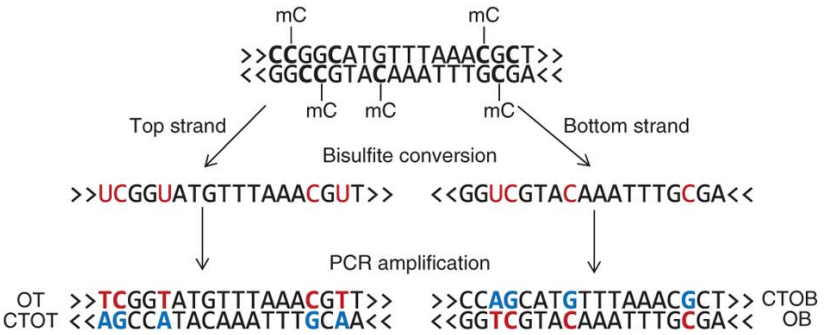


Figure 2.11: Calling methylation in bisulfite sequencing. After bisulfite application, all unmethylated cytosines are converted to uracils. Methylated cytosines (tagged with mC) stay unchanged. After polymerase chain reaction (PCR), uracil is replaced by thymine. Hence, in the PCR products, unmethylated cytosine can be identified according to the presence of thymine in the original forward and reverse strands (OT, OB) as well as the presence of adenine in the complementary strands (CTOT, CTOB).
Source: Krueger et al. [39]

leads to changes in both, 5'-to-3' and 3'-to-5' reads, resulting in C→T nucleotide exchanges in one strand and in G→A exchanges in the complementary strand. In contrast, directional bisulfite treatment protocols, only affect 5'-to-3' reads and exclude the complementary strand's guanine opposing the cytosine from conversion. Hence, unmethylated positions are subject to C→T nucleotide exchanges only.

Introduction to Methods for Profiling of DNA Methylation

Approaches aimed at profiling DNA methylation can be broadly categorized into two categories: methods that rely on sequencing of DNA and microarray-based ones. The Infinium HumanMethylation450 BeadChip is the leading array-based approach. After bisulfite-treatment and whole-genome amplification, sample DNA is applied to a chip with more than 485,000 position-specific CpG probes. In a hybridization reaction, the fluorescently marked probes bind to the target DNA and the locus-specific methylation level is determined by analyzing the emitted light intensities.

The sequencing methods mainly differ in which regional aspects of the genome they analyze. Methods such as reduced-representation bisulfite sequencing (RRBS) and methylated DNA immunoprecipitation sequencing (MeDIP-seq) extract only a subset of the full genome. WGBS, on the other hand, determines methylation levels for every possible position in the genome. In RRBS, by using CpG-sensitive restriction enzymes, merely genome areas with a high CpG content are extracted and methylation levels are called via bisulfite conversion. Due to the reduced representation in

RRBS, solely 1% of the nucleotides in the genome have to be sequenced and it is possible to achieve a higher coverage in CPIs [51]. MeDIP-seq [18] relies on the purification of methylated DNA. For this, an antibody is raised against 5-methylcytosine and, with immunoprecipitation, only stretches of methylated DNA are extracted. After DNA extraction, the purified segments are further fragmented and then sequenced in order to determine their position in relation to a reference genome. Disadvantages of this approach include the low resolution of methylation measurement and experimental limitations of immunoprecipitation, e.g., antibody cross-reactivity.

In the next sections, we will discuss the two methods this thesis is based on, the Infinium HumanMethylation450 BeadChip and WGBS.

2.2.3 The Infinium HumanMethylation450 BeadChip

The Infinium HumanMethylation450 BeadChip (in the following referred to as Infinium 450K) is a microarray designed to measure methylation levels for more than $485,000$ sites in the human genome. The chip covers 99% of RefSeq genes [63]. Additionally, it contains all of the content categories defined by an expert consortium, among them 96% of CPIs, CpG sites outside CPIs, DNase hypersensitive sites, and differentially methylated sites in tumors [31]. Features such as its low-cost, fast application, and comprehensive selection of CpG sites make the Infinium 450K array a well-established tool in the community.

Methodology

The Infinium 450K chip is based on the principle of hybridization: Two single-stranded copies of DNA can form double-stranded DNA when their sequences are complementary to each other, allowing for hydrogen-bonding. The microarray contains thousands of small cells, each filled with probes specific to certain CpG loci. After bisulfite conversion of input DNA, hybridization is reported through probe extension with fluorescently labeled nucleotides. The emitted light intensities are then used to determine the methylation of the i-th interrogated CpG as

$$\beta_{450K}(i) = \frac{\max(y_{i,\text{methy}}, 0)}{\max(y_{i,\text{unmethy}}, 0) + \max(y_{i,\text{methy}}, 0) + \alpha},$$

where $y_{i,\text{methy}}$ and $y_{i,\text{unmethy}}$ are the intensities measured by the i-th methylated and unmethylated probes, respectively [19]. Negative values are reset to 0. The variable α (recommended default $\alpha = 100$) is a normalization constant that regularizes β when both intensities are low. The name of the beta-value comes from its distribution: It is beta-distributed when probe distributions are assumed to have gamma distributions.

The microarray is based on beads, globular structures covered with multiple copies of a specific oligonucleotide (50 bp), which are used to capture

target DNA. For the Infinium chip, there exist two types of assay designs, giving rise to different bead types. The type of assay used depends on the CpG in question: About 135,000 sites are covered by Infinium I beads and about 350,000 positions by Infinium II beads.

The Infinium I Assay For Infinium I sites, there are two bead types per locus (Fig. 2.12). One bead type (U) contains probes that bind to unmethylated sites, while the other (M) contains probes that bind to methylated sites. Both probe types rely on the same mechanism. In the case of hybridization, single nucleotide extension can take place, leading to the emission of light through green Cy3 (extension with C or G) or red Cy5 dye (extension with A or T). If probe and sample are not complementary, extension cannot occur and hence there is no signal. It is necessary to differentiate the methylated and unmethylated signal according to bead location on the chip, because both bead types use the same fluorescent dyes. When a probe covers more

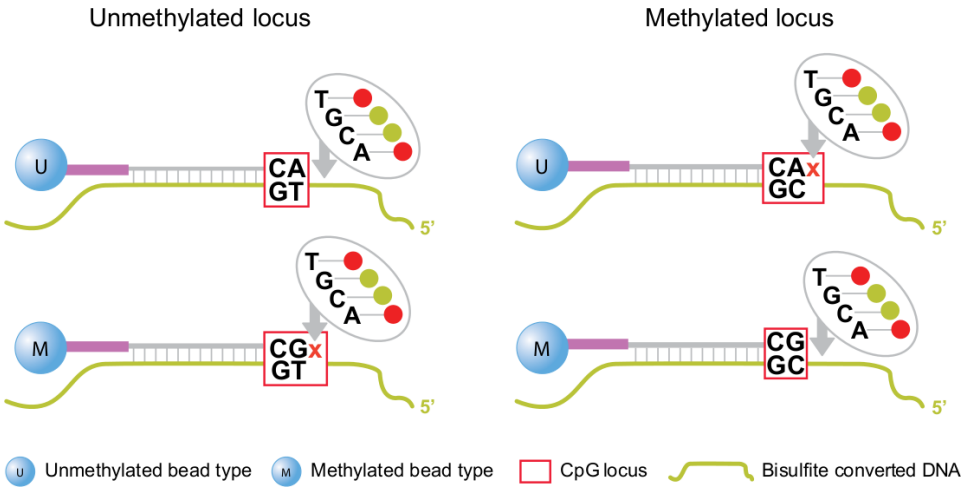


Figure 2.12: Chemistry of the Infinium I assay. Infinium I probes are 50 bp oligonucleotides with two bead types. One terminates across methylated cytosines, whereas the other terminates across non-methylated cytosines. Base extension can only occur when sample and probe have hybridized.
Source: Illumina [31]

than a single CpG, the Infinium I assay assumes that the methylation states of all these CpGs agree with each other. This assumption is reasonable considering that Eckhardt et al. [20] showed that 90% of CpG sites within 50 bases had the same methylation status for chromosomes 6, 20, and 22.

The Infinium II Assay For Infinium II sites, only a single bead type is employed (Fig. 2.13). This assay does not directly match the full CpG dinucleotide of the target DNA but rather only the guanine. Hence, depending

on the extended base, we either measure methylation (incorporation of C or G tagged with Cy3 dye) or no methylation (incorporation of A or T tagged with Cy5 dye). The methylation signal is formed by considering the ratio of Cy3 methylation and Cy5 non-methylation signal for every locus. To deal

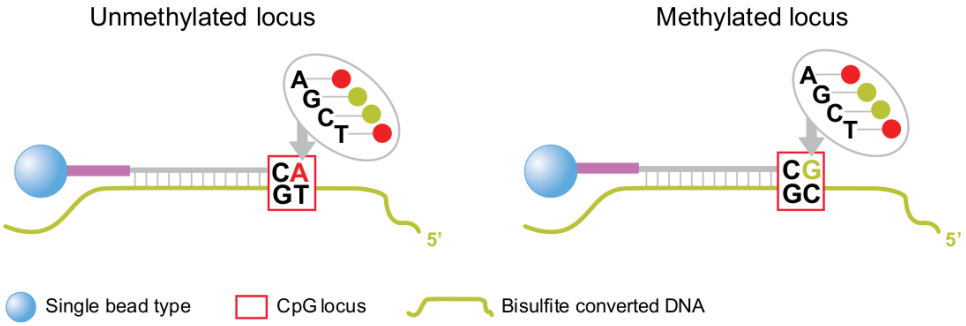


Figure 2.13: Chemistry of the Infinium II assay. A single 49-base oligonucleotide is used to bind sample bisulfite-converted DNA. Probe binding is independent of methylation because it occurs upstream of the interrogated CpG. Only the base extension step determines the methylation status.
Source: Illumina [31]

with multiple CpGs on one probe, degenerate bases are used to support hybridization independent of the sample DNA's methylation state. According to Illumina [31], a probe can contain up to three CpGs without compromising data quality. Nevertheless, the combined usage of two dyes in this assay can lead to dye bias, especially when considering data from different batches.

Assay Comparison Due to the different chemistries involved in the two assays, they exhibit distinct distributions of beta-values (Fig. 2.14). In an in-depth analysis, Dedeurwaerder et al. [15] showed that Infinium I probes exhibit a higher dynamic range than Infinium II probes, that is, a greater ratio between largest and smallest beta-values. In addition, it was shown that beta-values from type II probes were less accurate and reproducible than those from type I probes. They concluded to either always consider type I and type II probe loci separately or to normalize the distributions of the two types in order to be able to pool the data. In the same paper, Dedeurwaerder et al. [15] introduced the first method for adjusting the type II probe distribution to that of type I, called peak-based correction (PBC). In this approach, the two peaks representing no methylation and full methylation in the density plot are superimposed. Although this method is simple, it already brings about a noteworthy improvement in the quality of probe II methylation calls. Since its development in 2011, however, other, more sophisticated normalization methods such as subset-quantile within array normalization (SWAN) [48] and beta mixture quantile dilation (BMIQ) [76] have been introduced, further improving probe normalization.

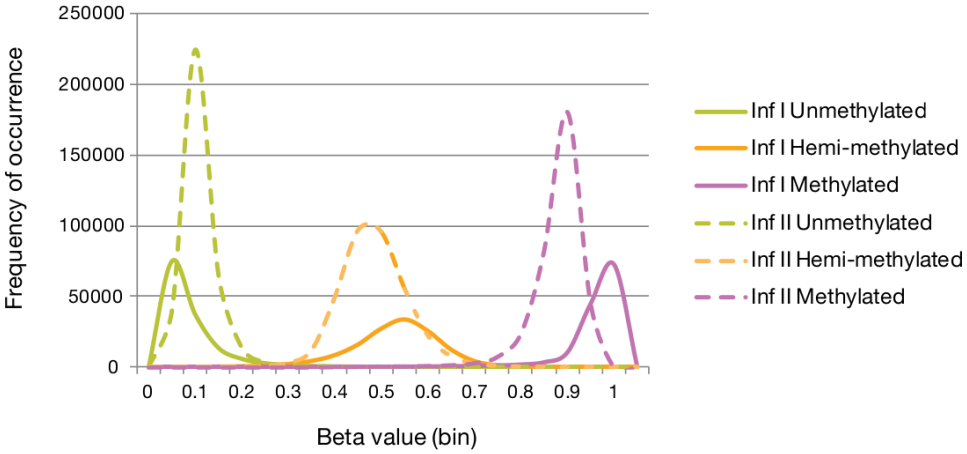


Figure 2.14: Distribution of type I and type II probe beta-values. Type II probes exhibit a smaller dynamic range than type I probes. About one quarter of the positions on the Infinium 450K chip is covered by type I probes, while about three quarters are covered by the less accurate type II probes.
Source: Illumina [31]

Accuracy and Limitations

It was shown that the Infinium 450K chip agrees well with other, established methods for measuring DNA methylation. For example, Illumina [31] reports an r^2 (the square of Pearson's correlation coefficient) between 0.915 and 0.931 for the agreement of methylation levels from 450K and WGBS. Still, there exist several technological limitations. First, to be able to compare signals from probes of type I and type II, it is necessary to normalize the distributions, which might lead to inaccuracies. Second, the chip contains cross-reactive probes, i.e., probes that can hybridize to different loci. In experiments, between 8.6% and 28% of probes mapped to multiple loci [61, 87]. Third, about 4.3% of probes contain SNPs at the CpG position [61]. The occurrence of a SNP constitutes a confounding factor for methylation measurement, since only measurements from individuals carrying the majority variant should be considered. Therefore, such positions should be filtered in inter-individual studies [16]. In addition, probes exhibiting high intensities seem to deviate more strongly from bisulfite-sequencing methylation calls [16] and type II probes with more than three CpGs incur inaccuracies [31].

2.2.4 Whole-Genome Bisulfite Sequencing

WGBS is able to afford the methylation level of every base in the genome with single nucleotide resolution. Its foundation lies in the combination of two technologies. In the first step, as explained in Section 2.2.2, DNA is

treated with bisulfite, which effects nucleotide exchanges in only the non-methylated bases. Then, in the second step, next-generation sequencing of the full genome's converted DNA is performed. Using a known reference genome, the methylation status of each base can be inferred by observation of base exchange frequencies. In the following, we will focus on whole-genome sequencing using the Illumina platform, which is frequently employed in BS-Seq studies [47].

Workflow for Illumina Sequencing

Library Preparation The first step in sequencing with the Illumina platform is the creation of a sequencing library. Since it is not possible to sequence an entire genome in one piece, it is initially necessary to shear the genome into smaller fragments. There exist multiple methods that are able to break DNA into smaller pieces. Two established methods for this task are digestion by restriction enzymes and sonication. In the Illumina protocol, ultrasonication, that is, the application of ultrasonic energy to agitate particles, is used. This mechanical shearing approach is unbiased and highly sensitive. After sonication, end damage to fragments is repaired and an extra adenosine is added to each fragment increasing ligation efficiency. Then, fragments with a size of about 200 to 300 bp are selected. Size selection is performed such that, later on, clusters of fragments do not overlap with each other. The last step in the preparation of the library is DNA amplification via PCR. The primers used for PCR include additional adapter-sequences that allow DNA to bind to Illumina's flow cell and form clusters.

Bridge Amplification After library preparation, the processed DNA is applied to the flow cell, a solid surface that facilitates DNA binding through the previously appended adapter-sequences. To amplify the attached DNA, a process called *bridge amplification* is initiated. This is a solid-phase PCR involving the generation of millions of clusters containing thousands of fragment copies each (Fig. 2.15). After double-stranded DNA has been formed by action of DNA polymerase, the strands are denatured. The single-stranded DNA then forms bridges by hybridizing with adjoining primers, thereby allowing DNA polymerase to bind again. When the complementary strand has been completed, the next PCR cycle is initiated. Multiple cycles of PCR gives rise to clusters of single-stranded DNA, each containing a multitude of DNA fragment copies.

Sequencing Illumina uses a sequencing by synthesis strategy to determine the sequence of bases encoded by the fragments. This approach relies on observing which bases are added to the newly constructed strand of DNA during double-strand formation. To determine the incorporated base, modified nucleotides are used. They exhibit two properties. First, they contain

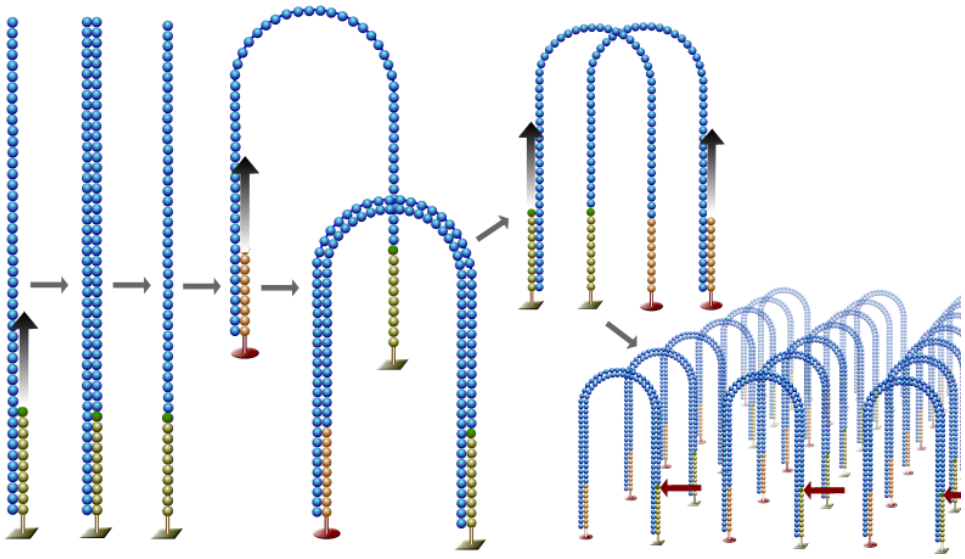


Figure 2.15: Illumina bridge amplification. As a solid-phase PCR, it consists of three stages. First, denaturation of DNA leads to single-stranded samples. Second, primer annealing, that is, hybridization of primers to single-stranded DNA. Third, primer extension, i.e., formation of double-stranded DNA through DNA polymerase. These steps are performed in repeated cycles to amplify DNA. The main characteristic of bridge amplification is that each sample is attached to the flow cell (solid-phase) and hence, single-stranded DNA forms bridge structures when annealing to primers, giving rise to the name of the method.
Source: Wikipedia [83]

a fluorescent dye. By tagging each of the four nucleotides with a different dye, it is possible to determine the incorporated base by the dye's emitted wavelength after excitation with a laser. Second, each nucleotide terminates the action of DNA polymerase temporarily. This allows the measurement of individual nucleotide extensions. One cycle of sequencing takes place as follows. First, a reversible dye-terminator nucleotide is incorporated by DNA polymerase. Due to the terminator included in the nucleotide, no further extension can occur. Then, non-incorporated nucleotides are washed from the flow cell, leaving only hybridized DNA behind. After dye laser-excitation, a camera takes an image of the integrated fluorochrome to determine the incorporated base. After removal of the dye from the incorporated nucleotide, DNA polymerase can extend the nascent DNA further, allowing the determination of the next base. The sequence of each fragment is determined by multiple cycles of nucleotide extension, washing, dye measurement, and cleavage. The final, contiguous DNA sequence is received either by assembly or, given an existing reference genome, mapping of reads.

Characteristics of Whole-Genome Bisulfite Sequencing

WGBS is mainly characterized by the used sequencing technology. At the moment, this is typically a next-generation sequencing approach, e.g., the Illumina platform. Advantages of next-generation sequencing technology include its high throughput and deep coverage. Depth of coverage refers to the number of reads that overlap with a single position. A high depth of coverage is desirable for two reasons. First, it makes measurements more robust with regard to errors. Second, a large sample allows the detection of low-frequency population variants, which was not possible with previous technologies, e.g., Sanger sequencing.

Multiple sources of errors exist for next-generation sequencing. Every sequencing-by-synthesis approach is subject to the inherent error rate of the used DNA polymerase. Furthermore, non-complementary bases can be incorporated before sequencing itself, e.g., in the amplification step required for library preparation, as well as during bridge amplification. One specific source of error for Illumina's sequencing platform is a phenomenon described as *lagging-strand dephasing* [55]. It arises when one of the strands in an ensemble of fragments is not extended during a cycle of sequencing, e.g., due to failed terminator cleavage in the previously incorporated nucleotide [35]. As a consequence, the affected strand is out of phase and its base readout lags behind the readout of the other strands. Lagging strands are a proposed explanation for the fact that Illumina's error rate increases towards the ends of reads [53]. Nakamura et al. [55] were able to show that specific sequence patterns can elicit lagging strands. Apart from lagging strands, it is also possible for a strand to pre-phase, that is, to be ahead of the other strands, e.g. when a non-terminating nucleotide is falsely incorporated, leading to multiple incorporations in a single cycle.

Another source of error is provided by cross-talk between dyes [45]. The dye frequencies typically overlap, so it is necessary to perform deconvolution and account for noise to determine the true signal. To weight individual color channels, a so-called *cross-talk matrix* is used. This matrix is estimated only from the first two sequencing cycles under the assumption that every nucleotide is equally likely. For Illumina, there seems to exist cycle-dependent cross-talk, which would require not a static but a dynamic cross-talk matrix for every cycle [35]. Bias in PCR amplification can also be a source of problems; certain fragments may be over-represented while others may be under-represented.

A limitation of bisulfite treatment is due to 5-hydroxymethylation of cytosine. Similarly to methylated cytosine, it also does not convert during standard bisulfite treatment [37]. Hence, with bisulfite-based methods, one does not actually measure only methylation but also hydroxymethylation levels. Another challenge is presented by errors in bisulfite conversion of nucleotides itself. There are two types of errors that can occur: overconversion and un-

derconversion. In overconversion, 5-methylcytosine is erroneously converted to thymine, while in underconversion unmethylated cytosine does not convert to uracil [28]. This can lead to false-negative and false-positive methylation calls, respectively. Warnecke et al. [81] were able to demonstrate several aspects of bisulfite treatment that can lead to conversion errors, among them characteristics in the sequence, which were already reported previously [30]. Apart from that, bisulfite treatment can be difficult with small samples of DNA because — with optimal settings for temperature and incubation time with regard to maximal conversion rate — DNA degradation rates of up to 96% were observed [29] but remedies exist. PBAT is an amplification-free sequencing approach tailored towards small DNA samples circumventing the bisulfite-induced loss of template DNA [54].

For post-processing of bisulfite-converted sequencing data, an accurate alignment of reads to the reference genome is of great importance. In contrast to regular sequencing reads, bisulfite-treated reads exhibit a greater variability due to bisulfite-induced nucleotide exchanges. It is therefore necessary to adjust mapping algorithms accordingly. There are several efficient mappers specifically tailored towards bisulfite-treated reads [9], for example, Bismark [38]. Additionally, proper care should be taken when calling methylation values. Here, differentiating between sequencing errors, SNPs, and actual methylation is one of the main challenges. Bis-SNP [47] is a tool that deals with these challenges and is discussed in more detail in Section 4.1.2.

Chapter 3

Materials

As material for this thesis, we were provided with two methylation data sets for the HepaRGd7R2 cell line, which were generated as part of Deutsches Epigenom Projekt (DEEP), the German contribution to the international human epigenome consortium (IHEC). HepaRGd cells are terminally differentiated hepatic cells derived from a human hepatic progenitor cell line. They are infinitely reproducible making them ideal for experimental research.

Of the two data sets, one was brought about by WGBS, while the other one was generated using the Infinium 450K chip. WGBS data was already aligned to the current human genome reference [43] (hg19) and provided in binary sequence alignment/map (BAM) format. The alignment was performed using GSNAP [85] with the *cmet-stranded* alignment mode, which is appropriate for directional bisulfite treatment protocols. For the microarray, the raw intensities for the green and red color channels were provided in *idat* format. For detailed information on how the two data sets differ in their measurements, we refer to Section 2.2 and, in the following, just give their main characteristics.

3.1 Properties of the Data Sets

3.1.1 CpG Coverage of WGBS Data

Each chromosome exhibited a similar CpG read coverage with medians around 10, except for chromosome Y, which had a smaller coverage (Fig. 3.1). However, this observation might be due to the small number of measured CpGs in this genomic region.

3.1.2 Methylation Measurements

Measurements from the Infinium 450K chip are in the interval $(0, 1)$ due to noise and α-normalization, while β_{WGBS} is in $[0, 1]$ and exhibits a greater dynamic range. As reported by Illumina [31], methylation measurements of

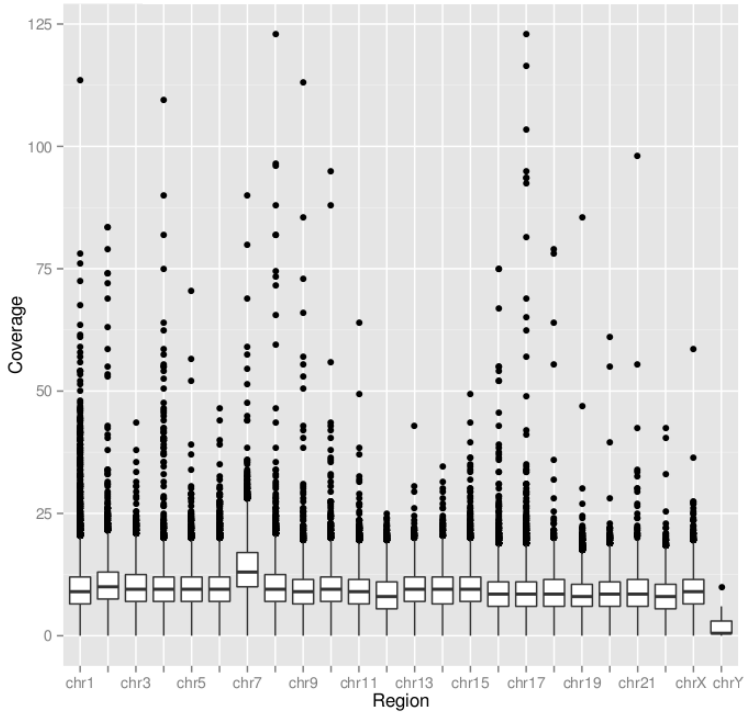


Figure 3.1: Chromosome-specific read coverage. The y-axis shows the number of reads covering each CpG dinucleotide in the data set for each chromosome, indicated on the x-axis.

the Infinium 450K chip agree well with those from WGBS. We verified their finding on our data and found a correlation of $r = 0.93$ between beta-values, see Fig. 3.3. We also analyzed the distribution of methylation values of non-normalized Infinium I and Infinium II probes (Fig. 3.2). It is evident that the beta-value distribution of the two Infinium 450K probe types exhibit substantial differences. However, since the assays measure different loci, there also exists an inherent difference in their underlying biological signals, as evidenced by differences in WGBS measurement distributions for Infinium I and Infinium II loci. We can confirm the higher dynamic range of Infinium I vs. Infinium II probes. However, there does not exist a substantial difference in agreement with WGBS in the light of a correlation of $r \approxeq 0.933$ for Infinium I probes and $r \approxeq 0.937$ for Infinium II probes.

Let us now consider the distribution of differences in methylation between data from the Infinium 450K chip and WGBS, as illustrated in Fig. 3.4a. Most observations (about 94%) possess a high concordance ($\Delta\beta \coloneqq \beta_{450K} - \beta_{\text{WGBS}} < 0.3$). There is a slight skew in the distribution of methylation differences, namely, there are more observations with $\Delta\beta > 0$ than with

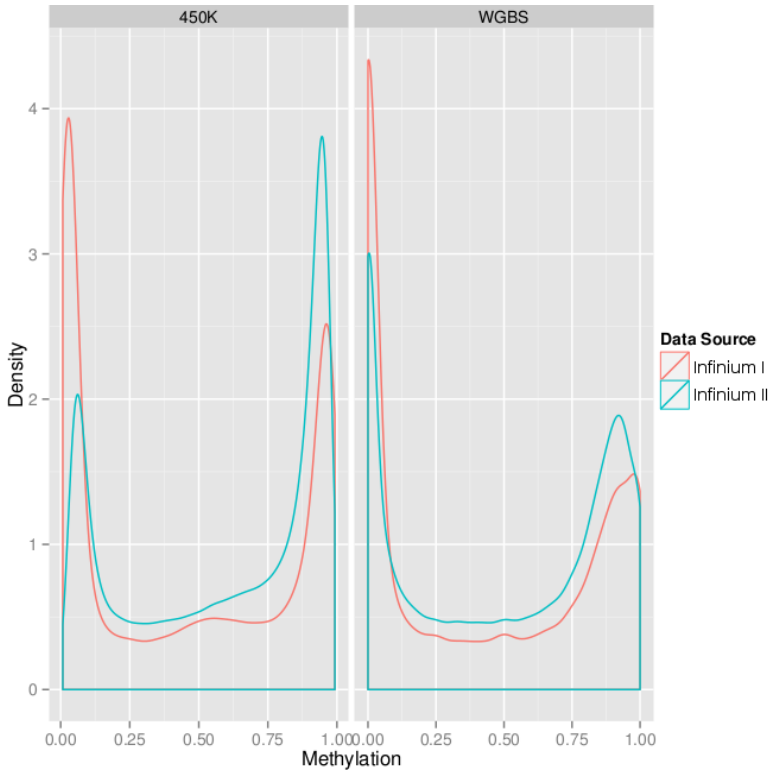


Figure 3.2: Infinium type I vs. type II methylation values. The left panel shows the beta-value density of the Infinium 450K chip, while the right panel gives the values of the corresponding loci for WGBS. The red curve indicates probes of the Infinium I assay, whereas the blue curve indicates probes of the Infinium II assay.

$\Delta\beta < 0$, see Fig. 3.4b and 1 in the appendix. The greatest divergence in methylation levels occurs at extreme values, that is, either for β close to 0 or 1. It seems that Infinium 450K technology is biased towards identifying loci as being fully methylated, while WGBS seems to be biased towards classifying loci as unmethylated.

We also investigated whether there exist certain regions in the genome for which methylation measurements were more inaccurate than for others (Fig. 3.5), but found no substantial region-specific effects for any chromosome. Consequently, we focused on the identification of local features predictive of $\Delta\beta$ in the following work.

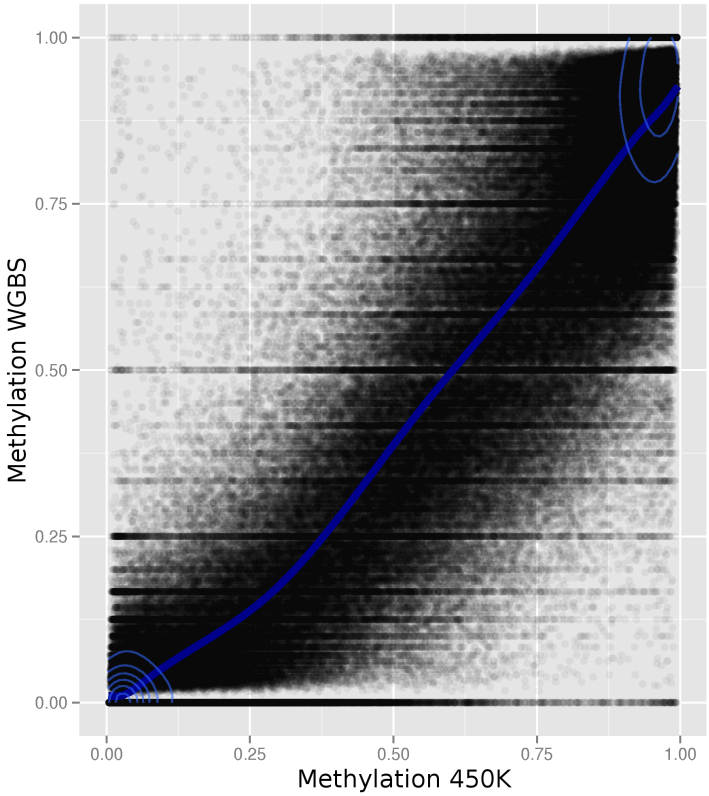
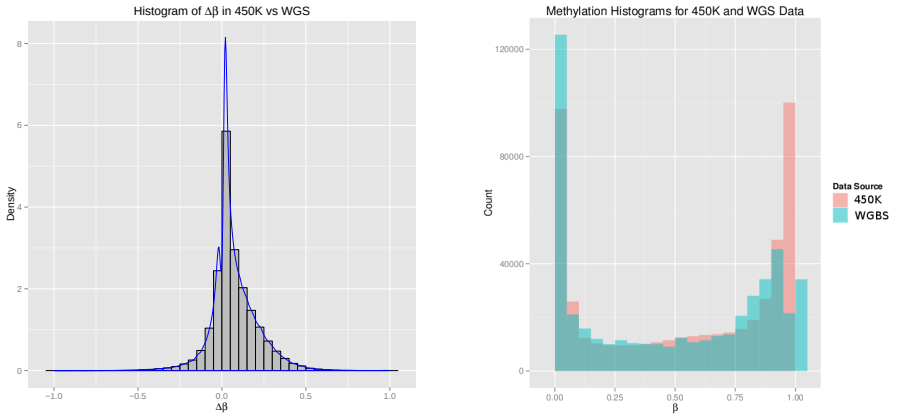


Figure 3.3: Comparison of beta-values between Infinium 450K and WGBS. The fitted density indicates a high correlation ($r = 0.93$). The elliptical 2D-densities indicate those regions in the plot, in which the majority of the points reside. Each level of these height-profiles indicates an area in the plot that has a certain number of data points associated with it. The largest number of observations are at the extreme ends of the beta-value range, indicating no methylation and full methylation, respectively.



(a) Frequencies of $\Delta\beta$ **(b)** Comparison of beta-value counts

Figure 3.4: Investigation of methylation differences between Infinium 450K and WGBS measurements.

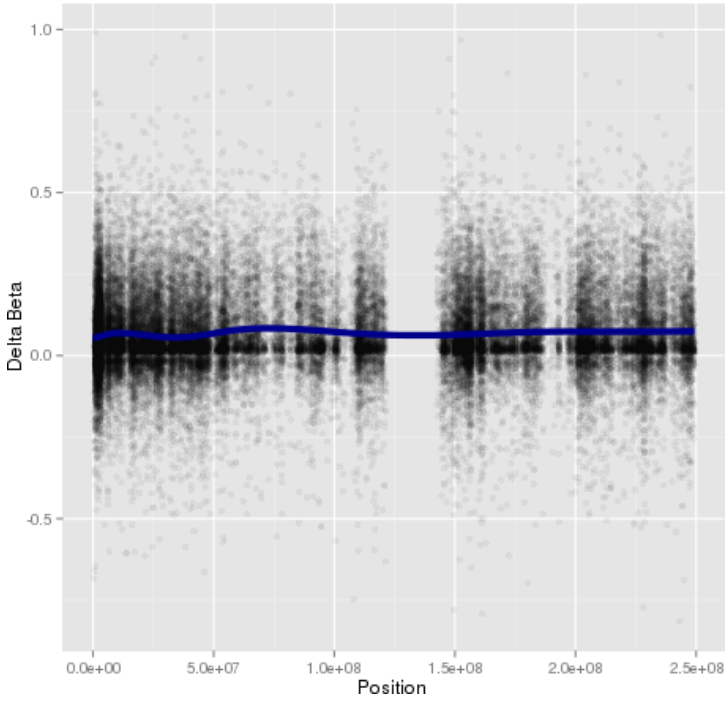


Figure 3.5: Methylation differences on chromosome 1 for 46, 866 positions.

Chapter 4

Methods

This chapter is organized into two parts. In the first, we give the methodological foundation of this project. First, we are concerned with the normalization of the two Infinium 450K assays. This step is necessary because probes of type Infinium I and Infinium II differ in their beta-value distributions. Next, we cover methylation calling for WGBS data, for which it is important to differentiate information about methylation status from SNPs. Afterwards, we give an overview of several machine learning concepts that were applied to analyze and predict methylation differences. Among them are SVMs, kernel functions, and cross-validation. Finally, we deal with methods for the interpretation of features and non-linear kernel functions. In the second part of this chapter, we present our own methodological contributions, namely the extension of existing string kernel functions into the numeric domain, which we refer to as *hybrid string kernels*.

4.1 Method Background

4.1.1 Infinium 450K Intra-Array Normalization

As discussed in Section 2.2.3, the two types of assays used in Infinium 450K chips, Infinium I and Infinium II, differ in their beta-value distributions. Hence, to simultaneously use observations from both assays in a single prediction model, it is necessary to normalize their distributions. BMIQ [76] adjusts the statistical distribution of beta-values of type II probes into a distribution characteristic of type I probes. To do so, a three-state beta-mixture model assigning probes to methylation states is employed. Then, probabilities are transformed into quantiles and, as a last step, a dilation transformation is applied to preserve the monotonicity and continuity of the data.

Three-State Beta-Mixture Model

To model fully methylated (M), hemimethylated (H), and unmethylated (U) positions, a three-state beta-mixture model is fitted to the probes of type I and type II individually. Since the beta distribution possesses two parameters, a and b, it is necessary to fit, for each probe type, three pairs of parameters, one pair for each of the three methylation states. Let $\{(a_U^I, b_U^I), (a_H^I, b_H^I), (a_M^I, b_M^I)\}$ be the estimated beta distribution parameters for type I probes and, analogously, $\{(a_U^{II}, b_U^{II}), (a_H^{II}, b_H^{II}), (a_M^{II}, b_M^{II})\}$, for type II probes. Membership of probes to states is determined by a maximum likelihood criterion.

Determination of Probabilities and Quantile Transformation

The probability of each beta-value β is modeled as

$$p(\beta^t) = \pi_U^t \mathcal{B}(\beta | a_U^t, b_U^t) + \pi_H^t \mathcal{B}(\beta | a_H^t, b_H^t) + \pi_M^t \mathcal{B}(\beta | a_M^t, b_M^t)$$

where $t \in \{I, II\}$ represents probe type and $\mathcal{B}$ is the probability density function of the beta distribution, which is given by $\mathcal{B}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$ where B, the beta function, normalizes the distribution. To determine the parameters (π, a, b), an expectation-maximization (EM) algorithm, which performs alternating steps of expectation and maximization, is used. In the expectation step, the expected log likelihood for the current parameter estimate is computed. In the maximization step, the parameters maximizing that likelihood are determined. By alternatingly executing the two steps in an iterative fashion one arrives at the maximum likelihood parameter estimates. Individual parameter sets are identified via (π_s^t, a_s^t, b_s^t) where t represents probe type and s methylation state. Since parameter estimates from EM are two-tailed, the authors consider values that lie on the left- and right hand side of the mean, which is given by

$$m_s^t = \frac{a_s^t}{a_s^t + b_s^t}.$$

Let U_{II}, H_{II}, and M_{II} indicate sets of type II probes assigned to unmethyalted, hemimethylated, and methylated states, respectively. Let U_{II}^L and U_{II}^R be the set of probes that fall on the left- and right hand side of m_U^{II}, respectively, and, correspondingly, set M_{II}^L and M_{II}^R for methylated probes. Then the probability of unmethylated type II probes left of the mean is

$$p = P(U | \beta_{U_{II}^L}^L) = F(\beta_{U_{II}^L} | a_U^{II}, b_U^{II}),$$

where F is the cumulative beta function. Computations for type I probes and probes on the right side of the mean are analogous and give, for each probe, its probability of belonging to a certain state.

Then BMIQ converts the probabilities of type II probes assigned to the methylated or unmethylated state to quantiles. The transformation is based on computing the inverse of the cumulative beta distribution using the corresponding type I parameters. For U-state probes, parameters (a_U^I, b_U^I) are used, while for M-state probes, parameters (a_M^I, b_M^I) are introduced. For example, to receive type I-adjusted quantiles for unmethylated probes on the left side of the mean, BMIQ sets the normalized β value to

$$\eta_{U_{II}^L} = F^{-1}(p|a_U^I, b_U^I).$$

For unmethylated probes on the right side of the mean, computations are similar ($1 - F$ instead of F). The operations for methylated probes are analogous.

Dilation

To guarantee the continuity of the normalized values, BMIQ still has to scale beta-values of hemimethylated probes to fit into the region defined by the endpoints $maxU = \max\{n_U^{II}\}$ and $minM = \min\{n_M^{II}\}$. Scaling is motivated by the fact that hemimethylated probes are not well-described by a beta-distribution. Since dilation is just based on simple observations about the hemimethylated regions and corresponding scaling, we do not give the details here but instead refer to the paper by Teschendorff et al. [76].

4.1.2 Calling Methylation for Bisulfite Sequencing Reads

Calling methylation for WGBS is based on the fact that, in bisulfite-converted DNA, unmethylated cytosine appears as thymine, while methylated cytosine remains unchanged. For a given position i in the genome, we consider its set of n reads, $\mathbf{D}(i) = \{D_1, \ldots, D_n\}$. The base of read j at genome position i is given by $D_j(i) \in \{C, T\}$. $D_M(i)$ and $D_U(i)$ with $\mathbf{D}(i) = D_M(i) \dot{\cup} D_U(i)$ are the reads with methylated and unmethylated bases, respectively. They are defined as

$$\begin{aligned} D_M(i) &= \{D_j | D_j(i) = C\} \quad \text{and} \\ D_U(i) &= \{D_j | D_j(i) = T\}. \end{aligned}$$

The beta-value for WGBS can be defined using the cardinalities of these sets, giving us

$$\beta_{\text{WGBS}}(i) = \frac{|D_M(i)|}{|D_M(i)| + |D_U(i)|}.$$

A limiting factor of this naive approach lies in the existence of SNPs and conversion errors during bisulfite treatment, which can blur the methylation signal. Hence, for accurate calling of methylation, one needs to extend this naive approach by account for these factors.

Bis-SNP: a Caller for Methylation and SNPs

Consider first the detection of SNPs in bisulfite sequencing data. Previous SNP callers for DNA relied on strand complementarity. However, bisulfite-converted reads are not complementary anymore when subjected to a directional protocol. Moreover, since bisulfite conversion results in C→T substitutions of unmethylated cytosines, they cannot be easily distinguished from evolutionary C→T SNPs, which account for nearly 80% of polymorphisms at CpG sites [78]. For accurate quantification of methylation levels, it is vital to differentiate the two events. Bis-SNP [47] is an established method for calling methylation and SNPs at the same time, with an accuracy of 96% for SNP detection.

Illumina sequencing employs a directional bisulfite treatment protocol and guanines opposite of cytosines are not subject to nucleotide substitutions resulting from cytosine conversions. Hence, the strand containing the guanine (G-strand) can be utilized as a reference for distinguishing SNPs and true methylation. Only reads containing the cytosine (C-strand) are used to determine methylation (Fig. 4.1). What sets Bis-SNP apart from previous, simpler approaches for determining methylation levels in whole-genome sequencing is its robust probabilistic framework.

Preprocessing Step In the first step, reads with mapping scores less than 30 and those mapped to multiple loci are excluded. For paired-end reads, only proper pairs are retained. Local realignment and sequence recalibration is afforded by a modified version of the genome analysis toolkit (GATK)[49], which is tailored towards bisulfite-converted DNA. By default, only positions with a recalibrated base score greater than 5 are used for SNP calling.

Probabilistic Model The foundation of Bis-SNP is GATK's Bayesian likelihood model with a number of bisulfite-specific adaptions. For a position i covered by n reads, let $\mathbf{D} = (D_1, D_2, \ldots, D_n)$ represent the corresponding base calls. The posterior probability of the underlying diploid genotype $G = AB$, with alleles A and B, is

$$\Pr(G|\mathbf{D}) = \frac{\pi(G)\Pr(\mathbf{D}|G)}{\Pr(\mathbf{D})}.$$

The prior probability of the genotype is given by $\pi(G)$. It is based on the genotype of the reference genome and population nucleotide substitution frequencies. The probability of the data is a sum over all possible genotypes AB,

$$\Pr(\mathbf{D}) = \sum_{AB} \pi(AB)\Pr(\mathbf{D}|AB).$$

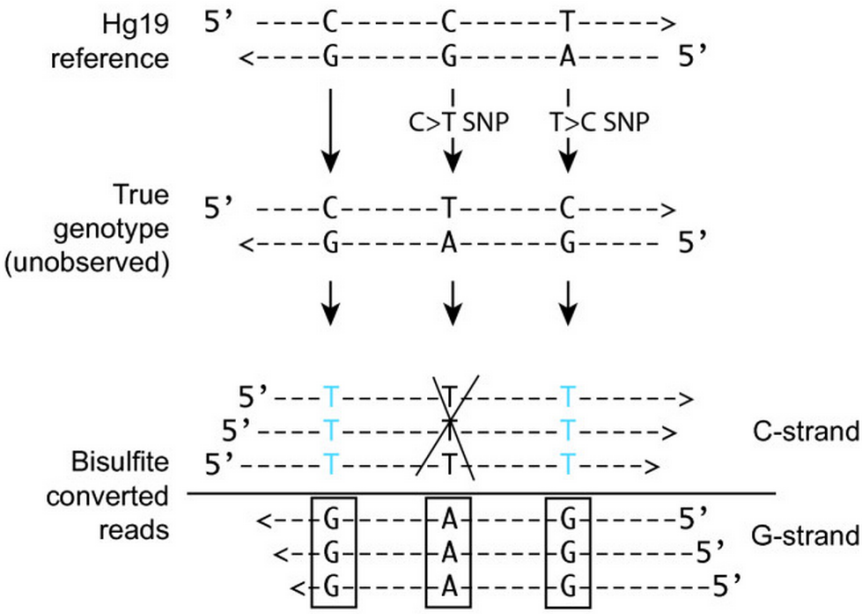


Figure 4.1: Bis-SNP methylation calls. Note that, for the third row representing the reads, only the C-strand was bisulfite converted and only reads from that strand are used to call methylation. The reads from the G-strand, which were not converted, are solely used for SNP calling. For the first C, we observe T/G, indicating an unmethylated cytosine since the reference also has a cytosine. For the second C, however, we observe T/A. The presence of A indicates that, we do not have an unmethylated cytosine but rather a C→T SNP. For the last C, we observe T/G, indicating a T→C SNP, whose cytosine was unmethylated and therefore bisulfite-converted. This is because the G-strand contains G, even though the reference has T/A.
Source: Liu et al. [47]

Assuming that each allele is equally probable, the likelihood of the data is the product of conditional probabilities,

$$\Pr(\mathbf{D}|G) = \prod_{j=1}^{n} \Pr(D_j|G),$$

where

$$\Pr(D_j|G = AB) = \frac{1}{2}\Pr(D_j|A) + \frac{1}{2}\Pr(D_j|B)$$

gives the probability of a base D_j given a genotype G.

Base quality scores are on a Phred scale and defined as,

$$Q = -10 \log_{10} \epsilon,$$

indicating the probability ϵ of a base call being erroneous. Consider a scenario with underlying alleles adenine ($A = a$) and thymine ($B = t$). Then,

bisulfite-conversion could not have occurred and the probability of the allele $B = t$ would be

$$\Pr(D_j|B=t) = \begin{cases} \frac{\epsilon_j}{3} & \text{if } D_j \neq t, \\ 1-\epsilon_j & \text{if } D_j = t \end{cases}$$

that is, either there was a sequencing error (probability $\frac{\epsilon_j}{3}$) if anything other than t was observed or the correct base t was observed (probability $1-\epsilon_j$).

For an underlying allele containing c or g, the probabilities become strand-specific, because the directional bisulfite protocol only affects one strand. It is necessary to consider two components. First, β_j, the probability that a position is methylated. Second, parameters for the efficiency of bisulfite conversion, namely the underconversion frequency α and the overconversion frequency γ. Let the C-strand be denoted by + and the G-strand by −. The full likelihood calculation for a cytosine is

$$\Pr(D_j|B=c) = \begin{cases} (1-\epsilon_j)[\beta_j(1-\gamma)+(1-\beta_j)\alpha] & \text{if } D_j = c^+ \quad (4.1) \\ \frac{\epsilon_j}{3} + (1-\epsilon_j)[\beta_j\gamma+(1-\beta_j)(1-\alpha)] & \text{if } D_j = t^+ \quad (4.2) \\ 1-\epsilon_j & \text{if } D_j = c^- \quad (4.3) \\ \frac{\epsilon_j}{3} & \text{otherwise} \quad (4.4) \end{cases}$$

In case 4.1 the correct base is observed $(1-\epsilon_j)$ and it is necessary to consider two scenarios. In the first, D_j is methylated (β_j) and therefore not converted $(1-\gamma)$ and in the second, D_j is unmethylated $(1-\beta)$ and unexpectedly does not convert (α). In case 4.2, since we observe a thymine in the C-strand, there could be an error $(\frac{\epsilon_j}{3})$ or not $(1-\epsilon_j)$. If it is not an error, we need to consider two possibilities. First, the cytosine could be methylated (β) and overconverted (γ) or it could be unmethylated $(1-\beta_j)$ and converted $(1-\alpha)$ as expected. In case 4.3 a cytosine is present on the non-converting strand. This is modeled by the complementary probability of a base miscall, $1-\epsilon_j$. The observation of any other base (case 4.4) hints at an error $(\frac{\epsilon_j}{3})$.

The final genotype G_{best} is that with maximum posterior $\Pr(G|\mathbf{D})$. The bisulfite parameters, α and γ, play only a minor one as they typically vary by less than 1% [80, 75, 81]. More important is the methylation rate β, which varies according to genomic context, organism, and cell type. Contexts can be specified by dinucleotides, e.g., CG and CH (where H stands for C, T, or A), for mammals. The final β_{WGBS} value is the number of C-strand reads with C divided by the number of C-strand reads with C or T. Bis-SNP's full workflow is delineated in Fig. 4.2.

In addition to the probabilistic framework, Bis-SNP also performs filtering steps for methylation and SNP calling. Non-conversion of unmethyalted cytosines is known to occur largely at the 5' end of Illumina reads. To deal with this problem, Bis-SNP walks along each read from 5' to 3' and discards

all cytosines that occur before the first C→T conversion. Furthermore, SNPs occurring in clusters are filtered and, to retain only bona fide SNPs, filters for coverage depth, strand bias, and quality are employed.

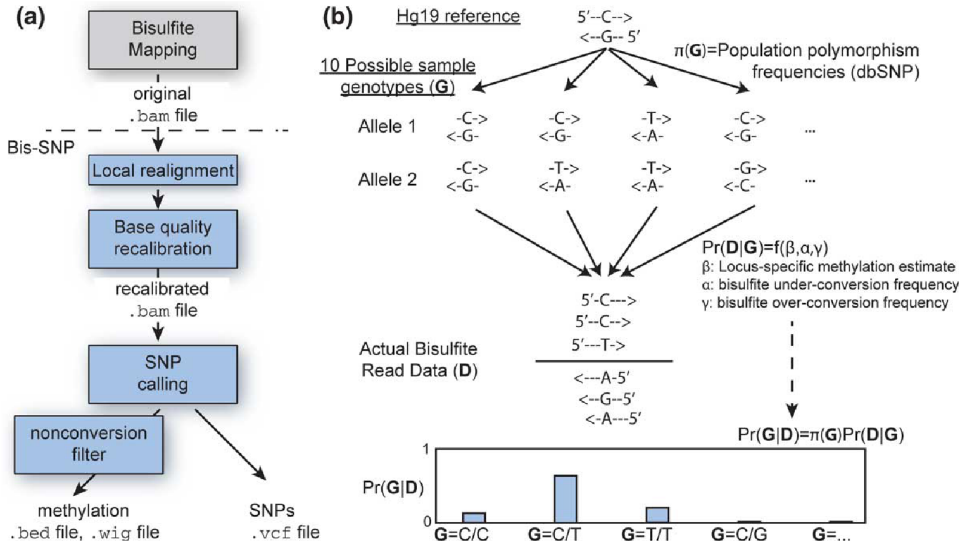


Figure 4.2: Bis-SNP workflow. (a) Program flow: after preprocessing steps such as local realignment and base quality recalibration, SNPs and methylation levels are outputted. (b) Probabilistic genotyping framework: for each locus, all possible genotypes G are considered. Then, using methylation estimate β, bisulfite conversion rates γ and α, and population polymorphism frequencies $\pi(G)$, likelihoods $\Pr(\mathbf{D}|G)$ representing the probability of the observed data, given a certain genotype G, are computed. Using the prior, likelihoods are transformed into posterior probabilities $\Pr(G|\mathbf{D})$ that represent the probability of a genotype given the reads. Finally, the genotype with the largest posterior is chosen.
Source: Liu et al. [47]

4.1.3 Supervised Statistical Learning and Support Vector Machines

Supervised statistical learning is a field concerned with learning from data with known outcomes, in order to train a model which is able to predict the outcomes of new data. At the core of supervised learning is the idea of learning general characteristics of the input data rather than its peculiarities, and to generalize learned patterns on yet unseen data. Depending on the type of outcome, we can distinguish two scenarios. In the first, the outcome is represented by a qualitative variable, which provides a label for each observation. Such a variable represents only a discrete set of classes, e.g., *blue*, *red*, or *green*, if one deals with colors. This scenario is called *classification*.

If the outcome is quantitative, one deals with regression. In this thesis, the outcome is $\Delta\beta$, the difference in methylation between Infinium 450K and WGBS, and hence we perform regression.

Input variables are represented by X and if X is a vector its components can be accessed by X_j. Outputs are represented by Y for quantitative variables. These upper case letters represent the generic aspects of variables. Lower-case letters, on the other hand, are used to represent observed variables, e.g., x_i is the i-th observation of X. Matrices are written in bold and upper-case, e.g., $\mathbf{X}$ indicates the $N \times p$ matrix of N observations of which every observation x_i is a vector representing p features. Vectors are always written in lower-case, except if they represent a full set of observations on a feature. For example, x_i is the p-vector for observation i, while $\mathbf{x}_i$ is the N-vector of all observations on variable X_i.

The learning task can be formulated in the following way. Given the value of an input vector X, make a good prediction of the output Y, denoted by $\hat{Y}$. To achieve this, we construct prediction rules using a set of training data consisting of pairs of measurements $(x_i, y_i), i = 1, \ldots, N$, where x_i is a p-vector of features and y_i gives the corresponding outcome. In our approach, we relied on SVMs, to afford accurate learning rules.

Support Vector Machines

Originally, SVMs were developed as a method for classification. They are based on finding the hyperplane that separates observations x_i belonging to two classes best. This is achieved by maximizing the margin, the area that separates points lying on one side of the hyperplane from those on the other. The SVM is based on the notion of support vectors, observations that lie on the border of the margin describing the decision boundary. If it is possible to separate the two classes perfectly, this is called the *separable case*. If, however, at least one observation is on the *wrong* side of the optimal hyperplane, this is the *non-separable case*. To deal with this, a new type of variable, the so-called *slack variable*, is introduced. Slack variables measure the extent to which an observation comes to lie on the wrong side of the hyperplane or the extent to which an observation on the correct side of the hyperplane comes to lie within the margin. Hence, to receive the best possible separation of data, one tries to minimize the slack, i.e., the extent of misclassifications, while maximizing the margin.

Support Vector Regression

Support vector regression (SVR) shares some similarities with SVMs. For example, in SVR there exists the equivalent of a margin. In contrast to classification, this margin is a consequence of the error function, Vapnik's

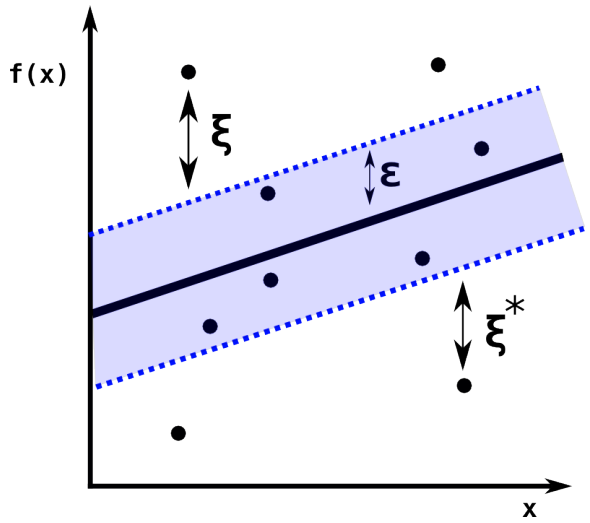


Figure 4.3: The epsilon-insensitive tube in support vector regression. Slack variables ξ are associated with observations that lie outside of the ϵ-tube. Observations that are located in the ϵ-region do not appear in the target function.

ϵ-insensitive loss,

$$|y - f(x)|_\epsilon := \max\{0, |y - f(x)| - \epsilon\},$$

for which x is the observed input data, $y \in \mathbb{R}$ is the observed output value, and $f(x)$ is the predicted value. Only observations with L_1-regularized losses $|y - f(x)|$ that exceed ϵ are part of the target function (Fig. 4.4). For linear regression, we have $f(x) = \langle \beta, x \rangle + \beta_0$ where $\beta \in \mathbb{R}^p$ is the vector of feature coefficients and β_0 gives the offset. One needs to minimize

$$\frac{1}{2}||\beta||^2 + C \sum_{i=1}^{N} |y_i - f(x_i)|_\epsilon$$

where C determines the extent to which prediction errors are penalized and β is L_2 regularized to restrict its magnitude. Another similarity between the SVM and SVR is the existence of slack variables in its constrained formulation. For SVR, two types of slack variables are introduced, one for each side of the ϵ-insensitive region (Fig. 4.3): ξ for cases with $f(x_i) - y_i > \epsilon$ and ξ^* for those observations with $y_i - f(x_i) > \epsilon$. The collection of all slack variables is referred to as $\xi^{(*)}$. The introduction of slack variables leads to

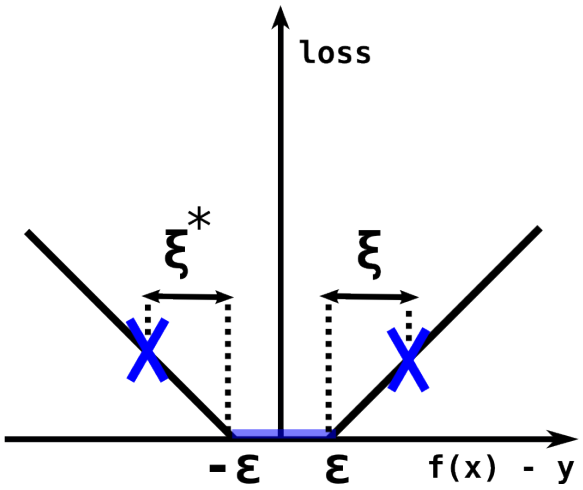


Figure 4.4: Vapnik's epsilon-insensitive loss. Only absolute deviations of predictions $f(x)$ from labels y larger than ϵ are penalized in support vector regression.

the following constrained optimization problem

$$\begin{aligned} \text{minimize}_{\beta,\xi^{(*)}} \quad & \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ \text{subject to} \quad & f(x_i) - y_i \leq \epsilon + \xi_i \\ & y_i - f(x_i) \leq \epsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \quad \forall i = 1, \ldots, N \end{aligned}$$

Here, only observations i with $\xi_i > 0$ lying outside the ϵ-region are part of the objective function and provide support vectors for SVR.

To generalize SVR for nonlinear scenarios, Lagrange multipliers are introduced giving rise to the following optimization problem

$$\begin{aligned} \underset{\alpha,\alpha^* \in \mathbb{R}^N}{\text{maximize}} \quad & -\epsilon\sum_{i=1}^{N}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)y_i \\ & -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle x_i, x_j\rangle \qquad (4.5) \\ \text{subject to} \quad & 0 \leq \alpha_i, \alpha_i^* \leq C \quad \forall i = 1, \ldots, N \qquad (4.6) \\ \text{and} \quad & \sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0. \qquad (4.7) \end{aligned}$$

The regression estimate takes the form

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i)\langle x_i, x\rangle + \beta_0. \tag{4.8}$$

This formulation of SVR is called ϵ-SVR, because the model is fully specified via the parameter ϵ, indicating the extent of the error-insensitive region. In practical applications, ϵ-SVR may be difficult to use, because it can be hard to find a well-fitting value for ϵ. A variant of SVR, called ν-SVR, allows an easier parameter setting.

ν-SVR The main practical advantage of ν-SVR versus ϵ-SVR is the better interpretability of ν in comparison to ϵ, allowing for straight-forward parameter selection via grid search. The hyperparameter ν provides an upper bound $0 \leq \nu \leq 1$ on the fraction of points allowed outside the ϵ-tube, which, asymptotically, bounds the number of support vectors. The parameter ϵ is implicitly set by the choice of ν. In ν-SVR, the following primal objective function is used

$$\frac{1}{2}||\beta||^2 + C\left(\nu N\epsilon + \sum_{i=1}^{N} |y_i - f(x_i)|_\epsilon\right),$$

where $\epsilon \geq 0$ is considered as a parameter over which one minimizes. Until now, only modeling of linear relationships via dot products was discussed. In the following, we will introduce kernel functions, one of the main assets of SVMs.

4.1.4 Kernel Functions

Kernel functions represent an inner product of two observations x_i and x_j in feature space by mapping data points to feature space through Φ:

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j)\rangle.$$

Using the feature space transformation, kernel functions may map data into a higher-dimensional space, thereby capturing non-linear data properties. The value of a kernel function represents the feature space similarity between x_i and x_j. The technique of replacing the feature space dot product in the optimization problem 4.5 and the SVR estimate $f(x)$ (Equation 4.8) with a kernel function is called the *kernel trick*. The new optimization function is

$$\begin{aligned}\underset{\alpha,\alpha^* \in \mathbb{R}^N}{\text{maximize}} \; & -\epsilon \sum_{i=1}^{N} (\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) y_i \\ & - \frac{1}{2} \sum_{i,j=1}^{N} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) k(x_i, x_j),\end{aligned}$$

with the same constraints as before and the new estimate

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \alpha_i) k(x_i, x) + \beta_0.$$

Kernel functions must be continuous, symmetric, and should result in a positive-semidefinite matrix. A kernel function whose matrix is positive-semidefinite is said to satisfy Mercer's theorem and exhibits non-negative eigenvalues, i.e., $Kv = \lambda v$ with $\lambda \geq 0$. The entry in row i and column j of the kernel matrix K is determined by $k(x_i, x_j)$, the similarity between the i-th and the j-th observation. The satisfaction of Mercer's theorem is important for optimization because it leads to a convex optimization problem with a unique solution.

Kernel functions enable the adaption of SVMs to the properties of the data set by affording similarity measures tailored towards observation-specific characteristics. In the following we will present the kernel functions we considered to deal with next-generation sequencing data.

Numeric Kernels

Numeric kernels compare observations $x, x' \in \mathbb{R}^p$. From this class of kernels, we employed the linear, polynomial, and Gaussian radial basis function kernels.

Linear Kernel The linear (*vanilla*) kernel,

$$k(x, x') = x^T x',$$

uses a regular dot product and gives linear SVR.

Polynomial Kernel The polynomial kernel,

$$k(x, x') = (s \cdot \langle x, x' \rangle + o)^d,$$

takes the d-th power of the dot product. In this way, interactions between up to d features can be modeled. Further adjustment of the kernel is possible by setting offset o and scalar s.

Gaussian radial basis function Kernel The Gaussian radial basis function (RBF) is an exponential function,

$$k(x, x') = \exp(-\sigma ||x - x'||^2),$$

dependent on a parameter σ fixing the width of the Gaussian. For small values of σ, the RBF has a narrow but peaky distribution, whereas for large values of σ, it has a wide but flat distribution.

String Kernels

The kernels presented in the previous section were all based on numeric input. There also exist kernels for other types of input. In this section, we deal with string kernels, which afford a measure of similarity for strings [6]. Their advantage over numeric kernels is that they are able to exploit specific properties of strings. For example, string kernels are able to compare subsequences of a certain length, allow for mismatches, or consider shifts in motifs (substrings) between two strings. They operate on strings $\mathcal{S}$ defined over an alphabet $\mathcal{A}$. Note that the adaptivity of string kernels comes at a greater computational cost in comparison to numeric kernels.

Weighted Degree String Kernel Sonnenburg et al. [71] introduced the weighted degree kernel (WDK) in 2005. It is based on comparing all substrings of observations x and x' of length less or equal than degree d and is defined as

$$k(x, x') = \sum_{k=1}^{d} \beta_k \sum_{i=1}^{|\mathcal{S}|-k+1} \mathbb{I}\{x[i : i+k] = x'[i : i+k]\}.$$

Matching substrings are scored using weights β_k depending on the length $k \leq d$ of matching substrings. The cardinality $|\mathcal{S}|$ represents the uniform sequence length and $\mathbb{I}$ is the indicator function,

$$\mathbb{I}(a = b) := \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

The subsequence starting from position i and ending (exclusively) at position j is indicated by $x[i : j]$. Consider Fig. 4.5 for a practical application of the WDK. The weights β_k are chosen such that the weight decreases for increasing k, because if matching substrings of size k are found, this implies having found matching substrings of length smaller than k already and one does not want to overemphasize long motifs. Hence, the suggested value for β_k is $2\frac{d-l+1}{d(d+1)}$.

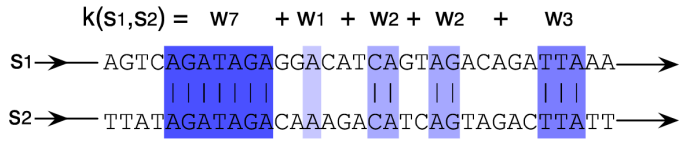


Figure 4.5: Weighted degree string kernel. Each matching substring makes a contribution whose weight depends on the match length. Here, w_i is used to indicate substring matches of length i.
Source: Sonnenburg et al. [72]

Weighted Degree String Kernel With Shifts The weighted degree kernel with shifts (WDKS) was introduced in 2005 by Rätsch et al. [64]. It is an extension of the WDK allowing matching of shifted substrings. For this, a new parameter s was introduced that describes the extent of the shift, giving rise to

$$k(x,x') = \sum_{k=1}^{d} \beta_k \sum_{i=1}^{|\mathcal{S}|-k+1} \sum_{\substack{s=0 \\ s+i\leq|\mathcal{S}|}}^{S} \delta_s \cdot \big(\mathbb{I}\{x[i:i+k] = x'[i+s:i+s+k]\} + \mathbb{I}\{x[i+s:i+s+k] = x'[i:i+k]\}\big) .$$

In the formula, the first indicator function term considers a shift in sequence x', while the second term considers a shift in sequence x. Once again, β_k represents the weight used to score matching substrings of length k. To differentiate between shift extents, a new weight, δ_s, is introduced. It is chosen such that large shifts result in smaller scores, as this indicates motifs that are further apart from each other. Hence, by default, we have $\delta_s = \frac{1}{2s+1}$. Fig. 4.6 gives an example for this kernel.

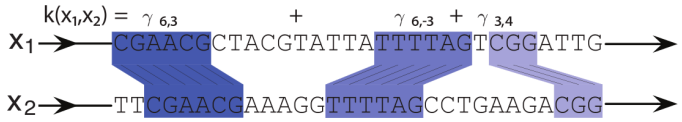


Figure 4.6: Weighted degree string kernel with shifts. Here, $\gamma_{i,j}$ is used to indicate the weight that is associated with matching substrings of length i with shift j.
Source: Sonnenburg et al. [72]

Please note that the implemented versions of the WDK and the WDKS, as well as their derived versions, also allow for a maximal number of mismatches mm between substrings.

Oligo String Kernel In the oligo kernel (OK) [50], each substring (oligomer) of a fixed length k is represented by a Gaussian function, see Fig. 4.7, giving rise to the kernel

$$k(x,x') = \sqrt{\pi}\sigma \sum_{\omega\in\mathcal{A}^k} \sum_{p\in S_\omega^x} \sum_{q\in S_\omega^{x'}} \exp\left(-\frac{1}{4\sigma^2}(p-q)^2\right) .$$

The sets S_ω^x and $S_\omega^{x'}$ give the occurrence positions of the motif ω in the sequences x and x', respectively. The subsequence ω has length k and is part of the set of all length k subsequences present in x and x', which is defined according to the alphabet $\mathcal{A}$ as $\mathcal{A}^k$. The value of the exponential function

depends mainly on the squared distance, $(p-q)^2$, between the motifs. The greater the squared distance between two motifs is, the smaller will be the values of the Gaussian function. Furthermore, positional uncertainty is in-

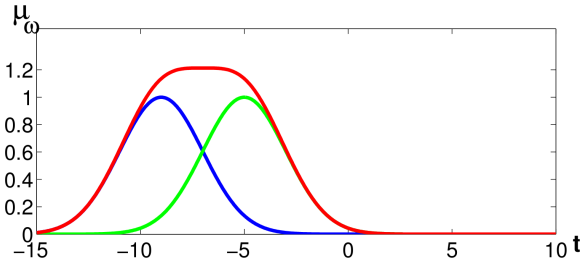


Figure 4.7: Positional uncertainty in oligo functions. An oligomer occurs at position -5 (green curve) and another one at position -9 (blue curve). The width of the two Gaussians, centered at the given positions, depends on the smoothing parameter σ. Adding the two Gaussians (red curve), the smoothness of the resulting function increases with increasing σ.
Source: Meinicke et al. [50]

troduced via the parameter σ. For large values of σ, Gaussians have a wide distribution (fuzzy assignment), while for small values of σ, they exhibit narrow distribution (exact assignment). This means that, in the OK, it is possible to fix to which extent one wants to penalize matching motifs on the basis of their distance in the sequence window.

Set Kernels

In multi-instance learning, a set of instances is associated with every observation. Set kernels, introduced in 2003 by Gärtner et al. [27], provide a similarity measure for sets of instances, in contrast to numeric kernel functions, which use feature vectors $x \in \mathbb{R}^p$ as input. They are well-suited for every type of scenario in which observations are structured into parts. In our scenario, every observation is a single CpG position whose set of instances consists of reads covering that CpG. The advantage of the set kernel is that we can directly work on the reads and do not have to use an intermediate step such as the consensus sequence. In this fashion, we hope to capture similarity more accurately than with the previously introduced kernels. Set kernels,

$$k_{\mathrm{s}}(Z, Z') = \sum_{\substack{x_i \in Z \\ x_j \in Z'}} k(x_i, x_j),$$

are defined over sets Z and Z' with instances $x_i \in Z$ and $x_j \in Z'$. To put it in a nutshell, set kernels compute the pairwise similarities of all observation instances using any type of base kernel function $k(x, x')$ and form their sum. A major asset of set kernels is the possibility of using any type of base kernel

function, which allows adaption to the data at hand. Since the size of the sets can vary, it is necessary to perform normalization. Otherwise, similarity would be biased according to set cardinalities, $|Z|$ and $|Z'|$. The normalized set kernel function is defined as

$$k_{\text{set}}(Z, Z') = \frac{k_s(Z, Z')}{f_{\text{norm}}(Z) \cdot f_{\text{norm}}(Z')},$$

where $f_{\text{norm}}(Z)$ refers to a normalization function. One can either perform normalization in feature space or opt for averaging. In our implementation, averaging was used, because there exist no significant performance differences between the two normalization methods [27]. In averaging, the normalization function is

$$f_{\text{norm}}(Z) = |Z|,$$

finally yielding the set kernel function,

$$k_{\text{set}}(Z, Z') = \frac{k_s(Z, Z')}{|Z| \cdot |Z'|}.$$

4.1.5 Validation and Cross-Validation

Validation

In validation, the performance of a trained model is evaluated empirically. To perform validation correctly, it is necessary to split the complete data set into two parts, training and test data. While models are formed on the training data, the test set is used to evaluate their performance. If model performance were not evaluated on an unseen test set, one would overestimate model performance, because the learned model was fitted to the same data. In addition to training and test set, it is necessary to consider another, separate data set, the *validation set*. This data set is used to estimate the parameters of the model. In our case, we need to consider just a single parameter for the SVM itself, namely ν. However, depending on the kernel function used, one needs to account for additional parameters, e.g., σ for the RBF kernel or motif length for string kernels. To determine these parameters, a technique called *grid search* is used. It is based on scanning a certain range of appropriate parameter values, rather than all possible values. For each kernel function, different parameter settings are used to train models on the training data. Then, the performance of these models with their specific parameter settings is determined on the validation set and the best parameter sets with regard to predictive performance are chosen. This process is called *parameter tuning*. Finally, the test set is used to determine the performance of each kernel with its selected parameters.

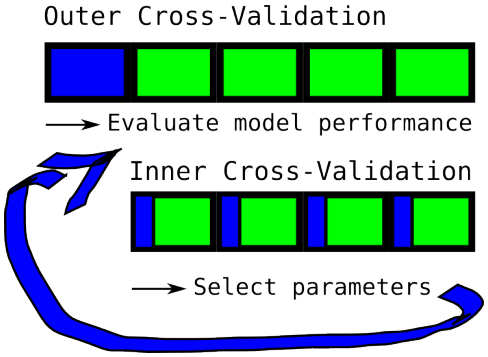


Figure 4.8: Nested cross-validation (CV). Folds in the data set are depicted by rectangular boxes. Whereas green boxes indicate the momentary training set, blue boxes show the current test set. In nested CV, the inner folds are used to perform parameter tuning, i.e., train a model and select that which performed best on the corresponding test sets. Then, in another run of CV, the outer folds are used to report the generalized performance of the model using the optimized parameters.

Cross-Validation

Cross-validation (CV) is used to accurately measure model performance without explicitly splitting the data into the three sets described above. Rather, the data set is divided into K folds. For each fold, the remaining folds are used to train a model, while the current fold is used to evaluate its performance. The CV error is given by the mean performance of all models. For parameter tuning, it is necessary to perform *nested CV*. Here, each of the K folds is further divided into K' folds. CV on the inner folds is used for parameter tuning. Then, using tuned parameters for each model, the outer folds are used to report the final model performance by performing another run of CV (Fig. 4.8).

Performance Measures for Regression

Two commonly used measures for predictive performance in regression are the mean squared error (MSE) and Pearson's coefficient of correlation, r. The MSE is given by

$$MSE = \frac{1}{N} \cdot \sum_{i=1}^{N} (f(x_i) - y_i)^2,$$

where $f(x_i)$ is the prediction for observation x_i, y_i is the observed outcome, and N is the total number of observation. A limitation of the MSE is its interpretability: Its range depends on the values of predictions and outcomes,

i.e., the range of the MSE is problem-specific. Pearson's correlation coefficient is in the range $[-1, 1]$ and gives the extent of the linear relationship shared by estimates and outcomes:

$$r = \frac{\sum_{i=1}^{N} \left(f(x_i) - \overline{f(x)}\right)(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N} \left(f(x_i) - \overline{f(x)}\right)^2} \sqrt{\sum_{i=1}^{N} (y_i - \overline{y})^2}},$$

where $\overline{f(x)}$ refers to the mean estimate and $\overline{y}$ to the mean outcome.

A correlation of 1 indicates perfect agreement, a correlation of 0 means that estimates and outcomes are not related at all, and a correlation of -1 reveals that outcomes and predictions share an inverse relationship. A benefit of correlation is that it is possible to assign p-values using Fisher's z-transformation. The transformation is given by $z = \frac{1}{2} \ln \frac{1+r}{1-r}$. Z-transformed correlations have a standard error of $SE = \frac{1}{\sqrt{N-3}}$. The test statistic, $Z = \frac{r_{\text{obs}} - r_{\text{test}}}{SE}$, can be used to verify whether the z-transformed observed correlation r_{obs} equals a specified z-transformed correlation r_{test}. By setting $r_{\text{test}} = 0$, one can determine the p-value for the hypothesis that an observed correlation was just due to chance.

4.1.6 Visual Analysis and Interpretation of Sequence Features

In section 5.1.4, we proposed that reads could play an important part in determining differences in measured methylation between Infinium 450K and WGBS technology. However, since we also considered other features, we tried to determine their individual impact on predictions using kernel-target alignment. To interpret which positions in the sequence play an important role for the prediction of $\Delta\beta$, we used Shannon entropy, motif discovery, and visualization of string kernels.

Kernel-Target Alignment

Kernel-target alignment is a method for multiple kernel learning (MKL), where one is concerned with the determination of coefficients weighting the contributions of individual kernel matrices to the result according to their predictive power. Usually, MKL is used for two reasons. First, by determining weights for individual kernel functions, it is possible to draw conclusions about the role each kernel plays for prediction. Second, MKL can possibly increase predictive power by upscaling the impact of important kernels with regard to prediction, while downscaling the importance of less important kernels. For this purpose, a weighted sum of kernel functions,

$$k(x_i, x_j) = \sum_{m=1}^{P} \mu_m \cdot k_m(x_i^m, x_j^m),$$

is used to represent similarity between observations. Here, P indicates the number of kernel functions considered, while μ_m represents the weight of kernel function k_m, which operates on feature vectors x_i^m and x_j^m. The weights μ are often regularized with L_1 or L_2 norms, but there are also general L_p-norm frameworks [36].

Kernel-target alignment was introduced by Cristianini et al. [14] in 2001. It is concerned with finding the similarity between a kernel and a target function to elucidate the agreement between a kernel and a learning task, finally giving rise to kernel weights μ_m. In the following, we introduce kernel-target alignment for classification first and then give the extension to regression.

Given labels $y \in \{-1, 1\}$, the optimal kernel function between two observations x and z is $k(x, z) = \mathbf{y}(x)\mathbf{y}(z)$. Hence, the alignment between a kernel K and the optimal kernel matrix $\mathbf{y}\mathbf{y}^T$ can be found by computing

$$\hat{A}(K, \mathbf{y}\mathbf{y}^T) = \frac{\langle K, \mathbf{y}\mathbf{y}^T \rangle_F}{\sqrt{\langle K, K \rangle_F \langle \mathbf{y}\mathbf{y}^T, \mathbf{y}\mathbf{y}^T \rangle_F}},$$

where the scalar matrix product is given by $\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^{N} K_1(x_i, x_j) \cdot K_2(x_i, x_j)$. The alignment $\hat{A}$ represents the cosine between K and the optimal kernel function.

For regression, one needs to modify the optimal kernel matrix $\mathbf{y}\mathbf{y}^T$ such that for each entry one has $\mathbf{y}'_i = \mathbf{y}_i - \overline{\mathbf{y}}$, where $\overline{\mathbf{y}}$ is the average outcome of the training set [34]. In addition, the generalization property needs to be verified. It can be shown that optimization of the alignment provides an upper bound on the ridge regression objective function, which forms a part of the generalization error [34, 13].

Shannon Entropy

Shannon entropy represents the expected value of the information contained in random variable observations [68]. It is typically measured in bits and, for a finite sample of a random variable X, given by

$$H(X) = -\sum_i \Pr(x_i) \log_2 \Pr(x_i).$$

To attain a better intuition about Shannon entropy, consider a coin toss scenario. If the coin is fair, that is, its probability for head and tail is 50% respectively, then the entropy is 1 bit per toss. If, however, one side is more likely than the other, entropy would decrease to less than 1 bit, indicating a reduction in uncertainty.

Motif Discovery with DREME

Discriminative DNA motif discovery (DREME) is a motif discovery tool based on Fisher's exact test [3]. Because we have already dealt with Fisher's

exact test in section 1.3.1, we will not give any more information here but refer to the publication instead.

Interpretation of String Kernels via Positional Oligomer Importance Matrices

One problem associated with non-linear kernel functions such as string kernels is that they are hardly interpretable. This is due to the application of the kernel trick, which affords only a weighting α over observations rather than features β. Positional oligomer importance matrices (POIMs) [73] overcome this shortcoming for string kernels by providing a powerful visualization tool similar to sequence logos. In contrast to just considering raw SVM weights α, POIMs consider the underlying correlation structure of k-mers to determine a more accurate notion of importance.

In SVMs, we have the classification function $\hat{y} = \text{sign}(f(x))$ where $f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b$ and α weighs individual observations. The main advantage of the SVM, namely its use of kernel functions to attain an implicit mapping into feature space, becomes its Achilles heel when it comes to the interpretation of features, for which an explicit mapping $\Phi(x)$ is necessary. However, utilizing string kernels, the explicit feature space mapping $\Phi(x)$ is available, e.g., for the weighted degree string kernel we have $f(x) = w \cdot \Phi(x) + b$ where $w = \sum_{i=1}^{N} \alpha_i y_i \Phi(x_i)$ is the normal vector of the separation in feature space. Each sequence can be mapped to a vector containing an entry for each pair of the form (k-mer, occurrence position) with a value of $\sqrt{\beta_k}$ if the k-mer exists at that position and a value of 0 if it does not at the corresponding position in the sequence.

Scoring System Each observation is a sequence $x \in \Sigma^k$ and formed according to the alphabet $\Sigma = \{A, C, G, T\}$. An oligomer (k-mer) is defined as $y \in \Sigma^k$. A positional oligomer (PO) is a pair of sequence and position, $(y, i) \in I := \cup_{k=1}^{K} \left(\Sigma^k \times \{1, \ldots, L - k + 1\}\right)$. A PO of order K ($k \leq K$) is scored with a weighting function $w : I \to \mathbb{R}$, giving rise to the score of an observation,

$$s(x) = \sum_{k=1}^{K} \sum_{i=1}^{L-k+1} w(x[i]^k, i) + b,$$

Here, $x[i]^k$ represents a subsequence of length k starting from position i. POIMs compute sequence likelihoods by using a zeroth-order Markov model approximation, in which all positions are independent of each other. To deal with overlaps, the importance of individual POs, $Q(z, j)$ is computed by considering the expected increase in score resulting from observing the oligomer z at position j.

Modes of Representation POIMs can be visualized in several ways, of which we will highlight two, POIM plots and differential POIMs. In positional oligomer importance matrix plots, POIMs of order $k \leq 3$ are visualized as heat maps. POs, (z, j), are illustrated by tabulating k-mers z as rows and positions j as columns. Cells are colored according to their importance, $Q(z, j)$. Since the number of k-mers explodes for large values of k, this type of illustration is only useful for small k.

Differential POIMs are used to show the gain in importance resulting from the consideration of longer k-mers. For this, one takes the maximum absolute importance over all k-mers of length k at position j. Then, the maximal importances of the two sets of $(k-1)$-mers covering the highest scoring k-mer at the current position are subtracted from its importance, giving rise to the differential POIM value.

4.2 Method Extensions

4.2.1 Hybrid String Kernels

In our string kernel computations, the input was solely based on the consensus sequences derived from aligned, windowed sequence reads. However, we have additional, numeric data that we would like to use to find a better representation of similarity between WGBS CpG positions in their sequence. Using frequency information it could be easier to find erroneous positions that otherwise vanish in the consensus. For a given position, if one observes cytosine eight times and adenine only once, this could be vital information, not represented by the consensus. Using the frequency of the consensus base $(\frac{8}{9})$, we obtain a measure of confidence for the called base.

The idea behind hybrid string kernels is to combine both features, numeric and string input. Since we already have many tools at our disposal to model the similarity of strings, we decided to scale string kernel weights according to the frequency profile of the sequence. To compare frequency profiles, cosine similarity, which will be introduced in the next section, seemed a useful measure. We pursued two approaches for hybrid string kernels. In the first, we just considered the frequency of the called consensus base among all observed base. In the second, we considered the full base frequency profile of the sequence, which is determined by the frequency of each possible base for every position in the sequence. Our expectation was that using the frequency information from non-consensus bases, we would be more sensitive to low-frequency base miscalls. Still, we were concerned that the consideration of all base frequencies could introduce too much noise. Consider matching consensus bases in two sequences with a frequency of $\frac{8}{9}$ for the consensus in each string. Now, assume that, at the given position, one of the strings has a frequency of $\frac{1}{9}$ for A and the other has a frequency of $\frac{1}{9}$ for C. In the first weighting approach, we would just use the frequency of the consensus, that

is, $\frac{8}{9}$. However, in the second weighting approach, we would also consider the frequencies of the other bases, i.e., $\frac{1}{9}$ for A in the first string and $\frac{1}{9}$ for C in the second string. While the first approach would yield a string similarity of 1, the second approach would just give us a similarity of approximately 0.98. If specific called bases other than the consensus impact the measurement of methylation, then the usage of all base frequencies should be superior to the mere usage of the consensus base frequency. On the other hand, if we would find that the first approach leads to better predictive models, we could conclude that specific base miscalls probably do not impact the emergency of methylation measurement errors.

Cosine Similarity

We use cosine similarity to represent the similarity between frequency profiles for two reasons. First, it is independent of vector magnitude. This guarantees that there does not exist a bias for sequences exhibiting large frequencies for most positions. Second, frequency vectors can have an angle of at most $\theta = 90°$ and, therefore, cosine similarity is in the range $[0, 1]$, which is a useful property for scaling string kernel weights. Cosine similarity,

$$s(x, x') = \cos(\theta) = \frac{\sum_{i=1}^{n} x_i x'_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} x'^2_i}}$$

represents the cosine of the angle between two real-valued vectors x and x'. Vectors pointing into similar directions exhibit a high cosine similarity and the maximum value is at $\cos 0° = 1$. Vectors pointing into different directions, on the other hand, have a larger angle, e.g., perpendicular vectors have $\cos 90° = 0$ (Fig. 4.9). In the following paragraphs, we will formalize

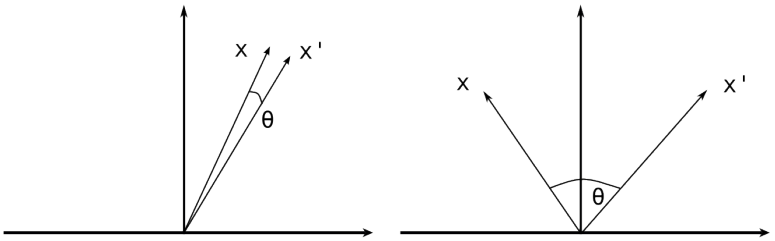


Figure 4.9: Cosine similarity. Vectors that are highly similar and thus point into similar directions have a small angle θ, resulting in high cosine similarity (close to 1). Vectors with low similarity, as portrayed on the right hand side point into different directions, resulting in a low cosine similarity (close to 0).

the concept of hybrid string kernels. Note that we do not explicitly distinguish between the two types of weighted approaches for hybrid string kernels.

Rather, frequency features $x_{\mathcal{F}}$ can be either vectors in $\mathbb{R}^{202}$ (consensus frequencies) or in $\mathbb{R}^{808}$ (frequencies of all bases). In addition, observations x are split into two parts; their sequence component $x_{\mathcal{S}}$ and their frequency component $x_{\mathcal{F}}$.

Hybrid Weighted Degree String Kernel

In the hybrid weighted degree kernel (HWDK),

$$k(x, x') = \sum_{k=1}^{d} \beta_k \sum_{i=1}^{|\mathcal{S}|-k+1} \mathbb{I}\left\{x_{\mathcal{S}}[i : i+k] = x'_{\mathcal{S}}[i : i+k]\right\} \cdot s\left(x_{\mathcal{F}}[i : i+k], x'_{\mathcal{F}}[i : i+k]\right),$$

we identify all matching motifs with lengths of $k \leq d$ and use weights β_k to decrease the impact of motifs according to their length k, as described by Sonnenburg et al. [71]. To integrate base frequencies into this kernel function, we scale the value of the indicator function according to the cosine similarity between the sequences' base frequencies at positions where we find matching motifs. Note that $x_{\mathcal{F}}[i : j]$ gives us the frequencies associated with positions i to j (exclusively), which can either be a vector in $\mathbb{R}^{j-i}$ for the consensus frequency approach or in $\mathbb{R}^{(j-i)\cdot 4}$ for the approach that considers all base frequencies.

Hybrid Weighted Degree String Kernel With Shifts

With the hybrid weighted degree kernel with shifts (HWDKS),

$$\begin{aligned} k(x, x') = \sum_{k=1}^{d} \beta_k \sum_{i=1}^{|\mathcal{S}|-k+1} \sum_{\substack{s=0 \\ s+i \leq |\mathcal{S}|}}^{S} \delta_s \cdot \big(&\mathbb{I}\{x_{\mathcal{S}}[i : i+k] = x'_{\mathcal{S}}[i+s : i+s+k]\} \\ &\cdot s(x_{\mathcal{F}}[i : i+k], x'_{\mathcal{F}}[i+s : i+s+k]) \\ &+ \mathbb{I}\{x_{\mathcal{S}}[i+s : i+s+k] = x'_{\mathcal{S}}[i : i+k]\} \\ &\cdot s(x_{\mathcal{F}}[i+s : i+s+k], x'_{\mathcal{F}}[i : i+k])\big) \end{aligned}$$

motifs with pairwise window distances $s \leq S$ can be found. All matching motifs between two consensus sequences $x_{\mathcal{S}}$ and $x'_{\mathcal{S}}$ with lengths $k \leq d$ are identified and motif length is scored using β_k, while shifts are penalized by δ_s. Frequency information is integrated by scaling the indicator function $\mathbb{I}$ for matching motifs according to the similarity of frequency profiles at the shifted positions.

Hybrid Oligo String Kernel

The hybrid oligo kernel (HOK) is given by

$$k(x, x') = \sqrt{\pi}\sigma \sum_{\omega \in \mathcal{A}^k} \sum_{p \in S_\omega^x} \sum_{q \in S_\omega^{x'}} \exp\left(-\frac{1}{4\sigma^2}(p-q)^2\right) \cdot s\left(x_{\mathcal{F}}[p : p+k], x'_{\mathcal{F}}[q : q+k]\right).$$

The set $\mathcal{A}^k$ gives all subsequences of length k present in the consensus sequences $x_{\mathcal{S}}$ and $x'_{\mathcal{S}}$. By setting ω, we fix a specific motif in that set and use the index variables p and q to specify the positions in $x_{\mathcal{S}}$ and $x'_{\mathcal{S}}$ at which ω occurs. The magnitude of the exponential function decreases with increasing squared distance $(p-q)^2$ between motifs and σ reflects the extent of positional uncertainty. In our hybrid extension, we additionally scale the Gaussian function according to the frequency profile agreement of the positions in the matching motifs such that motifs with divergent frequency profiles are downweighted.

Chapter 5

Workflow

In this chapter, we delineate the workflow of this project including data preprocessing, generation of data sets, and the application of statistical methods. In preprocessing we deal with selecting only those reads covering Infinium 450K positions, calling methylation, and normalization. We then consider the different data sets generated by sampling and the prediction models. Finally, we cover strategies to interpret these models, among them kernel-target alignment, Shannon entropy, motif discovery, and POIMs.

5.1 Overview of Data Preprocessing

Data from WGBS and Infinium 450K technology is different in various aspects. A fundamental difference between the two types of measurements is data size. Whereas WGBS measures the full genome (about 3.2 Mb), the chip measures merely a selected set of CpG positions (about 0.485 Mb). For supervised statistical learning we are required to have measurements from both methods. Hence, we retain only WGBS measurements that are also available on the Infinium 450K chip. Features were derived from WGBS data, while outcomes were constructed from the measured methylation levels of both methods.

Infinium measurements represent raw color intensities, while WGBS data consists of aligned reads. To obtain beta-values, it is necessary to perform methylation calling for each of the methods. In addition, for Infinium 450K data, normalization between probe types is necessary to adjust for the smaller dynamic range and inaccuracy of type II probes. Filtering of defective probes, e.g., those with SNPs, extreme intensities, or a large number of CpGs, might be helpful. Furthermore, it may be beneficial to fit the Infinium beta-value distribution to that of WGBS for better comparability. Finally, after low-level processing, data were converted to a machine learning-suitable format, i.e., a feature matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$ with N observations and p features.

5.1.1 Intersecting Infinium 450K and WGBS Loci

Since we are only interested in data points for which we have both, the value from WGBS and Infinium 450K, we extracted data from WGBS, retaining only those WGBS reads overlapping with at least one Infinium 450K position (Fig. 5.1). To achieve this, using the R package RnBeads (`http://rnbeads.mpi-inf.mpg.de/`), we generated a BED file indicating all positions measured by the chip. We then applied *samtools 0.1.18-dev* to generate the WGBS BAM file of intersecting reads. In the next sections, we

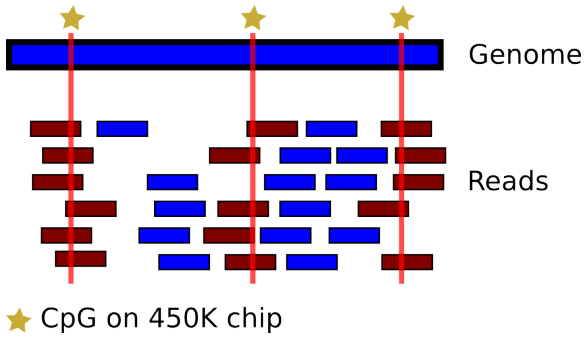


Figure 5.1: Intersection of WGBS and Infinium 450K data. The blue rectangle at the top represents a genome excerpt in which CpG positions measured by the array are tagged by stars. Below the genome are the WGBS reads. Reads that did not overlap with Infinium 450K positions (blue) were discarded. Only reads overlapping with Infinium 450K sites (red) were retained.

describe the processing steps required to obtain beta-values for both data sets, starting with Infinium 450K first.

5.1.2 Processing Infinium 450K Data

For Infinium 450K data, beta-values are determined from color intensities. In addition, it is important to correct for the different methylation distributions of probe types I and II. Adjustment to the WGBS beta-value distribution and filtering according to probe intensities, number of CpGs per probe, and SNPs are further, possibly helpful steps.

Methylation for the Infinium 450K chip was called using RnBeads 0.99.10 yielding β_{450K} as defined in Section 2.2.3. We used BMIQ [76] to adjust the distribution of type II probes to that of type I probes, because it compared favourably with other normalization strategies and is assumption free [16, 76].

Fitting Infinium 450K to WGBS Beta-Values

While methylation values reported by Infinium 450K are in the bounded range $(0, 1)$, WGBS values are in the range $[0, 1]$. Normalization was used to deal with these distribution differences. Due to the sound statistical framework of BMIQ we were able to fit the distribution of 450K beta-values to that of WGBS by treating them as probe type II and I, respectively. Specifically, we set $t = (\text{WGBS}, 450K)$ instead of $t = (I, II)$ in the BMIQ procedure outlined in Section 4.1.1.

Filtering of Inaccurate Infinium 450K Probes

Filtering of 450K probes was done according to three criteria. The first was probe intensity: Values with extreme intensities are known to often be erroneous and should therefore be excluded. The second criterion concerns SNPs: Probes containing SNPs should be excluded because SNPs can obscure the true methylation signal when they occur at a CpG position and can exacerbate probe hybridization. The third criterion was CpGs: Probe II types containing multiple CpGs encompass degenerate bases to allow hybridization independent of methylation status. However, if more than three CpGs are present on a single probe, binding affinity is reduced to a critical level. To handle the three aspects, we used RnBeads' greedy cut procedure to eliminate probes with extreme intensities, removed all probes containing SNPs, and filtered probe II reads containing more than three CpGs.

Please note that these two preprocessing steps were only performed for some data sets, which are discussed in more detail in Section 5.2.2.

5.1.3 Processing WGBS Data

WGBS data were processed with Bis-SNP 0.81.2 to call methylation and determine SNPs with the the options *-stand_call_conf 20*, *-stand_emit_conf 0*, *-mmq 30*, *-mbq 0*, and *-S LENIENT*. As a reference genome, we used the current version of the human genome, hg19, and, as a reference for SNPs, we deployed the current dbSNP release, dbSNP 137 [69]. A listing of those read IDs involved in calling methylation for individual CpGs was generated via the *-cpgreads* and *-notEncrypt* options in Bis-SNP.

5.1.4 Defining the Outcome and Feature Extraction

For supervised machine learning, the input data set needs to be in a specific format, in which every observation consists of an outcome and an associated set of features. In the following paragraphs, we describe how we defined the outcome in this learning scenario and how features were extracted from the data set.

Defining the Outcome

For every CpG, the outcome was formed by setting

$$\Delta\beta = \beta_{450K} - \beta_{\text{WGBS}}$$

to represent the difference in methylation between the Infinium 450K chip and WGBS.

Mapping Reads to Infinium 450K CpGs and Windowing of CpGs

Using the list of reads employed for calling methylation for CpGs by Bis-SNP, we associated every CpG on the chip with its set of overlapping reads, see Fig. 5.2. We then dealt with finding a uniform representation of CpGs according to reads. For this, we considered windows spanning the neighborhood of the CpG. Since, for our data set, the maximum read length was $n_{\text{max}} = 101$, we considered windows of size $n_{\text{win}} = 202$ to capture reads extending to the left starting from the cytosine position as well as reads extending to the right starting at the guanine position (Fig. 5.3). As we saw

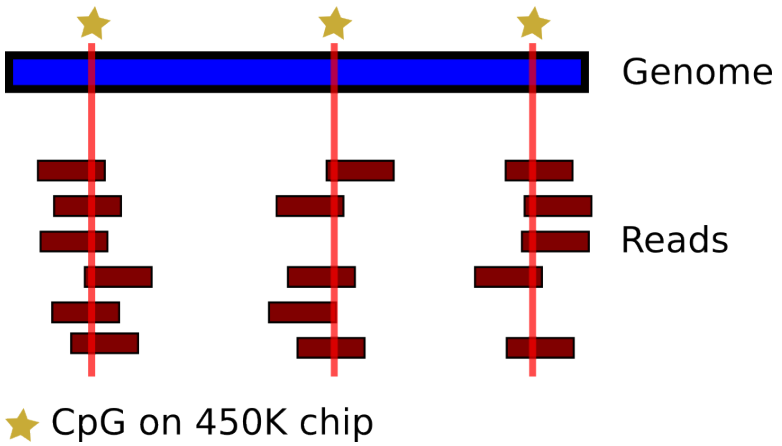


Figure 5.2: Mapping WGBS reads to 450K CpGs. Stars represent loci in a genome for which measurements of the Infinium 450K chip exist. For each of these positions, we gathered all overlapping WGBS reads that were used in calling methylation via Bis-SNP.

in Section 2.2.4, WGBS reads exhibit certain error patterns, e.g., caused by lagging strands or conversion errors due to bisulfite treatment. We therefore propose that certain characteristics in the sequence impact the accuracy with which methylation is detected by next-generation sequencing technology, thereby possibly impacting the outcome $\Delta\beta$. To use sequences as a feature, it was necessary to align reads and take their compact idiosyncratic gapped alignment report (CIGAR) strings into account.

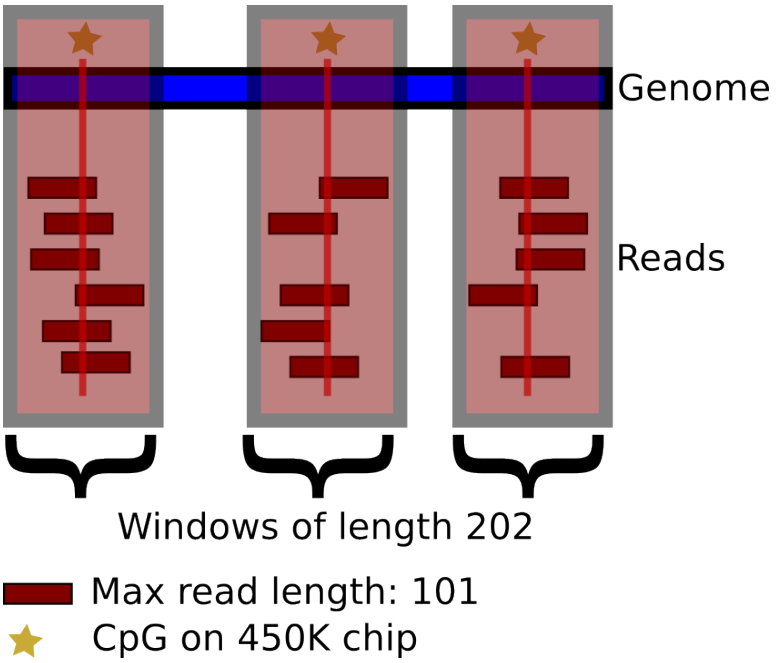


Figure 5.3: Windowing of CpGs. For each set of WGBS reads, we generated sequence windows of 202 in order to arrive at a uniform representation of sequence features. To attain these windows, reads were aligned according to their CIGAR strings and gap characters were introduced for positions that were not covered by a read.

CIGAR Strings CIGAR strings provide a method for annotating read alignment with respect to a reference genome. Standard CIGAR strings are composed of just three events: M, for a match or mismatch, I, for an insertion, and D for a deletion. A CIGAR string describes the full alignment of a read to the reference by prefixing each of the event symbols, M, I, and D, with a value indicating the number of bases subject to that event. Fig. 5.4 exemplifies the format of CIGAR strings. Extended CIGAR strings also include the event symbols N for skipped bases, S for soft clipping, H for hard clipping, and P for padding. Clipping can occur during the Smith-Waterman algorithm when a sequence cannot be fully aligned such that subsequences at the end are clipped off. Soft clipping refers to non-agreeing bases that were clipped from the alignment, but are still contained in the sequence of the read. Hard clipping, on the other hand, refers to bases that were fully removed from the alignment. Padding is a special operation that is used to represent the alignment of inserted regions and is mainly used for de novo sequencing, which is why we are not concerned with this event.

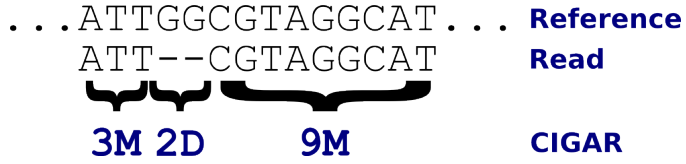


Figure 5.4: Format of CIGAR strings. The sequence at the top represent a reference genome excerpt; the sequence below represents an aligned read. Matches and mismatches are indicated by M, while deletions in the read with respect to the reference are indicated by D. Each event is prefixed by its number of occurrences, giving rise to the CIGAR string $3M2D9M$.

Read Alignment For read alignment, let i_{450K} represent the position of the Infinium 450K CpG, more exactly, the position of the cytosine. Furthermore, let i_{WGBS} represent the start position of a read overlapping with the current CpG. To determine the start of the read with respect to the window, we form an offset of the form $o = i_{\text{WGBS}} - i_{450K}$, which represents the number of positions that the read protrudes from the CpG. If the read starts at the cytosine, the offset will be 0, if it starts on the guanine, it will be 1, and otherwise it has a negative value ($\geq 1 - n_{\max}$). The start position of the read in the aligned window (starting from 1) is given by $i'_{\text{WGBS}} = n_{\max} + o$.

For the rest of the alignment, we need to consider the CIGAR string of the read. Let $x = i'_{\text{WGBS}}$ be the current position in the window and let $y = 1$ be the current position in the read. We iterate over the CIGAR string of the read and set the sequence window entry w for that read according to the events in the CIGAR string. For the window, let $w[a : b]$ represent the subsequence of w that starts at a and ends before b. Furthermore, let r represent the sequence of the read. Window sequences w are made up of characters from the alphabet $\mathcal{A} = \{A, C, G, T, -\}$, where $-$ is used to represent gaps. Initially, we set $w[1 : 203] = -$ such that positions before the start of the read and after the end of the read are represented by gap characters. For read alignment we differentiated between the different CIGAR events:

- n matches/mismatches (M): set $w[x : x + n] = r[y : y + n]$, $x = x + n$, $y = y + n$
- n deletions (D): set $w[x : x + n] = -$, $x = x + n$
- n insertions (I): ignore the following bases by setting $y = y + n$
- n soft skips (S): set $w[x : x + n] = r[y : y + n]$, $x = x + n$, $y = y + n$
- n hard skips (H): set $w[x : x + n] = -$, $x = x + n$
- n skipped bases (N): set $w[x : x + n] = -$, $x = x + n$

This procedure was performed until the end of every CIGAR string was reached. Note that we treated deletions, hard skips, and skipped bases in the same way by introducing gaps, because skipped bases are not present in the reads. Soft skips, on the other hand, are contained in the read sequence and were therefore taken into account. Since we required to have observations of the same length for usage in a statistical framework, we had to discard all inserted bases. Alignment of reads yielded, for each CpG, a matrix $W \in \mathcal{A}^{n \times n_{\text{win}}}$, where n is the number of reads overlapping with a given CpG. Since conventional learning approaches require features to be vectors rather than matrices, another conversion step was necessary.

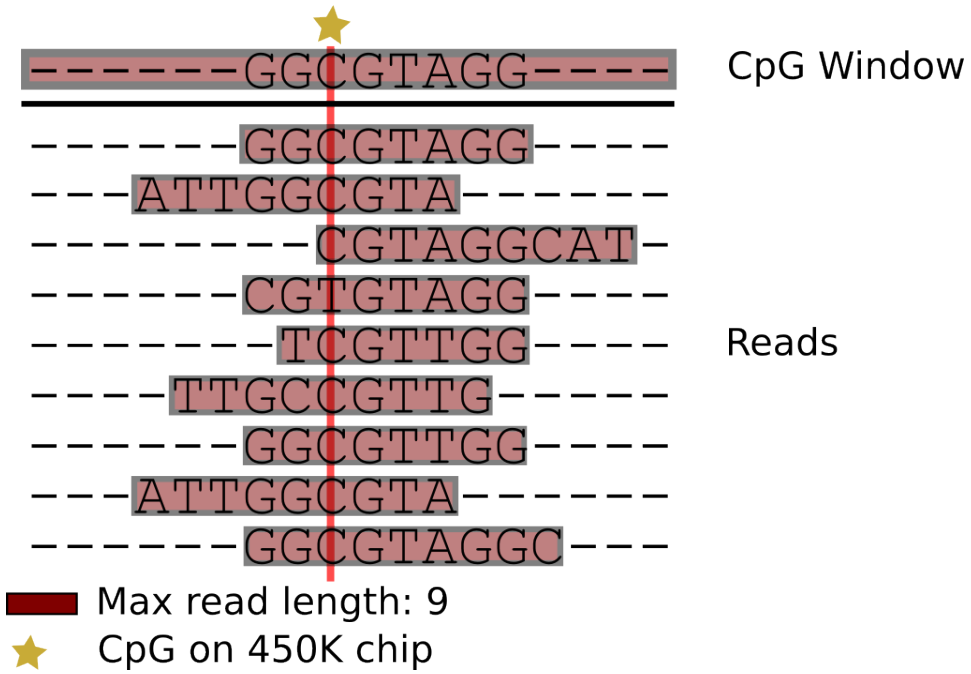


Figure 5.5: Consensus sequence formation. For each position, the most frequent character is chosen to form the consensus sequence, which is indicated at the top.

Forming Consensus Sequences Consensus sequences provide a method of condensing the information contained in a set of sequences $\{W_i\}$ into a single sequence. The consensus sequence is formed by determining, for each position, the character with the highest frequency (Fig. 5.5). A numeric representation of consensus sequences was obtained by binary-coding of bases, resulting in a feature vector of length $4 \cdot n_{\text{max}}$. In this encoding, gaps were not modeled explicitly, but rather represented by setting the frequency of every base to 0. The structure of the resulting windowed sequences is illustrated in Fig. 5.6.

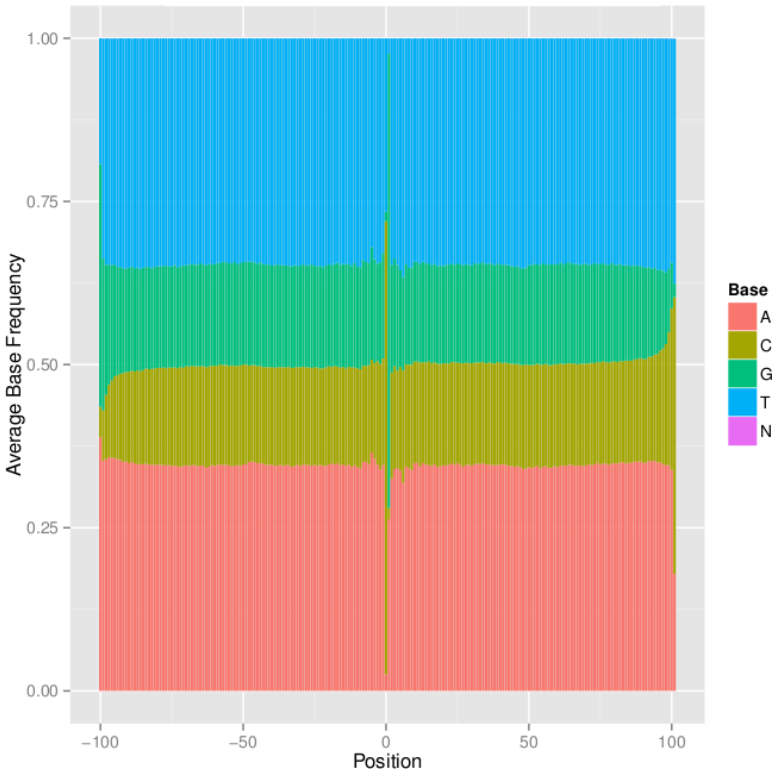


Figure 5.6: Structure of sequence windows. For every position in the consensus sequence window, the corresponding base frequency is indicated by stacked bars. The CpG is located at positions 0 and 1 on the x-axis and is characterized by an overrepresentation of C (methylated) and T (unmethylated) at the cytosine position and an overrepresentation of G (methylated) and A (unmethylated) at the guanine position. The other positions show typical distributions for the most part. The overall lower frequency for C and G is characteristic for the known underrepresentation of CpG dinucleotides in vertebrate genomes. The biased tails at the borders of the frequency distribution could be a consequence of library preparation, e.g., due to cleaving of adapter sequences.

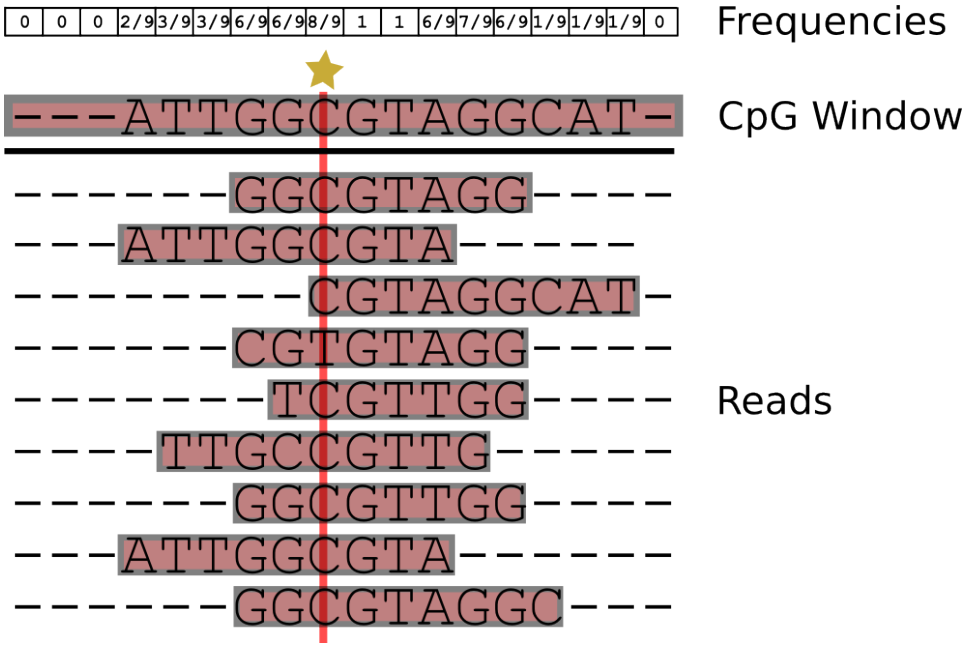


Figure 5.7: Consensus-frequency approach. Usage of frequency information for the consensus base and modified formation of the consensus sequence. Here, gap character are only introduced if no other characters are present at a position.

Encoding Frequency Information In addition to the sequences themselves, we were also interested in modeling the frequency of each base, for which we considered two scenarios. In the first, we just extracted the frequencies of the bases in the consensus sequence, resulting in an $\mathbb{R}^{n_{\max}} = \mathbb{R}^{202}$ vector. In the second, we consider the frequencies of every possible base, resulting in an $\mathbb{R}^{4 \cdot n_{\max}} = \mathbb{R}^{808}$ vector. The first approach corresponds to setting the frequencies of the minor bases (those that were not the most frequent) to 0.

In cases where we considered frequency information in conjunction with sequence information, we use a modified consensus approach. Instead of picking the most frequent character (including the gap), we selected the most frequent base. This is because bases confer more information about the characteristics of the sequence than gaps and the availability of base frequencies allows appropriate weighting of motifs. See Fig. 5.7 and 5.8 for an illustration of the two frequency-based approaches and their consensus formation.

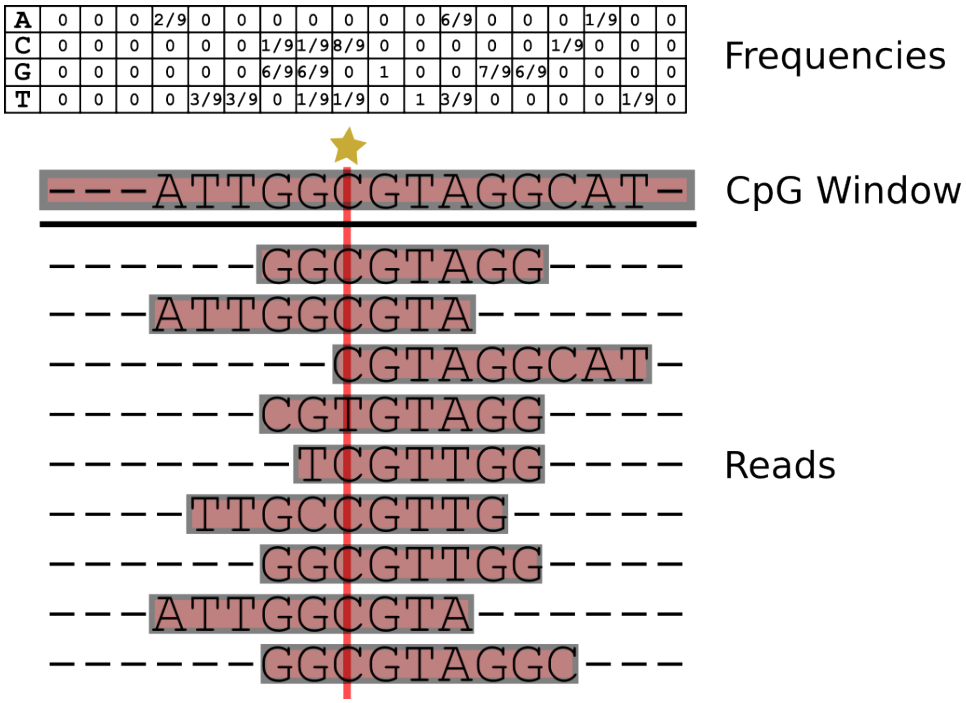


Figure 5.8: Base frequency approach. Usage of positional frequencies for each base in addition to the consensus sequence. Consensus formation was the same as for the consensus-frequency approach.

Masking the CpG position To examine the impact of the CpG dinucleotide position on the prediction of $\Delta\beta$, we masked these positions by setting their frequencies to 0 and introducing gaps in the windowed sequences.

Non-Sequence Features

Beside the read sequences, we considered global properties of reads as well as positional depth of coverage, positional base quality, and functional annotations as features for predicting $\Delta\beta$.

Features were regarded as *global* when they did not represent position-specific information, but information integrated over the whole extent of reads overlapping with a CpG. This set of features was comprised of mean depth of coverage, mean base quality, and ratio of forward reads among all reads, all combined in a single vector, in which each feature was rescaled to the range $[0, 1]$ ensuring comparability.

In addition to the global properties, we also studied position-specific features, namely depth of coverage, base qualities, and functional annotations. These annotations are categorical features from ENSEMBL [25], which were computed using hidden Markov models (HMMs) [22, 23], as well as CPI annotations from RnBeads. The following classes were used: active promoter,

weak promoter, poised promoter, strong enhancer, weak enhancer, insulator, transcription transition, transcription elongation, weak transcription, repressed, heterochromatin, repetitive/copy number variation (CNV), and CPI/no CPI.

We expected depth of coverage to play a role in the emergence of methylation measurement inconsistencies, because a low sampling rate increases the probability of sequencing errors and might lead to underrepresentation of certain alleles. Base quality indicates the probability of a miscall and, hence, low quality positions might imply incorrect methylation calls. Furthermore, due to the strong impact of context-sensitive errors exposed in the work from Allhoff et al. [1], we also considered the ratio of forward reads because we would expect more errors in positions with biased ratios exhibiting motifs associated with CSEs.

5.2 Generation of Data Sets

5.2.1 Sampling

The full data set comprised more than $485,000$ observations. Hence, due to computational limitations, it was necessary to perform sampling. To build a representative model that is able to deal with inputs of varying characteristics, we used stratification in combination with principal component analysis (PCA).

Stratification

In random sampling, one samples from the whole population. In stratified sampling, on the other hand, a sample is drawn for every subpopulation. Such a subpopulation is called a *stratum*, giving rise to the procedure's name. Stratified sampling is justified when one expects subpopulations to exhibit characteristic distributions, while ensuring that enough samples of each distribution are retained. We defined strata according to differences in methylation, $\Delta\beta$, between 450K and WGBS. Since $\Delta\beta$ is bounded by $[-1, 1]$, the strata were formed by subdividing this interval into 20 stretches. For each of the 20 strata we then performed PCA on the sequence features to ensure that we did not randomly exclude observations characteristic of differences in methylation.

PCA and Stratified Sampling with PCA

PCA is a statistical technique to reduce the dimensionality of multi-variable observations. Using an orthogonal, linear transformation in the input space, one obtains linearly uncorrelated variables, called principal components (PCs). Each PC represents the highest variance in the data under the constraint that

it is orthogonal to all preceding PCs, i.e., the first PC accounts for the greatest degree of variability, while the second PC accounts for the second largest degree of variability, and so on.

For sampling with PCA, we first selected, for every stratum, a number of PCs. We included PCs until we exceeded 95% of explained variance, but considered no more than five PCs at a time. We then sampled, for each stratum's selected PCs, the observations with the largest absolute projected values. The number of observations sampled per PC was determined by its contribution to the total explained variance of all selected PCs.

5.2.2 Data Set Overview

Based on the described sampling approaches, we generated several data sets on which we trained models. For our first analysis, we performed stratified sampling via PCA on the binary-coded consensus sequences to generate a data set of $12,598$ observations (*Baseline*). Next, we generated three data sets of sizes $9,098$, $11,660$, and $10,347$(*FNS-C*, *FNS-F*, *FNS-AF*), each of which was sampled using PCA. This time, however, PCA was used on binary-coded consensus sequences, consensus base frequencies, and all base frequencies, respectively. In contrary to the baseline data set, the *FNS* data sets were filtered according to the criteria described in Section 5.1.2 and labels were adjusted according to normalized Infinium 450K methylation values, see Section 5.1.2. Furthermore, for the full base frequency data set, we considered non-normalized Infinium 450K probes (*I-FNS-AF*, *II-FNS-AF*) and built prediction models for each probe type. To evaluate the performance of set kernels, we generated a data set (*Set*) in which each observation represents an individual read, rather than a consensus sequence by treating every row W_i in the alignment matrices as an individual data point, leading to a data set of $67,359$ reads representing $3,825$ CpGs. For this data set we did not use PCA for sampling. Table 5.1 gives an overview of all data sets and used kernel functions.

5.3 Application of Statistical Methods

Our motivation for using SVMs as the predictive model is that they are flexible and well-suited for dealing with high-dimensional data, e.g., for vectors in $\mathbb{R}^{808}$. Both properties are a consequence of the usage of kernel functions, which afford the mapping of input data into a higher-dimensional space, the *feature space*, which allows capturing of non-linear relationships. Since we aim at predicting $\Delta\beta$, the difference in methylation between the Infinium 450K chip and WGBS, our outcome is not categorical but quantitative, i.e., real-valued. Therefore, we employed SVR, which is contingent on numerical outputs y_i.

5.3.1 Overview of Trained Models

Using ν-SVR, we trained models employing different kernel functions and features. With regard to kernel functions, we can differentiate between four categories: numeric kernels (linear, polynomial, and RBF), string kernels (WDK, WDKS, and OK), hybrid versions of string kernels (HWDK, HWDKS, and HOK), and set kernels, which make use of the mentioned numeric and string kernels. See Table 5.1 for a full overview of applied models and their corresponding data sets. Concerning features, we used depth of coverage (*Cov*), functional annotations (*Fun*), base qualities (*Q*), and sequence features (consensus *C*, consensus frequency *F*, all frequencies *AF*), for the *Baseline* data set. All of the other models are just based on sequence features. For filtered, normalized (Infinium 450K to WGBS), and sampled (feature-specific PCA) data sets *FNS* (*FNS-C*, *FNS-F*, *FNS-AF*), we applied numerical kernels and string kernels, as well as their hybrid versions. To evaluate the predictive performance of the set kernel *S*, we used numeric and string kernels as base kernels and applied them to the *Set* data set. Impact of intra-array normalization with regard to probe type was studied using numeric and hybrid string kernels on the two probe type-specific data sets *I-FNS-AF* and *II-FNS-AF*, giving rise to the two approaches *I-AF* and *II-AF*. The impact of the CpG dinucleotide position on predictive performance was studied on the *Set* and three *FNS* data sets, leading to the masked approaches *M-S*, *M-C*, *M-F*, and *M-AF*. For the three masked *FNS* approaches we used all numerical kernels but only the best performing string kernel (to limit runtime), while for *M-S* we used all of the numeric and string kernels.

Grid search was used to determine the best kernels and SVR parameter settings, for which we let ν range from 0.2 to 0.8. Table 5.2 gives an overview of the used kernel functions and their considered parameters. The used set kernels are not listed in this able, since their base kernels employed the same parameter ranges as the corresponding numeric and string kernels. The predictive accuracy of the learned SVR models was evaluated using 10 runs of nested cross-validation with 10 folds.

Table 5.2: Kernel functions and parameters. The parameter σ gives the width of the RBF and oligo kernels, d gives the maximum substring length for weighted string kernels and the exact substring length for the oligo kernels, mm is the maximal number of allowed mismatches for weighted kernels, and s indicates the shift of the WDKS and HWDKS.

Type	Kernel	σ	d	mm	s
Numeric	Linear	-	-	-	-
	Poly	-	$2, 3, 4$	-	-
	RBF	Fitted	-	-	-
String	WDK	-	$3, 6, 10$	$0, 1, 2$	-
	WDKS	-	$3, 6, 10$	$0, 1, 2$	$0, 1, 2, 3$
	OK	$0.5, 1.2, 2, 3, 4$	$3, 4, 5, 6$	-	-
Hybrid	HWDK	-	$3, 6, 10$	$0, 1, 2$	-
	HWDKS	-	$3, 6, 10$	$0, 1, 2$	$0, 1, 2, 3$
	HOK	$0.5, 1.2, 2, 3, 4$	$3, 4, 5, 6$	-	-

Training and evaluation of models was performed with R 2.15.1. Kernel matrices for string kernels, hybrid string kernels, and set kernels were computed using Python 2.6.6 with a modified version of the Shogun Machine Learning Toolbox 2.0.0, written in C++. Modifications to Shogun include the incorporation of the gap character in the alphabet, hybrid string kernels, and an implementation of set kernels. To deal with gaps, every substring containing a gap character was ignored, because we did not expect the location of reads in the sequence window to play a role in determining similarity of observations with regard to $\Delta\beta$.

5.3.2 Interpretation of Features and Models

To determine the role of individual features, we used kernel-target alignment. Interpretation of sequences was facilitated by Shannon entropy, motif discovery, and visualization through POIMs.

Kernel-Target Alignment

In one of our initial analyses we considered multiple features to predict differences in methylation, $\Delta\beta$, namely sequence features, global features, qualities, coverage, and functional annotations. To determine the role of every feature, we used kernel-target alignment to afford a feature weighting by using the best-performing kernel function for each feature. We employed the WDKS with degree 3, maximum mismatch 1, and shift 2 for consensus sequences, the polynomial kernel of degree 2 for quality and global features, and the RBF kernel for coverage as well as functional annotations. As a second consideration, we hoped that the combined kernel matrix derived

from kernel-target alignment might increase the predictive performance of the resulting SVR model. We used the kernel-target alignment implementation from the openkernel package (`http://www.openkernel.org/`) with the *-alignf* option, for which we had to convert R kernel matrices to libSVM format using the *R.matlab* package and Matlab [8]. We chose kernel-target alignment as a means for MKL, because it performed comparable to other methods and was readily available [12, 40].

Computing Shannon Entropy for Sequences

To obtain the Shannon entropy for each position in the sequence window, we set $\Pr(x_i)$ as the frequency of the consensus base and computed the mean entropy for each position.

Motif Discovery in Sequences with DREME

We used the DREME online platform `http://meme.nbcr.net` to discover motifs that were enriched in loci associated with consistent vs. inconsistent methylation measurements. Since DREME is based on differentiating two classes, we needed to convert from a regression to a classification problem. All CpG loci with $|\Delta\beta| < 0.3$ were assigned to the class of consistent measurements, while those with $|\Delta\beta| \geq 0.3$ were assigned to the other class, representing inconsistent measurements. The class cutoff value was chosen in analogy to a paper from Sproul et al. [74], where genes were considered unmethylated with $\beta \leq 0.3$ and methylated with $\beta \geq 0.7$.

Visualization of String Kernels with POIMs

To visualize important positions for predicting $\Delta\beta$ with string kernels, POIMs were used. The benefit of using shorter or longer motif sizes was studied with differential POIMs. To employ POIMs, we transformed our problem into the classification domain in the same way as for DREME, explained in the previous paragraph.

Table 5.1: Overview of data sets and trained models. *Cat* refers to a category of approaches that use similar data or features. *Data* refers to the used data set. *Approach* indicates the features used for training models. Tick marks indicate used SVR kernel functions, while crosses label kernel functions that have not been used. Note that the *Set* kernel function is a short hand for set kernels with base kernels of either type linear (Lin), polynomial (Poly), radial basis function (RBF), weighted degree kernel (WDK), weighted degree kernel with shifts (WDKS), or oligo kernel (OK). Hybrid weighted degree kernel (HWDK), hybrid weighted degree kernel with shifts (HWDKS), and hybrid oligo kernel (HOK) refer to hybrid versions of the three string kernels. In the *Approach* column, we have the features *G* (global properties), *Cvg* (depth of coverage), *Q* (base qualities), *Fun* (functional annotations), *C* (consensus), *F* (consensus frequencies), and *AF* (base frequencies). For data sets, *Baseline* describes our base data set, while the *FNS* data sets represent filtered, normalized, and sampled (feature-specific PCA) data sets. Approaches labeled with *I* and *II* refer to models that are based on non-normalized Infinium probe I and probe II data, respectively. Approaches prefixed with *M* did not facilitate the CpG dinucleotide position for predictions. *Set* data sets have individual reads as observations rather than CpGs.

			Kernel Function									
Cat	Data	Approach	Lin	Poly	RBF	WDK	WDKS	OK	HWDK	HWDKS	HOK	Set
Base	Baseline	G	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
	Baseline	Cvg	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
	Baseline	Q	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
	Baseline	Fun	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
	Baseline	C	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
	Baseline	F	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
	Baseline	AF	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
FNS	FNS-C	FNS-C	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗
	FNS-F	FNS-F	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
	FNS-AF	FNS-AF	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
I/II	I-FNS-AF	I-AF	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
	II-FNS-AF	II-AF	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗
Masked	FNS-C	M-C	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
	FNS-F	M-F	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
	FNS-AF	M-AF	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
	Set	M-S	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Set	Set	S	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓

Chapter 6

Results

In this chapter, we present the results of the analyses described in Chapter 5. We deal with findings from preprocessing, performance of trained ν-SVR models, and the upshot of feature interpretation and visualization.

6.1 Preprocessing

Preprocessing of data entailed BMIQ-normalization [76] of probe types for Infinium 450K data, adjustment of the 450K β-distribution to that of WGBS, and filtering of 450K probes and PCA-based sampling.

The BMIQ-normalized beta-value densities for Infinium 450K are illustrated in Fig. 3 in the appendix. Normalization slightly improved the correlation of Infinium type II loci with WGBS from $r \approxeq 0.937$ to $r \approxeq 0.939$. We found that filtering probes according to the criteria discussed in Section 5.1.2 did not increase correlation of Infinium 450K beta-values with those from WGBS. See Fig. 4 in the appendix for the BMIQ-normalized Infinium 450K methylation values with regard to WGBS.

Regarding sampling, PCA did not find a reasonable dimension reduction considering that the first five PCs usually captured only 10% of the total variance, see Fig. 5 in the appendix. This finding was independent of the features used for PCA, that is, either binary-coded sequences, consensus frequencies, or all base frequencies.

6.2 Model Performance

In the following, we report the predictive performance of ν-SVR models for the individual feature sets with regard to $\Delta\beta$, the difference in measured methylation between Infinium 450K and WGBS. We selected the best-performing models according to the correlation r between outcomes and predictions in 10 runs of 10-fold cross-validation. We first deal with the performance of sequence features and then evaluate non-sequence features.

Table 6.1: Result overview. *Data* gives the data set used to build models, *Approach* refers to the set of features, and *Model* gives the best performing ν-SVR model (according to Pearson correlation r) for the corresponding set of features. The given p-values reflect the probability that the true correlation between predictions and outcomes is 0. In the *Approach* column, we have the features G (global properties), *Cvg* (depth of coverage), Q (base qualities), *Fun* (functional annotations), C (consensus), F (consensus frequencies), and AF (all base frequencies). For data sets, *Baseline* describes the base data set, while *FNS* represents filtered, normalized, and PCA-sampled data sets. Approaches labeled with either I or II refer to models based on non-normalized Infinium probe I and probe II measurements, respectively. Approaches prefixed with M did not facilitate the CpG dinucleotide position for predictions. *Set* data sets have individual reads as observations rather than CpGs.

Data	Approach	Model	ν	r	p-value
Baseline	G	RBF ($\sigma \approxeq 0.21$)	0.6	0.095	$< 2.2 \cdot 10^{-16}$
Baseline	Cvg	RBF ($\sigma \approxeq 0.001$)	0.6	0.07	$2.776 \cdot 10^{-15}$
Baseline	Q	Polynomial ($d = 2$)	0.4	0.055	0.1315
Baseline	Fun	Polynomial ($d = 2$)	0.6	0.15	$< 2.2 \cdot 10^{-16}$
Baseline	C	WDKS ($d = 3$, $mm = 1$, $s = 2$)	0.6	0.41	$< 2.2 \cdot 10^{-16}$
Baseline	F	RBF ($\sigma \approxeq 0.002$)	0.4	0.62	$< 2.2 \cdot 10^{-16}$
Baseline	AF	RBF ($\sigma \approxeq 0.0005$)	0.3	0.65	$< 2.2 \cdot 10^{-16}$
FNS-C	FNS-C	WDKS($d = 3$, $mm = 1$, $s = 2$)	0.6	0.41	$< 2.2 \cdot 10^{-16}$
FNS-F	FNS-F	Polynomial ($d = 2$)	0.5	0.62	$< 2.2 \cdot 10^{-16}$
FNS-AF	FNS-AF	Polynomial ($d = 2$)	0.7	0.66	$< 2.2 \cdot 10^{-16}$
I-FNS-AF	I-AF	Polynomial ($d = 2$)	0.7	0.66	$< 2.2 \cdot 10^{-16}$
II-FNS-AF	II-AF	RBF ($\sigma \approxeq 0.0004$)	0.6	0.61	$< 2.2 \cdot 10^{-16}$
FNS-C	M-C	WDKS ($d = 10$, $mm = 0$, $s = 3$)	0.7	0.22	$< 2.2 \cdot 10^{-16}$
FNS-F	M-F	HWDK ($d = 10$, $mm = 1$)	0.7	0.2	$< 2.2 \cdot 10^{-16}$
FNS-AF	M-AF	HWDK ($d = 10$, $mm = 0$)	0.4	0.23	$< 2.2 \cdot 10^{-16}$
Set	M-S	WDKS ($d = 10$, $mm = 1$, $s = 3$)	0.3	0.37	$< 2.2 \cdot 10^{-16}$
Set	S	WDKS ($d = 10$, $mm = 2$, $s = 3$)	0.2	0.73	$< 2.2 \cdot 10^{-16}$

Table 6.1 gives an overview of the best-performing model for each feature and data set. One thing to note about this table are the p-values for correlations. They reflect the likelihood that the true correlation between predictions and outcomes is 0, see Fisher's Z-transformation in Section 4.1.5. It is evident that for large data sets, leading to high z-distribution standard errors, the test statistic Z will be large. Additionally, since we have $r_{\text{test}} = 0$, this leads to even larger values for Z and the strength of its associated p-value, explaining the significant correlation found for nearly all sets of features. More detailed results for individual feature and data sets can be found in the appendix, in Tables 1 (approach *G*), 2 (approach *Cvg*), 3 (approach *Q*), 4 (approach *Fun*), 5 (approach *C*), 6 (approach *F*), 7 (approach *AF*), 8 (approach *FNS-C*), 9 (approach *FNS-F*), 10 (approach *FNS-AF*), 11 (approach *I-AF*), 12 (approach *II-AF*), 13 (approach *M-C*), 14 (approach *M-F*), 15 (approach *M-AF*), 16 (approach *M-S*), and 17 (approach *S*).

6.2.1 Performance of Sequence Features

We evaluated the performance of models based on sequence features with different modes of representation, namely consensus sequences, sequences with consensus base frequencies, sequences with all base frequencies, and, finally, individual read sequences.

Consensus Sequence

On the consensus sequence, for the *Baseline* data set, the best string kernel (WDKS) achieved $r \approxeq 0.41$, while numeric kernels had $r < 0.02$. For the *FNS-C* data set, string kernels also outperformed numeric kernels by far. While the performance of string kernels ranged from $r \approxeq 0.33$ (OK with maximal degree 6, width 0.5, and $\nu = 0.2$) to $r \approxeq 0.41$ (WDKS with degree 3, maximal mismatch 1, maximal shift 2, and $\nu = 0.6$), numeric kernels using binary-coded consensus sequences had $r < 0.016$.

Weighted Sequence

For consensus-weighted sequences (*Baseline* data set), numeric kernels outperformed hybrid string kernels. While the best models based on numeric kernels all obtained $r \approxeq 0.62$, the best hybrid string kernel models only achieved r ranging from 0.36 (HOK) to 0.45 (HWDK). This finding was validated on the *FNS-F* data set, where again numeric kernels outperformed hybrid string kernels with similar performances.

Fully Weighted Sequence

Using all base frequencies (*AF* data set), numeric kernels (r ranging from 0.64 to 0.66) outperformed the hybrid string kernels (r ranging from 0.35

to 0.64). The Gaussian RBF with $\sigma \approxeq 0.0005$ and $\nu = 0.3$ performed best for numeric kernels. For hybrid string kernels, the HWDK with degree 3, maximal mismatch 1 and $\nu = 0.4$ performed best. The other two hybrid string kernels did not perform as well; their most predictive regression models achieved only $r \approxeq 0.36$ (HWDKS with degree 3, maximal mismatch 0, shift 1, and $\nu = 0.5$) and $r \approxeq 0.42$ (HOK with degree 4, width 0.5, and $\nu = 0.2$).

For the *FNS-AF* data set, numeric kernels also outperformed hybrid string kernels. While the best models using numeric kernels had $r \approxeq 0.64$ to $r \approxeq 0.66$, hybrid string kernels were more inconsistent in their performance; for example, the best HWDK model had $r \approxeq 0.62$, but the best HWDKS model had $r \approxeq 0.32$.

Performance of Probe-Type-Specific Models

The best SVR model for Infinium probes of type I (data set *I-FNS-AF*) used a polynomial kernel of degree 2 with $\nu = 0.7$ and achieved $r \approxeq 0.66$. The best model for Infinium II probes (data set *II-FNS-AF*) was based on a Gaussian RBF kernel ($\sigma \approxeq 0.0004$ with $\nu = 0.6$) and achieved a performance of $r \approxeq 0.61$.

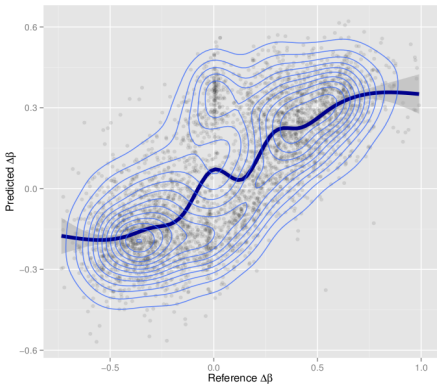
Set Kernels

For SVR models based on set kernels, we could observe that performances for numeric kernels using all base frequencies (r ranging from 0.69 to 0.71) were similar to those of string kernels (r ranging from 0.67 to 0.73). For numeric base kernels the Gaussian RBF with $\sigma = 1,070$ and $\nu = 0.4$ performed best with $r \approxeq 0.71$. For string kernels, the WDKS with degree 10, maximal mismatch 2, and shift 3 with $\nu = 0.2$ performed best ($r \approxeq 0.73$).

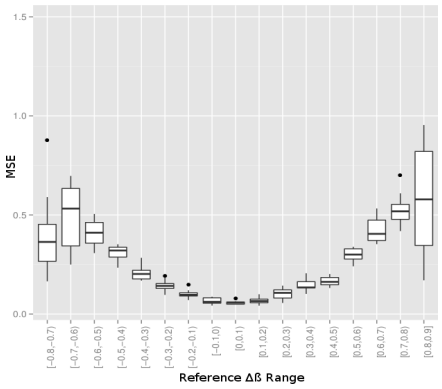
Fig. 6.1 gives the predictive performance of the set kernel using the previously mentioned WDKS and found two things. First, predictive performance was better for observations with $\Delta\beta > 0$ than for those with $\Delta\beta < 0$. Second, prediction was more difficult for extreme values of $\Delta\beta$ as indicated by larger median MSEs and larger variance in such regions. Note that these findings were not particular to set kernels, but were also obtained for other kernels and data sets.

Masked CpG Experiments

Masking the CpG position in sequences resulted in a reduced performance for all approaches compared to non-masked sequences. In the following, we give the best-performing kernel function and its correlation for each approach. For set kernels (Fig. 6.2), string kernels ($r \approxeq 0.35$ and $r \approxeq 0.37$ for the WDK without and with shifts, respectively) outperformed the linear and RBF kernels considerably with a correlation between predictions and outcomes of



(a) Scatter plot of predicted vs. reference $\Delta\beta$ values. The blue curve is a fitted density indicating the trend in predicted values. The blue ellipses are 2D-densities indicating those regions where most of the data points reside.



(b) MSEs for individual ranges of $\Delta\beta$ values in steps of 0.1.

Figure 6.1: Predictive performance ($r \approxeq 0.73$) of SVR with set kernels using the weighted degree string kernel with degree 10, maximal mismatch 2, shift 3, and $\nu = 0.2$.

$r \approxeq 0.23$ and $r \approxeq 0.26$ for their best models, respectively. The polynomial set kernel with degree 3 and $\nu = 0.7$ almost reached the performance of the best string kernel for sets, achieving $r \approxeq 0.36$. On the other data sets (*FNS-C*, *FNS-F*, *FNS-AF*), hybrid string kernels outperformed numeric kernels considerably. The best numeric kernels attained only $r \approxeq 0.03$ for each of the data sets, while the best hybrid string kernels had correlations of approximately 0.22 (*FNS-C*), 0.2 (*FNS-F*), and 0.23 (*FNS-AF*).

6.2.2 Performance of Non-Sequence Features

Apart from the sequence, we were also concerned with determining the influence of non-sequence features on differences in methylation measurements between Infinium 450K and WGBS. In the following we show properties of these features and report the performances of their SVR models.

As global features we considered a vector consisting of coverage depth, base qualities, and strand directionality ratios in order to test whether a combination of several features might lead to a good predictor. However, the best model based on the Gaussian RBF kernel, achieved only $r \approxeq 0.095$ using $\nu = 0.6$ with $\sigma \approxeq 0.21$.

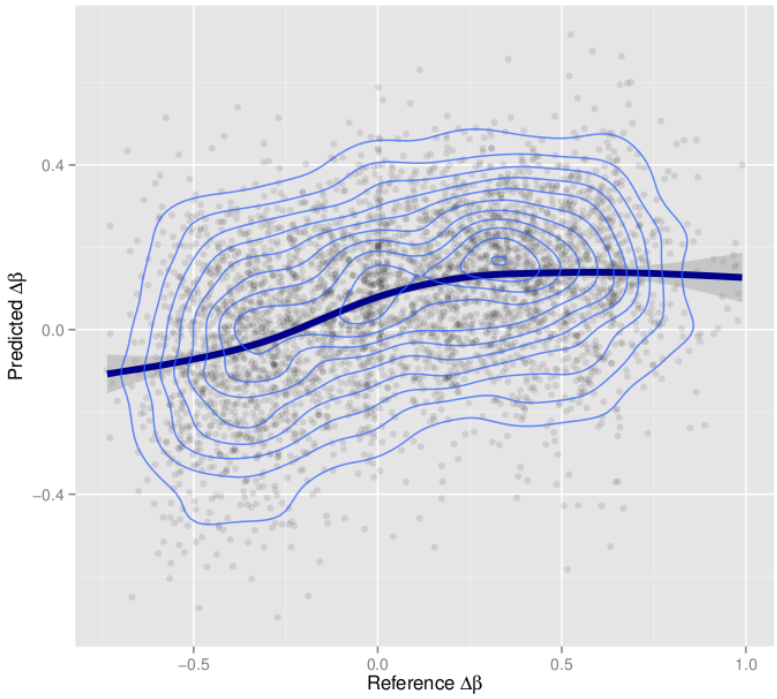


Figure 6.2: Predictive performance ($r \approxeq 0.37$) of SVR with the set kernel using the WDKS with degree 10, maximal shift 3, maximal mismatch 1, and $\nu = 0.3$ on masked sequences. The blue curve is a fitted density indicating the trend in predicted values. The blue ellipses are 2D-densities indicating those regions where most of the data points reside.

Depth of Coverage

In sequencing, depth of coverage is one of the most important aspects when considering data quality. The depth of coverage is determined by the number of reads that are observed at a given position in a genome. In this thesis, we considered the depth of coverage of CpG dinucleotide sites. A high depth of coverage is desirable because, among others, it reduces the impact of sequencing errors. Therefore, we expected CpGs with a low depth of coverage to exhibit greater values of $|\Delta\beta|$. To investigate this suspicion, we considered the relation of $\Delta\beta$ and depth of coverage (Fig. 6.3) and found that there is no noteworthy relationship. It seems that depth of coverage is neither related to the median difference in methylation nor to variance in $\Delta\beta$. Consequentially, the trained SVR models based on depth of coverage did not perform well. Even the best SVR model with $\nu = 0.6$ using a Gaussian RBF kernel with $\sigma \approxeq 0.001$ achieved only $r \approxeq 0.07$. Please note that the small sample sizes for low coverage and high coverage regions might have biased these findings.

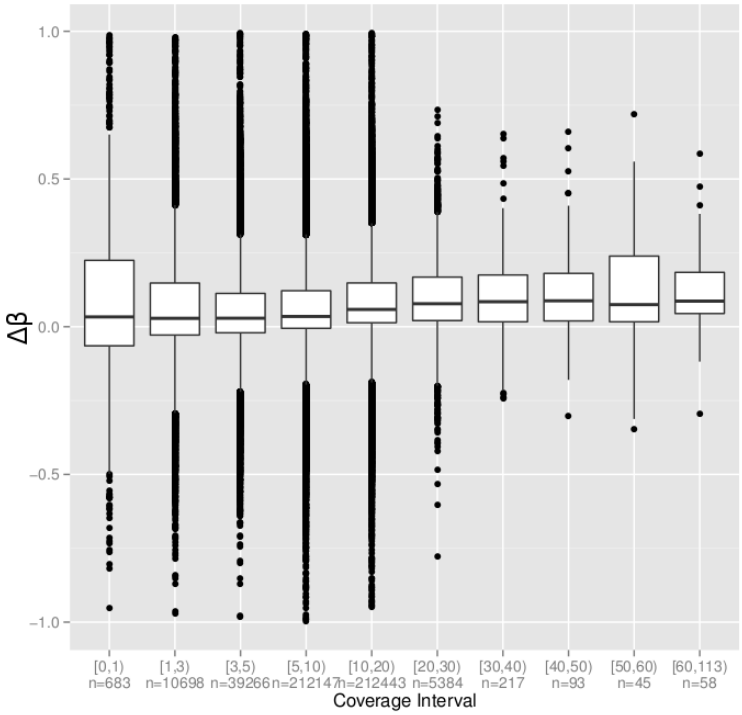


Figure 6.3: Methylation differences with respect to coverage. The x-axis represents sets of reads within a given depth of coverage range (e.g., the range $[3, 5)$ indicates all reads with a coverage of 3 or 4). Additionally, the size of each coverage range (i.e., its number of reads), is indicated below the range. For example, $n = 39,266$ indicates that $39,266$ reads had a coverage in the range $[3, 5)$.

Base Qualities

We observed that base qualities around the CpG dinucleotide locus exhibited greater variance than those in other regions, see Fig. 6.4. Note that the black areas in the plot are a consequence of outlier overplotting and one should rather focus on the boxes and their whiskers, which lie in the 50-70 Phred quality range indicating error probabilities between 0.001 and 0.00001 for most observations. Interestingly, the tails of CpG windows exhibited both, greater median base quality and variance than the other positions. Nevertheless, the trained SVR models based on positional base qualities did not perform well; the best-performing model with $\nu = 0.4$ and a polynomial kernel with degree 2 just had a non-significant correlation of $r \approxeq 0.055$ (p-value $= 0.1315$) at a significance level of $\alpha = 0.05$.

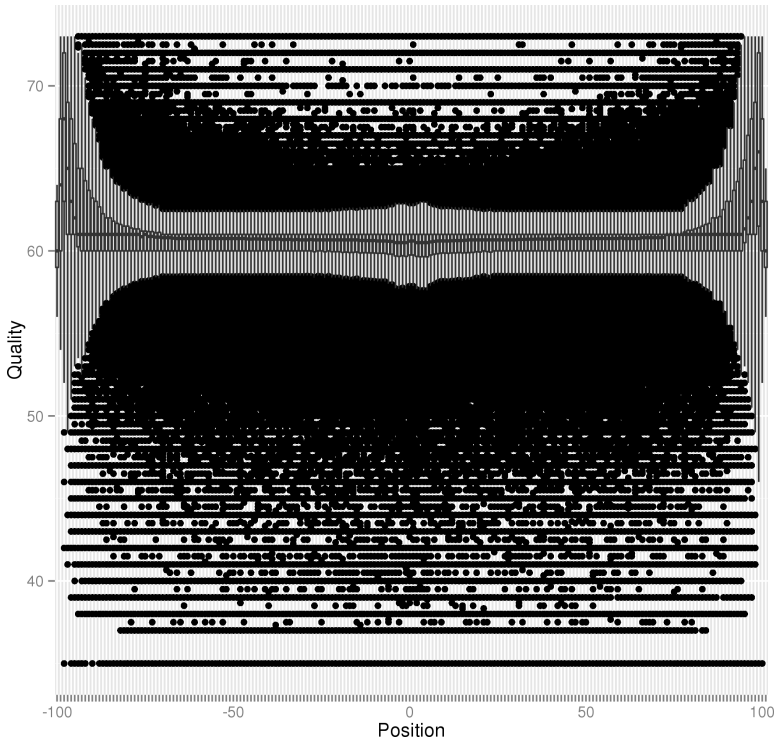


Figure 6.4: Base quality versus sequence window position.

Functional Annotations

The best-performing model based on functional annotations of CpG positions had $\nu = 0.6$ and used a polynomial kernel with degree 2, but only attained a correlation of $r \approxeq 0.148$ with the outcome $\Delta\beta$.

6.3 Interpretation and Visualization of Features

Since interpretation of models based on non-linear kernel functions is complicated, we employed several approaches, among them Shanon entropy, kernel-target alignment, motif discovery, and POIM visualization, to get a better understanding of feature importance.

6.3.1 Kernel-Target Alignment

To study the role of individual features for prediction and to improve predictive power by combining several features, we performed kernel-target alignment of the best kernel matrices for consensus sequences (WDKS with degree 3, maximal mismatch 1, and shift 2), global features (Gaussian RBF kernel with $\sigma \approxeq 0.21$), base qualities (polynomial kernel with degree 2), depth

of coverage (Gaussian RBF kernel with $\sigma \approxeq 0.001$), and functional annotations (polynomial kernel with degree 2), giving rise to feature weights 0.18, 0.05, 0, 0, and 0.98, respectively. We compared the performance of a combined kernel using uniform weights ($\frac{1}{5}$ for each kernel matrix) with that of the combined kernel using weights from kernel-target alignment. We found that the SVR model using the uniform-weight matrix achieved a correlation of approximately 0.466, while the model with optimized weights achieved a correlation of approximately 0.47 with the outcome.

6.3.2 Shannon Entropy

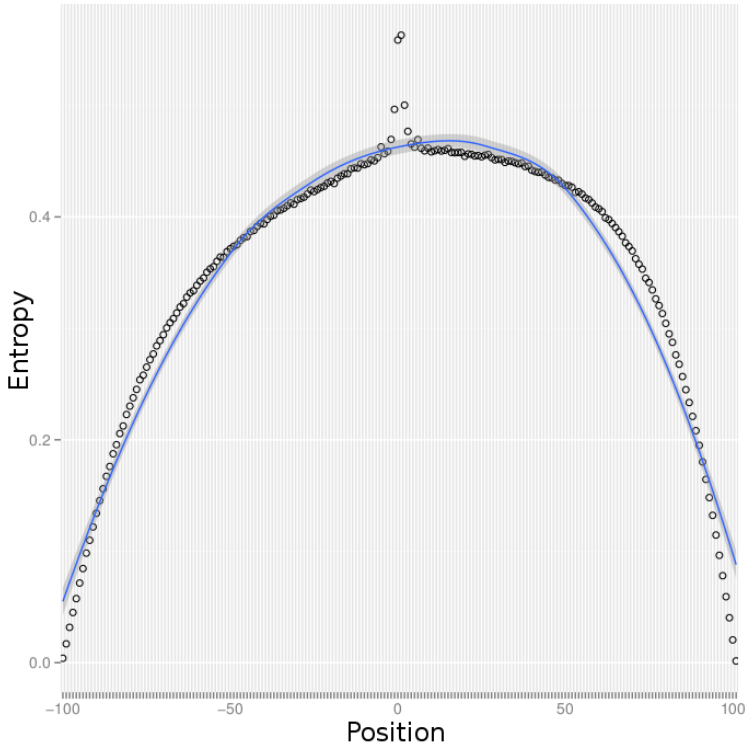


Figure 6.5: Shannon entropy per sequence window position. The CpG dinucleotide locus is given by positions 0 and 1. The blue curve is a density fitted to the positional Shannon entropies indicated by hollow circles.

The information content of the CpG position was studied by computing the Shannon entropy for every position in the sequence window (Fig. 6.5). The average unpredictability of the CpG dinucleotide locus and neighboring positions was much higher in comparison to the other positions exhibiting entropies larger than 0.5 bit. The parabolic nature of Shannon entropy in the sequence window is a consequence of the the CpG dinucleotide in the

window occurring at positions 0 and 1. When we align reads, it is much more probable that gaps are introduced at the window tails, as they are less likely to be covered by any reads. Ergo, the surprisal drops when we move from the center of the window towards the tails, where we mostly observe gap characters.

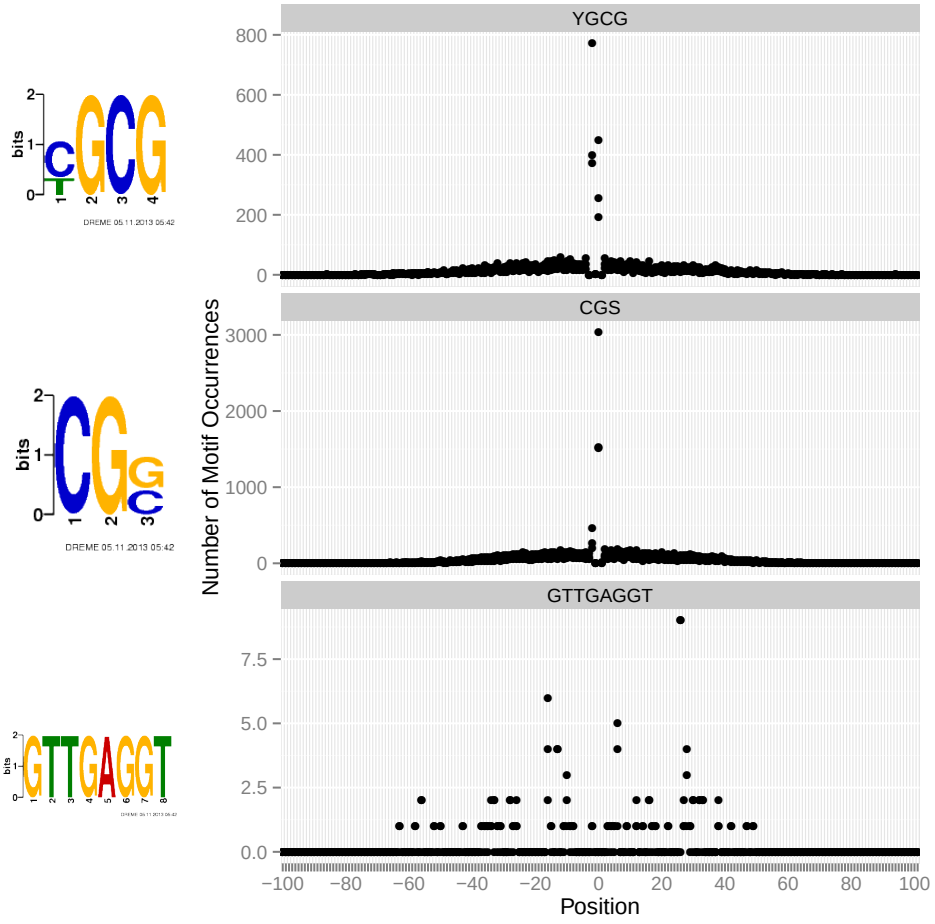


Figure 6.6: DREME motif discovery. Motifs are illustrated on the left. On the right side, we see the number of motif occurrences in the CpG-aligned reads. Positions 0 and 1 represent the CpG dinucleotide locus.

6.3.3 Motif Discovery

Using DREME, three significant motifs were found, namely YGCG (Y indicates cytosine or thymine, p-value 1^{-11}), CGS (S indicates cytosine or guanine, p-value 1.3^{-7}), and GTTGAGGT (p-value 1.9^{-7}). Fig. 6.6 illustrates the motifs and their positions in the reads. Positions around the CpG dinucleotide seem to play a major role in differentiating beta-value agreement

and disagreement, considering their strong impact on the first two motifs, YGCG and CGS.

6.3.4 Positional Oligomer Importance Matrices

The importance of the CpG position was further underlined by POIM visualization. Here, k-mers were scored according to their importance in predicting $\Delta\beta$, see Fig. 6.7. Positions other than the CpG dinucleotide seem to convey only weak signals for predicting $\Delta\beta$. Looking at differential POIMs (Fig. 6.8), with k ranging from 1 to 10, we can see that, except for the CpG position, there is no clear signal as to which value for k is best-suited for every position. Only, for the CpG dinucleotide, obviously, we have a considerable peak indicating an improvement when considering k-mers of size 2. The illustrated POIMs are based on the best-performing string kernel for the consensus sequence, which was the WDKS with degree 3, shift 2, and a maximal mismatch of 1.

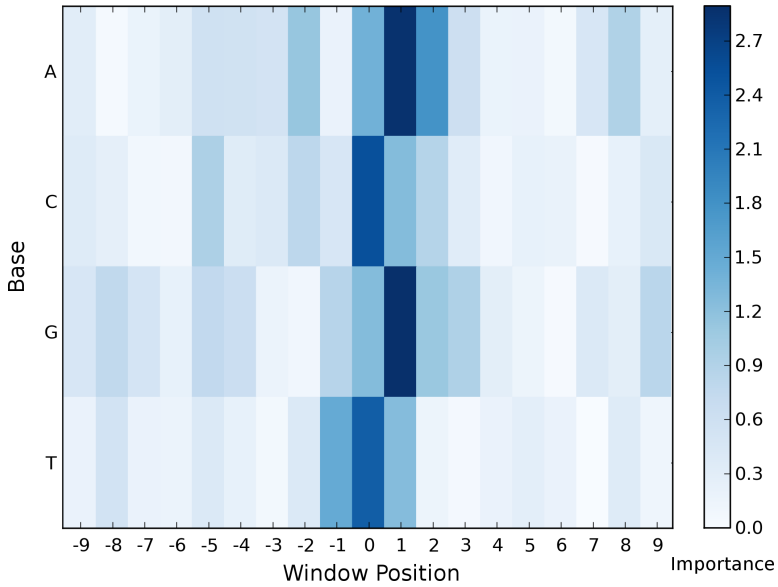


Figure 6.7: Excerpt of a positional oligomer importance matrix for individual bases ($k = 1$). Only the area around the CpG position, which represents window positions 0 and 1, is shown. The importance of individual bases for predicting $\Delta\beta$ is proportional to color intensity.

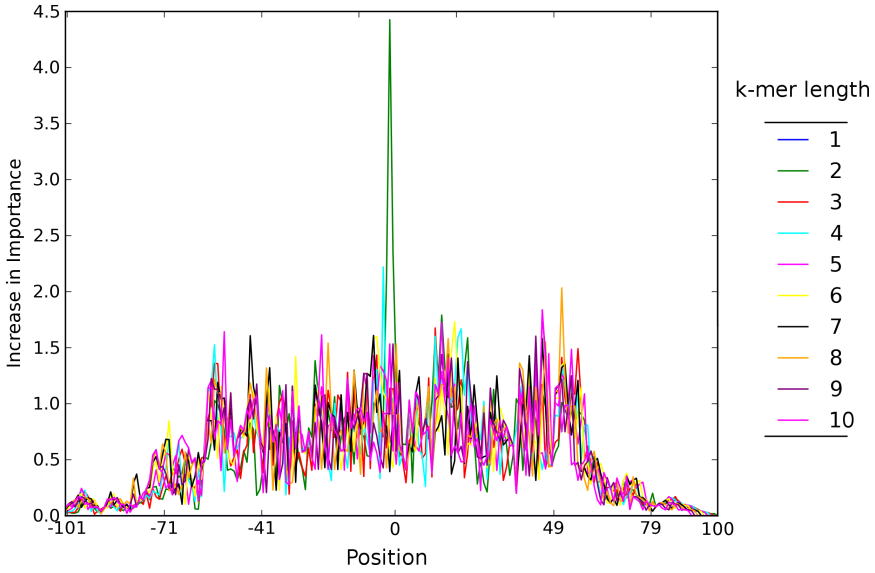


Figure 6.8: Differential positional oligomer importance matrices for $k = 1, 2, \ldots, 10$. Each plot indicates the increase in positional importance attained by considering oligomers of increasing length, e.g., the green plot indicates the increase in importance associated with incrementing k-mer length from 1 to 2.

Chapter 7

Discussion

Here, we discuss the role of individual features for prediction and whether kernel-target alignment is worthwhile in our application scenario. Furthermore, we cover the role of sequence features for predicting $\Delta\beta$, and, finally, the ramifications of the CpG position on predictions.

7.1 Evaluation of Kernel Target Alignment and Impact of Individual Features

We found that, in our scenario, MKL using kernel-target alignment did not lead to a significantly better (p-value = 0.68 at a significance level of $\alpha = 0.05$) predictive model ($r \approxeq 0.47$) when compared with the performance of the combined kernel matrix based on uniform weights ($r \approxeq 0.466$). This did not come as a surprise, considering that the weights found by kernel-target alignment did not seem well-balanced. The vast amount of weight (0.98) was associated with functional annotations, while sequence and base qualities received small weights (0.18 and 0.05, respectively), while global properties and coverage were completely discarded. We would have expected most of the weight to be assigned to sequence features ($r \approxeq 0.413$ for the best model), which outperformed functional annotations considerably ($r \approxeq 0.148$ for the best model). However, functional annotations had been shown to be the most predictive feature among non-sequence features and therefore rightfully received the largest amount of weight among these features. One possible reason why functional annotations were determined so important, could again be the implicit coding of methylation state, for example due to annotations of CPIs or euchromatin and heterochromatin.

Still, we found that by combining various features, namely, sequence, global properties, coverage, quality, and functional annotations, it is possible to increase predictive power in comparison with individual features. The SVR model using the weights determined by kernel-target alignment performed better ($r \approxeq 0.469$) than the individual underlying matrices ($r \approxeq$

0.41 for sequences, $r \approxeq 0.095$ for global properties, $r \approxeq 0.07$ for coverage, $r \approxeq 0.055$ for quality, and $r \approxeq 0.148$ for functional annotations) leading to a significant increase in correlation when compared to the correlation of the best consensus model (p-value $< 2.2 \cdot 10^{-16}$). This finding was not surprising for two reasons. First, in Chapter 6 we had already demonstrated that there was a high inter-feature variance regarding the predictive power for $\Delta\beta$. Accordingly, we were able to significantly improve the correlation with the outcome by finding optimal kernel matrix weights using kernel-target alignment in comparison to individual models. This finding agrees with previous work, which had already shown that MKL can improve performance, even though it does not always leads to considerably better results than individual kernels or the uniform-weight combined kernel [40, 4, 86]. Additionally, it is well known that a uniformly weighted combined kernel is often able to compete with current MKL methods when it comes to predictive performance [12]. On the other hand, note that MKL does not always perform better than models based on individual features: Cortes et al. [11] were able to confirm the results by Lanckriet et al. [42], showing that, for a specific set of features, MKL never performed significantly better than learners based on individual features. For our analysis, we only considered the best kernel matrix as determined by previous evaluations. It is possible that we might have attained better MKL results if we had let MKL choose the best kernel function in relation to the other features. However, this potential increase in performance would have come at the cost of greater runtime and since we did not see any substantial increase in performance, we did not deal with further, time-consuming analyses of MKL models.

7.2 On the Role of Sequence Features

We found that sequences are the most helpful feature for predicting $\Delta\beta$. Hence, in the following we will discuss our findings regarding the performance of the evaluated kernel functions and the the role of the different data sets we considered.

7.2.1 Comparison of Kernel Performances for Sequences

In our analyses, we measured the performance of several kernel functions in various data scenarios. We found that set kernels performed best at predicting $\Delta\beta$, both for regular and masked sequences. This makes sense, because in our prediction problem there is an inherent structure in which every CpG can be represented by its set of overlapping reads. While the other kernel functions work on consensus sequences and, as such, have to discard some information, set kernels can consider even the slightest differences between individual reads. Additionally, we are able to maintain an appropriate similarity measure in this multi-instance learning scenario by using string kernels.

The hybrid string kernels we devised to deal with a combination of string and numeric input yielded better predictive performances compared to the plain sequence kernels. For the *Baseline* data set, the best string kernel had a correlation of approximately 0.413 (approach C), whereas the best hybrid string kernels for the consensus frequency had correlations of roughly 0.449 and 0.641, for approaches F and AF, respectively. This indicates that in our scenario, where base frequency information is available, hybrid string kernels provide a more appropriate measure of similarity than conventional string kernels. The similar results for approaches *FNS-C*, *FNS-F*, and *FNS-AF* validated this finding.

However, hybrid string kernels typically did not exceed the predictive performance of the numeric kernels. The lesser predictive performance of hybrid string kernels might be caused by their usage of the consensus sequence, which already gives a blurred signal. Since matching of strings is based on the consensus, we could find matching substrings not fully supported by their frequency profiles, which are therefore scaled down according to cosine similarity, leaving only few high-impact motifs behind. Another problem could be the usage of cosine similarity for the comparison of frequency profiles. While string kernels implement positional uncertainty via weights associated with shifts (WDKS) or Gaussian functions (OK), cosine similarity is a position-specific measure. Hence, it might be necessary to decrease the role of frequency profiles for motifs that are further apart from each other. Another possibility would be the introduction of an additional kernel parameter setting the general influence of the frequency profiles. In this way, one could obtain kernel functions that are a mixture of conventional string kernels and the hybrid string kernels defined in this thesis. By tuning this parameter according to the data at hand, predictive performance could be further increased, even though one should keep model complexity in mind.

Nevertheless, hybrid string kernels outperformed numeric kernels on sequences in which the CpG position was masked. A possible explanation for this is that numeric kernels could merely pick up the single signal at the CpG position, but may not able to capture more complex patterns in sequences. This would also explain the substantial drop in performance of numeric kernels from regular sequences (e.g. $r \approxeq 0.66$ for the *FNS-AF* approach) to masked sequences (e.g. $r \approxeq 0.037$ for the *M-AF* approach). While numeric kernels, in particular, the polynomial and RBF kernels, seem to be well-suited at picking up the signal from the CpG position, they are barely able to pick up any signal when that position is removed. This finding underlines the benefit of string kernels over numeric kernels and also shows that hybrid string kernels can have an advantage over numeric kernels in some scenarios. On the other hand, the best-performing hybrid string kernel in the masked *M-AF* scenario ($r \approxeq 0.23$) was still worse than the best-performing set kernel ($r \approxeq 0.37$), which illustrates the power of set kernels.

We found that, in general, predictive performance of models improved

with increasing amount of information available for learning. For example, models based on consensus sequences (*FNS-C*), consensus sequences with frequencies of the consensus-base (*FNS-F*), and consensus sequences with all base frequencies (*FNS-AF*) had a correlation of 0.41, 0.62, and 0.66, respectively. We expected to see an improvement when going from consensus sequences to weighted sequences. However, we did not necessarily expect that the performance of models using all base frequencies would be better than the performance of those relying on consensus-frequencies only. This further improvement shows that the complete nucleotide profile plays a role in measurement errors, but also that SVR is well capable of handling such high-dimensional data.

Another major aspect of our analyses was the determination of the best-suited string kernel for the comparison of bisulfite-converted reads. We found that, for our data, the WDKS performed best in most scenarios. Interestingly, its non-shifted counterpart (WDK), which does not account for positional uncertainty, usually performed better than the OK, which allows for positional uncertainty. A possible explanation for this is that the width parameter σ for the Gaussians used in the OK is a more volatile parameter and could have necessitated finer grid search sampling. Regarding the choice of parameters for the weighted degree string kernel, we found that typically, degrees of 3 or 10 worked best. In addition, all well-performing kernels allowed for shifts and also, to a certain extent, for mismatches. Seemingly, it is also necessary to consider longer shifts, e.g., shifts of length 3 for string kernels of higher degree, e.g., for degree 10.

In general, predictive performance was better in the $\Delta\beta > 0$ than the $\Delta\beta < 0$ region. We first thought that we could solve this problem by normalizing Infinium 450K probes to WGBS measurements, because there is a trend for Infinium 450K methylation measurements to be larger than those from sequencing. However, this was not the case. This could indicate that, for some reason, overconversion of CpGs is more easily detected in the sequence than underconversion. We also found that the predictive performance decreased for more extreme observed values of $\Delta\beta$ (e.g., values close to either 1 or -1). This might have been a consequence of the few existing samples with large values of $|\Delta\beta|$, as well as larger possible differences between predicted and reference values for such observations, leading to larger MSEs.

7.2.2 Impact of Forming Probe-Type-Specific Models

We split the *FNS-AF* data set into a set containing only Infinium I probes and another containing only Infinium II probes to determine whether intra-array normalization of probes has a detrimental effect on predictive power. As expected, the predictive performance of the SVR model based on Infinium I probes was better than that of models based on Infinium II probes, which correlates with previous findings indicating that measurements from Infinium

II probes are less reliable than those from Infinium I [15]. However, we also found that the performance of the best models weighted according to the chip's relative probe type frequency ($\overline{r} \approxeq 0.72 \cdot 0.66 + 0.28 \cdot 0.61 \approxeq 0.65$) was comparable to the best model using both probe types simultaneously ($r \approxeq 0.66$ for the *FNS-AF* data set). This shows that intra-array normalization with BMIQ accurately adjusts beta-values of Infinium II probes to those of Infinium I probes and that it is not necessary to consider probe-type-specific prediction models in this scenario.

7.2.3 The Impact of PCA-Sampling, Filtering, and Normalization of Infinium 450K Positions to WGBS Values

The predictive performance of models based on data sets generated via PCA-sampling, filtering, and normalization of Infinium 450K to WGBS values (*FNS-C*, *FNS-F*, *FNS-AF*) did not deviate considerably from the performance of the same models on the *Baseline* data set (*C*, *F*, *AF*), which was not subjected to the three processing steps. Furthermore, we saw that PCA did not identify clear structures in sequences and that normalization of Infinium 450K methylation values does not improve predictive performance. Hence, to obtain reasonable prediction models it is not necessary to perform either filtering, normalization of Infinium 450K to WGBS measurements, or sampling with PCA. Since filtering did not improve predictive performance, this shows that the generated models are robust with regard to dubious probes from Infinium 450K.

7.3 Role of the CpG Position

Shannon entropy was largest for the CpG dinucleotide locus, two of three motifs found via DREME overlapped with the CpG locus, and the performance of predictive models dropped considerably when masking the CpG position. All of these findings suggest that the CpG dinucleotide is the most important feature in differentiating whether measurements agree between WGBS and the Infinium 450K chip or not. Additionally, the best performing models employing the OK or HOK exhibited small values for σ (e.g. $\sigma = 0.5$) throughout all approaches that made use of the CpG locus. On the other hand, models that only used masked sequences exhibited much larger values for σ (e.g. $\sigma = 4$). Remember that small values for σ allow for little positional uncertainty, while large values allow for greater uncertainty. Therefore, the difference between σ values in the two approaches might also indicate that in the non-masked scenario, the signal is mainly based on the positionally constrained CpG locus.

One possible explanation for this finding is the correlation ($r \approx -0.23$) between the WGBS methylation level and $\Delta\beta$. Also note that the correlation of $\Delta\beta$ and WGBS methylation is much stronger for sampled data sets

due to stratified sampling according to $\Delta\beta$, e.g., $r \approxeq -0.72$ for the *Baseline* data set, see Fig. 2 in the appendix. This relationship has impacts on both, interpretation of features and learned models. Let us first deal with the impact of the correlation between $\Delta\beta$ and β_{WGBS} on finding relevant sequence positions.

For motif discovery and POIM visualization of the sequence kernel, we have split observations into two classes, the set of positions for which we consider methylation to agree ($|\Delta\beta| < 0.3$) and the set of observations with inconsistent measurements ($|\Delta\beta| \geq 0.3$). Since $\Delta\beta = \beta 450K - \beta_{\text{WGBS}}$, the first set can only exhibit β_{WGBS} in the range $[0, 0.3)$, whereas the second set has β_{WGBS} in the range $[0.3, 1]$. This means that in the second set, it is more likely for the WGBS CpG to be methylated, while in the first set it would rather be unmethylated. Hence, it makes sense that the CpG position, which determines the methylation state, plays such an important role in differentiating the two sets with regard to $\Delta\beta$.

We were not able to determine whether there exist effects specific to the CpG position predictive of $\Delta\beta$ that are not solely based on the correlation between $\Delta\beta$ and β_{WGBS}. However, we saw in Section 6.2.1 that all of the models based on masked CpG positions performed drastically worse than their regular counterparts, which underlines the relevance of the CpG position for predictions and validates the findings from motif discovery and POIM visualization. However, using only masked sequences, we were still able to predict $\Delta\beta$ to some extent and have found a significant, non-CpG motif (GTTGAGGT) indicative of inconsistent methylation measurements, pointing to the existence of a signal outside the CpG locus.

Chapter 8

Conclusion and Outlook

In this chapter, we conclude our findings on multiple kernel learning and non-sequence features, the role of the CpG position and the choice of kernel functions for sequences. Then, we give a brief project summary, entailing achievements, conclusions, and remaining challenges in predicting methylation measurement differences between Infinium 450K and WGBS.

8.1 On Multiple Kernel Learning and Non-Sequence Features

While MKL is certainly able to significantly increase predictive performance in some scenarios, we did not find a substantial effect for the features we considered. Hence, one should not consider MKL a be-all and end-all solution for improving predictive performance in multi-feature scenarios, but first think about the properties of individual features and how they might interact with each other. In particular, MKL, from our experience, will not be able to increase performance substantially if individual features are only hardly predictive of the outcome. Additionally, it seems that kernel matrices resulting from unpredictive features and/or inappropriate kernel functions exhibit small variances, while those of predictive features and/or suitable kernel functions exhibit high variances. This explains why uniform-weight kernel combinations often achieve similar performances to combinations using weights optimized via MKL; when kernels are averaged, the impact of unpredictive, low-variance matrices is small, while that of predictive, high-variance matrices is large.

Even though we found that MKL led to a slight boost in predictive performance when compared with the performance of models based solely on individual features, it would first be necessary to focus more on the sequences of reads rather than their other characteristics. While non-sequence features play a role, albeit limited, it seems that the sequence has the greatest impact on $\Delta\beta$.

8.2 On the Role of the CpG Position

It is reasonable to assume that the predictivity of the CpG position for $\Delta\beta$, from the results and discussion provided in Sections 6.2 and 7.3, is largely based on its implicit representation of methylation state, which can be facilitated to bound $\Delta\beta$. A statistical model for the prediction of differences in methylation measurements from Infinium 450K chips and next-generation sequencing should, however, not be based on the current methylation state of the training data, but rather capture effects that are based on other characteristics, such as the sequence itself. If a model were to hinge on the methylation state, it would not be applicable to data exhibiting a differing epigenetic state. Therefore, to obtain high model generalizability, our recommendation would be to remove any signal from features indicative of methylation state, e.g., by masking CpG positions.

8.3 On the Choice of Kernel Functions

We saw that set kernels outperformed all other kernel functions, in particular when dealing with sequences in which the CpG position was masked. Hence, we would recommend the usage of set kernels utilizing string kernels (e.g., the WDKS, for the comparison of individual CpGs represented by bisulfite-converted reads). In addition, it would be interesting to determine whether it is possible to increase the performance of hybrid string kernels by choosing a more appropriate scaling factor for string kernel weights rather than cosine similarity. Here, it could be useful to introduce a similarity measure that accounts for positional uncertainty.

8.4 Summary

The goal of this thesis was to work towards assigning confidence values representing the accuracy of methylation measurements to individual CpG positions using WGBS measurements as input data. We were able to show that it is indeed possible to predict differences in methylation — with significant correlation — between Infinium 450K technology and WGBS, which was not done before. We found that non-sequence features do not play a major role in the emergence of inconsistent methylation measurements, but found that sequences are highly predictive. We introduced hybrid string kernels, new kernel functions that are appropriate for observations consisting of both, sequences and numeric data. They performed better than conventional string kernels and also outperformed numeric kernels for observations with masked CpG positions. Using comprehensive analyses, we studied the impact of various kernel functions on the predictive power of SVR models and found that set kernels are best-suited for predicting differences in methylation. We

discovered that the CpG position is highly predictive of $\Delta\beta$, but should be excluded for predictions as it reflects methylation state.

In conclusion, we found that it is possible to predict differences in methylation measurements between Infinium 450K and WGBS based on WGBS features. We could imagine the usage of our SVR model in epigenetic genome-wide association studies in order to differentiate between CpG positions for which WGBS methylation measurements are highly accurate and those that are not. For this, however, one would first need to consider whether the current predictive model performance (e.g., $r \approxeq 0.37$ for masked sequences) is already good enough for practical purposes. In addition, it would be favorable to first deal with the challenges delineated in the following paragraph, before making such a model publicly available and integrating it into the corresponding pipelines.

One remaining challenge is the actual indication of confidence values for methylation values of CpG positions, rather than differences in methylation. This should be manageable by relating individual predictions to the observed distribution of differences in methylation. Furthermore, it would be beneficial to put more effort into the statistical analysis of sequences (e.g., by clustering) to get a better understanding of the signals responsible for the predictive power of sequences. Further validation on different test data would be advantageous to compare generalizability of models based on regular and masked sequences, e.g. for data from a different tissue. Additionally, evaluation on different training data would be useful to get a better understanding of model bias.

Appendices

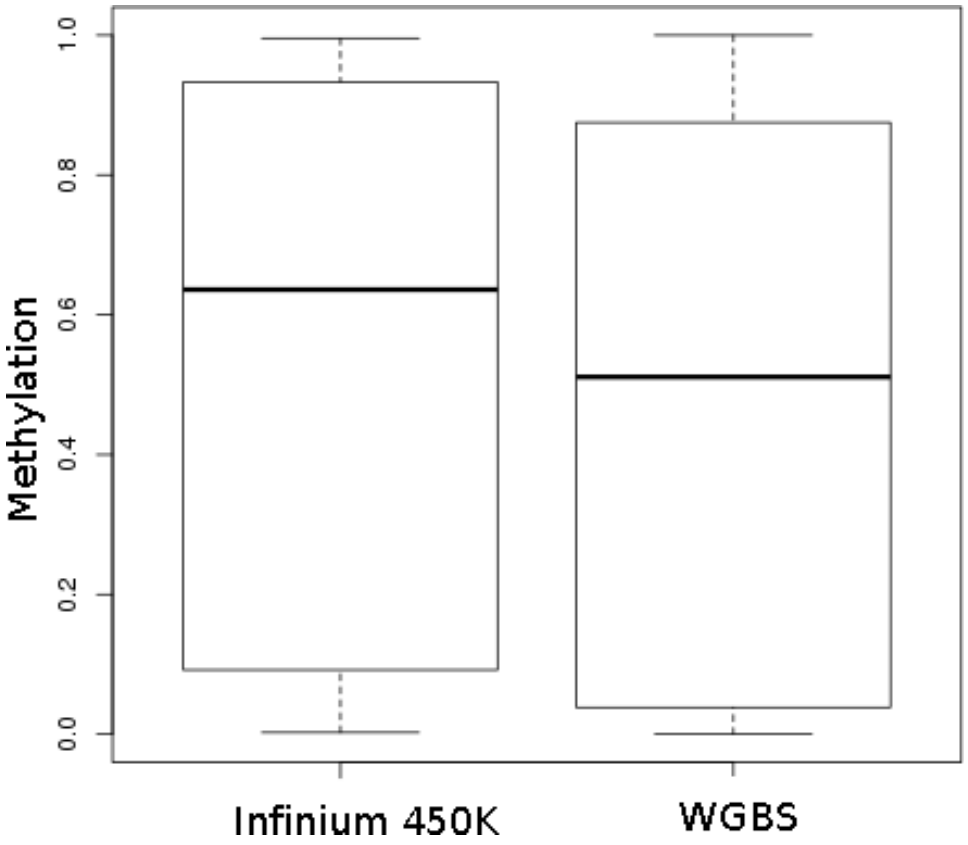


Figure 1: Comparison of beta-values for Infinium 450K and WGBS. The median beta-value of Infinium 450K is about 0.1 higher than for WGBS. The variances of both approaches are similar.

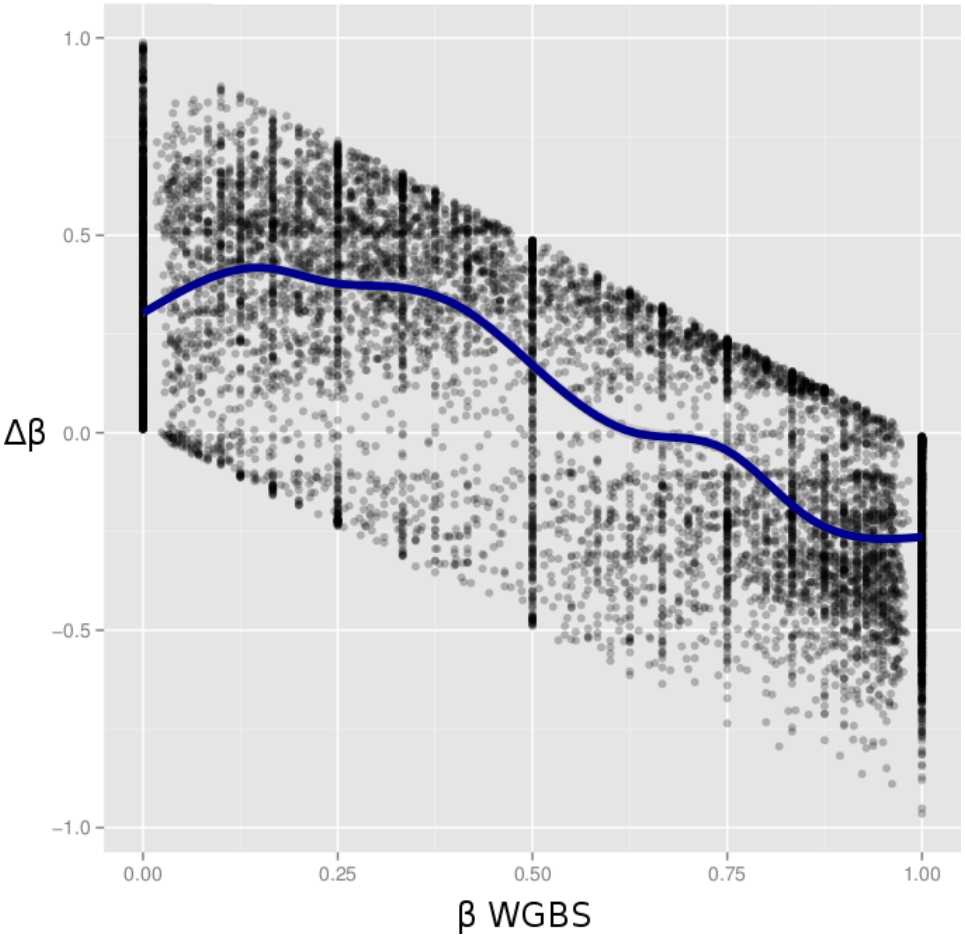


Figure 2: Correlation between WGBS methylation and methylation differences. The blue curve indicates a fitted density.

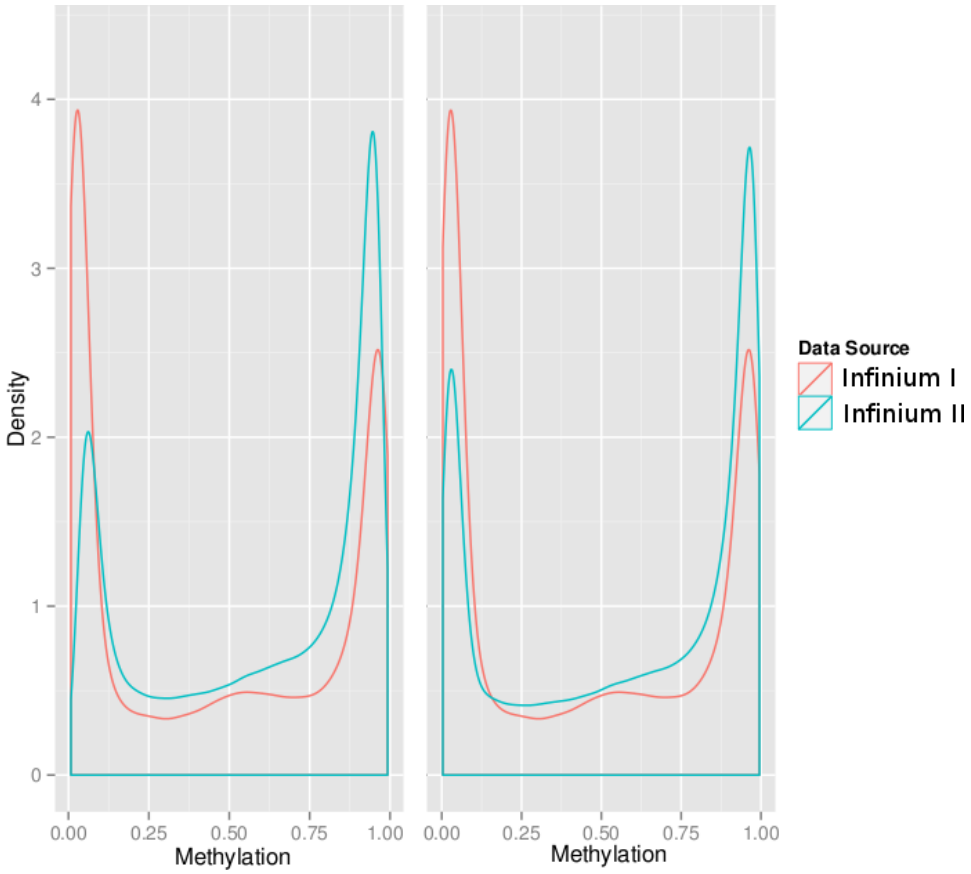


Figure 3: Densities of Infinium 450K probes before (left plot) and after normalization (right plot).

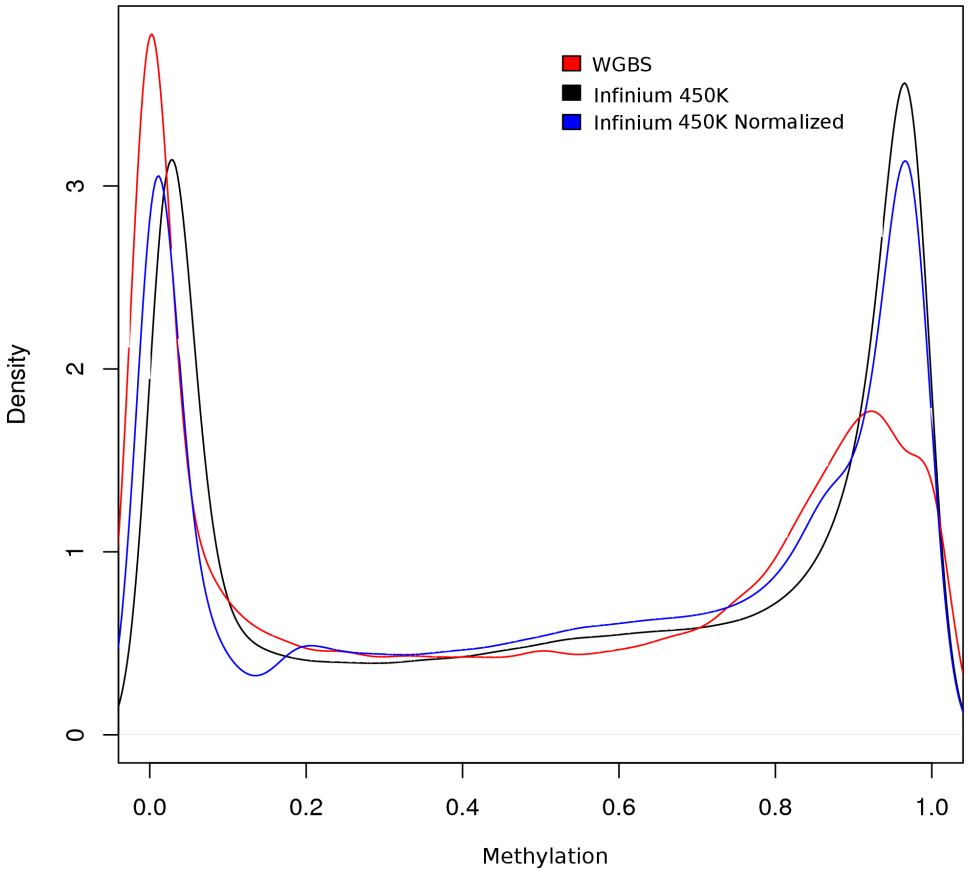


Figure 4: Normalization of Infinium 450K to WGBS methylation. The red curve gives the density for WGBS methylation values, while the blue and black curve give the methylation density for Infinium 450K measurements fitted and not adjusted to WGBS measurements, respectively.

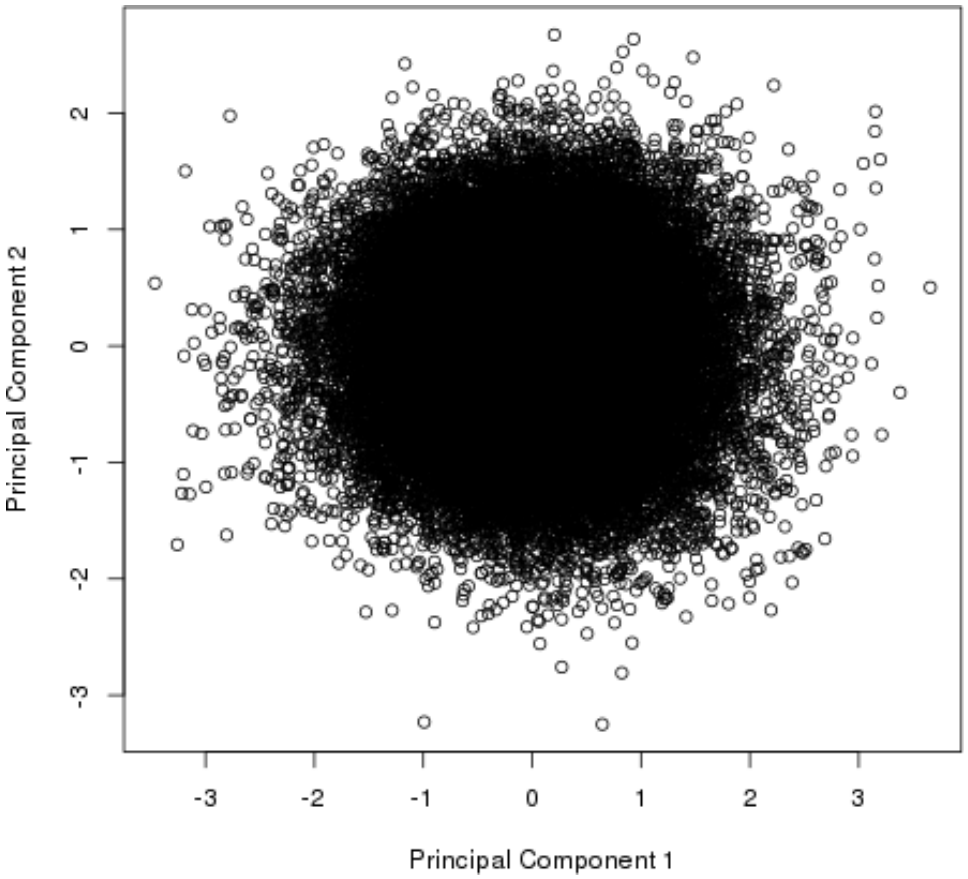


Figure 5: Representation of sequences in the first two PCs according to their real-valued representation using all base frequencies, for the stratum with $\Delta\beta \in [-0.1, 1)$.

Kernel	Parameters	MSE	r
RBF	$\sigma \approxeq 0.21$, $\nu = 0.6$	0.136	0.093
Poly	$d = 2$, $\nu = 0.5$	0.136	0.085
Linear	$\nu = 0.6$	0.137	0.022

Table 1: Best performing ν-SVR models with global features (approach G) for the *Baseline* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.6$	0.138	0.07
RBF	$\sigma \approxeq 0.001$, $\nu = 0.6$	0.137	0.054
Linear	$\nu = 0.7$	0.138	0.048

Table 2: Best performing ν-SVR models with coverage features (approach C) for the *Baseline* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.4$	0.138	0.057
RBF	$\sigma \approxeq 0.002$, $\nu = 0.7$	6983.304	0.018

Table 3: Best performing ν-SVR models with base call quality features (approach Q) for the *Baseline* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.6$	0.135	0.148
RBF	$\sigma \approxeq 0.018$, $\nu = 0.5$	0.134	0.148
Linear	$\nu = 0.6$	0.135	0.138

Table 4: Best performing ν-SVR models with functional annotations (approach F) for the *Baseline* data set.

Kernel	Parameters	MSE	r
WDKS	$d = 3$, $mm = 1$, $s = 2$, $\nu = 0.6$	0.118	0.413
WDK	$d = 10$, $mm = 2$, $\nu = 0.2$	0.117	0.394
OK	$d = 6$, $\sigma = 0.5$, $\nu = 0.2$	0.122	0.336
Poly	$d = 2$, $\nu = 0.7$	0.138	0.02
Linear	$\nu = 0.7$	0.154	0.018
RBF	$\sigma \approxeq 0.001$, $\nu = 0.7$	0.145	0.014

Table 5: Best performing ν-SVR models with consensus sequences (approach C) for the *Baseline* data set.

Kernel	Parameters	MSE	r
RBF	$\sigma \approxeq 0.002$, $\nu = 0.4$	0.083	0.631
Linear	$\nu = 0.4$	0.084	0.625
Poly	$d = 2$, $\nu = 0.3$	0.085	0.621
HWDK	$d = 3$, $mm = 1$, $\nu = 0.4$	0.112	0.449
HWDKS	$d = 10$, $mm = 1$, $s = 2$, $\nu = 0.4$	0.112	0.446
HOK	$d = 5$, $\sigma = 0.5$, $\nu = 0.2$	0.121	0.357

Table 6: Best performing ν-SVR models with weighted consensus sequences (approach F) for the *Baseline* data set.

Kernel	Parameters	MSE	r
RBF	$\sigma \approxeq 0.0$, $\nu = 0.3$	0.08	0.658
Linear	$\nu = 0.3$	0.08	0.649
HWDK	$d = 3$, $mm = 1$, $\nu = 0.3$	0.081	0.641
Poly	$d = 2$, $\nu = 0.2$	0.083	0.633
HOK	$d = 4$, $\sigma = 0.5$, $\nu = 0.2$	0.115	0.423
HWDKS	$d = 3$, $mm = 0$, $s = 1$, $\nu = 0.5$	0.124	0.355

Table 7: Best performing ν-SVR models with fully weighted consensus sequences (approach AF) for the *Baseline* data set.

Kernel	Parameters	MSE	r
WDKS	$d = 3$, $mm = 1$, $s = 2$, $\nu = 0.7$	0.126	0.415
WDK	$d = 10$, $mm = 2$, $\nu = 0.4$	0.125	0.397
OK	$d = 4$, $\sigma = 0.5$, $\nu = 0.2$	0.292	0.132
RBF	$\sigma \approxeq 0.001$, $\nu = 0.2$	0.147	0.016
Linear	$\nu = 0.2$	0.153	0.011
Poly	$d = 2$, $\nu = 0.4$	0.144	0.005

Table 8: Best performing ν-SVR models with consensus sequences for the *FNS-C* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.5$	0.082	0.613
RBF	$\sigma \approxeq 0.001$, $\nu = 0.7$	0.083	0.606
Linear	$\nu = 0.7$	0.084	0.599
HWDKS	$d = 3$, $mm = 0$, $s = 3$, $\nu = 0.7$	0.112	0.434
HWDK	$d = 10$, $mm = 2$, $\nu = 0.5$	0.111	0.427
HOK	$d = 3$, $\sigma = 0.5$, $\nu = 0.3$	0.514	0.121

Table 9: Best performing ν-SVR models with weighted consensus sequences for the *FNS-F* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.7$	0.078	0.662
Linear	$\nu = 0.6$	0.08	0.651
RBF	$\sigma \approxeq 0.0$, $\nu = 0.6$	0.081	0.646
HWDK	$d = 3$, $mm = 1$, $\nu = 0.6$	0.086	0.626
WDKS	$d = 10$, $mm = 1$, $s = 3$, $\nu = 0.5$	0.129	0.328
HOK	$d = 3$, $\sigma = 0.5$, $\nu = 0.4$	0.572	0.173

Table 10: Best performing ν-SVR models with fully weighted consensus sequences for the *FNS-AF* data set.

Kernel	Parameters	MSE	r
Poly	$d = 2$, $\nu = 0.7$	0.098	0.655
RBF	$\sigma \approxeq 0.0$, $\nu = 0.6$	0.105	0.619
Linear	$\nu = 0.6$	0.105	0.618
HWDK	$d = 3$, $mm = 1$, $\nu = 0.6$	0.111	0.593
HWDKS	$d = 6$, $mm = 0$, $s = 3$, $\nu = 0.6$	0.169	0.054
HOK	$d = 3$, $\sigma = 0.5$, $\nu = 0.4$	0.148	0.433

Table 11: Best performing ν-SVR models with fully weighted consensus sequences for the *I-FNS-AF* data set.

Kernel	Parameters	MSE	r
RBF	$\sigma \approxeq 0.0$, $\nu = 0.6$	0.067	0.605
Linear	$\nu = 0.5$	0.068	0.597
Poly	$d = 2$, $\nu = 0.6$	0.069	0.592
HWDK	$d = 3$, $mm = 1$, $\nu = 0.5$	0.077	0.537
HOK	$d = 3$, $\sigma = 1.2$, $\nu = 0.3$	0.5	0.118
HWDKS	$d = 3$, $mm = 0$, $s = 1$, $\nu = 0.5$	0.107	0.044

Table 12: Best performing ν-SVR models with fully weighted consensus sequences for the *II-FNS-AF* data set.

Kernel	Parameters	MSE	r
WDKS	$d = 10$, $mm = 0$, $s = 3$, $\nu = 0.7$	0.153	0.22
Poly	$d = 2$, $\nu = 0.3$	0.144	0.037
RBF	$\sigma \approxeq 0.001$, $\nu = 0.2$	0.147	0.035
Linear	$\nu = 0.2$	0.153	0.032

Table 13: Best performing ν-SVR models with masked consensus sequences (approach *M-C*) for the *FNS-C* data set.

Kernel	Parameters	MSE	r
HWDK	$d = 10$, $mm = 1$, $\nu = 0.7$	0.141	0.203
Poly	$d = 2$, $\nu = 0.3$	0.132	0.033
RBF	$\sigma \approxeq 0.007$, $\nu = 0.2$	0.136	0.032
Linear	$\nu = 0.2$	0.138	0.022

Table 14: Best performing ν-SVR models with masked weighted consensus sequences (approach *M-F*) for the *FNS-F* data set.

Kernel	Parameters	MSE	r
HWDK	$d = 10$, $mm = 0$, $\nu = 0.4$	0.138	0.234
Poly	$d = 2$, $\nu = 0.3$	0.139	0.032
RBF	$\sigma \approxeq 0.001$, $\nu = 0.2$	0.141	0.029
Linear	$\nu = 0.6$	0.154	0.034

Table 15: Best performing ν-SVR models with masked fully weighted consensus sequences (approach *M-AF*) for the *FNS-AF* data set.

Kernel	Parameters	MSE	r
WDKS	$d = 10$, $mm = 1$, $s = 3$, $\nu = 0.4$	0.179	0.334
OK	$d = 3$, $\sigma = 4$, $\nu = 0.2$	0.189	0.277
Poly	$d = 2$, $\nu = 0.7$	0.195	0.275
WDK	$d = 6$, $mm = 1$, $\nu = 0.6$	0.204	0.245
Linear	$\nu = 0.7$	0.225	0.241
RBF	$\sigma \approxeq 1070.58$, $\nu = 0.7$	0.196	0.204

Table 16: Best performing ν-SVR models with masked sequences (approach *M-S*) for the *Set* data set.

Kernel	Parameters	MSE	r
WDKS	$d = 10$, $mm = 2$, $s = 3$, $\nu = 0.2$	0.067	0.727
RBF	$\sigma \approxeq 1070.58$, $\nu = 0.4$	0.076	0.715
WDK	$d = 10$, $mm = 1$, $\nu = 0.2$	0.069	0.715
Poly	$d = 2$, $\nu = 0.2$	0.072	0.702
Linear	$\nu = 0.3$	0.073	0.693
OK	$d = 4$, $\sigma = 0.5$, $\nu = 0.2$	0.076	0.677

Table 17: Best performing ν-SVR models using the set kernel sequence representation from the *Set* data set.

Bibliography

[1] Manuel Allhoff, Alexander Schönhuth, Marcel Martin, Ivan G Costa, Sven Rahmann, and Tobias Marschall. Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1, 2013.

[2] Matthew D Anway, Andrea S Cupp, Mehmet Uzumcu, and Michael K Skinner. Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science*, 308(5727):1466–1469, 2005.

[3] Timothy L Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[4] Luke Barrington, Mehrdad Yazdani, Douglas Turnbull, and Gert RG Lanckriet. Combining Feature Kernels for Semantic Music Retrieval. In *ISMIR*, pages 614–619, 2008.

[5] Stephen B Baylin and Joyce E Ohm. Epigenetic gene silencing in cancer–a mechanism for early oncogenic pathway addiction? *Nature Reviews Cancer*, 6(2):107–116, 2006.

[6] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, 4(10):e1000173, 2008.

[7] Adrian Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16(1):6–21, 2002.

[8] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[9] Aniruddha Chatterjee, Peter A Stockwell, Euan J Rodger, and Ian M Morison. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, 40(10):e79–e79, 2012.

[10] Melissa Conerly and William M Grady. Insights into the role of DNA methylation in disease through the use of mouse models. *Disease Models & Mechanisms*, 3(5-6):290–297, 2010.

[11] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. L2 regularization for learning kernels. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 109–116. AUAI Press, 2009.

[12] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Two-stage learning kernel algorithms. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 239–246, 2010.

[13] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[14] Nello Cristianini, John Shawe-Taylor, Andre Elisseeff, and Jaz Kandola. On kernel target alignment. In *NIPS*, volume 2, page 4, 2001.

[15] 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.

[16] Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, page bbt054, 2013.

[17] Brian G Dias and Kerry J Ressler. Parental olfactory experience influences behavior and neural structure in subsequent generations. *Nature Neuroscience*, 17(1):89–96, 2014.

[18] Thomas A Down, Vardhman K Rakyan, Daniel J Turner, Paul Flicek, Heng Li, Eugene Kulesha, Stefan Graef, Nathan Johnson, Javier Herrero, Eleni M Tomazou, et al. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotechnology*, 26(7):779–785, 2008.

[19] 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.

[20] Florian Eckhardt, Joern Lewin, Rene Cortese, Vardhman K Rakyan, John Attwood, Matthias Burger, John Burton, Tony V Cox, Rob Davies, Thomas A Down, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378–1385, 2006.

[21] Melanie Ehrlich, Miguel A Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C Kuo, Roy A McCune, and Charles Gehrke. Amount

and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research*, 10(8):2709–2721, 1982.

[22] Jason Ernst and Manolis Kellis. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825, 2010.

[23] Jason Ernst, Pouya Kheradpour, Tarjei S Mikkelsen, Noam Shoresh, Lucas D Ward, Charles B Epstein, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael Coyne, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

[24] Mehrnaz Fatemi, Martha M Pao, Shinwu Jeong, Einav Nili Gal-Yam, Gerda Egger, Daniel J Weisenberger, and Peter A Jones. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Research*, 33(20):e176–e176, 2005.

[25] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755, 2014.

[26] M Gardiner-Garden and M Frommer. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282, 1987.

[27] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alex J Smola. Multi-Instance Kernels. In *ICML*, volume 2, pages 179–186, 2002.

[28] Diane P Genereux, Winslow C Johnson, Alice F Burden, Reinhard Stöger, and Charles D Laird. Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies. *Nucleic Acids Research*, 36(22):e150–e150, 2008.

[29] Christoph Grunau, Susan J Clark, and André Rosenthal. Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. *Nucleic Acids Research*, 29(13):e65–e65, 2001.

[30] Janet Harrison, Clare Stirzaker, and Susan J Clark. Cytosines adjacent to methylated CpG sites can be partially resistant to conversion in genomic bisulfite sequencing leading to methylation artifacts. *Analytical biochemistry*, 264(1):129–132, 1998.

[31] Illumina. HumanMethylation450 BeadChip Achieves Breadth of Coverage Using Two Infinium Chemistries. *Technical Report*, 2012.

[32] An Jansen and Kevin J Verstrepen. Nucleosome positioning in Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, 75(2): 301–320, 2011.

[33] Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.

[34] Jaz Kandola, John Shawe-Taylor, and Nello Cristianini. On the extensions of kernel alignment. Technical report, University of Southampton, 2002.

[35] Martin Kircher, Udo Stenzel, Janet Kelso, et al. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol*, 10(8):R83, 2009.

[36] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *The Journal of Machine Learning Research*, 12:953–997, 2011.

[37] Skirmantas Kriaucionis and Nathaniel Heintz. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324(5929):929–930, 2009.

[38] Felix Krueger and Simon R Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11): 1571–1572, 2011.

[39] Felix Krueger, Benjamin Kreck, Andre Franke, and Simon R Andrews. DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–151, 2012.

[40] Abhishek Kumar, Alexandru Niculescu-Mizil, Koray Kavukcuoglu, and Hal Daume III. A binary classification framework for two-stage multiple kernel learning. *arXiv preprint arXiv:1206.6428*, 2012.

[41] Peter W Laird. Cancer epigenetics. *Human Molecular Genetics*, 14 (suppl 1):R65–R76, 2005.

[42] Gert RG Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[43] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[44] Heidi Ledford. Language: disputed definitions. *Nature*, 455(7216):1023–1028, 2008.

[45] Lei Li and Terence P Speed. An estimate of the crosstalk matrix in four-dye fluorescence-based DNA sequencing. *Electrophoresis*, 20(7): 1433–1442, 1999.

[46] Paul Lichtenstein, Niels V Holm, Pia K Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, 343(2):78–85, 2000.

[47] Yaping Liu, Kimberly D Siegmund, Peter W Laird, Benjamin P Berman, et al. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol*, 13(7):R61, 2012.

[48] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol*, 13(6):R44, 2012.

[49] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010.

[50] Peter Meinicke, Maike Tech, Burkhard Morgenstern, and Rainer Merkl. Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 5(1):169, 2004.

[51] Alexander Meissner, Andreas Gnirke, George W Bell, Bernard Ramsahoye, Eric S Lander, and Rudolf Jaenisch. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877, 2005.

[52] Erica L Mersfelder and Mark R Parthun. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic Acids Research*, 34(9):2653–2662, 2006.

[53] Michael L Metzker. Sequencing technologies: the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[54] Fumihito Miura, Yusuke Enomoto, Ryo Dairiki, and Takashi Ito. Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, 40(17):e136–e136, 2012.

[55] Kensuke Nakamura, Taku Oshima, Takuya Morimoto, Shun Ikeda, Hirofumi Yoshikawa, Yuh Shiwa, Shu Ishikawa, Margaret C Linak, Aki

Hirai, Hiroki Takahashi, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, 39(13):e90–e90, 2011.

[56] National Institute of General Medical Sciences. The New Genetics. `http://publications.nigms.nih.gov/thenewgenetics/chapter1.html`, 2014. [Online; accessed 24-March-2014].

[57] Kisho Ohtani and Stefanie Dimmeler. Epigenetic regulation of cardiovascular differentiation. *Cardiovascular Research*, 90(3):404–412, 2011.

[58] Marcus E Pembrey, Lars Olov Bygren, Gunnar Kaati, Sören Edvinsson, Kate Northstone, Michael Sjöström, and Jean Golding. Sex-specific, male-line transgenerational responses in humans. *European Journal of Human Genetics*, 14(2):159–166, 2006.

[59] Arturas Petronis. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*, 465(7299):721–727, 2010.

[60] Theresa Phillips. The role of methylation in gene expression. *Nature Education*, 1(1):116, 2008.

[61] E Magda Price, Allison M Cotton, Lucia L Lam, Pau Farré, Eldon Emberly, Carolyn J Brown, Wendy P Robinson, Michael S Kobor, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics Chromatin*, 6(1):4–4, 2013.

[62] Martin L Privalsky. Depudecin makes a debut. *Proceedings of the National Academy of Sciences*, 95(7):3335–3337, 1998.

[63] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1): D61–D65, 2007.

[64] Gunnar Rätsch, Sören Sonnenburg, and Bernhard Schölkopf. RASE: recognition of alternatively spliced exons in C. elegans. *Bioinformatics*, 21(suppl 1):i369–i377, 2005.

[65] Eric J Richards. Inherited epigenetic variation: revisiting soft inheritance. *Nature Reviews Genetics*, 7(5):395–401, 2006.

[66] Vincenzo EA Russo, Robert A Martienssen, Arthur D Riggs, et al. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

[67] Yoshimasa Saito, Hidetsugu Saito, et al. Role of CTCF in the regulation of microRNA expression. *Frontiers in Genetics*, 3:186–186, 2011.

[68] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5 (1):3–55, 2001.

[69] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.

[70] Andrea Smallwood, Pierre-Olivier Estève, Sriharsa Pradhan, and Michael Carey. Functional cooperation between HP1 and DNMT1 mediates gene silencing. *Genes & Development*, 21(10):1169–1178, 2007.

[71] Sören Sonnenburg, Gunnar Rätsch, and Christin Schäfer. Learning interpretable SVMs for biological sequence classification. In *Research in Computational Molecular Biology*, pages 389–407. Springer, 2005.

[72] Sören Sonnenburg, Gunnar Rätsch, and Konrad Rieck. Large scale learning with string kernels. In *Large Scale Kernel Machines*. MIT Press, 2007.

[73] Sören Sonnenburg, Alexander Zien, Petra Philips, and Gunnar Rätsch. POIMs: positional oligomer importance matrices - understanding support vector machine-based signal detectors. *Bioinformatics*, 24(13):i6–i14, 2008.

[74] Duncan Sproul, Colm Nestor, Jayne Culley, Jacqueline H Dickson, J Michael Dixon, David J Harrison, Richard R Meehan, Andrew H Sims, and Bernard H Ramsahoye. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proceedings of the National Academy of Sciences*, 108 (11):4364–4369, 2011.

[75] Kristen H Taylor, Robin S Kramer, J Wade Davis, Juyuan Guo, Deiter J Duff, Dong Xu, Charles W Caldwell, and Huidong Shi. Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer research*, 67(18):8511–8518, 2007.

[76] 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 450K DNA methylation data. *Bioinformatics*, 29(2):189–196, 2013.

[77] Trygve Tollefsbol. *Handbook of epigenetics: the new molecular and medical genetics*. Academic Press, 2010.

[78] Daniel J Tomso and Douglas A Bell. Sequence context at human single nucleotide polymorphisms: overrepresentation of cpg dinucleotide at polymorphic sites and suppression of variation in cpg islands. *Journal of molecular biology*, 327(2):303–308, 2003.

[79] Timothy J Triche, Daniel J Weisenberger, David Van Den Berg, Peter W Laird, and Kimberly D Siegmund. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, 41(7): e90–e90, 2013.

[80] Toshikazu Ushijima, Naoko Watanabe, Eriko Okochi, Atsushi Kaneda, Takashi Sugimura, and Kazuaki Miyamoto. Fidelity of the methylation pattern and its variation in the genome. *Genome research*, 13(5):868–874, 2003.

[81] Peter M Warnecke, Clare Stirzaker, Jenny Song, Christoph Grunau, John R Melki, and Susan J Clark. Identification and resolution of artifacts in bisulfite sequencing. *Methods*, 27(2):101–107, 2002.

[82] Ian CG Weaver, Nadia Cervoni, Frances A Champagne, Ana C D'Alessio, Shakti Sharma, Jonathan R Seckl, Sergiy Dymov, Moshe Szyf, and Michael J Meaney. Epigenetic programming by maternal behavior. *Nature Neuroscience*, 7(8):847–854, 2004.

[83] Wikipedia. Illumina bridge amplification. `http://en.wikipedia.org/wiki/File:DNA_Sequencing_Bridge_Amplification.png`, 2014. [Online; accessed 24-March-2014].

[84] Wikipedia. 5-Methylcytosine. `http://upload.wikimedia.org/wikipedia/commons/7/7b/Cytosine_5-methylation.png`, 2014. [Online; accessed 24-March-2014].

[85] Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7): 873–881, 2010.

[86] Sharon Wulff and Cheng Soon Ong. Analytic center cutting plane method for multiple kernel learning. *Annals of Mathematics and Artificial Intelligence*, pages 1–17, 2013.

[87] Xu Zhang, Wenbo Mu, and Wei Zhang. On the analysis of the Illumina 450K array data: probes ambiguously mapped to the human genome. *arXiv preprint arXiv:1308.1912*, 2013.

Abbreviations

BAM binary sequence alignment/map.

BMIQ beta mixture quantile dilation.

CIGAR compact idiosyncratic gapped alignment report.

CNV copy number variation.

CpG cytosine-phosphate-guanine.

CPI CpG island.

CSE context-sensitive error.

CTCF CCCTC-binding factor.

CV cross-validation.

DEEP Deutsches Epigenom Projekt.

DNA deoxyribonucleic acid.

DNMT DNA methyltransferase.

DREME discriminative DNA motif discovery.

EM expectation-maximization.

GATK genome analysis toolkit.

HAT histone acetyl transferase.

HDAC histone deacetylase.

HMM hidden Markov model.

HMT histone methyltransferase.

HOK hybrid oligo kernel.

HP1 heterochromatin protein 1.

HWDK hybrid weighted degree kernel.

HWDKS hybrid weighted degree kernel with shifts.

IHEC international human epigenome consortium.

MBD methylated-CpG-binding domain protein.

MeDIP-seq methylated DNA immunoprecipitation sequencing.

MKL multiple kernel learning.

MSE mean squared error.

OK oligo kernel.

PBC peak-based correction.

PC principal component.

PCA principal component analysis.

PCR polymerase chain reaction.

PO positional oligomer.

POIM positional oligomer importance matrix.

RBF radial basis function.

RNA ribonucleic acid.

RRBS reduced-representation bisulfite sequencing.

SAM S-adenosyl methionine.

SNP single nucleotide polymorphism.

SVM support vector machine.

SVR support vector regression.

SWAN subset-quantile within array normalization.

WDK weighted degree kernel.

WDKS weighted degree kernel with shifts.

WGBS whole-genome bisulfite sequencing.

Index