The statistical analysis of high-throughput assays for studying DNA methylation

Peter Francis Hickey

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

May 2015

Department of Mathematics and Statistics
The University of Melbourne

Bioinformatics Division
The Walter and Eliza Hall Institute of Medical Research

Abstract

DNA methylation is an epigenetic modification that plays an important role in X-chromosome inactivation, genomic imprinting and the repression of repetitive elements in the genome. It must be tightly regulated for normal mammalian development and aberrant DNA methylation is strongly associated with many forms of cancer.

This thesis examines the statistical and computational challenges raised by high-throughput assays of DNA methylation, particularly the current gold standard assay of whole-genome bisulfite-sequencing. Using whole-genome bisulfite-sequencing, we can now measure DNA methylation at individual nucleotides across entire genomes. These experiments produce vast amounts of data that require new methods and software to analyse.

The first half of the thesis outlines the biological questions of interest in studying DNA methylation, the bioinformatics analysis of these data, and the statistical questions we seek to address. In discussing these bioinformatics challenges, we develop software to facilitate novel analyses of these data. We pay particular attention to analyses of methylation patterns along individual DNA fragments, a novel feature of sequencing-based assays.

The second half of the thesis focuses on *co-methylation*, the spatial dependence of DNA methylation along the genome. We demonstrate that previous analyses of co-methylation have been limited by inadequate data and deficiencies in the applied statistical methods. This motivates a study of co-methylation from 40 whole-genome bisulfite-sequencing samples. These 40 samples represent a diverse range of tissues, from embryonic and induced pluripotent stem cells, through to somatic cells and tumours. Making use of software developed in the first half of the thesis, we explore different measures of co-methylation and relate these to one another. We identify genomic features that influence co-methylation and how it varies between different tissues.

In the final chapter, we develop a framework for simulating whole-genome bisulfite-sequencing data. Simulation software is valuable when developing new analysis methods since it can generate data on which to assess the performance of the method and benchmark it against competing methods. Our simulation model is informed by our analyses of the 40 whole-genome bisulfite-sequencing samples and our study of co-methylation.

Declaration

This is to certify that

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface,

(ii) due acknowledgement has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed,

Preface

The datasets used in this thesis were provided by the investigators listed below.

Aaron Statham and Sue Clark, Garvan Institute of Medical Research; Peter Molloy, CSIRO. The *EPISCOPE* dataset.

Several publicly available datasets were also used and are described in Chapter 3.

Acknowledgments

I would like to thank my PhD supervisors, Terry Speed and Peter Hall. Peter, thank you for supporting my decision to radically change tack very early on in my PhD from a topic in theoretical statistics to a heavily computational and interdisciplinary problem. Terry, thank you for agreeing to take me on as a student after this change of heart. I have learnt much from working with you and greatly appreciate your generosity and guidance over the course of my PhD. Thanks also to Gordon Smyth, the chair of my PhD committee, for always taking the time to help me.

Thank you to the Walter and Eliza Hall Institute for being such a great place to do research. In particular, thank you to my wonderful friends and colleagues in the Bioinformatics Division for your support and advice. A special thanks to Keith Satterley and the IT department who both tolerated and helped fix my computational blunders.

My PhD would not be possible without the people who provided me with data to analyse. I am grateful to Sue Clark, Peter Molloy, Aaron Statham, Emma Whitelaw and Harry Oey for the opportunity to collaborate with and learn from you. Thanks also to Ryan Lister, Kasper Hansen and Felix Krueger for making their data publicly available and answering all my nitpicky questions. Open science is the best science.

Equally, my PhD would not be possible without the countless people who contribute to the software I use in my daily research. A special thanks to those who helped me as I learnt to write software of my own: Felix Krueger, Toby Sargeant, Martin Morgan and Hervé Pagès. Open software is the best software.

I am grateful to the Australian tax payer for supporting me financially during my studies with an Australian Postgraduate Award. I also thank the Victoria Life Sciences Computing Initiative for additional funding and Melanie Bahlo for allowing me a very flexible work arrangement as a research assistant and her encouragement to pursue a PhD.

I owe an enormous debt of gratitude to my thesis writing buddies, Hannah Vanyai and Darcy Moore, who kept me going during tough times and to my thesis proofreaders, Belinda Phipson, Saskia Freytag, Goknur Giner, and Dineika Chandrananda who have greatly improved this thesis.

Finally, and most of all, thank you to my parents, Gay and John, to my brothers, Jack, Luke and Daniel, and to my friends. Your support, love, and encouragement kept me going.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction to genomics and DNA methylation

Overview

This chapter summarises the basic biology necessary for understanding this thesis. It introduces DNA methylation, describes some of its biological roles, and introduce assays for studying DNA methylation, with a particular focus on bisulfite-sequencing.

1.1 From DNA to genomes

The genome is the genetic material of an organism. In most organisms, the genome is encoded by deoxyribonucleic acid, *DNA*. In eukaryotes[1], which includes plants, fungi, and animals, the DNA is wound around repeating units of eight *histone* protein cores to form *nucleosomes*, which are the fundamental unit of *eukaryotic chromatin*. Chromatin compactly packages the DNA into *chromosomes*, so that the organism's complete nuclear DNA, its nuclear genome, might fit into the nucleus of its cells[2].

[1] Eukaryotes are organisms composed of one or more cells with a distinct nucleus and cytoplasm. Most of a eukaryotic cell's DNA is contained in the nucleus [Alberts *et al.* 2007].

[2] Not all of an organism's genome is present in the nucleus of a cell. Important exceptions are mitochondrial DNA (mtDNA) found in animals, plants and fungi and chloroplast DNA (ctDNA) found in plants.

1.1.1 DNA

The double helix is the most common, and most famous, structure of DNA. In the double helix, the two strands of DNA run in opposite directions to each other and are therefore *anti-parallel*. One strand is called the 5' strand (pronounced "5 prime strand" and also known as the sense strand, Crick strand or top strand) and the other strand is called the 3' strand (pronounced "3 prime strand" and also known as the antisense strand, Watson strand or bottom strand). Along each strand of the double helix are the four DNA nucleobases (*bases*): adenine (A), cytosine (C), guanine (G) and thymine (T). These bases form *complementary* base pairings, A with T and C with G, along the DNA double helix. This is illustrated in Figure 1.1.

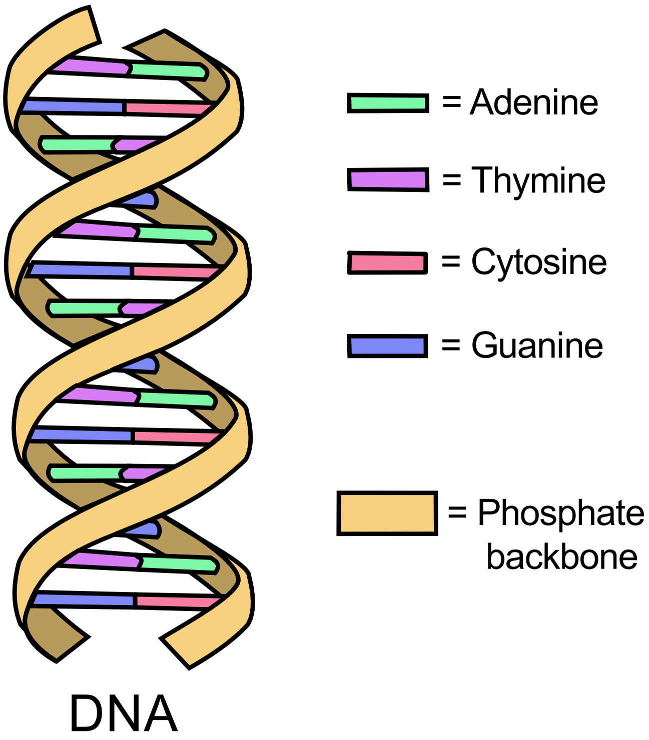


Figure 1.1: Simple diagram of double-stranded DNA showing complementary base pairing. By Forluvoft (Own work) [Public domain], via Wikimedia Commons `http://commons.wikimedia.org/wiki/File%3ADNA_simple2.svg`

A gene is a sequence of DNA that is *transcribed* to produce a functional product in the form of ribonucleic acid, *RNA*. RNA may in turn be translated into a protein sequence or perform other roles in the regulation of gene expression. It is important to note here than not all DNA is transcribed into RNA, which is not to say that untranscribed DNA is unimportant. For instance, there are untranscribed *regulatory sequences* of DNA that determine whether a nearby gene is transcribed. There is also *junk DNA* that is of little consequence to the organism [Alberts *et al.* 2007]. Conversely, not all transcribed DNA is a gene. DNA transcription is *permissive* and there are many DNA sequences that are transcribed by accident or in error.

DNA is able to self-replicate. This means that eukaryotic cells created during *mitosis* contain the same DNA as the 'parent' cell. During DNA replication, the two strands are separated and each strand's complementary DNA sequence is copied by an enzyme called *DNA polymerase*. It is because the two strands of DNA are complementary that ensures the daughter cell contains the same DNA sequence as the parent cell[3].

1.1.2 Nucleosomes and chromatin

The core of a nucleosome consists of four pairs of histones, H2A, H2B, H3 and H4, which are consequently known as the *core histones*. A fifth histone, H1/H5, is known as the *linker histone*. Each of these histones has a 'tail' consisting of a string of amino acids. These tails can undergo *post-translational modifications*, such as methylation, acetylation and phosphorylation, which can alter their interactions with DNA and nuclear proteins [Alberts *et al.* 2007]. Histone modifications are discussed in Section 1.2.

The nucleosomes are interconnected by *linker DNA* to form the macromolecule called chromatin. The linker sequences are between 20 to 60 base pairs (bp) of DNA in length, while approximately 147 bp are wrapped around each nucleosome and a further 20 bp wrapped around each additional H1/H5 histone [Annunziato 2008].

Chromatin is often described as either 'closed' or 'open'. Closed chromatin, *heterochromatin*, is more tightly packed than the open *euchromatin*. Heterochromatin is associated with transcriptionally repressed regions of the genome because the machinery required to translate DNA to RNA is less able to physically access the DNA. In contrast, euchromatin

[3] This of course ignores errors in the replication process. Such errors are very rare events but because DNA replication happens so frequently these events do occur. There are error-correcting processes that reduce the chance that such an error is retained in the daughter sequence, however, these are not perfect. Hence errors in DNA replication are one source of what are known as *mutations* in the DNA.

is associated with transcriptionally active regions of the genome since its more open nature allows easier access for the translational machinery.

The same region of an individual's genome may vary between heterochromatin and euchromatin states at different stages of the organism's life and in different cells of the organism. This is one mechanism by which gene expression is regulated.

Basic descriptions of nucleosomes and chromatin are ripe with analogies. In one such popular analogy, the histones are the 'spool' around which the DNA 'thread' is wrapped to form nucleosomes. The chromatin then has the appearance of "beads on a string" when viewed under an electron microscope [Alberts *et al.* 2007]. Chromatin is further coiled up into various literally-named structures, such as the 30-nanometre and 250-nanometre fibres, and, ultimately, packaged into *chromosomes*. The set of chromosomes makes up an individual's *genome*.

1.1.3 Genomes

As is clear from the above description, the genome is a complex three-dimensional structure. Nonetheless, in bioinformatics and computational biology, the genome is mostly considered as a single-stranded, one-dimensional string of the bases *A*, *C*, *G* and *T*.

While eukaryotic genomes share the above-described features, and indeed share many regions of common DNA sequence, eukaryotic genomes come in many shapes and sizes. I only discuss the genomes of two species relevant to my thesis: *Homo sapiens* (human) and *Mus musculus* (house mouse).

The human genome

Humans are *diploid* organisms, meaning that we have two copies of each chromosome in a typical cell[4]. We inherit one chromosome of each pair from our mother and one from our father. A typical human cell has 23 pairs of nuclear chromosomes, 22 *autosomes* and 1 pair of *sex chromosomes*, as well as hundreds or thousands of copy of the small mitochondrial chromosome[5].

[4] A sperm or egg cell is haploid and has a single (recombined) copy of each chromosome.

[5] The mitochondrial DNA is maternally inherited.

The length of a chromosome is typically reported as the number of DNA base pairs in a single copy of that chromosome and the *haploid* length of a genome is the sum of these chromosome lengths. The haploid human genome is approximately 3 billion base pairs long (Golden Path Length `http://asia.ensembl.org/Homo_sapiens/Location/Genome?r=1`).

A human *reference genome* was jointly completed by the International Human Genome Consortium and Celera Genomics in 2003 [Venter *et al.* 2001, Lander *et al.* 2001]. This reference genome does not represent the genome of any one human since it uses DNA donated by several different people [Venter *et al.* 2001, Lander *et al.* 2001]. Rather, a reference genome is a kind of map or scaffold that can be used to identify similarities and differences between individual genomes.

The human genome has obvious uses in medical research and biotechnology, but is also used to learn about evolution and human history, such as human migration patterns [Hellenthal *et al.* 2014]. Every person, even a *monozygotic* ('identical') twin, has their own unique genome [Bruder *et al.* 2008]. However, genomes of any two randomly selected people are identical at approximately 99.9% of sites. Furthermore, the vast majority of the human genome, 98% by some estimates [Elgar and Vavouri 2008], is made up of non-coding DNA and upwards of 50% is repetitive sequence [Treangen and Salzberg 2012].

The mouse genome

Mice are also diploid organisms, but have 19 pairs of autosomes, one pair of sex chromosomes and a mitochondrial chromosome. The mouse genome is slightly smaller than the human genome, at 2.7 billion base pairs long (Golden Path Length `http://asia.ensembl.org/Mus_musculus/Location/Genome?r=1`). Like the human genome, there is a mouse reference genome [Mouse Genome Sequencing Consortium *et al.* 2002]. It is based on several female mice from the *C57BL/6J* strain, an important strain of mouse widely used in medical research.

Mouse strains used in medical research are highly inbred due to years of concerted mating programs. This reduced genetic variability, and the control that researchers have over it, make these mice a very powerful tool in identifying the biological cause of a diverse range of phenotypes.

1.1.4 Genetic variation

A key question in biology, perhaps *the* key question, is what determines an individual's *phenotype*. An individual's phenotype is the set of its observable characteristics resulting from the interaction of its genotype with the environment. Two simple examples of phenotypes are height and weight. Both have a genetic component, e.g., the offspring of tall parents are on average taller than the offspring of short parents, but also have environmental components, such as the contribution of diet to weight. In medical research, a person's phenotype might be whether she is affected by a particular disease. It might also be some proxy, such as her blood pressure or the expression levels of particular genes.

Variation between individual's genomes, be it at single base[6] or across larger regions[7], is one important source of phenotypic variation. Importantly, genetically-driven phenotypic variation is frequently *heritable*, meaning that phenotypes can be passed on from one generation to the generation via the genome.

Environmental variation has clear influence on certain phenotypes. However, it often can be difficult to determine whether phenotypic variation is driven by genetic variation or environment variation, particularly in humans where genetically similar individuals typically also grow up in similar environments.

The above discussion has been about phenotypic variation in a population. But there is also phenotypic variation within the individual. If that sounds strange, consider the fact that in your body a neuron, a leukocyte (a white blood cell), and a cone cell (a photoreceptor in the retina) all have identical genomes. In fact, all cells in an organism, excluding the gametes, have an identical genome[8], yet play very different biological roles. This is due to different genes being active in different cells.

[6] A base position that is variable in the population is called a single nucleotide variant (*SNV*). A SNV that is frequently variable in the population, say at least 1% frequency, is called a single nucleotide polymorphism (SNP). All SNPs are SNVs but the converse is not true.

[7] An example of a larger genetic variant is an *indel*, which is a short insertion or deletion of sequence in an organism's DNA, usually with respect to a reference genome.

[8] Even this is a simplification. For example, it ignores somatic mutations (the occurrence of a mutation in the somatic tissue of an organism, resulting in a genetically mosaic individual) and V(D)J recombination (which occurs in lymphocytes and is vital for antibody diversity).

Epigenetics, which I describe in the next section, plays a role in determining the between-individual phenotypic variation, as well as the within-individual phenotypic variation. DNA methylation, the focus of my thesis, is the prototypical and most well-studied epigenetic modification.

1.2 Epigenetics

Interest in epigenetics has grown remarkably in recent years. However, *epigenetics* is also a real Humpty Dumpty phrase; each author seems to believe that, “When I use [the] word, . . . it means just what I choose it to mean — neither more nor less” [Carroll and Tenniel 1897]. As noted by Deans and Maggert [2015], “the unfortunate fact is that the increased use of the term *epigenetics* is likely due more to inconsistencies in its definition than to a consensus of interest among scientists”.

Conrad Waddington coined the phrase in 1942 as a portmanteau of the words ‘epigenesis’ and ‘genetics’ [Waddington 2012]. Waddington meant epigenetics as the study of how “processes involved in the mechanism by which the genes of the genotype bring about phenotypic effects” [Waddington 2012].

A popular contemporary definition of epigenetics is attributed to the epigeneticist Arthur Riggs — epigenetics is “the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence[9]” [Russo *et al.* 1996, pp. 1]. The *epi* prefix, derived from the Greek word for ‘upon’, ‘near to’, or ‘in addition’, emphasises the idea that epigenetics encodes information ‘on top of’ the DNA sequence. However, it is quite different to Waddington’s original definition.

More recently, the definition of epigenetics has taken on a “more biochemical flavour” [Daxinger and Whitelaw 2010] to include marks whose heritability is yet to be established. The heritability, or lack thereof, of histone modifications means that many epigeneticists do not consider these to be truly epigenetic [Berger *et al.* 2009], and describing them as such is a sure-fire way to annoy a good percentage of your audience [Ledford 2008].

Sir Adrian Bird, an esteemed British geneticist, attempts to unite these definitions

[9] Mitotically heritable means heritable during cell division and meitotically heritable means heritable during sexual reproduction.

[Bird 2007]:

> "The following could be a unifying definition of epigenetic events: the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states. This definition is inclusive of chromosomal marks [e.g., histone modifications], because transient modifications associated with both DNA repair or cell-cycle phases and stable changes maintained across multiple cell generations qualify."

Regardless of which definition you subscribe to, an *epigenetic mark* is the modification that causes this 'epigenetic change'. In fact, 'causes' may be too strong a claim, as much of current epigenetics research is in identifying associations rather than causations, and the question of whether the epigenetic mark is a cause or consequence of the ascribed function is oft-debated.

The *epigenome* of a cell is the set of epigenetic marks present on the cell's genome. In contrast to the genome, which is identical between cells within an individual, the epigenome is highly variable between cells within an individual. Indeed, we can identify variation for a single epigenetic mark within cells of the same cell type from the same individual. *Epigenomics* is the study of the epigenome, analogous to genomics being the study of the genome. However, one can rarely study the epigenome in isolation from the underlying genetic sequence, as there is evidence that the epigenetic variation is associated with genetic variation [Zhang *et al.* 2010, Bell *et al.* 2011, van Eijk *et al.* 2012, McVicker *et al.* 2013]

One mark that most authors agree is an epigenetic mark is DNA methylation, which I describe in the next section and the study of which is the focus of my thesis.

1.3 DNA methylation

DNA methylation is a chemical modification of DNA that can impart information on top of the DNA sequence. It is heritable during mitotic cell division, which means that it is faithfully copied across to the daughter cell during cell division[10]. It therefore fits into

[10] In practice this copying is not as faithful as, say, the copying of DNA from the parent to the daughter strand. Furthermore, the faithfulness of this copying will be different in different conditions, such as in

Arthur Riggs' aforementioned definition of an epigenetic modification. DNA methylation is found in bacteria, fungi, plants and animals.

A major reason that DNA methylation is studied is that it is essential for normal development in mammals [Li *et al.* 1992]. It is also involved in many key biological processes related to human development and health, such as the regulation of gene expression [Razin and Cedar 1991], silencing of transposable elements [Jones and Takai 2001], X chromosome inactivation [Mohandas *et al.* 1981] and tumorigenesis [Ehrlich 2002].

However, DNA methylation is not found in all organisms[11], and so is not essential for life. Furthermore, the level of DNA methylation varies widely amongst different organisms, with many having very low levels of methylation [Capuano *et al.* 2014].

When people speak of DNA methylation they are generally referring to methylation of the cytosine base. Even more specifically, they are referring to the molecule *5-methylcytosine*. 5-methylcytosine, abbreviated as *5mC*, is by far the most common form of DNA methylation in the animal and plant kingdoms[12]. However, I will continue to use the term *DNA methylation* to describe the more specific 5mC, as is standard in the literature.

A German chemist, W.G. Ruppel, first identified a methylated nucleic acid in 1898. Ruppel was studying *tuberculinic acid*, the poison of *Mycobacterium tuberculosis*[13] , and discovered that it contained a methylated base [Ruppel 1899]. In 1925, Johnson and Coghill isolated 5-methylcytosine as a product of hydrolysis of tuberculinic acid, the nucleic acid of *Mycobacterium tuberculosis* [Johnson and Coghill 1925]. However, Johnson and Coghill's results were disputed for over twenty years by other researchers who were unable to replicate the original findings [Vischer *et al.* 1949].

In 1945, Hotchkiss ultimately proved Johnson and Coghill correct when he isolated 5-methylcytosine from nucleic acid prepared from cow thymus [Hotchkiss 1948]. Using paper chromatography, Hotchkiss demonstrated that methylated cytosine existed and was

a healthy liver cell compared to a cancerous liver cell. Nevertheless, the copying of DNA methylation is faithful enough for most biologists to consider it as a mitotically heritable mark, most of the time. The enzymes responsible the replication of DNA methylation, the DNA methyltransferases, are discussed in Section 1.3.6.

[11] For example, DNA methylation not detectable in yeast [Capuano *et al.* 2014]

[12] Two additional examples of DNA methylation are N6-methyladenine (m6A) and N4-methylcytosine (m4C). N6-methyladenine is a methylated form of adenine, which is found in mRNA [e.g., Adams and Cory 1975] and DNA, although the latter only in bacterial DNA [Ratel *et al.* 2006]. N4-methylcytosine has also been detected in bacterial DNA [e.g., Ehrlich *et al.* 1985, Ratel *et al.* 2006].

[13] *Mycobacterium tuberculosis* was then known as *Tubercle bacillus.*

distinct from conventional cytosine and uracil.

The typical site of DNA methylation is at the C5 carbon position of a cytosine base, hence 5-methylcytosine. Figure 1.2 shows the structure of 5mC.

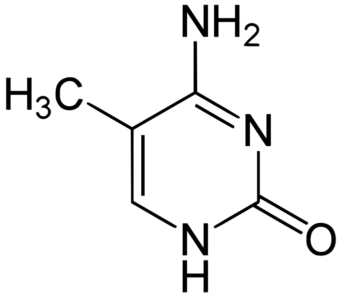


Figure 1.2: Chemical structure of 5-methylcytosine. “5-methylcytosine”. By Yikrazuul (Own work) [Public domain], via Wikimedia Commons `http://commons.wikimedia.org/wiki/File:5-Methylcytosine.svg`

A cytosine may be described being ‘methylated’ or ‘unmethylated’, however, care must be taken when using these terms. At the lowest level, the level of single-stranded DNA, methylation is a binary event: a cytosine is either methylated or unmethylated. Double-stranded DNA, at least at *palindromic* methylation loci[14], is generally symmetrically methylated, i.e. the loci on each strand are both methylated or both unmethylated. However, *hemimethylation*, where the methylation loci on one strand is methylated and its partner on the opposite strand is unmethylated, can and does occur.

Within a diploid cell, a particular cytosine may be unmethylated or methylated on both homologous chromosomes or methylated on one chromosome and unmethylated on the other. While the former is more common, examples of the latter case, such as *allele-specific methylation* [Shoemaker *et al.* 2010] and *genomic imprinting* [Li *et al.* 1993], are important epigenetic phenomena.

1.3.1 DNA methylation in mammals

The importance of 5-methylcytosine in mammalian genomes is such it has been dubbed the “fifth base” of the DNA code [Lister and Ecker 2009]. In mammalian genomes, most

[14] A palindromic DNA sequence is one that is identical when read in the 5’ to 3’ direction on both the original strand and the complementary strand of the double helix. For example, *CG* is a palindromic sequence.

cytosines are unmethylated except for those at *CpG dinucleotides*. A CpG dinucleotide, or more simply a *CpG*, is a cytosine followed by a guanine in the linear DNA sequence. The 'p' stands for the phosphate backbone of DNA and some authors omit it in favour of simply calling it CG methylation. A CpG is a palindromic sequence and is generally symmetrically methylated. Approximately 70% of CpGs are methylated in mammals [Laird 2003], meaning that the cytosine in the CpG dinucleotide is a 5-methylcytosine.

1.3.2 CpG dinucleotides

CpGs are underrepresented in the human genome. The *GC-content* of the human genome, which is defined as the percentage of bases that are either guanines or cytosines [Benjamini and Speed 2012], is approximately 41%. If these bases were uniformly distributed across the genome then we would expect about 4.1% of dinucleotides to be CpGs. Instead, only 1% of dinucleotides are CpGs.

One reason for the relative scarcity of CpGs is that methylated cytosines can spontaneously deaminate to thymines [Scarano *et al.* 1967]. Thus, over time, many methylated CpGs will become TpGs, leading to a genome-wide reduction in the proportion of CpGs and a genome-wide increase in the proportion of TpGs (see Figure 1.3). There are many other evolutionary pressures on the distribution of bases in a genome. One effect of this is that the distribution of CpGs is far from uniform. In fact, CpGs tend to form clusters, which are termed CpG islands.

1.3.3 CpG islands and other sandy metaphors

One way to explore the distribution of CpGs in the human genome is to look at the distribution of distances from one CpG to the next, the *intra-pair distances* (IPDs). Figure 1.4 is a plot of the empirical cumulative distribution function of CpG IPDs for the human reference genome (hg19). We see that approximately 70% of CpGs are within 100 bp of the next CpG. Figure 1.4 also shows the expected IPD distribution under a model where CpGs are uniformly distributed along the genome with probability equal to the observed frequency of CpGs on each chromosome. By comparing the observed IPDs to the expected IPDs we see that the distribution of distances between CpGs has more 'close' pairs than

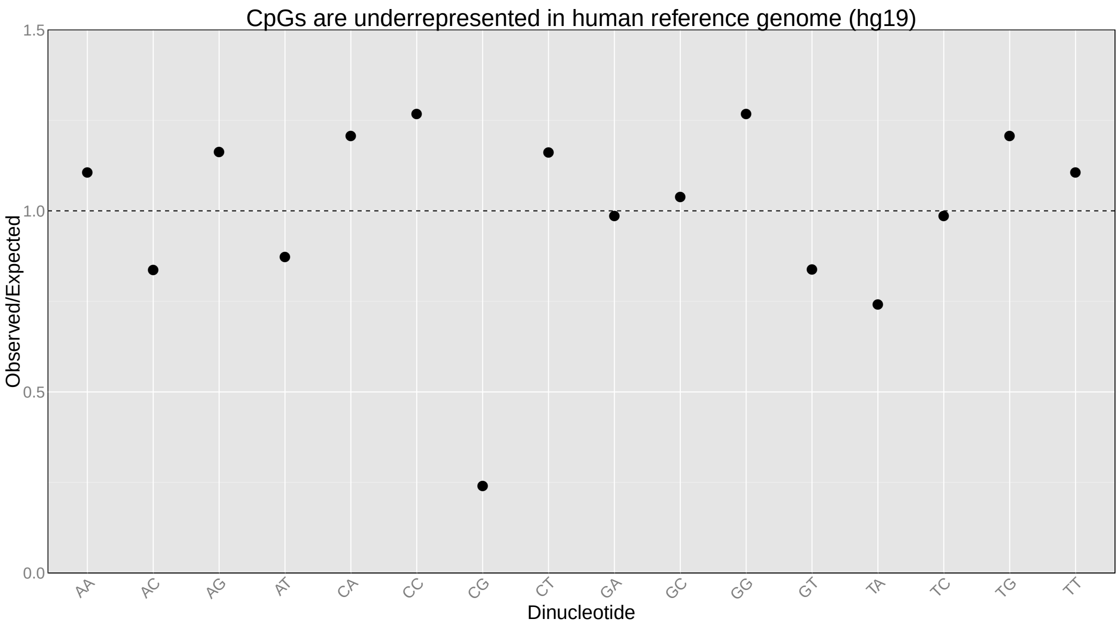


Figure 1.3: Observed to expected ratio of dinucleotides in the human reference genome (hg19). The expected frequency is computed under an 'independence' model, based on the observed frequencies of each base.

we would expect by chance.

Figure 1.5 is an alternative way to visualise these data by plotting the percentage of pairs of CpGs with a given IPD. Figure 1.5 shows that there is a cluster of CpGs with $IPD < 10$. These largely correspond to CpGs that lie within what are called *CpG islands* (CGIs).

CpG islands contain the 20% to 40% of CpGs that are frequently unmethylated in mammalian genomes. CpG islands are important regulatory elements in the genome and are where most differences in DNA methylation between different cell types are found [Wu *et al.* 2010]. The classical definition of a CpG island, given by Gardiner-Garden and Frommer [1987], and used by the popular UCSC genome browser (`http://genome.ucsc.edu/cgi-bin/hgGateway`), is a region of the genome where the following conditions are satisfied:

1. The (moving) average of GC-content is greater than 50%, and
2. The observed-to-expected ratio of CpGs is greater than 0.6, and
3. The region is longer than 200 bp.

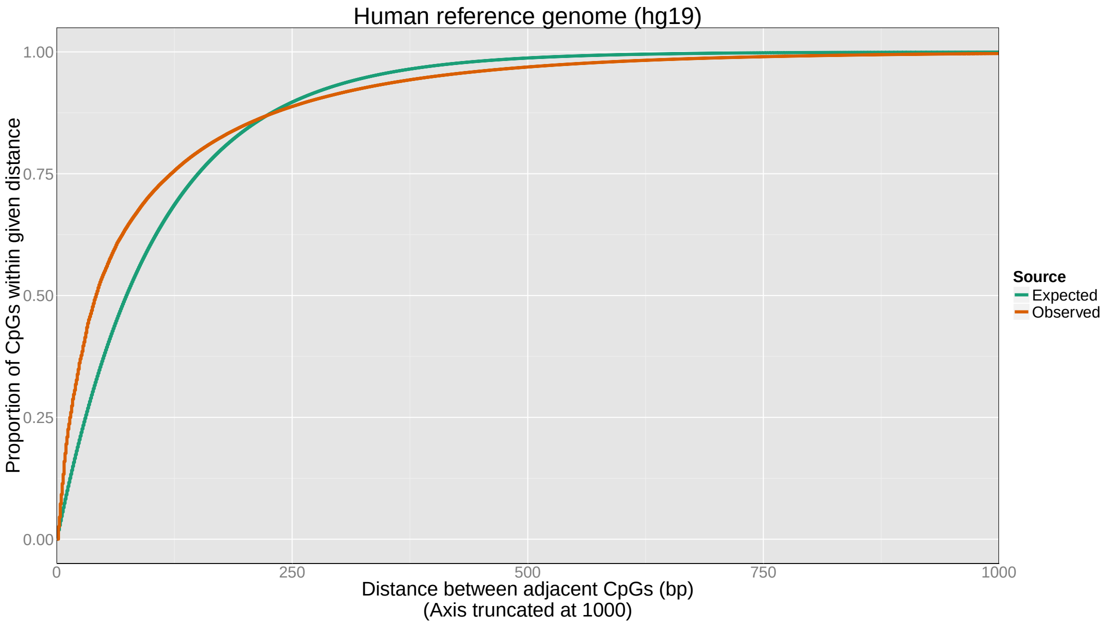


Figure 1.4: Plot of the empirical cumulative distribution function of distance between adjacent CpGs in the human reference genome (hg19). The observed distances are contrasted with those under the 'expected' model whereby CpGs are distributed uniformly at random with probability equal to the observed frequency of CpGs on each chromosome.

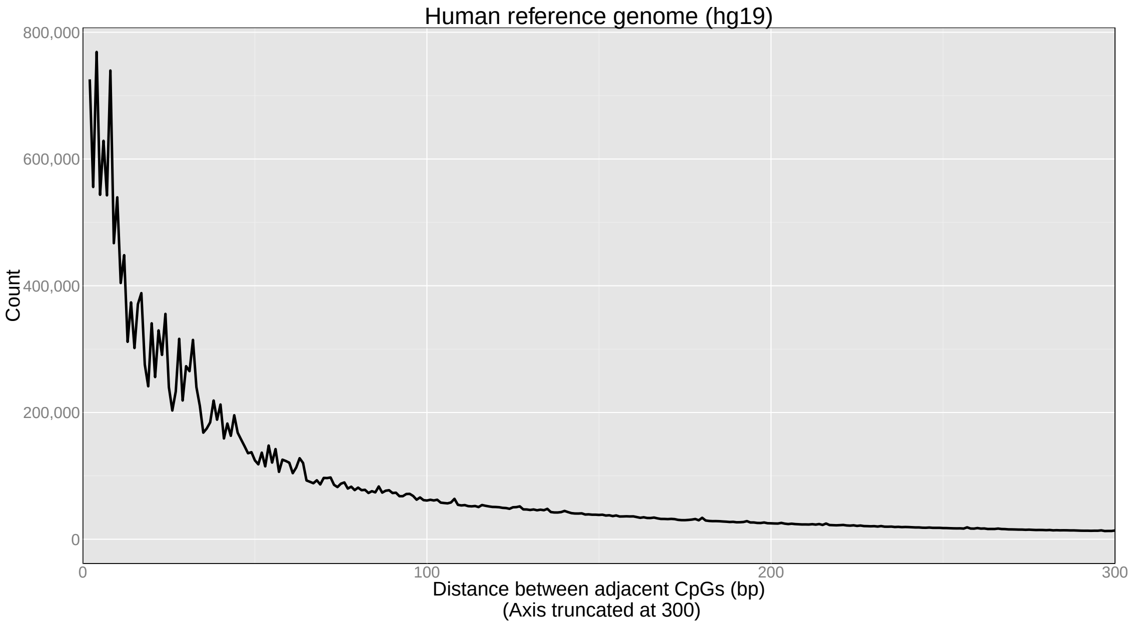


Figure 1.5: The frequency of the observed distances between adjacent CpGs in the human reference genome (hg19).

This definition was refined by Takai and Jones [2002] to exclude *Alu*-repetitive elements, which are otherwise misclassified as *bona fide* CpG islands. More recently, Wu *et al.* [2010] developed a hidden Markov model to predict CpG islands based on the CpG density and GC-content of the region; this definition is used in the remainder of my thesis.

An alternative definition, and one that pre-dates the definition of Gardiner-Garden and Frommer [1987], is based on the identification of unmethylated regions of the genome, which are typically CpG-dense. Such regions were previously called HpaII tiny fragment islands, or *HTF islands*, and named after the restriction enzyme used to identify them [Cooper *et al.* 1983, Bird *et al.* 1985].

These 'sandy/beachy' metaphors have been continued (i.e. stretched to breaking point) with various authors defining CpG island shores, CpG island shelves, CpGs in the open sea, CpG deserts and CpG canyons. CpG island shores, shelves and the open sea are all defined with respect to CpG islands:

- CpG island shores are regions within 2 kb of CpG islands. These have been demonstrated to have an increased variability of CpG methylation [Irizarry *et al.* 2009].
- CpG island shelves are defined as regions within 2 kb of a CpG island shore [Bibikova *et al.* 2011].
- The open sea contains those CpGs not classed as being in a CpG island, CpG island shore or CpG island shelf [Sandoval *et al.* 2011].

Other metaphors, these based on methylation levels rather than CpG density, include methylation deserts [Li *et al.* 2012] and CpG canyons [Jeong *et al.* 2014]. Because these regions are defined with respect to methylation levels rather than DNA sequence, these are generally identified in a tissue-specific manner.

1.3.4 Non-CpG methylation

In humans, cytosine methylation in most cell types is found almost exclusively at CpGs [Jones 2012]. There are, however, certain cell types with widespread non-CpG methylation. Non-CpG methylation is often classified as CHH methylation or CHG methylation, where H is the IUPAC code for any base except G (`http://www.bioinformatics.org/sms2/`

`iupac.html`). The rule-of-thumb for mammalian genomes is that non-CpG methylation is rare in somatic cells but common in pluripotent cells. Of course, there are exceptions to every rule, especially in biology.

To give a few examples, Lister *et al.* [2009] found that in a *fibroblast* cell line that 99.98% of cytosines that displayed statistically significant evidence of methylation occured at CpGs. In contrast, in an embryonic stem cell line, they found that 24.5% of cytosines that displayed statistically significant evidence of methylation occured in a non-CpG context. However, it should also be noted that these non-CpG loci that were methylated had, on average, a much lower level of methylation than their CpG counterparts.

A subsequent paper from the same group extended this result. Lister *et al.* [2011] reported that, more generally, non-CpG cytosines account for 20% to 30% of cytosines with statistically significant evidence of methylation in *pluripotent* cell lines. Pluripotent cells includes embryonic stem (*ES*) cells and induced pluripotent stem (*iPS*) cell lines.

An exception to the rule that non-CpG methylation is largely restricted to pluripotent cells is provided by Lister *et al.* [2013], who found that neurons also have non-CpG methylation, albeit at a lower level (1.3% to 1.5% of all non-CpG cytosines were methylated).

Overall, non-CpG methylation in humans is less well studied and less well understood than CpG methylation. This is partly due to sampling bias since commonly used assays for studying DNA methylation, such as the Illumina 27k and 450k microarrays, measure almost exclusively CpG methylation. However, recent technological advances mean that cytosine methylation can be routinely assayed regardless of the sequence context (see Section 1.4).

Non-CpG methylation is very common in other organisms, such as plants. For example, Lister *et al.* [2008] found that in the widely-studied *Arabidopsis thaliana* that 45% of cytosines that displayed statistically significant evidence of methylation are at CHG or CHH loci. They also found that the level of methylation at non-CpG loci, however, is typically lower than that observed at CpG dinucleotides.

1.3.5 Modifications of a modification

Methylation is not the only chemical modification of cytosines, although it is by far the most common. Listed from most frequent to least frequent, these are 5-hydroxymethylcytosine

(*5hmC*), 5-formylcytosine (*5fC*) and 5-carboxylcytosine (*5caC*) [Plongthongkum *et al.* 2014]. The biological significance of these marks is still being determined, in part because the assays for studying these are still in development and because their relatively scarcity in the genome means that experiments to detect these modifications are more difficult and expensive.

One genome-wide study of 5hmC found that less than 1% of all assayed cytosines in mouse fetal cortex and adult cortex cells displayed any statistically significant evidence of hydroxymethylation [Lister *et al.* 2013]. These cytosines appeared to be restricted to the CpG context and had only very low levels of 5hmC.

Kriaucionis and Heintz [2009] and Tahiliani *et al.* [2009] discovered that the TET enzymes can convert 5mC to 5hmC, 5hmC to 5fC and 5fC to hcaC. This suggests a role for 5hmC, 5fC and 5caC in the process of removing 5mC marks.

1.3.6 Writers, readers, and erasers

A frequently used analogy when describing epigentic marks refers to 'writers', 'readers' and 'erasers' [e.g., Moore *et al.* 2013]. In the case of DNA methylation, writers catalyse the methyl group onto the DNA, readers recognise methylated DNA, and erasers remove the methyl group from the DNA.

In mammalian cells, the writers are the DNA *methyltransferase* (DNMT) enzymes. The DNMTs are commonly split into two groups, namely the maintenance methyltransferases and the *de novo* methyltransferases.

DNA methylation is not preserved by the DNA replication machinery and so it is the role of the maintainence methyltransferases to restore the methylation pattern on the daughter strand of DNA following DNA replication. In mammals, *DNMT1* is known as the maintenance DNA methyltransferase.

DNMT3a and *DNMT3b* are known as the *de novo* methyltransferases, although these are also required for the maintenance of DNA methylation [Jones and Liang 2009]. Both *DNMT1* and *DNMT3b* appear to be essential for mammalian development since mouse knockouts[15] for either gene are embryonically lethal [Li *et al.* 1992]. In contrast, mouse

[15] A knockout mouse for gene X is a mouse that has been genetically engineered to remove or otherwise

knockouts for *DNMT3a* are runted but survive for approximately 4 weeks after birth [Li *et al.* 1992].

DNMT2, now known as *TRDMT1*, was once thought to be a DNA methyltransferase but was shown to in fact methylate a small RNA and not DNA [Goll *et al.* 2006]. Another protein, *DNMT3L*, is homologous to *DNMT3a* and *DNMT3b* but does contain catalytic domain that is necessary for methyltransferase activity. Instead, *DNMT3L* is thought to stimulate the activity of *DNMT3a* and *DNMT3b* [Jurkowska *et al.* 2011].

The readers of DNA methylation recognise methylated DNA. These readers can recruit additional proteins to the site of the methylated cytosine to perform a variety of functions related to gene expression. For example, the methyl-CpG-binding domain (*MBD*) group of proteins bind to DNA containing a methylated CpG, which can then suppress gene expression by preventing transcription factor binding at that site [Nan *et al.* 1993]. Another group, the ubiquitin-like, containing PHD and RING finger domain (*UHRF*) proteins, help *DNMT1* methylate hemimethylated DNA, such as the daughter strand created during DNA replication [Sharif *et al.* 2007, Bostick *et al.* 2007].

The removal or erasure of DNA methylation, called *demethylation*, may be characterised as *passive* loss or *active* removal. Passive loss occurs when the maintenance methyltransferases do not efficiently perform their role of restoring DNA methylation following cell division. This leads to a gradual, stochastic, and genome-wide loss of DNA methylation after multiple cell divisions. This form of passive demethylation, sometimes called replication-dependent demethylation, cannot explain observations of local tissue-specific differences in DNA methylation [Irizarry *et al.* 2009] nor the two stages of rapid global demethylation that occur during development [Wu and Zhang 2014].

Active demethylation is currently an active area in epigenetics research. Multiple mechanisms have been proposed, and it is indeed likely that there are multiple ways to achieve active demethylation. These mechanisms were recently reviewed by Wu and Zhang [2014], which I briefly summarise:

1. The direct removal of the methyl group from 5mC is considered unlikely due to the strong carbon-carbon bond between the methyl group and the cytosine.

inactivate gene X. Mouse knockouts can be either heterozygous knockouts (one copy still of the gene is still present/active) or homozygous knockouts (both copies of gene absent/inactive).

2. There is evidence that the DNA repair machinery can be co-opted to remove a methylated base or the surrounding region. The excised base or region is then repaired with unmethylated cytosines replacing 5mCs.
3. 5mC oxidation-dependent active DNA demethylation. This follows from the observation that the TET enzymes can iteratively oxidate a 5mC → 5hmC → 5fC → 5caC reaction. The removal of 5hmC, 5fC or 5caC is biochemically 'easier' than the removal of 5mC and could occur via a more efficient form of replication-dependent demethylation, direct removal of the oxidized methyl group or through the DNA repair machinery.

One question raised by the third point is whether 5hmC, 5fC and 5caC are simply intermediate products in an active demethylation cycle or if they themselves are bona fide epigenetic marks. This is an active area of research.

1.4 Assays for studying DNA methylation

A challenge to measuring DNA methylation is that it is erased by standard molecular biology techniques, such as the *polymerase chain reaction* (PCR) and bacterial cloning, and it is not revealed by DNA hybridization assays [Laird 2010]. Therefore, almost all assays of DNA methylation require one of the following *pre-treatments* of the DNA:

1. Enzyme digestion
2. Affinity enrichment
3. Sodium bisulfite conversion

Following a pre-treatment, DNA methylation can be assayed using standard techniques such as:

1. Gel-based analysis
2. Sanger sequencing
3. Microarray hybridisation
4. Massively parallel sequencing

An exception to this classification scheme are a new class of assays that seek to directly 'read' whether a position is methylated or unmethylated without requiring a pre-treatment of the DNA. For example, both Laszlo *et al.* [2013] and Schreiber *et al.* [2013] measure the change in current as a DNA molecule passes through a nanopore to infer whether a cytosine is methylated.

Almost all assays of DNA methylation measure a population average from a pool of hundreds or thousands of cells. For a diploid organism, this is an average over several distinct levels: the two DNA strands, the two homologous chromosomes within a diploid cell and the hundreds or thousand of cells used in the assay. Hundreds or thousands of cells are required in order to have sufficient material as input for the assay. Assays that require only a single cell as input do exist [e.g., Smallwood *et al.* 2014, Guo *et al.* 2013] but are still in development and not yet in widespread use.

The *resolution* and *throughput* of an assay are two key variables when choosing which to use for an experiment. The resolution of an assay is the scale on which DNA methylation can be measured[16]. For example, a high resolution assay allows a researcher to quantify the level of DNA methylation at a single base whereas a low resolution assay might only allow for qualitative assessment (i.e. presence or absence) of DNA methylation at larger regions, such as CpG islands.

The throughput of an assay can be quantified in two ways. The first is per-sample throughput, which is how many measurements of DNA methylation are made per-sample[17]. This is typically what people mean when they describe an assay as being 'high-throughput' or 'low-throughput' and is the definition I use in the title of my thesis. Depending on your definition of 'high', a high-througput assay will produce on the order of tens of thousands to billions of measurements per sample. The second definition of throughput is related to cost, be it money or time, i.e. 'how many samples can I afford to analyse?'.

The choice of which assay to use for an experiment is a trade-off between resolution, per-sample throughput, and per-cost throughput. Experiments that use an assay with high resolution and high per-sample throughput generally have fewer samples (due to the

[16] Depending on the experiment and its aims, the resolution of an assay might instead be defined as the scale on which DNA methylation can be quantified or the scale at which allows inference to address a specific hypothesis.

[17] This might reasonably be argued as being a definition of resolution.

associated higher costs) than experiments using a lower resolution assay or an assay with lower per-sample throughput.

In this section I describe each of these pre-treatments but focus on the bisulfite-conversion assays. In particular, I describe in detail the 'gold standard' assay of DNA methylation, *whole-genome bisulfite-sequencing*, that combines the sodium bisulfite conversion pre-treatment with massively parallel sequencing to produce whole-genome maps of DNA methylation at single-base resolution.

1.4.1 Enzyme digestion assays

Restriction endonucleases are an important technique in molecular biology. These enzymes can preferentially 'cut', 'cleave', or 'digest' DNA at or near to particular sequence motifs. The motif at which a restriction enzyme cleaves DNA is called the *recognition motif* or *restriction sequence.* The methylation of a position in the recognition motif can inhibit a restriction enzyme from cleaving the DNA. This can be used to design an assay to infer the methylation state of a DNA fragment.

For example, the recognition site of the restriction enzyme *HpaII* is `CCGC`. However, *HpaII* will only digest DNA when the second cytosine in the motif is unmethylated. The *HELP* (*HpaII* tiny fragment enrichment by ligation-mediated PCR) assay compares DNA digested by *HpaII* to one digested with another restriction enzyme that has the same recognition motif but is methylation-insensitive (*MspI*) to identify *hypomethylated* regions of a genome [Khulan *et al.* 2006].

Assays based on restriction enzymes were some of the first developed for studying DNA methylation. These were initially developed for studying a small number of loci although they have been extended to genome-scale analysis approaches [Laird 2010]. One such genome-wide assays is *CHARM*, 'comprehensive high-throughput arrays for relative methylation' [Irizarry *et al.* 2008]). CHARM combines a methylation fractionation step (be it *MeDIP*, *HpaII*, or, as in the original publication, *McrBC*) with a tiling array and analysis techniques that leverage regional DNA methylation levels.

While restriction enzyme assays do not typically provide single-base resolution data, the `methylCRF` software [Stevens *et al.* 2013] is able to infer single CpG methylation levels by combining data from a restriction enzyme based assay (MRE-seq) with one based on affinity enrichment (MeDIP-seq) in a sophisticated statistical analysis.

1.4.2 Affinity enrichment assays

Affinity enrichment assays compare measurements between an 'enriched' version and an 'input' (control) version of the same sample to infer the presence or absence of DNA methylation. This may use antibody immunoprecipitation or methyl-binding proteins. Some examples of affinity enrichment assays for DNA methylation are the microarray-based MeDIP, mDIP and mCIP and their sequencing-based relatives, MeDIP-seq and mDIP-seq.

These are all low resolution assays since they are based on the enrichment of regional differences between the enriched and input samples. Furthermore, the bioinformatic analysis of data from these assays is complicated by the varying CpG density along the genome, which leads to different enrichment affinities for different regions of the genome. However, these assays can provide a relatively cheap and efficient genome-wide assessment of DNA methylation [Laird 2010].

1.4.3 Sodium bisulfite conversion assays

In the 1980s, two research groups independently discovered that when DNA is treated with sodium bisulfite ($NaHSO_3$), unmethylated cytosines deaminate to uracils much faster than do methylated cytosines [Shapiro *et al.* 1970, Hayatsu *et al.* 1970]. The methylated cytosines are said to be 'protected' from conversion to uracils. This discovery led to the development of assays for studying studying cytosine methylation based on the pre-treatment of DNA with sodium bisulfite [Frommer *et al.* 1992, Clark *et al.* 1994], which are referred to as bisulfite-conversion assays.

When bisulfite-treated DNA is amplified by PCR, the uracils are converted to thymines. Therefore, these bisulfite-conversion assays are all based on the idea of comparing the sequence of the untreated DNA to the sequence of the bisulfite treated DNA to infer the methylation state of all cytosines in the sequence by whether or not they were converted to uracil/thymine following the bisulfite treatment (Figure 1.6).

Initial experiments based on the sodium bisulfite pre-treatment of DNA used Sanger

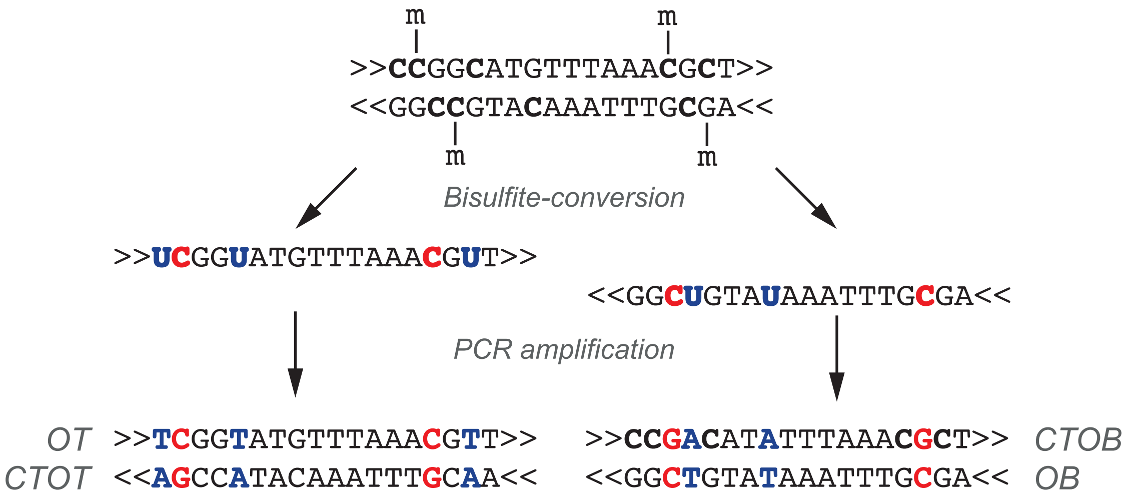


Figure 1.6: The effect of bisulfite-treatment of DNA. The double-stranded DNA is denatured and each strand undergoes bisulfite-treatment. Methylated cytosines remain as cytosines while unmethylated cytosines become uraciles. These bisulfite-converted DNA strands then undergo PCR amplification which converts the uracils to thymines. Note that while there are four possible PCR products, some bisulfite-sequencing protocols do not sample PCR products from the *CTOT* or *CTOB* strands. *OT* = original top strand; *OB* = original bottom strand; *CTOT* = complementary to the original top strand; *CTOB* = complementary to the original bottom strand. This figure is adapted from Krueger *et al.* [2012].

sequencing of cloned PCR products, a very laborious task that restricted experiments to studying a limited number of short segments of DNA. Although subsequent enhancements in the automation of Sanger sequencing improved the throughput of these assays, it was never going to be able to deliver a cost-effective, genome-scale assay of DNA methylation. The development of hybridisation microarrays provided cheap, genome-wide measurements of DNA methylation from bisulfite-treated DNA.

Microarrays contain thousands, even millions, of short *oligonucleotide* probes. Each probe is is designed to hybridise to a particular DNA sequence and emits a fluorescent signal that can be measured to infer the strength of the hybridisation. Therefore, an (idealised) way to analyse DNA methylation with a microarray is to hybridise bisulfite-converted DNA to a microarray that contains probes for both the methylated and unmethylated versions of all sequences of interest. The relative methylation of each sequence can be inferred from the relative intensities of the 'methylated probe' to the 'unmethylated probe'. Such an idealised experiment brushes over many complications including [Laird 2010]:

- The reduced complexity of bisulfite-converted DNA (from a 4-base alphabet to a mostly 3-base alphabet) leads to decreased hybridisation specificity
- Sequences containing multiple cytosines require multiple probe versions in order to assay all possible methylation patterns
- This approach requires the design of organism-specific microarrays

The Illumina Infinium HumanMethylation450 BeadChip (*Illumina 450k array*) provides a modern implementation of this type of assay for studying human DNA. This array assays $482,422$ cytosines, 99.3% of these CpGs, across a wide variety of genomic features [Stirzaker *et al.* 2014]. There are well-established methods and software for analysing Illumina 450k data, which were recently reviewed by Dedeurwaerder *et al.* [2014].

Genomics research was revolutionised by the development of cheap high-throughput sequencing technology, and the study of DNA methylation was no exception. In 2008, two papers were published describing methods for whole-genome shotgun sequencing of bisulfite-converted DNA using the nascent Solexa/Illumina sequencing technology [Cokus *et al.* 2008, Lister *et al.* 2008]. Whole-genome bisulfite-sequencing remains the gold standard assay for measuring genome-wide DNA methylation data.

Cokus *et al.* [2008] termed their approach *BS-seq* while Lister *et al.* [2008] called their method *methylC-seq*. From a bioinformatics perspective, the main difference is that the BS-seq protocol produces sequencing reads from four bisulfite-converted DNA strands — the original top strand (OT), the complementary strand to the original top strand ($CTOT$), the original bottom strand (OB) and the complementary strand to the original bottom strand ($CTOB$) — which require mapping to four different *in silico* converted reference genomes, followed by a merge of the alignment results. In contrast, the methylC-seq protocol only produces sequencing reads from two bisulfite-converted strands — OT and OB — and so only requires mapping to two different *in silico* converted reference genomes, followed by a merge. The methylC-seq protocol is now the standard whole-genome bisulfite-sequencing protocol, due in part to the simpler bioinformatics analysis.

Whole-genome bisulfite-sequencing remains an expensive assay, which limits its use in studies involving large numbers of samples. Furthermore, depending on the choice of several sequencing parameters, approximately 35% to 72% of reads will not contain any CpGs (Figure 1.7), which might be considered a gross waste of resources for some experiments.

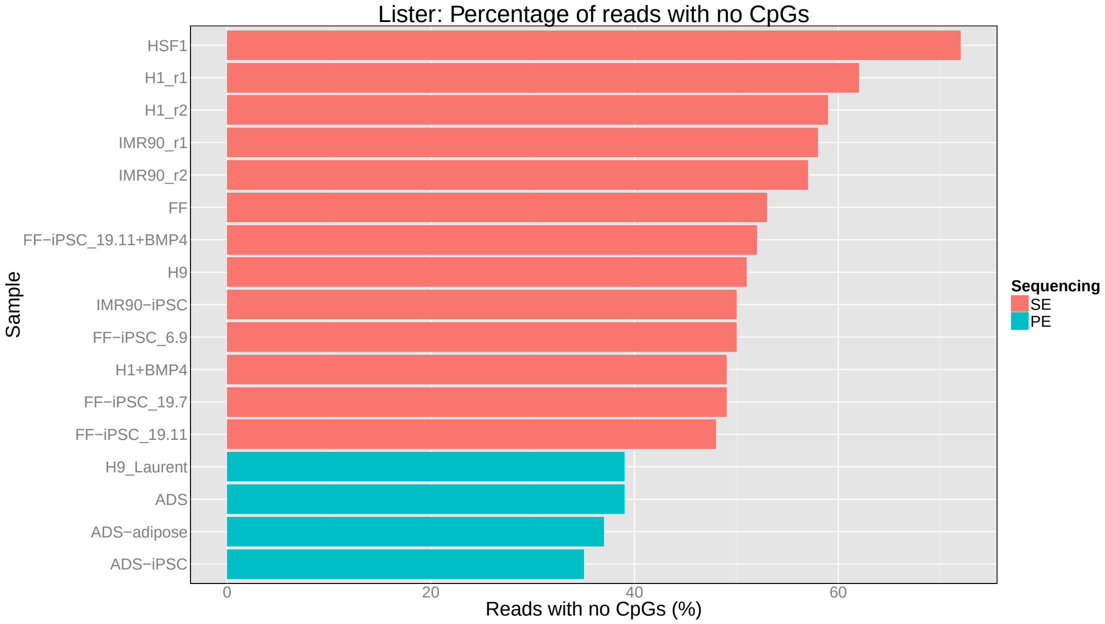


Figure 1.7: The percentage of reads that do not contain any CpGs for each sample in the *Lister* dataset. Paired-end (PE) reads are counted as a single unit. Therefore, samples sequenced with paired-end reads have a lower percentage of reads without CpGs than do samples sequenced with single-end (SE) reads. See Chapter 3 for a description of the *Lister* dataset.

Several assays have been developed to perform targeted high-throughput bisulfite-sequencing. The targeted nature of these assays are their obvious advantage and disadvantage; only a subset of the genome needs be sequenced but you only obtain information about methylation for that subset. Depending on the experiment, this tradeoff may be worthwhile, and these assays have been successfully used in a number of studies such as the BLUEPRINT Epigenome project (`http://www.blueprint-epigenome.eu`) and the NIH Roadmap Epigenomics Mapping Consortium (`http://www.roadmapepigenomics.org`)

Some examples of these targeted bisulfite-sequencing are:

- Reduced representation bisulfite-sequencing (*RRBS*, Meissner *et al.* [2005]): RRBS uses restriction enzymes to first select regions of the genome with a high CpG density (based on the Msp-I cleavage motif), which are subsequently treated with sodium-bisulfite and sequenced.

- Extended reduced representation bisulfite-sequencing, also known as enhanced reduced representation bisulfite-sequencing (eRRBS, Akalin *et al.* [2012a]). A modified version of the RRBS protocol.
- NimbleGen's SeqCap Epi Enrichment System and Agilent's SureSelectXT Human Methyl-Seq: Both of these commercial products use DNA hybridisation to enrich the sequencing library for pre-defined regions of interest. This enriched library is then bisulfite-converted and sequenced.

One final class of bisulfite-conversion assays does not use microarrays or high-throughput sequencing. Sequenom's EpiTYPER uses mass spectrometry to analyse DNA methylation from bisulfite-converted DNA [Ehrich *et al.* 2006]. This platform can provide quantitative measurements of CpG methylation across hundreds of loci and multiple samples and may be used to validate findings discovered using other platforms [Laird 2010].

Pros and cons of bisulfite-conversion assays

Bisulfite-conversion assays are considered the gold standard for studying DNA methylation since cytosine methylation can be detected at single-base resolution [Stirzaker *et al.* 2014]. In fact, single-molecule, single-base resolution is even possible for short DNA sequences when bisulfite-treated DNA is analysed with sequencing[18].

Almost all bisulfite-based assays require a considerable amount of DNA that is extracted from a population of cells (e.g., 1 to 5 μg for whole-genome bisulfite-sequencing). However, the minimal amount of DNA is being reduced with each technological advance. Recently, Guo *et al.* [2013] and Smallwood *et al.* [2014] published single-cell bisulfite-sequencing assays, although these are not yet a commercially available sequencing assay. Furthermore, these are not yet proper genome-wide assays. The technique of Guo *et al.* [2013] is adapted from RRBS and so only assays a small percentage of cytosines in the genome. And the technique of Smallwood *et al.* [2014], while intended as a genome-wide assay, can reportedly only "accurately measure DNA methylation at up to 48.4% of CpG sites" [Smallwood *et al.* 2014].

[18] Microarray hybridisation assays can provide single-base resolution but not single-molecule resolution. The signal from a microarray-based experiment is a sample-wide average since, for each locus, the signal is relative to the proportion of DNA fragments in the sample methylated at that locus.

A recently discovered disadvantage of bisulfite-conversion assays is that they are unable to distinguish 5hmC from 5mC; 5-hydroxymethyl, like 5-methyl, similarly protects a cytosine from deamination to uracil. In effect, the detection of 5mC, and all subsequent inference, is confounded with that of 5hmC. For most experiments this isn't much of a problem — most cells have very low levels of 5hmC and so there is little confounding — however, in certain experiments this needs a more careful approach. To address this issue, Booth *et al.* [2012] developed *oxidative bisulfite-sequencing* (oxBS-seq) and Yu *et al.* [2012] developed *Tet-assisted bisulfite sequencing* (TAB-seq) for separate 5mC and 5hmC detection.

oxBS-seq specifically measures 5mC. The input DNA is oxidated by potassium perruthenate ($RKRuO_4$), which converts 5hmC to 5fC, prior to bisulfite-treatment. Only 5mC is protected from conversion during the bisulfite-treatment, which effectively means that only 5mC remains to be detected at the sequencing stage. The level of 5hmC can be estimated by performing traditional bisulfite-sequencing and then 'subtracting' the oxBS-seq signal (5mC) from the bisulfite-sequencing signal (5mC + 5hmC).

TAB-seq takes the opposite approach to oxBS-seq by specifically measuring 5hmC. The input DNA is treated with a β-glucosyltransferase, which converts 5hmC to β-glucosyl-5-hydroxymethylcytosine (5gmC), followed by TET oxidation. Only 5gmC is protected from TET oxidation, which effectively means that only 5hmC remains to be detected at the sequencing stage. The level of 5mC can be estimated by performing traditional bisulfite-sequencing and then 'subtracting' the TAB-seq signal (5hmC) from the bisulfite-sequencing signal (5mC + 5hmC).

There is also potential confounding of 5C, 5mC and 5hmC with 5fC and 5caC, although this has received less attention since 5fC and 5caC are believed to exist at far lower quantities than 5mC and 5hmC. Nonetheless, sequencing assays of 5fC and 5caC exist based on the idea of treating the DNA with a chemical and performing 'signal subtraction' with traditional bisulfite-sequencing or another assay [Wu *et al.* 2014].

Another disadvantage of bisulfite-conversion assays is that they require knowledge of the underlying DNA sequence in order to infer the methylation states of cytosines. This requires either a parallel experiment to sequence the target region(s) or reliance on a reference genome. When relying on a reference genome, the inference of the methylation state can be confounded by the DNA sequence of the sample [Liu *et al.* 2012]. Figure 1.8 illustrates such an example.

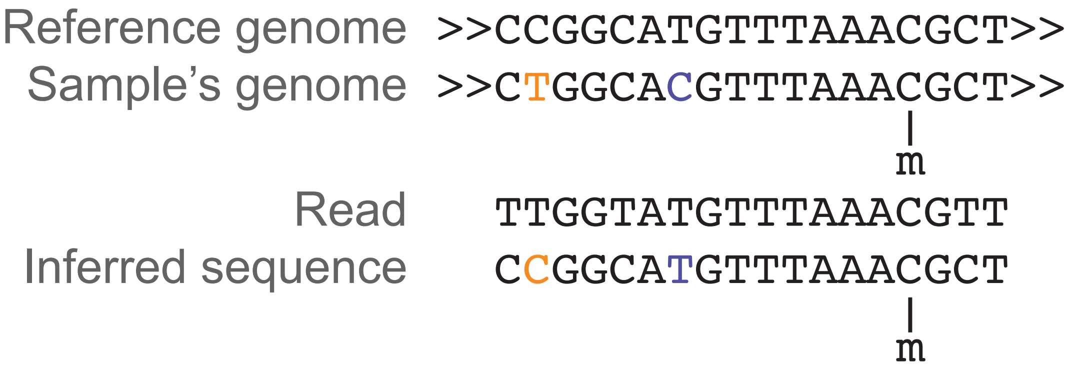

Reference genome >>CCGGCATGTTTAAACGCT>>
Sample's genome >>CTGGCACGTTTAAACGCT>>
m
Read TTGGTATGTTTAAACGTT
Inferred sequence CCGGCATGTTTAAACGCT
m

Figure 1.8: Sequence variation between the reference genome and the sample's genome can result in incorrect inference about the methylation state of the sample's genome. The locus in orange is a cytosine in the reference genome but a thymine in the sample's genome. Because the read is compared against the reference genome, it may be incorrectly inferred to be an unmethylated cytosine. The locus in purple is a thymine in the reference genome but an unmethylated cytosine in the sample's genome. Because the read is compared against the reference genome, it may be incorrectly inferred to be a thymine.

The bisulfite-treatment of DNA can introduce biases and other problems [Warnecke *et al.* 2002]. Four examples are *PCR-bias*, *incomplete bisulfite-conversion*, *bisulfite over-conversion*, and *DNA degradation*. PCR-bias is the difference in amplification efficiency of methylated and unmethylated versions of the same DNA sequence [Warnecke *et al.* 1997]. Incomplete bisulfite-conversion leads to cytosines being incorrectly inferred as 5mC since they cytosines were not converted to uracils by the bisulfite-treatment. Conversely, bisulfite over-conversion results in a methylated cytosine incorrectly being inferred to be an unmethylated cytosine, although this is uncommon than incomplete bisulfite-conversion [Warnecke *et al.* 2002]. DNA degredation occurs because sodium bisulfite damages DNA, resulting in the fragmentation of long molecules. This limits the size of the fragments that can be studied using bisulfite-conversion assays to approximately 500 bp [Ecker 2010]. The influence of these biases can depend on the experimental setup and their potential effects should be born in mind when interpreting results.

1.4.4 DNA kinetics assays

Ideally, an assay of DNA methylation could simply 'read' methylated cytosines as a distinct signal from unmethylated cytosines. This idea forms the basis of assays using the kinetics of DNA to infer the presence of DNA methylation and other DNA modifications. While very exciting, these assays do not yet scale to studying genome-wide DNA methylation levels in mammalian-sized genomes.

The only commercially available assay in this class is the Pacific Biosciences SMRT technology [Flusberg *et al.* 2010]. It infers the presence of DNA modifications by comparing the time it takes to 'read' the modified form of a base, such as 5mC, to the time it takes to read its unmodified form. This does not require that the DNA is bisulfite-converted prior to sequencing. Because the DNA does not undergo bisulfite-conversion (nor the attendant short fragmentation of the DNA), it is in theory possible to analyse DNA methylation from individual, long DNA molecules using Pacific Biosciences SMRT technology[19]. However, due to the error rate and cost of SMRT sequencing, it is currently all but unfeasible to study genome-wide DNA methylation in mammalian-sized genomes.

The error rate of Pacific Biosciences SMRT sequencing is currently higher than that of Illumina sequencing. This means it is more difficult to make reliable inferences on DNA methylation from individual reads. Given sufficient sequencing coverage, it would be possible to reliably estimate the average level of methylation at a given cytosine, however, the cost of SMRT sequencing all but prohibits high-coverage sequencing of mammalian-sized genomes.

1.5 Summary

DNA methylation is an epigenetic modification with important roles in many biological systems. In mammals, CpGs are frequently methylated while non-CpG cytosines are less frequently methylated. Most unmethylated CpGs are found in CpG islands, which are important regulatory elements in the genome.

Whole-genome bisulfite-sequencing provides single-base resolution data and is the gold

[19] While bisulfite-treated DNA could be sequenced using SMRT sequencing, this would eliminate the real advantage of the technology, namely, the long reads generated by this sequencer.

standard assay for studying DNA methylation. Methods for analysing whole-genome bisulfite-sequencing data are the focus of my thesis. The following chapter describes the bioinformatics analysis of a bisulfite-sequencing experiment, including its many statistical and computational challenges.

Chapter 2

Bioinformatics analysis of whole-genome bisulfite-sequencing data

Overview

This chapter explains the bioinformatics analysis of whole-genome bisulfite-sequencing data, concentrating on the most widely used methylC-seq protocol. All data used in my thesis were generated using this protocol.

There are four fundamental steps in the analysis of bisulfite-sequencing data:

1. Data quality control checks
2. Read mapping and post-processing of mapped reads
3. Methylation calling
4. Downstream analyses

This chapter focuses on steps 1-3, while Chapter 5 addresses the wide variety of analyses available at Step 4. Steps 1 and 2 will be familiar to anyone who has analysed high-throughput sequencing data, but each requires a twist to work with bisulfite-sequencing data. Step 3 is obviously unique to assays of DNA methylation, but there are similarities to variant calling from DNA sequencing.

The chapter concludes by introducing the `methtuple` software that I wrote, a unique methylation caller for extracting methylation patterns at tuples of methylation loci. `methtuple` is critical for work in later chapters on co-methylation (Chapter 7) but has wider application in facilitating downstream analyses of bisulfite-sequencing data.

2.1 Data quality control checks

The first step in any analysis of high-throughput sequencing data is to perform a quality control check of the data. Much of this is done visually by comparing summary graphs of the current sample(s) to previous 'good' samples. As such, much of data quality control checking relies on the judgement of the analyst.

The `FastQC` software is a very useful tool for performing this first step (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). It produces summary graphs of many key measures such as base quality scores, read length distribution and sequence contamination. `FastQC` is a general purpose tool for performing quality control checks of high-throughput sequencing data. This means that some of its output must be interpreted with caution for bisulfite-sequencing data. For example, `FastQC` will report a warning (resp. error) if the relative frequency of the four bases differ by more than 10% (resp. 20%). As noted in the `FastQC` documentation, such a warning/error should be ignored for bisulfite-sequencing data, owing to the inherent bias in its sequence composition.

Perhaps the most important quality control of bisulfite-sequencing data is the identification and removal of contaminating sequences. `FastQC` will screen a subset of the reads against a list of known, common contaminants such as adapter sequences. When sequencing is performed using the widely used Illumina technology, adapter sequences must be ligated to the ends of each DNA molecule in the library. The adapters do not contain the biological sequence of interest, however, the sequencer can 'read into' the adapter sequence, particularly when using paired-end sequencing of short DNA fragments such as those frequently created in bisulfite-sequencing libraries. This means that some reads are chimeras that contain the biological sequence of interest (from the sample) and junk sequence (from the adapters). This contamination needs to be removed for two reasons:

1. Reads containing adapter contamination will generally not map to the reference genome, meaning these reads are needlessly wasted.
2. If they do map, then this will result incorrect inferences; the ‘garbage in, garbage out’ maxim.

Using a tool such as `Trim Galore!` (`http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) or `cutadapt` [Martin 2011], the reads can be *trimmed* to remove these contaminants. Reads might also be trimmed to remove low quality sequencing cycles, which are common at the 3’ end of reads, although this isn’t as essential as trimming to remove contaminants.

2.2 Read mapping and post-processing of mapped reads

Read mapping is complicated by the bisulfite-treatment of the DNA. Following bisulfite-treatment, the DNA fragments are mostly composed of three bases rather than four, which means there are many more sequence mismatches between a read and its true mapping location. Simply using standard read mapping software and allowing for more mismatches would result in many reads mapping to multiple locations in the reference genome. Instead, a field of read mapping software dedicated to bisulfite-sequencing data has developed. Several review articles have summarised and compared the various approaches [Chatterjee *et al.* 2012, Krueger *et al.* 2012, Kunde-Ramamoorthy *et al.* 2014].

These bisulfite-sequencing read mappers take one of two approaches:

1. Methylation-aware mismatch penalties.
2. *In silico* bisulfite-conversion of reads and reference genomes.

While methylation-aware mappers provide the highest efficiency, these suffer from a bias whereby methylated reads are preferentially mapped over unmethylated reads [Krueger *et al.* 2012]. This biases downstream inference and means that these mappers are generally less popular.

In silico bisulfite-conversion mappers convert all cytosines to thymines (resp. guanines to adenines) of the forward (resp. reverse) strand from the reference genome. They then take each read and create two *in silico* bisulfite-converted versions of it[1]: the CT-read replaces all residual thymines with cytosines and the GA-read replaces all residual guanines with adenines. The CT-read is mapped against the CT-genome and the GA-read is mapped against the GA-genome using a standard mapping tools such as `Bowtie1` [Langmead *et al.* 2009], `Bowtie2` [Langmead and Salzberg 2012] or `bwa` [Li and Durbin 2009, 2010][2].

Depending on the exact settings used, the mapper reports the 'best' location of each read with respect to the two reference genomes. It reports the original sequence of the read in the output file so that the methylation status of each position can be inferred by comparing it to the corresponding reference sequence.

In silico bisulfite-conversion mappers avoid the bias inherent in the methylation-aware mappers because all reads, regardless of methylation status, 'look the same' to the mapper. However, they do suffer from a slight loss in mapping efficiency [Krueger *et al.* 2012].

Table 2.1 list some popular bisulfite-sequencing read mappers, which have been selected to highlight the variety of underlying mapping software used by these tools.

Table 2.1: Four popular bisulfite-sequencing read mappers, selected to highlight the variety of underlying mapping software used by these tools.

Name	Reference	Underlying mapping software
Bismark	Krueger and Andrews [2011]	Bowtie1 or Bowtie2
bwa-meth	Pedersen *et al.* [2014]	bwa-mem
BSMAP	Xi and Li [2009]	SOAP
Novoalign	`http://www.novocraft.com/products/novoalign/`	Novoalign

Each of these aligners can report the output in the standard Sequence Alignment/Map format, `SAM`, or its binary equivalent, `BAM` [Li *et al.* 2009]. However, there is no agreed upon standard in the `SAM` specification for encoding the data specific to bisulfite-sequencing, which means that each mapper does this in a slightly different way. This complexity

[1] Two versions are made because we don't know *a priori* from which of the two strands the read originated.

[2] If the data were generated using a non-directional protocol, then each read of the CT-read and the GA-read are mapped to both of the CT-genome and GA-genome, resulting in four mapping steps per read.

makes it difficult for downstream analysis tools to support the output of different mapping software.

Read mapping is not perfect and produces both false positive and false negative results. False positives are due to reads mapped to the wrong location and reads mapped to multiple locations with equal mapping scores. False negatives are reads that are not mapped to any location; these reads are effectively lost from any downstream analysis. The parameter settings used by the mapping software determine the false positive and false negative rates.

There are biological and technical reasons why mapping against a reference genome can produce these errors. Biologically, if the sample contains sequences that are too genetically divergent from the reference genome then these sequences will be difficult, even impossible, to map. A particularly problematic class of sequences are those from repetitive regions of the genome. These repetitive sequences will map to multiple locations in the reference genome equally well. Furthermore, the number of times these repetitive sequences occur differs between the reference genome and the sample's genome.

Reads from Illumina sequencing are often too short to resolve the mapping location of these repetitive sequences. Resolving the mapping location of repetitive sequences can be achieved by using other sequencing technologies, such as Pacific Biosciences SMRT technology [Flusberg *et al.* 2010], which produces longer reads.

Another source of technical error in read mapping is really due to sequencing error. A sequencing error can transform a uniquely mapping read to one that maps equally well to multiple locations or, worse still, a read that maps uniquely to a single, but incorrect, location. Sequencing errors can also corrupt a read so badly that it no longer can be mapped. In practice, most people try to mitigate these problems through their choice of parameters used by the read mapping software.

Ideally, mapping software assigns the degree of confidence it has that the read is correctly mapped via a mapping quality score (`mapQ`). In theory, reads might be down-weighted in downstream analyses based on the mapping quality score. However, these mapping quality scores are often poorly calibrated, particularly for methylC-seq data, which makes them less useful. `Bismark` [Krueger and Andrews 2011], a popular bisulfite-sequencing mapping software, only recently introduced mapping quality scores (`v0.12.1`, released in

April 2014).

The above problems are general challenges of read mapping and are not specific to bisulfite-sequencing data, although the reduced complexity of bisulfite-sequencing reads exacerbates these issues. The difficulty of mapping to repetitive regions of the genome is a particularly frustrating one for bisulfite-sequencing data. Repetitive sequences, such as LINEs and SINEs, are typically methylated in order to prevent transcription and are often of interest to researchers studying DNA methylation. The low mapping efficiencies of these regions means that there is often limited or less reliable data for these elements from bisulfite-sequencing data.

2.2.1 PCR duplicates

PCR amplification of the input DNA is a common step in creating a library for high-throughput sequencing. PCR amplification is often required to ensure that there is a sufficient amount of DNA for the sequencer to properly work. Unfortunately, it can introduce biases into the library that results in some molecules being over-represented or under-represented compared to their true frequency. This means that when we sequence the library that we might sequence multiple fragments that are all copies of the same original piece of DNA, which gives a biased sampling of our sample's genome. These multiply-sequenced fragments are called *PCR duplicates*.

In bisulfite-sequencing data, PCR duplicates containing a methylation locus can result in a biased estimate of the methylation level at that locus. This is because the sequenced reads do not accurately represent the true methylation levels of the sample.

There is generally no way to tell based on sequencing data if a read is truly a PCR duplicate. However, it is relatively easy to identify suspected PCR duplicates (which are almost always referred to as 'PCR duplicates'[3]). Software to identify PCR duplicates includes the `MarkDuplicates` function that is a part of the `Picard` software (`http://broadinstitute.github.io/picard/`), the `rmdup` function that is a part of the `SAMtools`

[3] The distinction between suspected PCR duplicates and true PCR duplicates is rarely made, possibly because the phrase is so clunky. Suspected PCR duplicates are almost always referred to as PCR duplicates with the implicit assumption that the reader is aware that these very likely include false positive calls. Consistent with the literature, I will use the term PCR duplicates when I refer to reads identified as PCR duplicates by some software. I will use *true* PCR duplicates when I need to distinguish the two concepts.

software [Li *et al.* 2009] and SAMBLASTER [Faust and Hall 2014]. These software all use mapped reads from a SAM or BAM file as input.

Roughly speaking, these tools will flag reads with identical start and end co-ordinates as being suspected PCR duplicates. This will inevitably lead to some false positives because reads may have identical co-ordinates yet not be true PCR duplicates. The false positive rate is a particular problem when a subset of the genome is sequenced at high coverage, such as in RRBS. This can be thought of as an example of the pigeonhole principle, which states that if we have m containers (positions where a read can start) and $n > m$ items (reads), then at least one container must contain more than one item (at least one position must have more than one read starting there).

We could make this mathematically more precise, but it doesn't give us a simple answer to the question, 'should we remove possible PCR duplicates from bisulfite-sequencing data?'. The unsatisfactory answer is, 'it depends'. A rule of thumb is that provided the average or median sequencing coverage of the 'genome' is less than the fragment length[4] then we expect few false positive calls.

In practice, for whole-genome sequencing data we can be fairly confident that suspected PCR duplicates are *true* PCR duplicates. However, for targetted sequencing, such as RRBS or amplicon sequencing, we are much less confident and may remove some of our signal if we remove possible PCR duplicates. Instead, for RRBS we might exclude regions with an "abnormally" high sequencing coverage [Krueger *et al.* 2012]. For amplicon sequencing we often can't afford to exclude possible PCR duplicates if, for example, the aim is to identify rare epialleles by very deep sequencing of a small region.

2.2.2 M-bias

Ideally, the probability that a base is called as methylated should be independent of the sequencing cycle. Hansen *et al.* [2011] found that this is not the case and that in fact there is considerable bias towards the start (5') and end (3') of reads. They called this *M-bias.*

M-bias can be identified by plotting the read-position methylation level (*rpml*), which is the proportion of reads that are methylated at each read-position, as a function of

[4] For single-end data, the 'fragment length' in these calculations is the read length.

read-position. These *rpml* are computed separately for each methylation type and, for paired-end sequencing data, separately for *read-1* and *read-2*. If there is no M-bias then this plot should be a horizontal line. A 'bend' or 'spike' in this plot is evidence of M-bias. Furthermore, these lines, which indicate the average level of methylation in the sample for that methylation type, should be at the same level for *read-1* and *read-2*, although we see this is often not quite the case. Figure 2.1 is the M-bias plot for the *ADS* sample from the *Lister* dataset[5], which shows significant CpG M-bias at the start of *read-2* and some noise at the start of *read-1*.

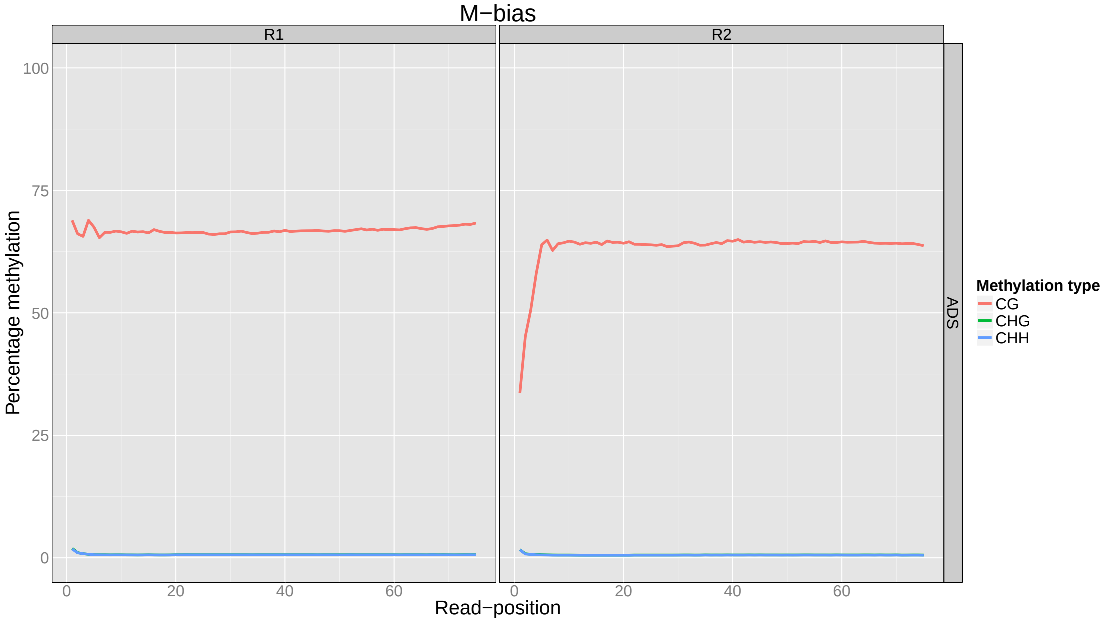


Figure 2.1: M-bias plot for the *ADS* sample from the *Lister* dataset. Each of *read-1* (R1) and *read-2* (R2) are plotted separately.

If a sample is processed over multiple batches, then M-bias estimation (and methylation calling) should be performed separately for each batch and then combined, or in some manner that is 'batch aware'. For example, two libraries with DNA derived from the same cell line, but with separate library preparations and sequencing runs, will likely suffer from batch effects due to differences in the library preparations or differences with the sequencing runs. Unfortunately, the person analysing the data doesn't always know all the sample processing steps that may have introduced such batch effects and so these can be hard to deal with in practice.

[5] See Chapter 3 for a description of this sample.

The strongest source of M-bias in Illumina whole-genome bisulfite-sequencing data is at the 5' end of *read-2*, which sequences the 3' end of the DNA fragment. Because the DNA fragment is often shorter than the sum of the read lengths, the 3' end of the fragment often contains adapter sequence and other 'junk' sequence. The adapter sequence may contain cytosine bases, which will be misinterpreted as evidence of methylation [Krueger *et al.* 2012]. Similarly, "fill-in cytosines" are used in the construction of RRBS libraries to repair the ends of DNA fragments after cleavage by MspI; these would also be misinterpreted as evidence of methylation [Krueger *et al.* 2012]. Another source of M-bias is incomplete or uneven bisulfite-conversion.

Estimating M-bias

Estimating M-bias and incorporating it into the methylation calling can be done using two different strategies:

1. Compute the M-bias from the aligned reads, then call methylation events. The methylation calling should include filters to remove the detected M-bias (along with any other additional filters). This strategy requires two passes over the `SAM/BAM` file, one to compute the M-bias and one to do the methylation calling. This is the approach used by `bismark_methylation_extractor` [Krueger and Andrews 2011] and `Bis-SNP` [Liu *et al.* 2012].
2. Call methylation events but retain the read-position of each methylation event. Compute the M-bias from this first file and then filter out methylation events that suffer from M-bias. This strategy requires only a single pass over the `SAM/BAM` file but requires additional information to be stored alongside the methylation calls which is then followed by a pass over the file containing the methylation calls. This is the approach taken by `BSmooth` [Hansen *et al.* 2012].

I find the first strategy conceptually simpler, and easier to program, so use it in my methylation calling software, `methtuple` (described in Section 2.4).

In theory, the M-bias could be estimated during the alignment step or during another processing step, such as sorting or marking PCR duplicates, to avoid an additional pass over the `SAM/BAM` file.

A somewhat subtle point is that M-bias should only be estimated from reads that are actually going to be used for methylation calling. Suppose that M-bias is highly correlated with some other quality metric, such as base quality, so that positions with M-bias also have low base quality[6]. If you already intend to ignore read-positions with a base quality less than some threshold in your methylation calling, then it makes sense to also ignore these positions when estimating M-bias, otherwise you will overestimate the effect of M-bias and unnecessarily exclude read-positions in your methylation calling. Unfortunately, perhaps the most widely used software for estimating M-bias, `bismark_methylation_extractor` Krueger and Andrews [2011], does not allow the user to exclude certain reads or read-positions when estimating M-bias.

Pre-trimming reads destroys the one-to-one relationship between read-position and sequencing cycle

Trimming reads prior to alignment (*pre-trimming*), such as using `Trim Galore!` (`http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`) to remove adapter sequence from reads, destroys the one-to-one relationship between the sequencing cycle and the read-position. This causes a minor problem when computing M-bias because we no longer know whether the read-position is identical to the sequencing cycle. Soft-clipping or hard-clipping reads of their adapter sequence *during* the alignment avoids this issue, because the clipping information (should be) preserved in the `CIGAR` string[7].

The M-bias plot is based on the read-position from the aligned data and not the sequencing cycle (which isn't directly available in the `SAM/BAM` file). If the reads have been pre-trimmed, then each read-position in the M-bias plot will therefore contain data from multiple sequencing cycles, which can amplify or mask the M-bias signal.

For example, suppose we performed 100 bp single-end sequencing and pre-trimmed the first 20 bp of 90% of the reads. Then, read-position 80 will comprise 10% sequencing cycle

[6] This is very often the case since the 3' end of reads are typically of lower quality and also frequently suffer from M-bias.

[7] Not all software properly handles the information in the `CIGAR` string, particularly for soft-clipped reads. This is a shortfall of the downstream tools and not of aligner-based clipping *per se*, but is nonetheless an issue in practice. `bwa-meth` [Pedersen *et al.* 2014] and `LAST` [Kiełbasa *et al.* 2011] both perform well without pre-trimming of reads because they can soft-clip reads on the fly whereas other bisulfite-sequencing aligners, such as `Bismark` [Krueger and Andrews 2011], cannot soft-clip reads and so require that reads are pre-trimmed.

80 and 90% sequencing cycle 100. It is very likely that sequencing cycle 100 suffers more from M-bias than does cycle 80, and so this will appear in the M-bias plot as M-bias at read-position 80.

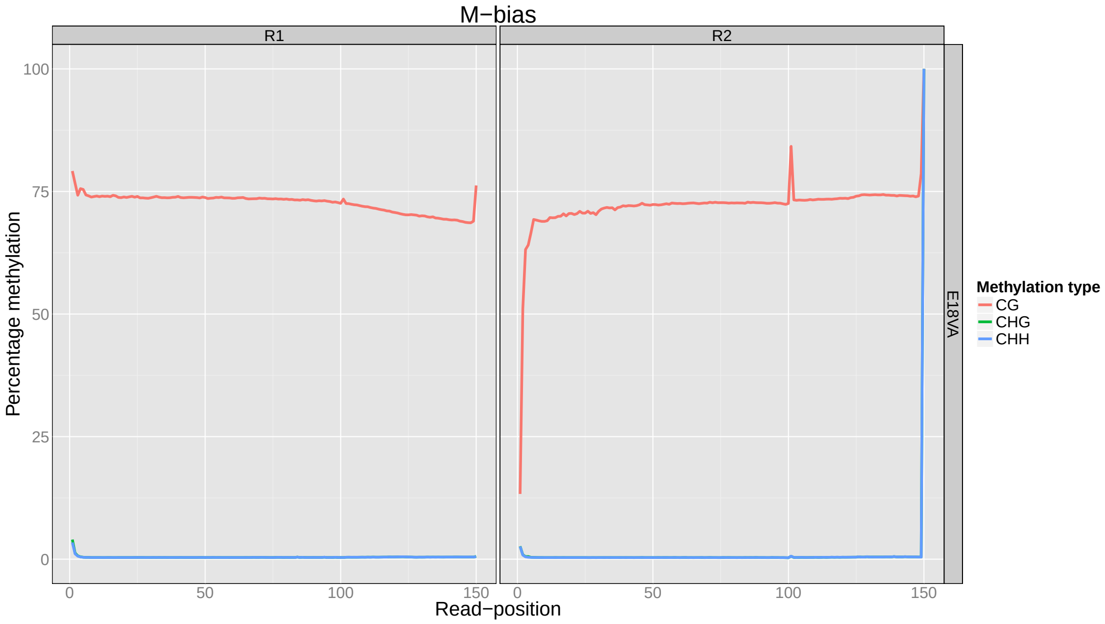


Figure 2.2: M-bias plot for the E18VA sample from the EPISCOPE dataset. Each of *read-1* (R1) and *read-2* (R2) are plotted separately. For CpGs in *read-1*, we see noise at the start of the read, followed by a downward slope in the M-bias, which ends with a spike. For CpGs in *read-2*, we see a downward spike at the start of the read following by a gradual increase in the M-bias curve, with a spike at read-position 101 and a spike at the last read-positions for all methylation types. The spike at read-position 101 is also evident, albeit to a lesser extent, in *read-1*. This position should be ignored in downstream analyses but we do not necessarily also want to ignore read-positions $102 - 150$ since this would remove one-third of the data.

The loss of the one-to-one relationship between sequencing cycle and read-position cannot be avoided if reads are pre-trimmed because the trimming information is not preserved. Hansen *et al.* [2012] suggest a separate M-bias plot for each read-length, which will help mitigate the effect of confounding between read-position and sequencing-cycle[8]. However, if trimming is performed during the alignment then all the necessary information is retained and the x-axis of M-bias plots would be ‘sequencing cycle’ rather than ‘read-position’, thus avoiding the issue entirely.

[8] This would also require that methylation calling is performed separately for read with different lengths because most methylation callers are unable to deal with different M-bias profiles for different read lengths.

Identifying M-bias

In practice, the M-bias curves for each sample are visually inspected to look for evidence of M-bias, i.e. read-positions whose methylation levels are 'too far away' from the majority of read-positions. It is rather a subjective decision to make. Two problems that I found when making these decisions were:

1. Maintaining consistency across samples.
2. Determining where to draw the line when there is no dramatic 'spike' in the M-bias.

For example, in Figure 2.2 it is clear that there are problems at the start and end of both *read-1* and *read-2*, as well as a big problem at read-position 101 in *read-2*. What is less clear is where to draw the line on the gradual decay toward the end of *read-1* and towards the start of *read-2*. This motivated me to write a few simple functions to perform more systematic processing of M-bias results. These are included in the `MethylationTuples` R package (described in Section 5.3) with the `MBias` class, its associated methods, `plot()` and `filter()`, and the helper function `readMBias()`.

For each sample, M-bias is computed separately for each methylation type because the level of methylation varies widely between CG and non-CG methylation types. For paired-end sequencing experiments, it is also done separately for each of *read-1* and *read-2* because M-bias is very different for these two mates and also because *read-2* often has a slightly lower average level of methylation than does *read-1* (e.g., see Figure 2.1).

In `MethylationTuples`, I use a simple normalisation of the read-position methylation levels (*rpml*). Specifically, the median level of methylation across all read-positions is subtracted from the read-position methylation level to create normalised read-position methylation levels, i.e. $nrpml_{CpG}^{read_1} = rpml_{CpG}^{read_1} - median(rpml_{CpG}^{read_1})$, for CpG methylation in *read-1*.

The `filter()` method identifies read-positions where the *rpml* differs by more than a given value (`threshold`) from *nrpml*. While it computes these statistics separately for each methylation type and read type, a common `threshold` is used for all methylation types and read types. I tend to use a value of `threshold = 3`, meaning that if the $median(rpml_{CpG}) = 75$, then any read-position with $rpml_{CpG} < 72$ or $rpml_{CpG} > 78$ is flagged as showing evidence of M-bias[9]. It is currently left to the user to decide whether a read-position should be excluded if it displays evidence of M-bias in a single methylation type or only if it displays evidence across all methylation types. I tend to exclude read-positions that display evidence of M-bias in the methylation type I am working with, which is typically CpG methylation.

What to do about M-bias?

Now that we've found read-positions with M-bias, what can we do about it? Typically, read-positions showing evidence of M-bias are excluded when calling methylation events. In fact, a slightly cruder procedure is typically used whereby the entire ends of reads are removed. For example, suppose we have 100 bp reads with M-bias observed at positions $1, 2, 3, 4, 9, 94, 96, 97, 98, 99, 100$, then read-positions $1 - 9$ and $94 - 100$ would be ignored when methylation calling. Both `bismark_methylation_extractor` and `Bis-SNP`, two popular methylation callers, use this method.

This strategy is generally sufficient because M-bias tends to occur as runs of read-positions at the 5' and 3' ends of reads. However, occasionally there are spikes in the M-bias plot, which indicate specific read-positions that we would like to exclude. Figure 2.2 shows an example of such a spike that occurred at read-position 101 in 150 bp reads. Using `bismark_methylation_extractor` we would be forced to either retain this position or to ignore read-positions $101 - 150$, effectively ignoring one third of the sequencing data, much of it unaffected by M-bias. The `methtuple` software avoids the unnecessary exclusion of those bases by allowing the user to specify the exact read-positions that she wants to exclude (discussed in Section 2.4).

In the datasets I have analysed, I have had to ignore up to 30 read-positions per read due to M-bias, a considerable loss of data. Ideally, we would be able to remove the effects of M-bias by accounting for it in methylation calling rather than by simply excluding those read-positions entirely. One idea is to inversely weight methylation calls by the level of M-bias. A downside to this approach is that this would turn an otherwise binary methylation call into a continuous value between 0 and 1, with an attendant loss in interpretability.

[9] The M-bias files created by `bismark_methylation_extractor --mbias_only` report the methylation levels as percentages rather than proportions.

However, we could still compute the ubiquitous β-values, traditionally defined as the proportion of reads that are methylated at a locus, and use these in downstream inferences, without much loss in interpretation. The bigger problem with this approach is the increased computational complexity and cost, and this is why I have not further pursued this idea.

2.2.3 Other biases

There are several other sources of potential bias in analysing bisulfite-sequencing data. These include sequencing and alignment errors, and sequence variation at, or nearby to, methylation loci. A particularly interesting source of bias is due to cellular heterogeneity.

Recent papers using single-cell bisulfite-sequencing [Guo *et al.* 2013, Smallwood *et al.* 2014] have investigated the extent of this cellular heterogeneity by comparing the methylomes of individuals cells that are nominally of the same 'type'. Cellular heterogeneity is particularly problematic when a sample contains multiple cell types.

For example, many studies of DNA methylation use whole blood as the sample tissue due to the ease with which it can be obtained. However, whole blood contains a mixture of cell types, each of which has a distinct methylation profile. This cellular heterogeneity can seriously bias downstream analyses and must be properly accounted for in any study exploring the relationship between differences in DNA methylation and a phenotype [Jaffe and Irizarry 2014, Houseman *et al.* 2014]. For example, Jaffe and Irizarry [2014] provide evidence that several reported relationships between age and DNA methylation are likely due to changes in the cell composition of whole blood with age and not due to DNA methylation changes *per se*.

Methods to estimate the cellular heterogeneity bias and adjust for it are available [e.g., Jaffe and Irizarry 2014, Houseman *et al.* 2014, Zou *et al.* 2014], although they have mostly been applied to DNA methylation arrays and not bisulfite-sequencing data. This is not to say that these problems don't exist for sequencing data, merely that these have not been as well-explored.

2.3 Methylation calling

Methylation calling is the process of calling each sequenced methylation locus as being either methylated or unmethylated[10], as well as determining the *context* or *type* of each methylation event (i.e. CpG, CHG or CHH) based on the sequencing data and a reference DNA sequence. In principle, this is a simple process, however, this belies some complications, which we discuss in this section.

Most bisulfite-sequencing alignment software either performs methylation calling during the alignment process, as done by `Bismark`[11], or as a separate step after the alignment and post-processing of the `SAM/BAM` file. An example of the latter is `Bis-SNP` [Liu *et al.* 2012], which performs methylation calling from bisulfite-sequencing data aligned with the user's choice of alignment software.

All bisulfite-sequencing assays use *reference-based* methylation calling. This means that they require the specification of a reference DNA sequence that the aligned bisulfite-sequencing data are compared against to infer the methylation state of each sequenced locus. Care must be taken to correctly handle the orientation and strand of the alignment.

When using reference-based methylation calling, the position of the methylation locus is with respect to the reference genome, since then all samples will use a common set of co-ordinates. Some methylation loci cannot be typed using a reference-based approach. For example, unless the genome of the sample is fully known, methylation loci in insertions cannot be distinguished from genetic variation since there is no reference sequence to compare them against.

2.3.1 Considerations

There are several issues that must be carefully considered when performing methylation calling, including filtering of reads and biases, choosing and refining the reference sequence,

[10] A third possibility is making the call that the 'methylation locus' is not in fact a methylation locus. For example, if the sequenced base at a cytosine in the reference sequence is an adenine or a guanine then this may be evidence that the position is not in fact a methylation locus.

[11] `Bismark` also includes a program called `bismark_methylation_extractor`, which, as the name suggests, extracts the methylation calls from the `SAM/BAM` file. So while `Bismark` annotates each base as methylated or unmethylated during the alignment, a secondary step using `bismark_methylation_extractor` is required to make the methylation calls.

and determining the context or methylation type of the cytosine.

Filtering of reads and bases

Prior to methylation calling, each read should be filtered to remove low-quality reads and low quality bases. When using a set of filters, at each step a read either 'survives', and is subjected to the proceeding filter, or 'dies', and is excluded from methylation calling[12]. Strictly speaking, each sequenced bases is assigned a weight in the filtering process, however, in practice, filters are normally first applied to reads and then to all bases within 'surviving' reads. In my own work analysing whole-genome bisulfite-sequencing data, I routinely use the following filters.

A read survives if:

1. The read is mapped (single-end or paired-end) and mapped in the expected orientation (paired-end only).
2. The read is not marked as a PCR duplicate.
3. The read has a mapping quality score greater than some threshold.

A sequenced base survives if:

1. The read-position of the base means that it is unlikely to be affected by M-bias.
2. The base quality score is greater than some threshold.
3. The base is a 'bisulfite mismatch' (e.g., the sequenced base is a C or T at a C in the reference sequence) and not a 'non-bisulfite mismatch' (e.g., the sequenced base is an A or a G at a C in the reference sequence).

Although incomplete bisulfite-conversion is a well-recognised issue, most analysis pipelines don't attempt to account for this during methylation calling[13]. An exception is `Bis-SNP` [Liu *et al.* 2012] which has an algorithm to exclude bases suspected of suffering from incomplete bisulfite-conversion at the 5' end of Illumina-generated reads.

[12] A read that is not used to estimate M and U may still be used in other analyses, such as estimating copy number variation.

[13] Instead, incomplete bisulfite-conversion is usually incorporated into calculations to estimate the methylation level at a given cytosine (see Section 4.4.2). The bisulfite-conversion rate may be estimated by analysing cytosines that are expected to be unmethylated, such as those in the chloroplast genome [Lister *et al.* 2008] or a spike-in control of lambda phage DNA [Lister *et al.* 2009].

Choosing and refining the reference sequence

The reference sequence is typically the reference genome used in the alignment step, in spite of the obvious differences between the sample's genome and the reference genome. This reference-based approach may be refined to incorporate genetic differences between the sample and the reference genome. This can be done in several ways:

1. Whole-genome sequencing or genotyping of the sample
2. Calling genetic variations directly from the bisulfite-sequencing data
3. Excluding sites of known genetic variation

The gold-standard is to perform whole-genome DNA sequencing of each sample. This data is then used to form a set of sample-specific methylation loci. This approach, however, is also very expensive due to the extra sequencing requirements. A cheaper alternative is to genotype the sample on a genome-wide SNP microarray. This will give very accurate, very cheap genotypes at a large number of loci ($500,000$ to $5,000,000$). However, it obviously cannot identify genetic differences that aren't on the array, such as novel sample-specific genetic variants.

The next best approach is that implemented in `Bis-SNP` [Liu *et al.* 2012], which is to call genetic variation from the bisulfite-sequence data itself and to then define a set of sample-specific methylation loci at which to call methylation events. `Bis-SNP` is designed for *directional* bisulfite-sequencing libraries such as the widely used Illumina whole-genome bisulfite-sequencing protocol.

Certain genetic variants, in particular heterozygous $C>T$ SNPs, are more difficult to accurately genotype than others. Unfortunately, $C>T$ SNPs are also quite important because they are the most common SNPs in mammals [Liu *et al.* 2012], mostly occur at CG dinucleotides and, as a result, are easily mis-called as unmethylated cytosines rather than as genetic variants. Fortunately, it is often possible to distinguish $C>T$ SNPs from unmethylated cytosines by examining the base on the opposing strand; if it is a G then the position must be a C, if it is a A then it must be a T (see Figure 2.3). Other base substitutions are more readily detectable, and insertion and deletion events (indels) may also be called.

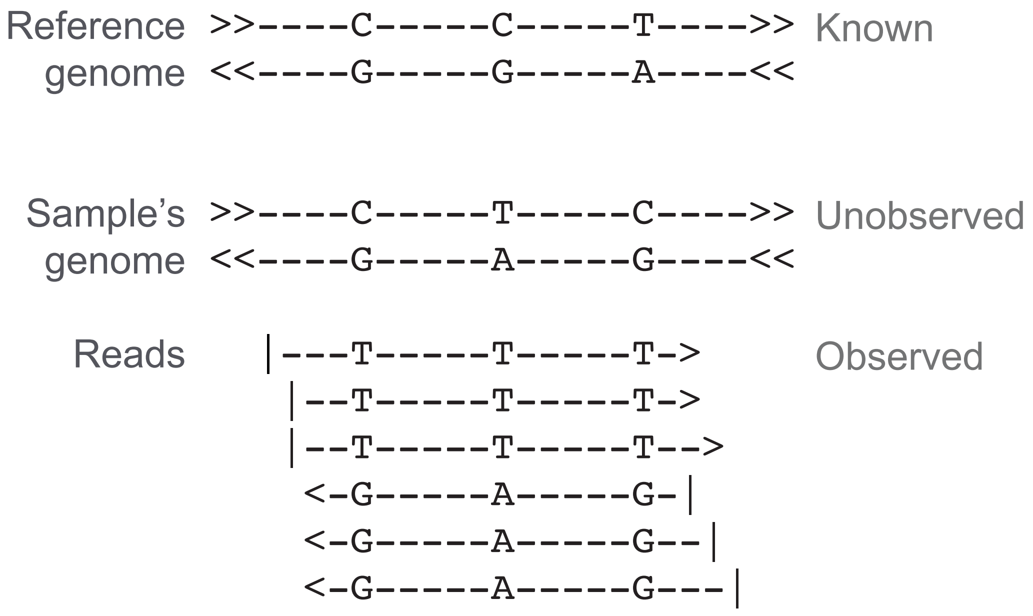


Figure 2.3: `Bis-SNP` is able to distinguish unmethylated cytosines (site 1), from cytosine to thymine genetic variants (site 2) and thymine to (unmethylated) cytosine genetic variants (site 3) by examining the reads mapped to the reverse strand. For all three loci, the reads mapped to the forward strand contain a thymine. However, it is the base on the reverse strand that reveals the true genotype. When combined with the reference genome it can be inferred whether the sample's genome, which isn't directly observed, has a genetic variant at that location. This is only possible with bisulfite-data generated using the directional protocol. This figure is adapted from Liu *et al.* [2012].

To emphasise, `Bis-SNP` provides three important pieces of information that make it almost as good as having whole genome DNA sequencing data on the same sample:

1. Reference-specific methylation loci, i.e. cytosines in the reference genome that are mutated to non-cytosine bases in the sample's genome.
2. Sample-specific methylation loci, i.e. cytosines in the sample's genome that are non-cytosine bases in the reference genome.
3. Other genetic variants that may be used in additional analyses, such as in identifying allele-specific methylation, or to refine the methylation *type* or *context* (see below).

Genotype calls made using `Bis-SNP` are less accurate than those from whole-genome DNA sequencing because of the reduced complexity of bisulfite-converted DNA. However, we essentially get to measure DNA variation 'for free' by using `Bis-SNP`, which makes it my preferred approach for incoporating genetic variation into methylation calling. The genetic variant calls made by `Bis-SNP` can also be used to *post-hoc* filter methylation calls made by other software. I use this approach to filter methylation calls made with `methtuple`.

The third approach, and arguably the bare minimum, is to call methylation events using the reference genome and to *post-hoc* exclude any loci that overlap sites of known genetic variation in the population. For example, we might exclude all cytosines in the reference genome that are also SNPs in dbSNP [Sherry *et al.* 2001].

This is a conservative approach, as it will remove loci regardless of whether the sample has a genetic variant at that position or not, but it may be a good enough method in some cases. It also obviously requires a database of known variation for the organism being studied, which is the case for commonly studied organisms such as humans and mice.

This approach can obviously only exclude sites of known variation from consideration, and cannot add sample-specific methylation loci. To remove those reference-specific methylation loci that are **not** found in databases of known genetic variation, we might identify loci in the sample that display a large number of non-C/T bases (resp. non-G/A bases) at a C (resp. G) on the forward (resp. reverse) strand of the reference genome[14]

Determining the context or methylation type

In addition to determining whether a cytosine is methylated or unmethylated, we also want to determine the *context* of the cytosine, also known as the *methylation type.* That is, we want to determine whether the cytosine is a CG, CHG or a CHH.

This is done by examining the two bases upstream of the cytosine. It can be done based on the reference sequence, as is done in `Bismark` and `methtuple`, or from the reads themselves. The obvious difficulty with using the reads themselves is if the cytosine occurs at the last or second last position of the read, in which case the context may not be unambiguously determined from the the read alone. Instead, the context may be refined by initially using the reference genome context and then correcting for any sample-specific genetic variants in the two downstream bases.

A further complication occurs when there is a genetic variant in the two bases upstream of the methylation locus. We would like to use the two upstream bases from the sample to

[14] This is like an *ad hoc* and limited version of `Bis-SNP`.

infer the sample-specific methylation context, either inferred from each read separately or from a variant calling procedure such as Bis-SNP. However, this further complicates the methylation calling and so tools such as Bismark derive the context from the reference genome alone.

2.4 methtuple

methtuple (https://github.com/PeteHaitch/methtuple) is software I wrote to perform methylation calling at genomic tuples that I call *m-tuples*. Before formally defining m-tuples, I first motivate the need for methtuple.

2.4.1 Motivation

Most methylation callers, such as bismark_methylation_extractor and Bis-SNP, perform methylation calling at single methylation loci, which I refer to as 1-tuples. The output file is a table, where each row records the co-ordinates of a cytosine and the number of methylated (M) and unmethylated (U) reads at that position. Table 2.2 is representative of the type of data returned by these programs. The file format is generally tab-delimited plain text, the Browser Extensible Data (BED) format or the Variant Call Format (VCF).

Chromosome	Strand	Position	M	U
chr1	+	100	7	1
chr1	−	101	5	2
chr2	+	400	0	3
chr2	+	450	1	2

Table 2.2: Example of output for methylation calling at 1-tuples. Each row records the the number of methylated (M) and unmethylated (U) reads at a 1-tuple. Loci may be stratified by strand, as is done here, in which case most CpGs will have measurements for both the positive and negative strands.

While 1-tuples are the basis of most analyses of bisulfite-sequencing data, they do not always give the complete picture of how DNA methylation is acting in the sample. To gain a clearer picture, we can leverage the fact that many bisulfite-sequencing reads contain multiple methylation loci (*m-tuples*) and that each read is from a single cell[15]. An example

[15] This ignores chimeric reads, which are created when two DNA fragments ligate to one another during

of where this is useful is shown in Figure 2.4 where we have two regions, each with four methylation loci, that have identical methylation calls at 1-tuples yet very different overall methylation patterns. Further examples of where m-tuples are useful are in studying *epialleles* and *epipolymorphisms* (Chapter 4) and *co-methylation* (Chapters 6 and 7).

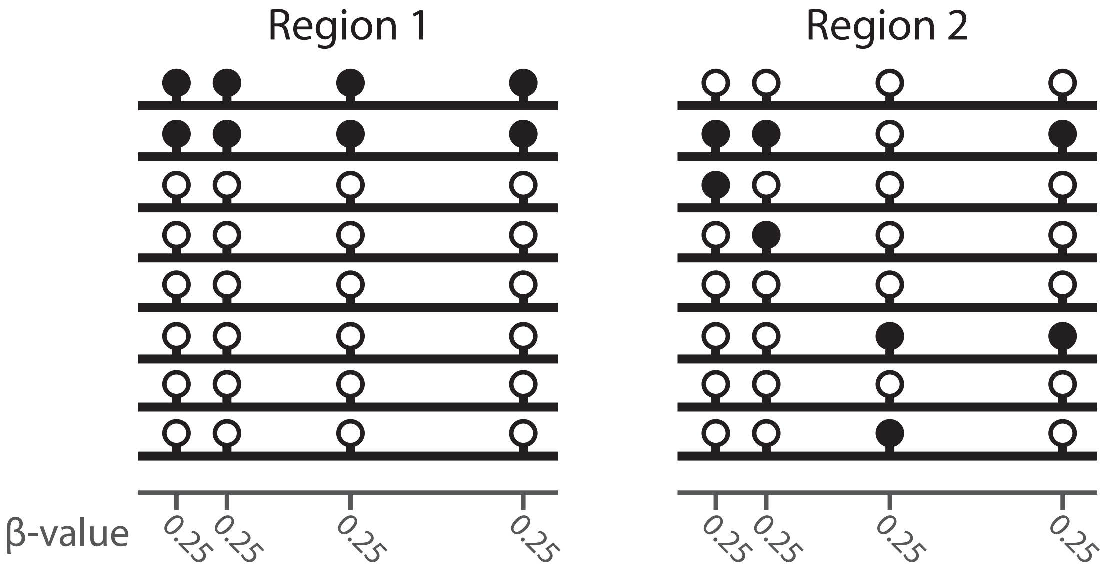


Figure 2.4: Two regions, each with four methylation loci that have identical β-values ($\beta = \frac{M}{M+U}$) at 1-tuples yet have very different overall methylation patterns. Each line is a read, a white circle is an unmethylated CpG and a black circle is a methylated CpG.

In order to study these phenomena, we firstly need software that can perform methylation calling at m-tuples, which is why I wrote `methtuple`. When I began my PhD, there was no software capable of calling methylation patterns at arbitrarily sized m-tuples from whole-genome bisulfite-sequencing data. Simultaneous with the development of `methtuple`, there have been some software published with similar functionality. However, none of these do exactly what I require and some have what I consider to be severe deficiencies (Table 2.3). To the best of my knowledge, `methtuple` is the only software that can perform methylation calling at m-tuples from whole-genome bisulfite-sequencing data.

the library preparation. Certain bisulfite-sequencing protocols frequently produce chimeric reads. For example, using the post-bisulfite adapter tagging (PBAT) protocol [Miura *et al.* 2012] with a low input amount of DNA results in a huge number of chimeric reads (personal communication from Felix Krueger). The standard whole-genome bisulfite-sequencing protocol is not known to suffer from this issue.

Table 2.3: Other software for methylation calling at m-tuples and their limitations. Abbreviations: `bme` = `bismark_methylation_extractor`.

Software	Reference	Input	Limitations
methclone	Li *et al.* [2014]	`Bismark BAM`	Unable to install
methpat	`https://github.com/bjpop/methpat`	Output of `bme`	Designed for amplicons not whole-genome data.
DMEAS	He *et al.* [2013]	Output of `bme`	Windows operating system only. Perl code only available as PDF file.

2.4.2 m-tuples

I define an m-tuple to be a tuple of $m = 1, 2, \ldots$ methylation loci. I refer to m as the *size* of the tuple. In principle, the m loci that make up an m-tuple could come from anywhere in the genome, but it makes most sense to require that the m loci be close to one another. In fact, I generally require that an m-tuple consists of m *adjacent* methylation loci[16]. An equivalent way of describing an m-tuple as comprising adjacent methylation loci is one where the number of intervening loci is zero ($NIL = 0$). There are three reasons that I focus on m-tuples with $NIL = 0$:

1. Quantity: From a sequence containing l methylation loci there are $l - \mathrm{m} + 1$ $NIL = 0$ m-tuples. In contrast, there are $\binom{l}{\mathrm{m}}$ $NIL \geqslant 0$ m-tuples. Obviously, $\binom{l}{\mathrm{m}} \geqslant l - \mathrm{m} + 1$, with strict inequality if $m \neq 1$ or $m \neq l$.
2. Interpretability: Results for m-tuples with $NIL = 0$ are simpler to interpret than when allowing $NIL \geqslant 0$. This is discussed in Chapter 7.
3. Measurability: We cannot observe methylation patterns from individual reads at m-tuples where the methylation loci are far apart due to the read length limitations of the Illumina sequencing technology. This is true even when $NIL = 0$ but is more of an issue if we allow $NIL \geqslant 0$.

[16] Two methylation loci are adjacent if there is no methylation loci in between the pair. For example, `CGCG` and `CGTTACG` both contain two adjacent CpGs (the intervening `TTA` in the second sequence does not include a CpG). In contrast, the first and last CpG in the sequence `CGTCGTCG` are **not** adjacent, since the intervening sequencing, `TCGT` include a CpG. Note that in situations where we are only interested in studying CpGs, we define 'methylation loci' to mean 'CpGs'. Therefore the sequence `CGTCTTCG` contains two adjacent methylation loci; while there is a `C` in the intervening sequence, `TCTT`, it is a CHH not a CpG.

When referring to m-tuples I implicitly mean those with $NIL = 0$; I will explicitly use the notation $NIL \geqslant 0$ when I wish to make clear that there may be intervening methylation loci in the m-tuple. The default option of `methtuple` is to produce m-tuples with $NIL = 0$ unless the `--all-combinations` flag is set[17].

I require that each methylation call at an m-tuple comes from a single read. There are 2^m possible methylation calls at an m-tuple. For example, at a 1-tuple there are 2^1 possible methylation calls — M or U; at a 3-tuple there are $2^3 = 8$ possible methylation calls — MMM, MMU, MUM, MUU, UMM, UMU, UUM or UUU.

For each m-tuple, I also define the intra-pair distance (*IPD*) as the vector containing the $(m - 1)$ pair-wise distances (measured in bp) between methylation loci in the m-tuple. For example, the 2-tuple (`chr7:+:145`, `chr7:+:163`) has $IPD = (163 - 145) = (18)$. The 5-tuple (`chr2:-:560`, `chr2:-:570`, `chr2:-:572`, `chr2:-:588`, `chr2:-:612`) has $IPD = (570 - 560, 572 - 570, 588 - 572, 612 - 588) = (10, 2, 16, 24)$. The IPD vector of a 1-tuple is undefined.

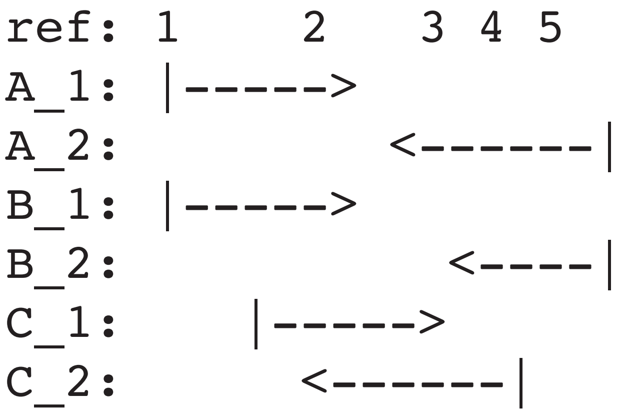
To illustrate several of the above-mentioned concepts, suppose we sequence a region of the genome containing five methylation loci with three paired-end reads (`A`, `B` and `C`), shown in Figure 2.5.

If we are interested in 1-tuples, Figure 2.6 shows what we would obtain from each read by running `methtuple`. The result is identical regardless of whether the `--all-combinations` flag is set.

If we are interested in 3-tuples, Figure 2.7 shows what we would obtain from each read by running `methtuple` in its default mode. A few things to note:

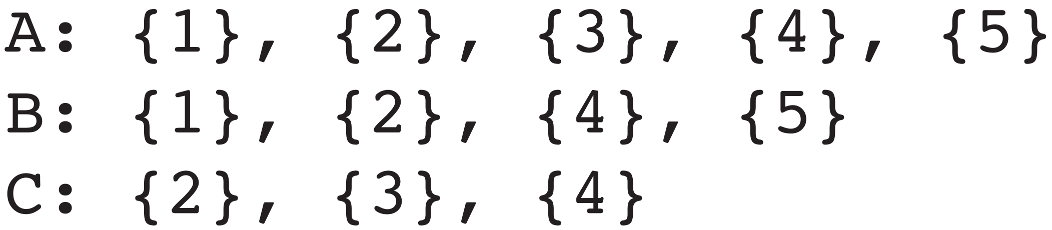
- Read-pair `A` sequences all three (= 5 - 3 + 1) adjacent 3-tuples
- Read-pair `B` sequences none of the adjacent 3-tuples but does ‘erroneously’ construct two 3-tuples from pairs of non-adjacent loci. This happens because m-tuples are created independently from each read-pair; effectively, read-pair `B` is unaware of methylation locus `3`. Depending on the downstream analysis, the user may wish to

[17] Actually, while `methtuple` tries to produce m-tuples with $NIL = 0$ it can't guarantee this because it would require looking up the reference genome sequence for each m-tuple (this is avoided for computational simplicity). This is only really an issue with paired-end sequencing, as is made clear in the examples of Figures 2.5, 2.6, 2.7 and 2.8. Some *post-hoc* filtering of the m-tuples will generally be required in order to remove those m-tuples with $NIL > 0$.



```
ref: 1    2   3 4 5
A_1: |----->
A_2:         <------|
B_1: |----->
B_2:           <----|
C_1:    |----->
C_2:      <------|
```

Figure 2.5: Diagram of three-paired end reads (`A`, `B` and `C`) mapping to a region containing five methylation loci (`1`, `2`, `3`, `4` and `5`). The suffix `_1` or `_2` indicates whether it is *read-1* or *read-2*, respectively.

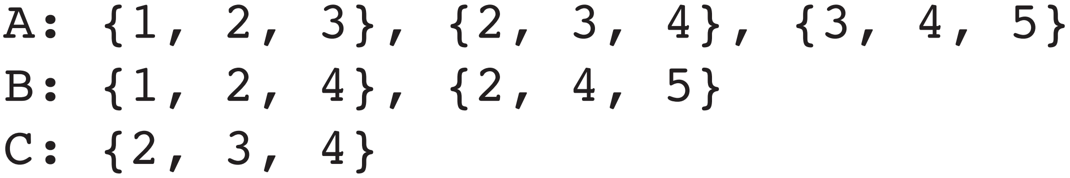


```
A: {1}, {2}, {3}, {4}, {5}
B: {1}, {2}, {4}, {5}
C: {2}, {3}, {4}
```

Figure 2.6: 1-tuples produced for each read for the toy example in Figure 2.6.

post-hoc filter out these m-tuples with non-adjacent loci.

- The twice-sequenced methylation loci in read-pair `C`, `2` and `3`, are only counted once.



```
A: {1, 2, 3}, {2, 3, 4}, {3, 4, 5}
B: {1, 2, 4}, {2, 4, 5}
C: {2, 3, 4}
```

Figure 2.7: 3-tuples produced for each read for the toy example in Figure 2.6.

Finally, Figure 2.8 shows the output if we were to analyse 3-tuples but with the `--all-combinations` flag set.

With current sequencing technology we are limited to extracting m-tuples that span no

```
A: {1, 2, 3}, {2, 3, 4}, {3, 4, 5}, {1, 2, 4}, {1, 2, 5},
   {1, 3, 4}, {1, 3, 5}, {1, 4, 5}, {2, 3, 5}, {2, 4, 5}
B: {1, 2, 4}, {2, 4, 5}, {1, 2, 5}, {1, 4, 5}
C: {2, 3, 4}
```

Figure 2.8: 3-tuples produced for each read for the toy example in Figure 2.6 when the `--all-combinations` flag is set.

more than 200 to 250 bp. This obviously affects the size of m-tuples that we can study. Figure 2.9 shows the number of CpGs per read for the Lister dataset (see Chapter 3 for a description of the Lister dataset). Longer reads, and paired-end reads, contain more methylation loci and so are more informative for analyses using m-tuples. This can be seen by comparing, for example, the *ADS* and *HSF1* samples. Samples sequenced more deeply will have more reads per m-tuple, although this can't be seen in these plots since they are normalised by sequencing depth.

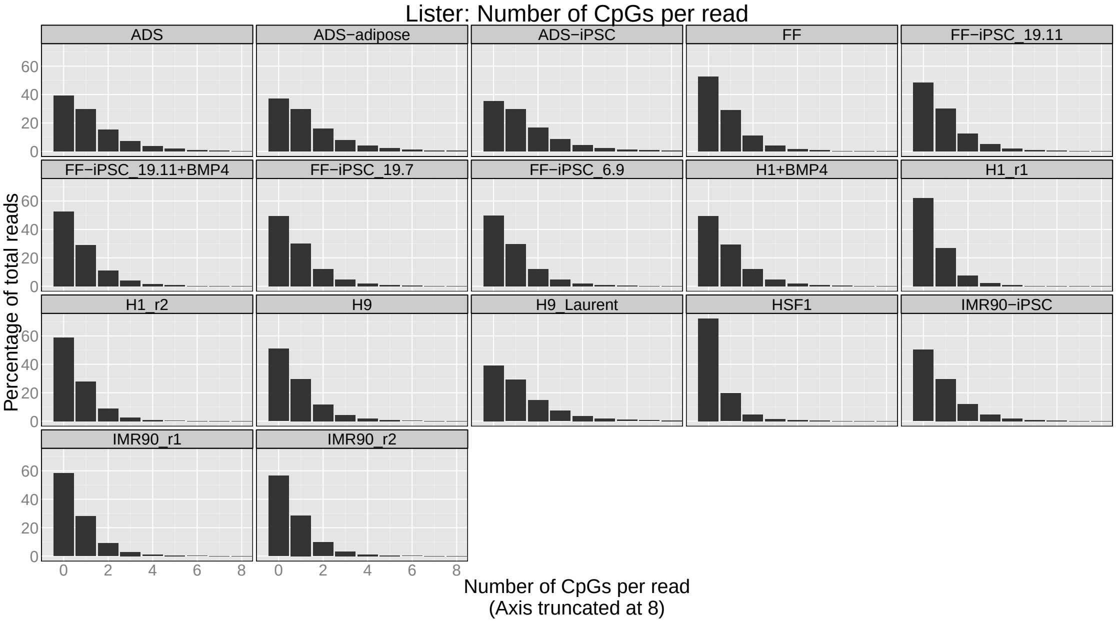


Figure 2.9: Number of CpGs per read for the Lister dataset.

2.4.3 Implementation

`methtuple` performs methylation calling for a single `BAM` file generated by `Bismark`. The user is required to specify the size of the tuples (`--m`), and the methylation type

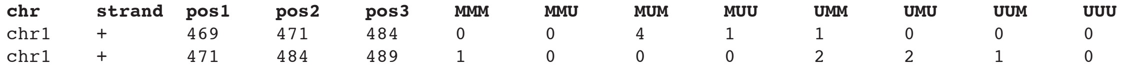
(`--methylationType`) for each run of the program. There are many useful options to filter reads and read-positions. Apart from the standard quality filters, `methtuple` is careful when processing paired-end reads to only count the base from one of the reads in any overlapping paired-end reads to avoid double-counting the bases in the overlapping region[18]. `methtuple` also allows the user to filter out specific read-positions rather than wholesale filtering of the ends of reads. This is particularly useful for samples where there is a 'spike' in the M-bias plot, such as that shown in Figure 2.2. Such a spike can be filtered out without also being forced to also filter out additional upstream read-positions that are not affected by M-bias.

`methtuple` is written in Python and uses the `pysam` (`https://github.com/pysam-developers/pysam/`) module to parse the `BAM` file. It is compatible with both Python2 and Python3. To improve performance, I provide a helper script to split the sample by chromosome and process each chromosome in parallel (`https://github.com/PeteHaitch/methtuple/blob/master/helper_scripts/run_methtuple.sh`). This helper script makes extensive use of `GNU parallel` [Tange 2011]. While Python-level parallel processing is desirable, this `GNU parallel`-based approach was simpler to implement and sufficient for my purposes.

`methtuple` is currently limited to processing files produced by `Bismark` due to its reliance on the `Bismark`-specific tags `XM`, the "methylation call string", `XR`, the "read conversion state for the alignment", and `XG`, the "genome conversion state for the alignment" (`http://www.bioinformatics.bbsrc.ac.uk/projects/bismark/Bismark_User_Guide.pdf`). It could be extended to work with other bisulfite-sequencing aligners. However, due to the eccentricities of each aligner, such an extension would have to be aligner-specific and is therefore a considerable undertaking. Each extension would require that tags analogous to the `XR`, `XG` and `XM` tags can be generated from the given `BAM` file. In the case of the `XM` tag, this would likely require that the reference genome is parsed in parallel with the `BAM` file, adding considerable computational overhead. Perhaps the easiest option would be to add a script that 'Bismark-ifies' the original `BAM` file. Since all my data are aligned with Bismark, or were converted to Bismark's `BAM` format, I have not yet had need to pursue this line of work.

[18] `methtuple` has several options for handling overlapping mates of paired-end reads via the `--overlap-filter` flag. The default method is to ignore any read-positions in the overlapping region where the methylation calls disagree.

The output format of `methtuple` is tab-delimited plain text, optionally compressed with `gzip` or `bzip2`. Figure 2.10 shows an example of the output for 3-tuples.



```
chr     strand  pos1    pos2    pos3    MMM     MMU     MUM     MUU     UMM     UMU     UUM     UUU
chr1    +       469     471     484     0       0       4       1       1       0       0       0
chr1    +       471     484     489     1       0       0       0       2       2       1       0
```

Figure 2.10: Example output of `methtuple` for 3-tuples.

2.4.4 Performance

I have used `methtuple` to perform methylation calling at CpG m-tuples, $m = 1, \ldots, 8$, for more than 40 whole-genome bisulfite-sequencing samples. Figure 2.11 shows the distribution of running times, Figure 2.12 the maximum memory usage across, and Figure 2.13 the output file sizes, for all the samples from the EPISCOPE, Lister, and Ziller datasets. For each sample, each chromosome was processed using a single core on one of the shared-use servers in the Bioinformatics Division (see Table A.1 in the Appendix for details of these machines).

The running time of `methtuple` is proportional to the number of reads mapped to the chromosome, which is proportional to the length of the chromosome and its copy number. The running time is largely independent of the tuple size (`-m`). The variation in running times within a chromosome is due to the number of reads generated per sample and the length of the reads, where the length of a paired-end read is defined as the sum of the mates' lengths. Samples with more reads take longer to process and samples sequenced with longer reads take longer because these contain more m-tuples.

The maximum memory usage is not strictly proportional to chromosome length. It is instead driven by the number and density of CpGs on the chromosome. For example, chromosome 19, which has the highest CpG density of all the autosomes in the human genome, requires far more memory than chromosome 18, which has less than half the CpG density of chromosome 19 (see Figure 2.14). The relationship between the maximum memory usage and the tuple size (`-m`) is complex; more data have to be retained as `-m` increases, thus increasing the memory usage, but fewer reads contain tuples of that size and so there aren't as many m-tuples or observations on these to count and retain. Memory usage is therefore relatively constant across values of `-m` for a given chromosome. The

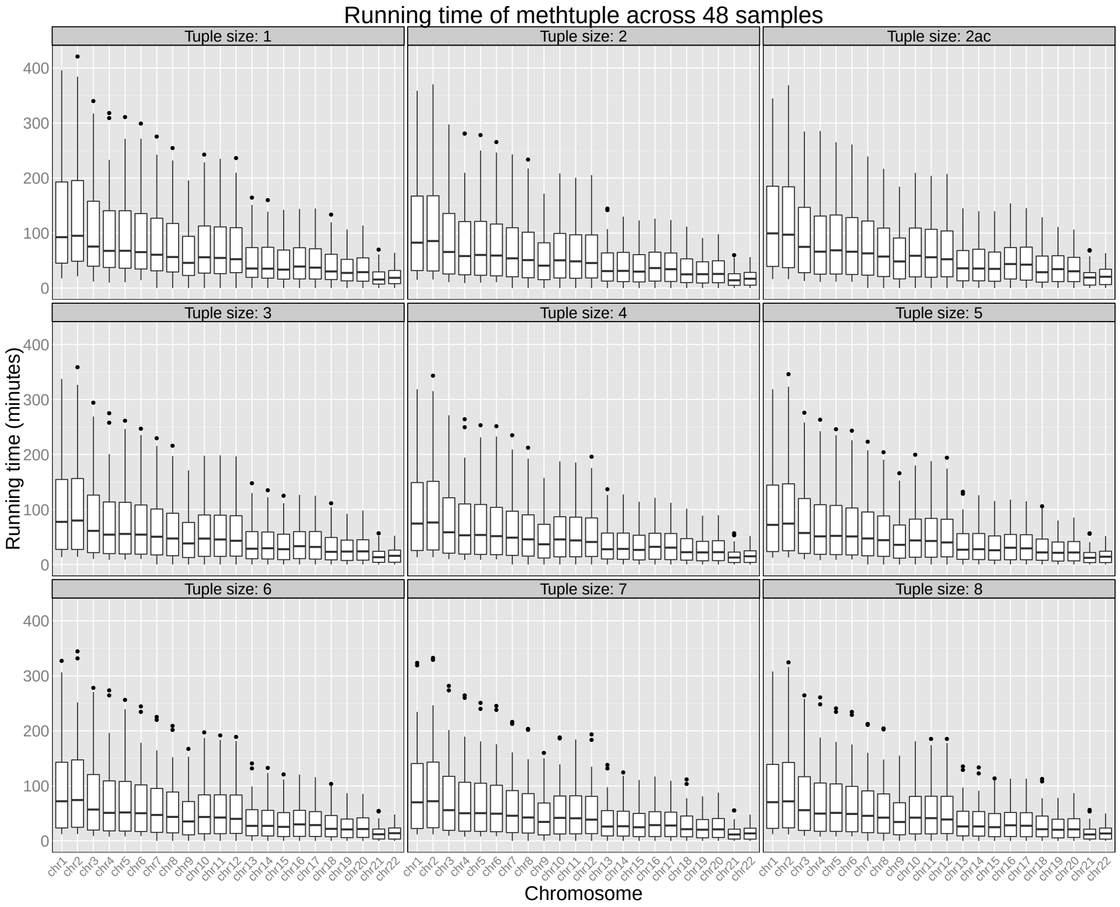


Figure 2.11: The running times are the 'User time' reported by `GNU time` converted from seconds to minutes. The suffix 'ac' on the tuple size means that the option `--all-combinations` was set. Note that the total number of samples is 48 because each of the Ziller sequencing runs is separately counted (see Chapter 3 for details).

obvious exception is for the results labelled `2ac`, which used the `--all-combinations` flag in conjunction with `-m 2`. This means that all 2-tuples with $NIL \geqslant 0$ were extracted and there are many, many more CpG 2-tuples with $NIL \geqslant 0$ than there are with $NIL = 0$, hence the increase in memory usage.

The regular structure of the output file means that these are particularly compressible. The size of the output file is almost always less than 1 GB when compressed with `gzip`.

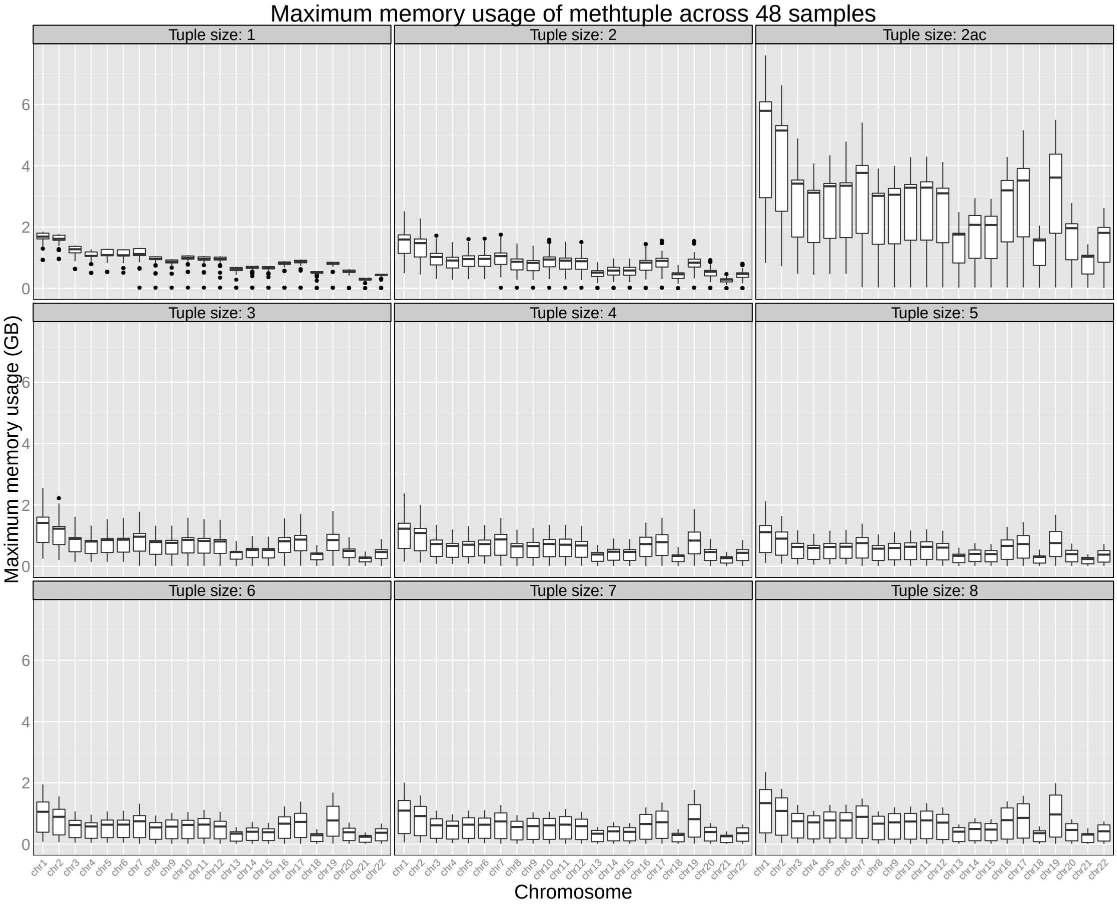


Figure 2.12: The maximum memory usage is the 'Maximum resident set size' reported by `GNU time` converted from kilobytes to gigabytes. These values are divided by four to fix bug in how `GNU time` reports the maximum memory usage (`https://bugzilla.redhat.com/show_bug.cgi?id=703865`). The suffix 'ac' means that the option `--all-combinations` was set. Note that the total number of samples is 48 because each of the Ziller sequencing runs is separetely counted (see Chapter 3 for details).

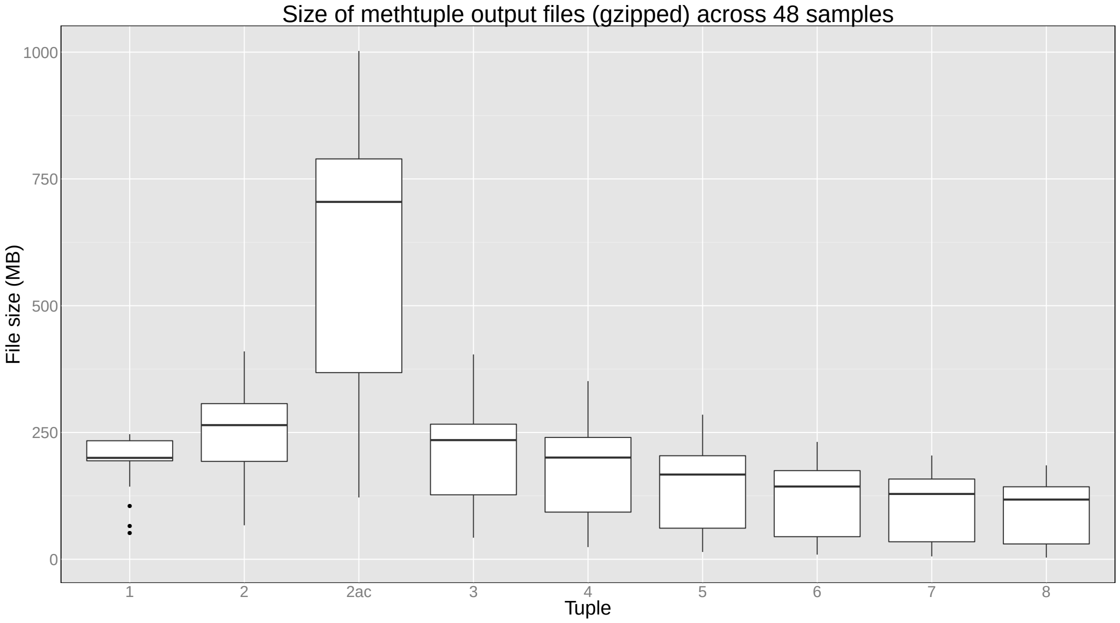


Figure 2.13: The sizes of the output file when compressed with gzip. The suffix 'ac' on the tuple size means that the option `--all-combinations` was set. Note that the total number of samples is 48 because each of the Ziller sequencing runs is separately counted (see Chapter 3 for details).

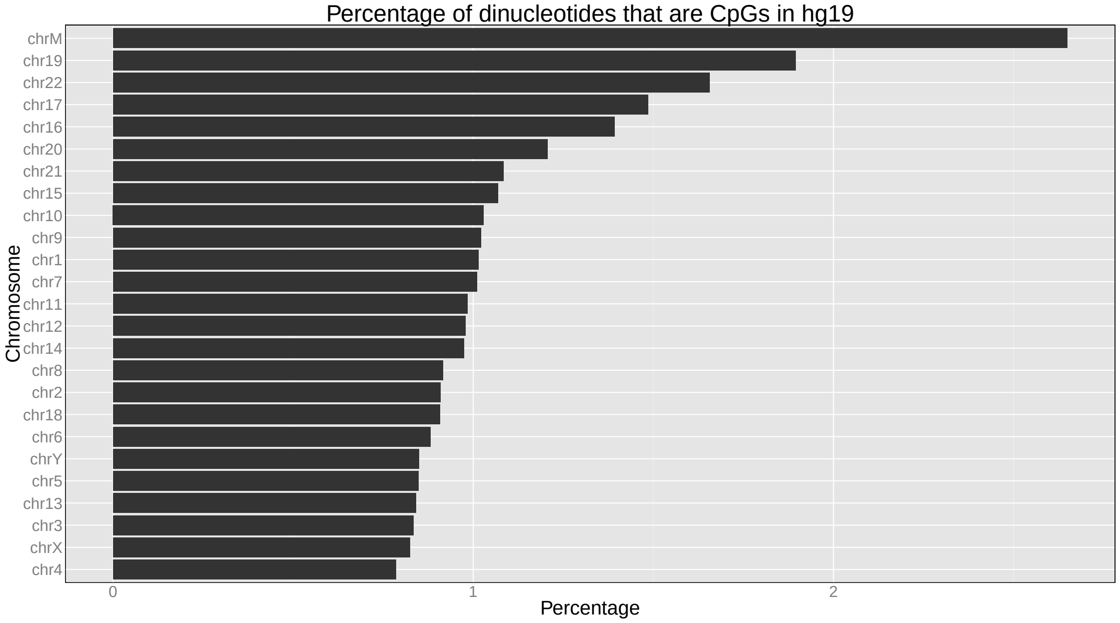


Figure 2.14: Percentage of dinucleotides that are CpGs for each chromosome in the human reference genome (hg19).

2.4.5 Availability

`methtuple` is open-source software released under the MIT licence and available from `https://github.com/PeteHaitch/methtuple`.

2.5 Summary

This chapter has detailed the first steps in a bioinformatics analysis of whole-genome bisulfite-sequencing data. There are many decisions to be made at each step and these will ultimately affect the quality of the data available for downstream analysis.

This chapter also introduced `methtuple`, the first of several pieces of software that were developed as part of my thesis. `methtuple` is software for calling methylation patterns at m-tuples from whole-genome bisulfite-sequencing data. It will be essential for our later analysis of within-fragment co-methylation (Chapter 7) and has wider utility in facilitating other downstream analyses based on methylation patterns at m-tuples (Chapter 5).

Chapter 3

Datasets used in thesis

3.1 Overview of data processing

This chapter briefly documents the 40 whole-genome bisulfite-sequencing samples that I use in my thesis. The `BAM` files containing the aligned reads for each sample underwent the same basic processing:

1. Genetic variants were called using the `bissnp_easy_usage.pl` script included with `Bis-SNP (v0.82.2)`.
2. M-bias was estimated using `bismark_methylation_extractor` with the `--mbias_only` flag set. These output files were then analysed using the `MethylationTuples` R package (see section 5.3) and all read-positions with an CpG normalised read-position methylation level (*nrpml*) more than 0.03 from the median, i.e. $|nrpml_{CpG} - median(nrpml_{CpG})| > 0.03$, were excluded from methylation calling (*read-1* and *read-2* analysed separately where applicable).
3. CpG methylation calling was performed using `methtuple (v1.4.0)` for m-tuples m $= 1, \ldots, 8$. CpG 2-tuples were called both with and without the `--all-combinations` flag; all other tuple sizes were called without the `--all-combinations` flag. The following `methtuple` flags were also used: `--methylation-type CG --ignore-duplicates --min-mapq 0 --overlap-filter XM_ol --ignore-duplicates`.
4. Sample-level m-tuples were combined at the dataset-level using the `MethylationTuples` R package (see Section 5.3). Specifically, a `MethPat` object was created for each of

the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets for 1-tuples, 2-tuples and 2ac-tuples (2-tuples with the `--all-combinations` flag in `methtuple` set), 3-tuples and 4-tuples. I did not create `MethPat` objects for m-tuples with m > 4 because the data are too sparse at this larger sizes to be generally useful.

The raw data for the *Lister*, *Seisenberger* and *EPISCOPE* datasets are all publicly available. The *EPISCOPE* data are not yet published and I do not have permission to make these publicly available. The scripts used to prepare the results for each chapter are available from `https://github.com/PeteHaitch/phd_thesis_analyses`. Further details of software used are available in Appendix A.3.

3.2 ***Lister*** dataset

The *Lister* dataset refers to whole-genome bisulfite-sequencing libraries used in Lister *et al.* [2009] and Lister *et al.* [2011]. The *Lister* data were the largest publicly available human whole-genome bisulfite-sequencing datasets until quite recently.

3.2.1 Sample descriptions

The methylC-seq libraries from the Lister *et al.* [2009] paper were the first published whole-genome bisulfite-sequencing libraries of mammalian DNA. A focus of this paper was comparing DNA methylation levels in a somatic tissue, fetal lung fibroblasts (*IMR90*), with those from a pluripotent tissue, embryonic stem cells (*H1*). Each tissue was run in duplicate. While Lister *et al.* [2009] refer to these "biological" replicates I believe that these are better described as technical replicates since each replicate is from the same cell line; what distinguishes the replicates are the number of cell passages and the subsequent library preparations and sequencing. In any case, the published analyses pool these duplicates, which ignores all between-replicate variability. These samples are detailed in Table 3.1.

The methylC-seq libraries from the Lister *et al.* [2011] include some created by the authors and some published by other groups. These samples include cell lines from differentiated cell lines, embryonic stem cell lines, pluripotent stem cell lines and *in vitro* differentiated from pluripotent stem cells. There are no replicates for any of the Lister

Table 3.1: Samples in *Lister* dataset from Lister *et al.* [2009]. The reported read lengths are prior to any trimming of the reads. Abbreviations: *ESC* = embryonic stem cell; *SE* = single-end sequencing.

Sample	Tissue type	Sequencing	Read length	Ave. coverage
IMR90_r1	Lung fibroblasts	SE	87 bp	14×
IMR90_r2	Lung fibroblasts	SE	87 bp	15×
H1_r1	ESC	SE	85 bp	15×
H1_r2	ESC	SE	85 bp	14×

et al. [2011] samples. These samples are detailed in Table 3.2.

Table 3.2: Samples in *Lister* dataset from Lister *et al.* [2011]. The reported read lengths are prior to any trimming of the reads. Abbreviations: *iPSC* = induced pluripotent stem cell; *ESC* = embryonic stem cell; *IVD* = *in vitro* differentiated from pluripotent cell line; *SE* = single-end sequencing; *PE* = paired-end sequencing.

Sample	Tissue type	Sequencing	Read length	Ave. coverage
ADS	Adipose	PE	75 bp	23×
ADS-adipose	Adipocytes from ADS	PE	75 bp	24×
ADS-iPSC	iPSC from ADS	PE	75 bp	26×
FF	Foreskin fibroblasts	SE	85 bp	16×
FF-iPSC_6.9	iPSC from FF	SE	85 bp	10×
FF-iPSC_19.7	iPSC from FF	SE	85 bp	9×
FF-iPSC_19.11	iPSC from FF	SE	85 bp	8×
FF-iPSC_19.11+BMP4	IVD from FF-iPSC_19.11	SE	85 bp	17×
IMR90-iPSC	iPSC from IMR90	SE	85 bp	9×
H1+BMP4	IVD from H1	SE	85 bp	33×
H9	ESC	SE	85 bp	9×
H9_Laurent	ESC	PE	75 bp	8×
HSF1	ESC	SE	47 bp	5×

There are four ‘mini datasets’ within the Lister data that I make some use of in my thesis. The first I refer to as the *Lister-ADS* data and includes samples *ADS*, *ADS-adipose* and *ADS-iPSC*, all from the 2011 paper. The *ADS* sample, a human adipose tissue cell line, is the ‘founder’ of this mini dataset. The *ADS-adipose* and *ADS-iPSC* are both derived from the *ADS* cell line. The *ADS-adipose* sample are “adipocytes derived from the *ADS* cells through adipogenic differentiation conditions”. The *ADS-iPSC* cell line is an induced pluripotent stem cell line derived from *ADS*.

The second mini dataset I refer to as the *Lister-FF* data and includes samples *FF*, *FF-iPSC_6.9*, *FF-iPSC_19.7*, *FF-iPSC_19.11* and *FF-iPSC_11.11+BMP4*, all from the 2011 paper. The *FF* sample, a foreskin fibroblast cell line, is the 'founder' of this this mini dataset. The *FF-iPSC_6.9*, *FF-iPSC_19.7* and *FF-iPSC_19.11* are all induced pluripotent stem cell lines derived from *FF*. In fact, *FF-iPSC_19.7* and *FF-iPSC_19.11* are subclones derived from *FF-iPSC_19*, whose methylome was not sequenced. I believe *FF-iPSC_6.9* is an independently derived iPSC cell line from *FF*, although this isn't made clear in the original publication. The *FF-iPSC_19.11+BMP4* sample is a trophoblast cell line derived by *in vitro* differentiating the *FF-iPSC_19.11* by growing a clone of it in bone morphogenic protein 4 (*BMP4*).

The third mini dataset I refer to as the *Lister-IMR90* data and includes samples *IMR90_r1* (2009), *IMR90_r2* (2009) and *IMR90-iPSC* (2011). The *IMR90-iPSC* sample is an induced pluripotent stem cell line derived from the *IMR90* cell line.

The final mini dataset I refer to as the *Lister-H1* data and includes samples *H1_r1* (2009), *H1_r2* (2009) and *H1+BMP4* (2011). The *H1+BMP4* sample is a trophoblast cell line derived by *in vitro* differentiating the *H1* by growing a clone of it in bone morphogenic protein 4 (BMP4).

3.2.2 Creation of `BAM` files

The aligned reads for the Lister *et al.* [2009] data were downloaded from `http://neomorph.salk.edu/human_methylome/`. The aligned reads for the Lister *et al.* [2011] data were downloaded from `http://neomorph.salk.edu/ips_methylomes/`. These samples had been aligned against the hg18 build of the human reference genome.

As the aligned reads were in a custom file format, I wrote Python scripts to convert these files to the canonical `SAM` format. These scripts are available from `https://github.com/PeteHaitch/Lister2BAM`. These `SAM` files were then converted to `BAM` files with `SAMtools` [Li *et al.* 2009] and duplicate reads were marked using Picard's `MarkDuplicates` routine (`http://broadinstitute.github.io/picard/`).

3.3 *EPISCOPE* dataset

The *EPISCOPE* data were kindly provided to me by Professor Susan Clark (Garvan Institute of Medical Research, Sydney) and Dr Peter Molloy (CSIRO Animal, Food and Health Sciences). This dataset is not yet published.

3.3.1 Sample descriptions

The data are from three human donors across four different tissues, for a total of 12 whole-genome bisulfite-sequencing libraries. The four tissues are:

- *BUF*: Buffy coat layer, which are leukocytes and platelets derived by centrifugation of a whole blood sample.
- *SA*: Subcutaneous adidose tissue, which is fat found just below the skin. Unlike visceral adipose tissue, subcutaneous adipose tissue is thought to be protective against obesity-related metabolic dysfunction [Chau *et al.* 2014].
- *VA*: Visceral adipocytes, which are derived from *VAT*.
- *VAT*: Visceral adipose tissue, which is located inside the abdominal cavity, packed between the organs and is associated with metabolic dysfunction [Chau *et al.* 2014].

The data are summarised in Table 3.3.

3.3.2 Creation of `BAM` files

The sequencing data for these 12 samples were processed and aligned by Aaron Statham (Garvan Institute of Medical Research, Sydney). Each sample was aligned to the human reference genome (hg19) using `Bismark` (`v0.8.3`) with the `Bowtie2` backend. The default alignment options were used, except that the maximum insert size for valid paired-end alignments was set to 1000 instead of 500 (`-X 1000`). Duplicate reads had already been removed from the `BAM` files that I received.

Table 3.3: Samples in *EPISCOPE* dataset. The reported read lengths are prior to any trimming of the reads. Abbreviations: *PE* = paired-end sequencing; *BUF* = buffy coat; *SA* = subcutaneous adipose; *VA* = visceral adipocytes; *VAT* = visceral adipose tissue.

Sample	Tissue	Sequencing	Read length	Ave. coverage
E13BUF	BUF	PE	101 bp	8×
E13SA	SA	PE	101 bp	28×
E13VA	VA	PE	150 bp	27×
E13VAT	VAT	PE	101 bp	25×
E18BUF	BUF	PE	101 bp	21×
E18SA	SA	PE	101 bp	25×
E18VA	VA	PE	150 bp	36×
E18VAT	VAT	PE	101 bp	26×
E23BUF	BUF	PE	101 bp	12×
E23SA	SA	PE	101 bp	29×
E23VA	VA	PE	101 bp	32×
E23VAT	VAT	PE	101 bp	31×

3.4 *Seisenberger* dataset

The *Seisenberger* data are from a study of the dynamics of DNA methylation reprogramming in mouse primordial germ cells [Seisenberger *et al.* 2012]. These were a convenience sample provided to me by a colleague, Felix Krueger (Babraham Institute). I thank Felix who sent me the `BAM` files containing processed and aligned reads.

3.4.1 Sample descriptions

I have the data for only three samples from the original publication, detailed in Table 3.4. The *J1_1* sample is from an embryonic stem cell line while both the *E6.5_epiblast_1* and *E16.5_male_1* samples are derived from pools of 10 to 30 embryos. Developmentally, the samples are ordered *J1_1* (embryonic stem cell), *E6.5_epiblast_1* (embryonic day 6.5 epiblast) and *E16.5_male_1* (embryonic day 16.5 male progenitor germ cells).

I believe that the samples I received labelled *J1_1* and *E16.5_male_1* in fact correspond to *J1_2* and *E16.5_male_2*, respectively, i.e. the second replicate rather than the first. The data I received are all 100 bp paired-end sequences, which matches replicate 2 rather than replicate 1 for both of these samples Seisenberger *et al.* [2012, Supplementary Table 1].

Table 3.4: Samples in *Seisenberger* dataset. The reported read lengths are prior to any trimming of the reads. All samples were first published in Seisenberger *et al.* [2012]. Abbreviations: *PE* = paired-end sequencing

Sample	Tissue	Sequencing	Read length	Ave. coverage
J1_1	ESC	PE	100 bp	12×
E6.5_epiblast_1	Epiblast	PE	100 bp	13×
E16.5_male_1	Male progenitor germ cells	PE	100 bp	12×

3.4.2 Creation of BAM files

The sequencing data for these 3 samples were processed and aligned by Felix Krueger (Babraham Institute). Each sample was aligned to the mouse reference genome (GRCm38/mm10) using `Bismark` (`v0.7.12`) with the `Bowtie1` backend. The default alignment options were used.

3.5 *Ziller* dataset

The *Ziller* data are a subset of the data used in Ziller *et al.* [2013]. Specifically, I use a convenience sample of 8 whole-genome bisulfite-sequencing libraries. These were made available to me by a collaborator, Aaron Statham (Garvan Institute of Medical Research, Sydney). I thank Aaron who sent me the `BAM` files containing processed and aligned reads.

3.5.1 Sample descriptions

The eight biological samples are as follows: frontal cortex from two 'normal' women donors (*Frontal_cortex_normal_1* and *Frontal_cortex_normal_2*) and from two women who had Alzheimer's disease (*Frontal_cortex_AD_1* and *Frontal_cortex_AD_2*); a sample from a human liver carcinoma cell line (*HepG2_cell_line*); a new sample from the *IMR90* lung fibroblast cell line (*IMR90_cell_line*); and samples from a colon cancer matched tumour-normal pair (*Colon_Tumor_Primary* and *Colon_Primary_Normal*).

Table 3.5 summarises the data for the 19 individual sequencing runs[1].

[1] The average sequencing coverage of the *post-hoc* merged samples are approximately the sums of the average sequencing coverage for the corresponding individual sequencing runs.

Table 3.5: Sequencing runs in *Ziller* dataset. The reported read lengths are prior to any trimming of the reads. All samples were first published in Ziller *et al.* [2013]. Abbreviations: *PE* = paired-end sequencing

Sample	Tissue	Sequencing	Read length	Ave. coverage
SRR949193	Frontal_cortex_normal_1	PE	101 bp	10×
SRR949194	Frontal_cortex_normal_1	PE	101 bp	10×
SRR949195	Frontal_cortex_normal_1	PE	101 bp	10×
SRR949196	Frontal_cortex_normal_2	PE	101 bp	9×
SRR949197	Frontal_cortex_normal_2	PE	101 bp	9×
SRR949198	Frontal_cortex_normal_2	PE	101 bp	9×
SRR949199	Frontal_cortex_AD_1	PE	101 bp	9×
SRR949201	Frontal_cortex_AD_1	PE	101 bp	9×
SRR949202	Frontal_cortex_AD_2	PE	101 bp	10×
SRR949203	Frontal_cortex_AD_2	PE	101 bp	10×
SRR949206	HepG2_cell_line	PE	101 bp	2×
SRR949207	HepG2_cell_line	PE	101 bp	1×
SRR949208	IMR90_cell_line	PE	101 bp	1×
SRR949209	IMR90_cell_line	PE	101 bp	3×
SRR949210	Colon_Tumor_Primary	PE	101 bp	8×
SRR949211	Colon_Tumor_Primary	PE	101 bp	8×
SRR949212	Colon_Tumor_Primary	PE	101 bp	9×
SRR949213	Colon_Tumor_Primary	PE	101 bp	8×
SRR949215	Colon_Primary_Normal	PE	101 bp	8×

3.5.2 Creation of `BAM` files

I received 19 `BAM` files from Aaron, which represent 19 sequencing runs of the eight biological samples. Each of the 19 `BAM` files was processed separately and I then *post hoc* merged the processed data from sequencing runs. This unfortunately reduces the power to detect genetic variants since the coverage of individual sequencing runs is lower than merged data, but does not adversely affect methylation calling since the number of methylated and unmethylated reads can be summed across sequencing runs.

3.6 CpG islands

I have used the CpG island definition from the hidden Markov model proposed by Wu *et al.* [2010] and implemented in the `makeCGI` R package (`v1.2`, `http://rafalab.jhsph.edu/CGI/`). The predicted CpG islands for the human reference genome (both hg18 and hg19) were downloaded from `http://rafalab.jhsph.edu/CGI/` on 29/10/2014. I used the `makeCGI` R package to create the predicted CpG islands for the mouse reference genome (mm10) since these were not available for download.

Chapter 4

A statistical framework for analysing whole-genome bisulfite-sequencing data

Overview

This chapter sets out a statistical framework for analysing bisulfite-sequencing data. The ideas here are simple, however, they have not yet been put into a unified mathematical framework. My intention in doing so is to clarify several subtleties that, in my experience, are potential sources of confusion.

Beginning with a single sample, I explain the various sources of variation in DNA methylation data and introduce the mathematical notation that I use throughout my thesis. I then extend this framework to multiple samples.

Finally, I describe key variables, common estimators of these and their statistical properties by analysing 40 whole-genome bisulfite-sequencing samples.

4.1 One sample

There are several levels of variation to consider in a bisulfite-sequencing experiment, even with only a single sample. I find it convenient to separate these into:

1. Pre-sequencing sources of variation
2. Sequencing and post-sequencing sources of variation.

In the following, I describe these sources of variation, which are illustrated in Figure 4.1.

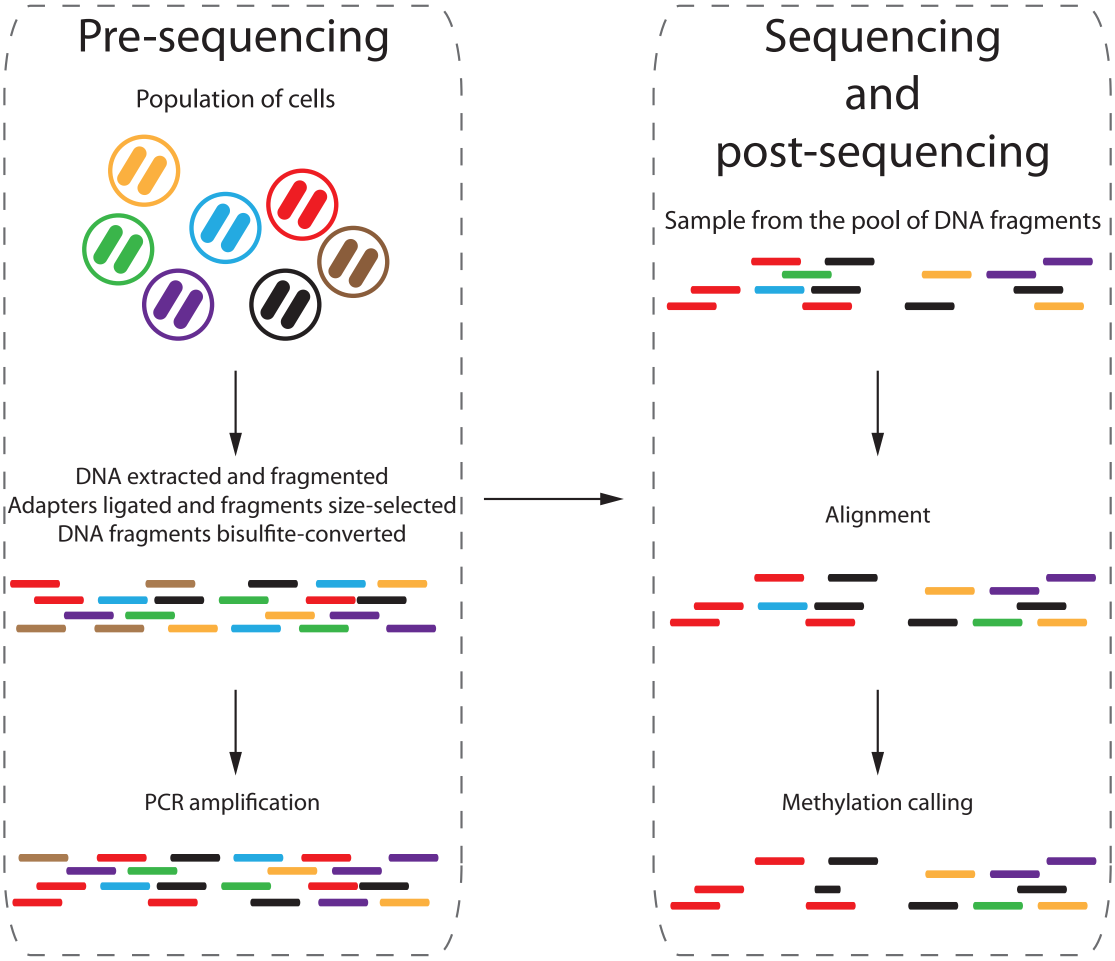

Pre-sequencing
Population of cells
DNA extracted and fragmented
Adapters ligated and fragments size-selected
DNA fragments bisulfite-converted
PCR amplification
Sequencing and post-sequencing
Sample from the pool of DNA fragments
Alignment
Methylation calling

Figure 4.1: A schematic illustrating several sources of variation in a bisulfite-sequencing experiment. Each sample starts as a population of cells, with potentially different methylation patterns. Each DNA fragment is coloured by its originating cell (although this is unknown in practice). Illustrated are PCR amplication bias (unequal representation of DNA from each cell following PCR amplification), sampling variation (not all DNA fragments are sequenced), alignment errors (not all sequenced fragments are mapped or may be mapped to the wrong location, as is the case for the green fragment), and filtering during methylation calling (not reads are used, the blue read, and reads may have read-positions removed, the black reads).

4.1.1 Pre-sequencing

A methylation locus is a single cytosine, that is, a CpG, CHG or CHH. The set of these loci is labelled $\mathcal{I} = \{pos_i : i = 1, \ldots, N_{loci}\}$, where pos_i is the genomic co-ordinates of the i^{th} locus, e.g., `chr1:+:666`. It will be convenient to refer to loci by the subscript i rather than by pos_i, although it is important to remember that the number of base pairs between the i^{th} and $(i+1)^{th}$ methylation loci varies along the genome. Furthermore, depending on whether the data are stranded, a pair of loci may be on opposite strands. In a small number of instances the i^{th} and $(i+1)^{th}$ methylation loci are on separate chromosomes, e.g., the last CpG on chromosome 1 and the first CpG on chromosome 2.

The methylation state of a locus can vary within a sample due to cell-to-cell heterogeneity of DNA methylation. A sample in a bisulfite-sequencing experiment contains DNA that is extracted from hundreds or thousands of cells and each cell may have a slightly different methylation profile. Furthermore, within a diploid cell there are two copies of each chromosome, and therefore two copies of each methylation locus, and these two copies can have different methylation states. It is also therefore necessary to consider not just the genomic co-ordinates of the locus but from which DNA fragment the methylation state originated.

Suppose that in the pool of DNA fragments for the sample that there are H_i fragments containing the i^{th} methylation locus. In general, H_i is unknown and will vary from locus to locus within a sample[1]. Note that the value of H_i is determined following the library preparation, including PCR amplification of the DNA; therefore, it can give a grossly distorted picture of the true representation of the cells. We denote by $\mathcal{H}_i$ the set of all fragments containing the i^{th} locus.

Although we do not know the number of fragments in the pool, we can define (and measure) the methylation state of a locus on a single DNA fragment. We denote by the indicator random variable, $Z_{h,i}$, the methylation state of i^{th} methylation locus on the h^{th} DNA fragment:

[1] Knowing H_i would require knowing: (1) the number of cells used as input (which might only be known to within an order of magnitude), (2) the ploidy of each cell (generally known) and (3) the number of PCR cycles (generally known). But the real problem is that none of the steps in creating the pool of DNA fragments is perfect. In particular, PCR introduces biases; some molecules are preferentially amplified while others 'drop out'. So even if we knew (1), (2) and (3) we cannot simply multiply these together to compute H_i, although this might at least give us a rough estimate.

$$Z_{h,i} = \begin{cases} 1 & \text{if methylated on the } h^{th} \text{ fragment} \\ 0 & \text{if unmethylated on the } h^{th} \text{ fragment} \end{cases}$$

By summing over the fragments containing the i^{th} locus, we obtain the number of fragments that are methylated at the i^{th} locus (M_i) and unmethylated at the i^{th} locus (U_i):

$$\begin{aligned} M_i &= \sum_{h=1}^{H_i} Z_{h,i} = |\{Z : Z \in \mathcal{H}_i, Z = 1\}| \\ U_i &= \sum_{h=1}^{H_i} (1 - Z_{h,i}) = |\{Z : Z \in \mathcal{H}_i, Z = 0\}| \end{aligned}$$

From these we can compute the proportion of fragments that are methylated at the i^{th} locus:

$$B_i = \frac{M_i}{M_i + U_i}$$

The above definitions can be extended from individual methylation loci, 1-tuples, to m-tuples. Mathematically, an m-tuple is denoted by a sequence of methylation loci, $(i, i+1, \ldots, i+m-1)$.

We denote the methylation pattern on the h^{th} DNA fragment containing the m-tuple $(i, i+1, \ldots, i+\mathrm{m}-1)$ by the vector of indicator random variables, $Z_{h,(i,i+1,\ldots,i+\mathrm{m}-1)}$,:

$$Z_{h,(i,i+1,\ldots,i+\mathrm{m}-1)} = \begin{cases} (0,0,\ldots,0) & \text{if unmethylated at the } \Big(i^{th},\ldots,(i+\mathrm{m}-1)^{th}\Big) \\ & \text{locus on the } h^{th} \text{ fragment} \\ (0,0,\ldots,1) & \text{if unmethylated at the } \Big(i^{th},\ldots,(i+\mathrm{m}-2)^{th}\Big) \\ & \text{locus and methylated at the } (i+\mathrm{m}-1)^{th} \\ & \text{locus on the } h^{th} \text{ fragment} \\ \vdots & \\ (1,1,\ldots,1) & \text{if methylated at the } \Big(i^{th},\ldots,(i+\mathrm{m}-1)^{th}\Big) \\ & \text{locus on the } h^{th} \text{ fragment} \end{cases}$$

We denote the set of all fragments containing the m-tuple $(i, i+1, i+m-1)$ by $\mathcal{H}_{(i,i+1,i+m-1)}$.

There are 2^{m} possible methylation patterns at an m-tuple. Rather than describe a methylation pattern by an m-vector of zeros and ones, I also write these using U and M; for example, the possible methylation patterns at a 2-tuple are MM, MU, UM and UU.

Analogous to the definition of M_i and U_i for 1-tuples (m = 1), we have when m = 2:

$$\begin{aligned} MM_{(i,i+1)} &= |\{Z : Z \in \mathcal{H}_{(i,i+1)}, Z = (1,1)\}| \\ MU_{(i,i+1)} &= |\{Z : Z \in \mathcal{H}_{(i,i+1)}, Z = (1,0)\}| \\ UM_{(i,i+1)} &= |\{Z : Z \in \mathcal{H}_{(i,i+1)}, Z = (0,1)\}| \\ UU_{(i,i+1)} &= |\{Z : Z \in \mathcal{H}_{(i,i+1)}, Z = (0,0)\}| \end{aligned}$$

We could extend the B_i values to m-tuples, although the intuitive interpretation of these as the average methylation level is lost. Instead, it reflects the relative frequencies of each methylation pattern. Here are the definitions for m = 2:

$$
\begin{aligned}
B_{(i,i+1)}^{MM} &= \frac{MM_{(i,i+1)}}{MM_{(i,i+1)} + MU_{(i,i+1)} + UM_{(i,i+1)} + UU_{(i,i+1)}} \\
B_{(i,i+1)}^{MU} &= \frac{MU_{(i,i+1)}}{MM_{(i,i+1)} + MU_{(i,i+1)} + UM_{(i,i+1)} + UU_{(i,i+1)}} \\
B_{(i,i+1)}^{UM} &= \frac{UM_{(i,i+1)}}{MM_{(i,i+1)} + MU_{(i,i+1)} + UM_{(i,i+1)} + UU_{(i,i+1)}} \\
B_{(i,i+1)}^{UU} &= \frac{UU_{(i,i+1)}}{MM_{(i,i+1)} + MU_{(i,i+1)} + UM_{(i,i+1)} + UU_{(i,i+1)}}
\end{aligned}
$$

The definitions for m > 2 follow in the obvious manner.

Again, I emphasise that $H_{(i,i+1,i+\mathrm{m}-1),j}, Z_{h,(i,i+1,i+\mathrm{m}-1),j}, B_{(i,i+1,i+\mathrm{m}-1),j}$ and the set of methylation patterns are unobservable. We aim to estimate these variables through sequencing the pool of DNA fragments.

4.1.2 Post-sequencing

A whole-genome bisulfite-sequencing experiment does not sequence every DNA fragment in the pool. Rather, sequencing can be thought of as sampling without replacement from the pool of DNA fragments. We have a large number (~10^{10}) of fragments in the pool and each methylation locus is only present on a small number of those fragments. We can therefore approximate this sampling by Poisson sampling, where the rate parameter for locus i is proportional to the number of fragments in the pool and inversely proportional to the number of fragmentation containing the i^{th} methylation locus, H_i.

We can ignore reads that do not contain any methylation loci as these are not relevant to ths discussion. We make three further simplifying assumptions:

1. Sequencing is performed without error.
2. Read mapping is perfect.
3. We perform single-end sequencing.

The effect of the first two assumptions are discussed in Section 2.2. The effect of the third assumption is minor. When using single-end sequencing, the methylation loci from a single read will always form a positively ordered sequence without gaps, i.e. $(i, i+1, i+2)$

and not, for example, $(i, i-1, i-2)$ nor $(i, i+1, i+3)$. However, when using paired-end sequencing, the methylation loci from a paired-end read will still be an ordered sequence but one of the following may occur:

1. There may be gaps due to the insert size being longer than the sum of the read lengths, e.g., $(i, i+1, i+3, i+4)$. In effect, we have missing data for any intervening methylation loci, the $(i+2)^{th}$ loci in this example.
2. Loci may be measured twice if the insert size is less than the sum of the read lengths, e.g. *read-1* gives us $(i, i+1)$ and *read-2* gives us $(i+1, i+2, i+3)$. In this example we must use only one of *read-1* or *read-2* as the measurement of the $(i+1)^{th}$ locus because otherwise we are 'double-counting'[2].

Each read measures the methylation state of one or more loci from a single DNA fragment. We denote by $\mathcal{R}_i$ the set of all mapped reads containing the i^{th} locus. The number of reads containing the i^{th} locus is referred to as the sequencing depth at the i^{th} locus, which we denote by $d_i = |\mathcal{R}_i|$, where $d_i \leqslant H_i$ with strict inequality for almost all i.

A single read containing the i^{th} locus is denoted $z : z \in \mathcal{R}_i$ and the observed methylation state is indicated by:

$$z : z \in \mathcal{R}_i = \begin{cases} 1 & \text{if methylated at the } i^{th} \text{ locus} \\ 0 & \text{if unmethylated at the } i^{th} \text{ locus} \end{cases}$$

By summing over the reads containing the i^{th} locus we obtain the number of reads that are methylated at the i^{th} locus (m_i) and unmethylated at the i^{th} locus (u_i):

[2] We should also check that the overlapping bases agree and, if not, either use the call with the highest base quality or ignore these overlapping positions in both reads. `methtuple` has several options for handling overlapping mates of paired-end reads via the `--overlap-filter` flag. The default method is to ignore any read-positions in the overlapping region where the methylation calls disagree.

$$
\begin{aligned}
m_i &= \sum_{r=1}^{d_i} z_{r,i} \\
&= |\{z : z \in \mathcal{R}_i, z = 1\}| \\
u_i &= \sum_{r=1}^{d_i} (1 - z_{r,i}) \\
&= |\{z : z \in \mathcal{R}_i, z = 0\}
\end{aligned}
$$

From these we can compute the proportion of reads that are methylated at the i^{th} locus as:

$$\beta_i = \frac{m_i}{d_i}$$

Here we have assumed that $d_i = m_i + u_i$, meaning that reads that do not have a methylation locus mapped to pos_i do not contribute[3].

This is the so-called β-value, which is commonly interpreted as an estimate of B_i, the proportion of cells in the sample that are methylated at the i^{th} locus. In Section 4.3.2 we discuss this interpretation and other estimators of the 'methylation level' at a locus.

These definitions can also be extended from 1-tuples to m-tuples. The set of all reads containing the m-tuple $(i, i+1, \ldots, i+\mathrm{m}-1)$ is denoted by $\mathcal{R}_{(i,i+1,\ldots,i+\mathrm{m}-1)}$ and has sequencing depth $d_{(i,i+1,\ldots,i+\mathrm{m}-1)} = |\mathcal{R}_{(i,i+1,\ldots,i+\mathrm{m}-1)}|$. A single read containing the m-tuple $(i, i+1, \ldots, i+\mathrm{m}-1)$ is denoted by $z : z \in \mathcal{R}_{(i,i+1,\ldots,i+\mathrm{m}-1)}$, and the observed methylation state is given by:

[3] Such reads can occur due to sequencing error, mapping error or genetically heterozygous methylation loci.

$$z : z \in \mathcal{R}_{(i,i+1,\ldots,i+\mathrm{m}-1)} = \begin{cases} (0,0,\ldots,0) & \text{if unmethylated at the} \\ & \left(i^{th}, (i+1)^{th}, \ldots, (i+\mathrm{m}-1)^{th}\right) \\ & \text{locus} \\ (0,0,\ldots,1) & \text{if unmethylated at the} \\ & \left(i^{th}, (i+1)^{th}, \ldots, (i+\mathrm{m}-2)^{th}\right) \\ & \text{locus} \\ & \text{and methlyated at the } (i+\mathrm{m}-1)^{th} \text{ locus} \\ \vdots & \\ (1,1,\ldots,1) & \text{if methylated at the} \\ & \left(i^{th}, (i+1)^{th}, \ldots, (i+\mathrm{m}-1)^{th}\right) \\ & \text{locus} \end{cases}$$

Note that we do not know from which DNA fragment (h) each read came from, only that all methylation loci in the read came from the same DNA fragment.

By summing over the reads containing the m-tuple, $(i, i+1, \ldots, i+\mathrm{m}-1)$, we obtain the number of reads containing each methylation pattern at that m-tuple. Here are the definitions for m = 2:

$$\begin{aligned} mm_{(i,i+1)} &= |\{z : z \in \mathcal{R}_{(i,i+1)}, z = (1,1)\}| \\ mu_{(i,i+1)} &= |\{z : z \in \mathcal{R}_{(i,i+1)}, z = (1,0)\}| \\ um_{(i,i+1)} &= |\{z : z \in \mathcal{R}_{(i,i+1)}, z = (0,1)\}| \\ uu_{(i,i+1)} &= |\{z : z \in \mathcal{R}_{(i,i+1)}, z = (0,0)\}| \end{aligned}$$

As we did for B-values, we can extend the β values to m-tuples. Here are the definitions for m = 2:

$$\beta^{mm}_{(i,i+1)} = \frac{mm_{(i,i+1)}}{d_{(i,i+1)}}$$
$$\beta^{mu}_{(i,i+1)} = \frac{mu_{(i,i+1)}}{d_{(i,i+1)}}$$
$$\beta^{um}_{(i,i+1)} = \frac{um_{(i,i+1)}}{d_{(i,i+1)}}$$
$$\beta^{uu}_{(i,i+1)} = \frac{uu_{(i,i+1)}}{d_{(i,i+1)}}$$

The definitions for m > 2 follow in the obvious manner.

One final definition is the average methylation level of each read, which is used by Landan *et al.* [2012]. For each read, $z \in \mathcal{R}_{(i,i+1,\ldots,i+\mathrm{m}-1)}$, the average methylation of the read, ζ_z, is defined as the proportion of methylation loci in the read that are methylated. Thus, $\zeta_z = 0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$.

4.1.3 Some complications

We now discuss some complications and how this framework might accommodate these issues in practice.

What is $\mathcal{I}$?

As mentioned in Chapter 2, studies using bisulfite-conversion assays rely on either a reference genome or, less commonly, separate DNA sequencing of the sample that is assayed. Different analysis strategies lead to different definitions of $\mathcal{I}$, which are approximations to the 'true' set of methylation loci in the sample, $\mathcal{I}^{truth}$. Listed here are definitions of $\mathcal{I}$ from least closely matching to most closely matching $\mathcal{I}^{truth}$:

1. $\mathcal{I}^{ref}$: Defined by the set of methylation loci in the reference genome. This ignore all genetic variation between the sample and the reference.
2. $\mathcal{I}^{refFilter}$: Defined by filtering out problematic loci from $\mathcal{I}^{ref}$. A conservative approach that removes many sites of genetic variation between the sample and the reference as well as sites that do not display genetic variation between the sample

and the reference. This approach cannot identify sample-specific methylation loci.

3. $\mathcal{I}^{Bis-SNP}$: Defined by calling genetic variants from the bisulfite-sequencing data using `Bis-SNP` [Liu *et al.* 2012]. Identifies sample-specific methylation loci and removes reference-specific methylation loci. This is the best approach if only bisulfite-sequencing data are available.
4. $\mathcal{I}^{WGS}$: Defined by identifying all methylation loci from *whole-genome sequencing* (WGS) of the sample's genome. The gold standard. All methylation loci are defined with respect to the sample's genome. The only differences between $\mathcal{I}^{WGS}$ and $\mathcal{I}^{truth}$ are due to sequencing errors, incomplete sequencing coverage of the sample's genome and variant calling errors.

Genetic heterozygosity at a methylation locus

The genome of a diploid organism has sites that are genetically heterozygous due to differences between the maternal and paternal chromosomes. Such heterozygous loci are sometimes also methylation loci; for example, a locus where the maternal chromosome is a CpG and the paternal chromosome is an ApG. In effect, the maternal and paternal chromosomes within the sample have different $\mathcal{I}^{truth}$.

The number of these genetically heterozygous methylation loci is often small enough not to worry about. However, in some studies, such as those of allele-specific methylation, these loci can be very important and should first be identified by calling heterozygous genetic variants using `Bis-SNP` or from whole-genome sequencing of the sample. In practice, the existence of such loci is often ignored.

4.2 Multiple samples

From a purely notational perspective, the move from a single sample to multiple samples simply requires an additional subscript, $j = 1, \ldots, N_{samples}$, where $N_{samples}$ is the number of samples. This defines the three levels in the hierarchy of a typical experiment: DNA fragments (h), methylation loci (i) and samples (j). For example, $Z_{h,i,j}$ is the methylation state on the h^{th} DNA fragment at the i^{th} methylation locus in the j^{th} sample and $\beta_{i,j}$ is the β-value for the i^{th} locus in the j^{th} sample.

A fourth level is how the samples relate to one another, such as the treatment group of each sample. This fourth level might be defined up-front, such as in a designed experiment looking for differences in methylation between tumour and normal tissue. Alternatively, the aim of the experiment might be to *discover* this grouping, such as in a clustering analysis.

We can define this fourth level using a design matrix $X = [X_j]$. For example, in a two-group experiment $X_j = 1$ if the sample is from group 1 and $X_j = 0$ if the sample is from group 2. We may also include covariates in the standard way by allowing X_j to be a row vector, $X_j = (x_{1,j}, \ldots, x_{P,j})$, where $x_{p,j}$ encodes the information on the p^{th} covariate for the j^{th} sample.

4.2.1 Some complications

In addition to the complications of the Section 4.1.3, we now have sample-to-sample variation that must be addressed within this framework.

What is $\mathcal{I}$?

Each sample has its own set of methylation loci, that is, $\mathcal{I}_j$ differs across j. Furthermore, sequencing coverage varies from sample-to-sample. This means that even if the samples have exactly the same $\mathcal{I}_j$, i.e. the samples are genetically identical, each sample will have a different set of loci with 'sufficient' sequencing coverage. Loci without sufficient sequencing coverage are effectively missing data.

In practice, we might choose to study $\mathcal{I}^{common} = \bigcap_j \mathcal{I}_j$ or some other suitably defined intersection of the $\mathcal{I}_j$, such as all methylation loci present in at least some fraction of the $N_{samples}$ samples.

A conservative analysis might only analyse those loci where at least some fraction of the samples have sufficient sequencing coverage. A less conservative analysis might try to impute the missing values based on methylation levels at neighbouring loci.

4.3 Parameter estimation

In this section we review various methods for estimating these key parameters. The main parameter of interest for each sample is the vector of methylation levels for each locus, $\boldsymbol{B_j} = (B_{1,j}, \ldots, B_{N_{loci},j})$. When necessary, I have 'translated' the original work into my notation to make these methods more readily comparable. I have suppressed j subscript when referring to a single sample.

4.3.1 Estimating M, U

It is rare to need direct estimates of M_i or U_i. In order to estimate M_i and U_i, the absolute number of methylated and unmethylated DNA fragments at the i^{th} locus, we would also require an estimate of the number of DNA fragments containing the i^{th} locus, H_i. Rather, we are generally interested in estimating the proportion of reads that are methylated, $B_i = \frac{M_i}{M_i+U_i}$, which does not require an estimate of H_i.

4.3.2 Estimating B

The simplest estimator of B_i is $\beta_i = \frac{m_i}{d_i}$, which has been widely used [e.g., Cokus *et al.* 2008, Lister *et al.* 2008, 2009, 2011]). The values of m_i and u_i are obtained by methylation calling and then counting the number of reads with each methylation state (see Section 2.3).

β_i is the maximum likelihood estimator of B_i under a (conditional) binomial model for the number of methylated reads at the i^{th} locus, $M_{i,j}|d_{i,j} \overset{d}{=} Binomial(d_{i,j}, B_{i,j})$.

More sophisticated methods have recently been proposed to estimate or model the average methylation level. These methods, which are still based on $m_{i,j}$ and $u_{i,j}$, include beta-binomial models [Feng *et al.* 2014, Sun *et al.* 2014, Dolzhenko and Smith 2014], and smoothing-methods [Hansen *et al.* 2011, 2012, Hebestreit *et al.* 2013].

Beta-binomial models

Several papers have proposed the beta-binomial distribution since, as noted by Dolzhenko and Smith [2014], it is "a natural model for describing methylation levels of an individual site across replicates" . The 'beta' component of the distribution models the underlying methylation level, $B_{i,j}$, while the 'binomial' component models the sampling of reads by sequencing. Another way to think of this is that the 'beta' component models the biological variability of the data, while the 'binomial' component models the sampling variability of sequencing. This separation of biological and technical variability has proven successful in detecting differential gene expression from RNA-seq data. For example, the `edgeR` software [Robinson *et al.* 2010] uses the negative binomial distribution, which can be thought of as a gamma-Poisson mixture distribution, to account for both the biological and sampling variability.

An attractive feature of the beta-binomial distribution is that it can be motivated by, and analysed with, Bayesian methods, including empirical Bayes methods, or frequentist techniques such as maximum likelihood. For example, the software `DSS` [Feng *et al.* 2014] and `MOABS` [Sun *et al.* 2014] both use the beta-binomial distribution in an empirical Bayes analysis of differential methylation from bisulfite-sequencing data. In contrast, `RADmeth` [Dolzhenko and Smith 2014] uses the beta-binomial model in a maximum likelihood framework to address the same problem.

Smoothing β-values

`BSmooth`, published in Hansen *et al.* [2011, 2012] and available in the R/Bioconductor package `bsseq`, and `BiSeq`, published in Hebestreit *et al.* [2013] and available in the R/Bioconductor package `BiSeq`, take a different approach to getting improved estimates of the $B_{i,j}$. Both `bsseq` and `BiSeq` use statistical smoothing of the 'raw' $\beta_{i,j} = \frac{m_{i,j}}{d_{i,j}}$. Smoothing is motivated and justified by the fact that the $B_{i,j}$ are spatially correlated within a sample. This phenomenon, called *co-methylation* is discussed and analysed in Chapters 6 and 7.

Smoothing is particularly powerful for loci with low sequencing coverage, where the denominator $m_{i,j} + u_{i,j}$ is small and the corresponding standard error of $\beta_{i,j}$ is large. The smoothed β-values, rather than the raw β-values, are then generally used in all downstream analyses.

Both `bsseq` and `BiSeq` use a binomial local likelihood smoother. In each case this smoother is chosen because BSmooth and BiSeq model the number of methylated reads at the i^{th} locus in the j^{th} sample by $M_{i,j}|d_{i,j} \overset{d}{=} Binomial(d_{i,j}, B_{i,j})$. The smoothing is 'local' to leverage co-methylation, which is considered a local phenomenon.

In both bsseq and BiSeq the raw β-values are weighted according to the binomial likelihood and a kernel function. The binomial likelihood weights β_i inversely to their standard error, $se(\beta_i)$, and the kernel gives greater weight to those β_i near the centre of the window. Lacey *et al.* [2013] note that loci with very high sequencing coverage will strongly influence the smoother, potentially biasing estimates at neighbouring loci with lower coverage.

bsseq assumes that for each sample that the underlying methylation level, $B_{i,j}$, is a smoothly varying function of the position in the genome, i. In contrast, BiSeq first creates clusters of CpGs and only assumes that the underlying methylation level is smooth at positions within each cluster.

Whenever smoothing is used, a key parameter is the bandwidth, which is the size of the window in which observations are included at each iteration of the smoother. bsseq uses a much larger window size than BiSeq; the default window size in bsseq is one that contains at least 70 CpGs and is at least 2000kb wide, whereas the default window size in BiSeq is 80bp, regardless of CpG-density. This is due to BiSeq being developed for RRBS data, which has a high CpG-density per window, whereas bsseq was developed for whole-genome data, which has a more variable, and lower on average, CpG density per window.

Another 'parameter' choice when smoothing is the choice of kernel, although this is generally less important than the choice of bandwidth. bsseq uses a tricube kernel and BiSeq uses a triangular kernel.

Hebestreit *et al.* [2013] and Lacey *et al.* [2013] compare the smoothing results of BiSeq to bsseq. Both Hebestreit *et al.* [2013] and Lacey *et al.* [2013] provide instances where they claim BiSeq gives more 'reasonable' smoothed values than bsseq. However, these comparison studies use RRBS data, which bsseq is not designed for‘, and their simulation so will favour methods designed for RRBS data[4].

[4] Both Hebestreit *et al.* [2013] and Lacey *et al.* [2013] altered the default bsseq parameters to try to make them comparable to BiSeq. Hebestreit *et al.* [2013] changed the default minimum window size to 80

4.4 Statistical properties of β-values

Under the (conditional) binomial model, $M_{i,j}|d_{i,j} \overset{d}{=} Binomial(d_{i,j}, B_{i,j})$, $\beta_{i,j} = \frac{m_{i,j}}{d_{i,j}}$ is an unbiased estimator of $B_{i,j}$ with standard error $se(\beta_{i,j}) = \sqrt{\frac{\beta_{i,j}(1-\beta_{i,j})}{d_{i,j}}}$ [Hansen *et al.* 2012]. The natural interpretation of $\beta_{i,j}$ is then as an estimator of the average level of methylation at the i^{th} locus in the j^{th} sample. In this section I discuss this interpretation and statistical properties of this estimator.

4.4.1 Empirical distributions of β-values

In a study involving multiple samples, the set of β-values can be summarised as a matrix where each row is a locus and each column is a sample. Some values will be missing, either because there was insufficient sequencing coverage to estimate a β-value or because that locus is not a cytosine for the sample in question. This matrix might be visualised to learn about the distribution of methylation levels, either row-wise (to learn about the variability across samples) or column-wise (to learn about the variability within samples).

Genome-wide distribution of β-values

Restricting our attention to CpGs, Figures 4.2, 4.3, 4.4 and 4.5 show the kernel density estimates of the genome-wide distributions of β-values, that is, the column-wise summaries, for each sample of the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets, respectively. Figures 4.6, 4.7, 4.8 and 4.9 show the same data but with the β-values grouped into 0.01-width bins and plotted against the percentages of CpGs that fall into each bin.

What is immediately clear is that these distributions are bimodal: most CpGs are highly methylated or lowly methylated. The exception is the *E16.5_male_1* sample from the *Seisenberger* data which is hypomethylated and with an enormous number of intermediately methylated CpGs. The *E16.5_male_1* sample is a progenitor germ cells from a pool of embryonic day 16.5 male mice. Between days E6.5 and E13.5, the mouse progenitor germ cells undergo global demethylation and it is only from day E16.5 onwards that they begin

bp but still required at least 20 CpGs per window. Lacey *et al.* [2013] kept the default minimum window size of $2,000$ bp but reduced the minimum number of CpGs per window to 50 from the default of 70. Nevertheless, the fact remains that `bsseq` is designed for analysing whole-genome bisulfite-sequencing data and not RRBS, which puts it at a disadvantage in these comparisons.

to be *de novo* methylated [Seisenberger *et al.* 2012], hence the wide variation in β-values.

Almost all the samples with significant intermediate methylation are either somatic cell lines (*ADS*, *ADS-adipose*, *FF*, *IMR90_r1* and *IMR90_r2* from the *Lister* dataset; *IMR90_cell_line* from the *Ziller* dataset) or cancer cells lines and tissue (*HepG2_cell_line*, *Colon_Tumor_Primary* and *Colon_Primary_Normal* from the *Ziller* dataset). Aside from the aforementioned *E16.5_male_1*, the *E6.5_epiblast_1* sample from the *Seisenberger* dataset also displays greater levels of intermediate methylation. This sample was also created by pooling DNA from multiple mice, which may explain the extra variability in the β-value distribution.

It has previously been observed that cancer samples have highly variable DNA methylation [Hansen *et al.* 2011], which, combined with the possibility of multiple sub-clones, explains these intermediate β-values in the cancer samples.

The explanation for the somatic cell lines is less clear. Notably, all of the 12 EPISCOPE samples, which are tissue samples rather than cell lines, have relatively low levels of intermediate methylation. Likewise, the various frontal cortex samples in the *Ziller* dataset, which includes both 'normal' and Alzheimer's samples (*Frontal_cortex_normal_1*, *Frontal_cortex_normal_2*, *Frontal_cortex_AD_1* and *Frontal_cortex_AD_2*), have very low levels of intermediate methylation. This raises the question as to whether the widespread partial methylation observed in the somatic samples from the *Lister* dataset is in fact a feature of somatic cell lines rather than somatic cells *per se*. Naively, a cell line is a 'pure' cell population, however, the DNA methylation data clearly reveal widespread cellular heterogeneity of DNA methylation.

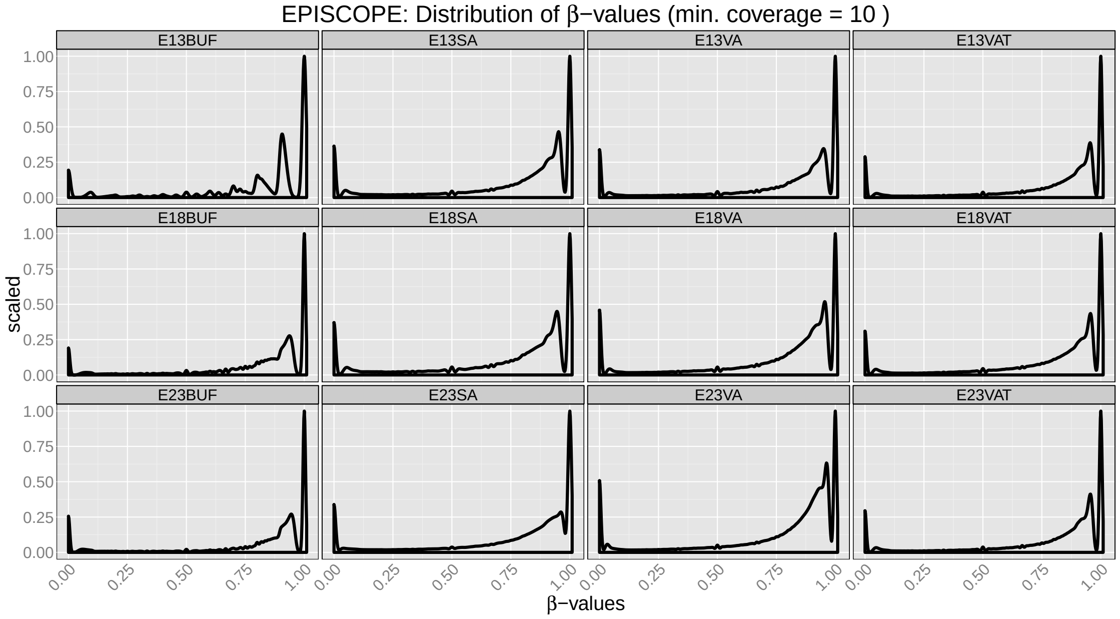


Figure 4.2: Kernel density estimates of the genome-wide distribution of CpG β-values for the *EPISCOPE* data. Densities are normalised so that the maximum value for each sample is 1. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

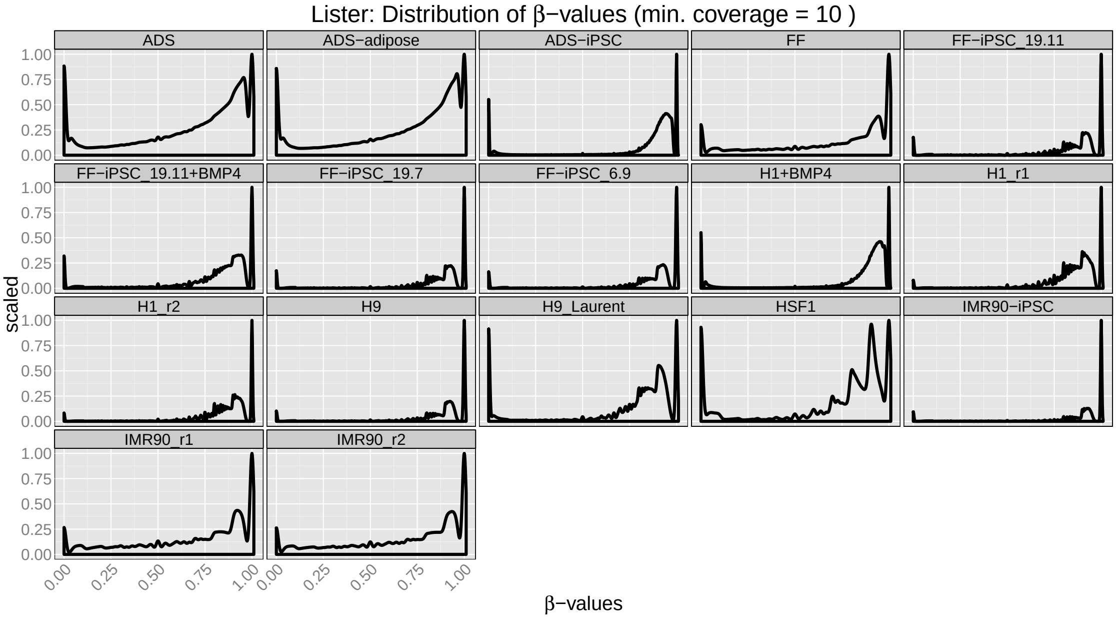


Figure 4.3: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Lister* data. Densities are normalised so that the maximum value for each sample is 1. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

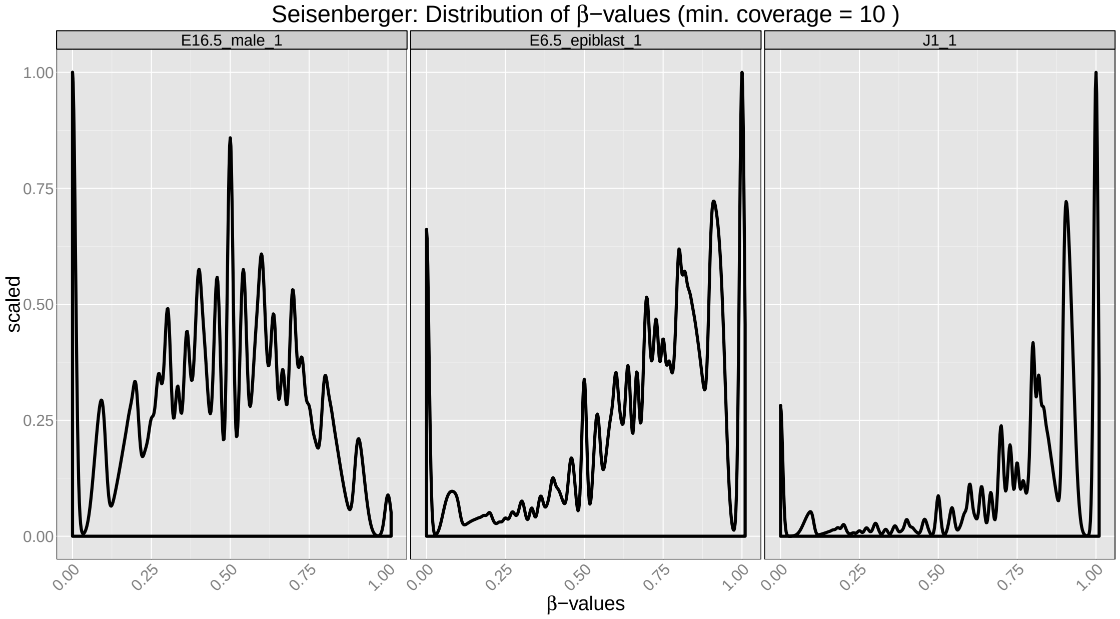


Figure 4.4: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Seisenberger* data. Densities are normalised so that the maximum value for each sample is 1. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

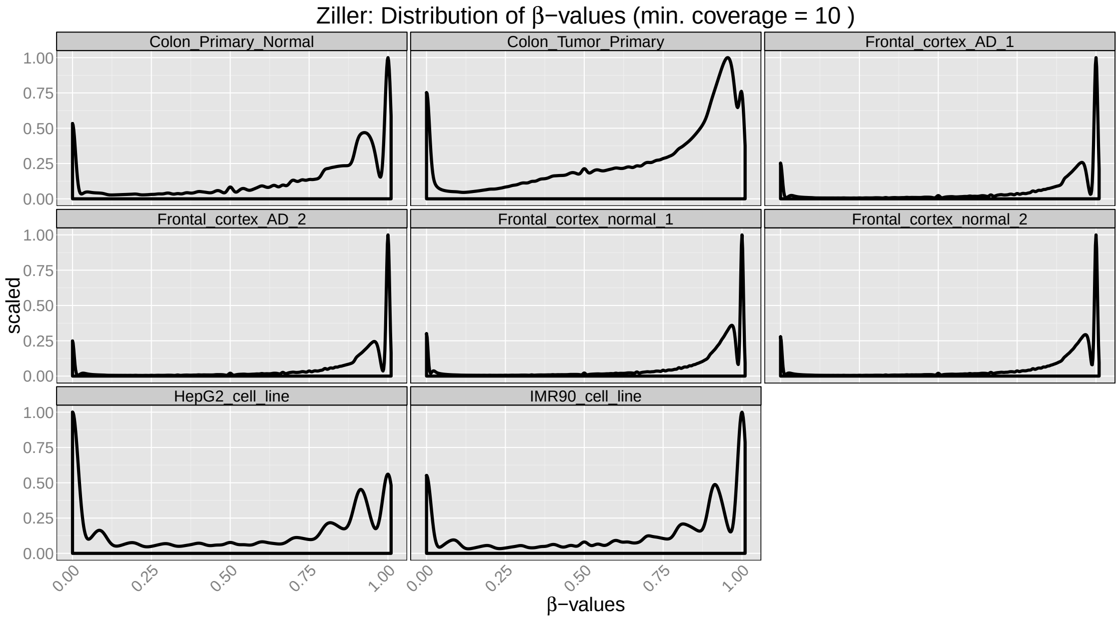


Figure 4.5: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Ziller* data. Densities are normalised so that the maximum value for each sample is 1. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included. ‘Spikes’ in the density estimate are due to the discreteness of β-values.

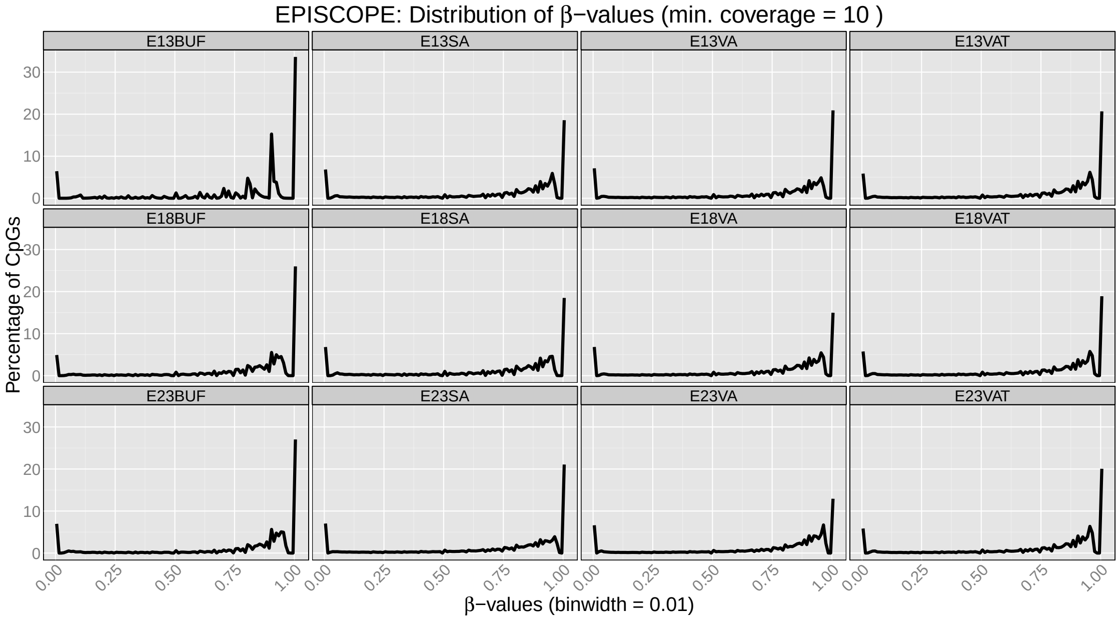


Figure 4.6: Frequency polygon of the genome-wide distribution of CpG β-values for the *EPISCOPE* data. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included.

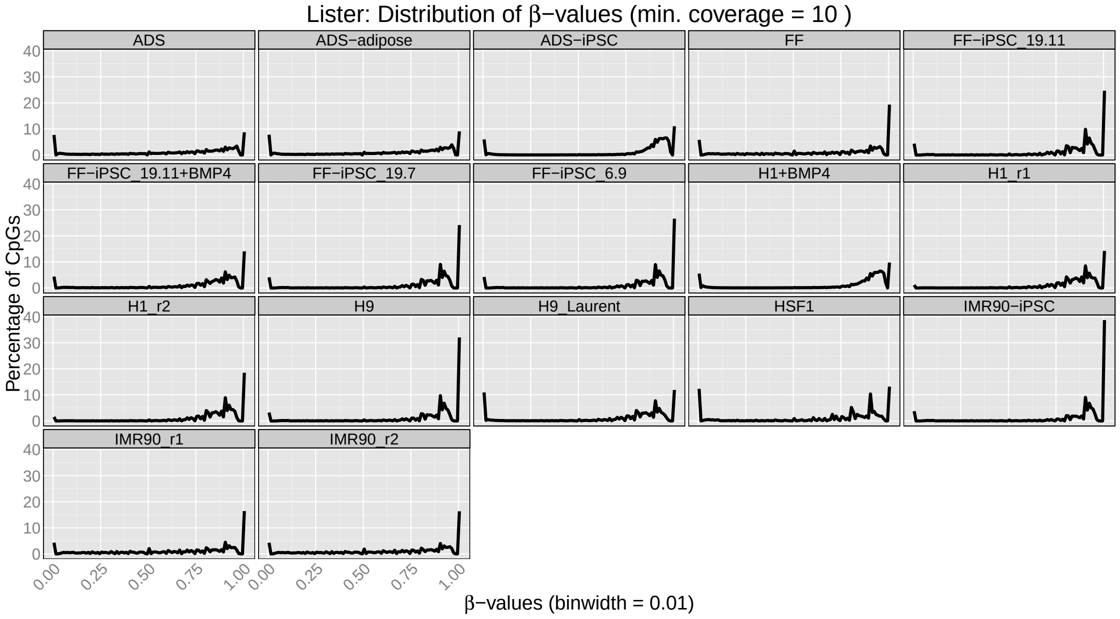


Figure 4.7: Frequency polygon of the genome-wide distribution of CpG β-values for the *Lister* data. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included.

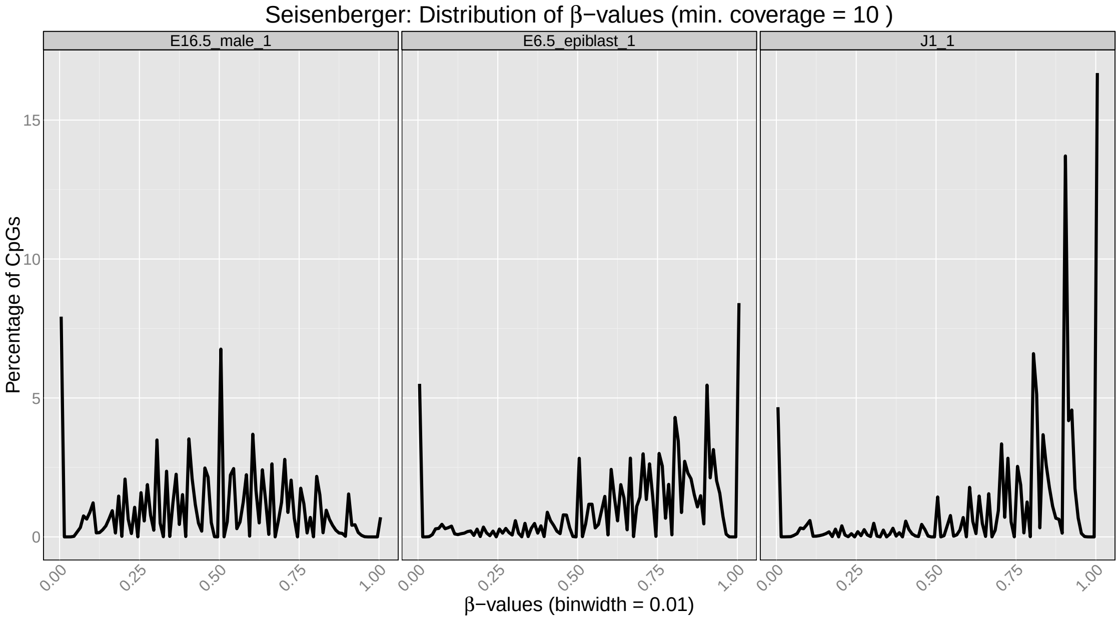


Figure 4.8: Frequency polygon of the genome-wide distribution of CpG β-values for the *Seisenberger* data. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included.

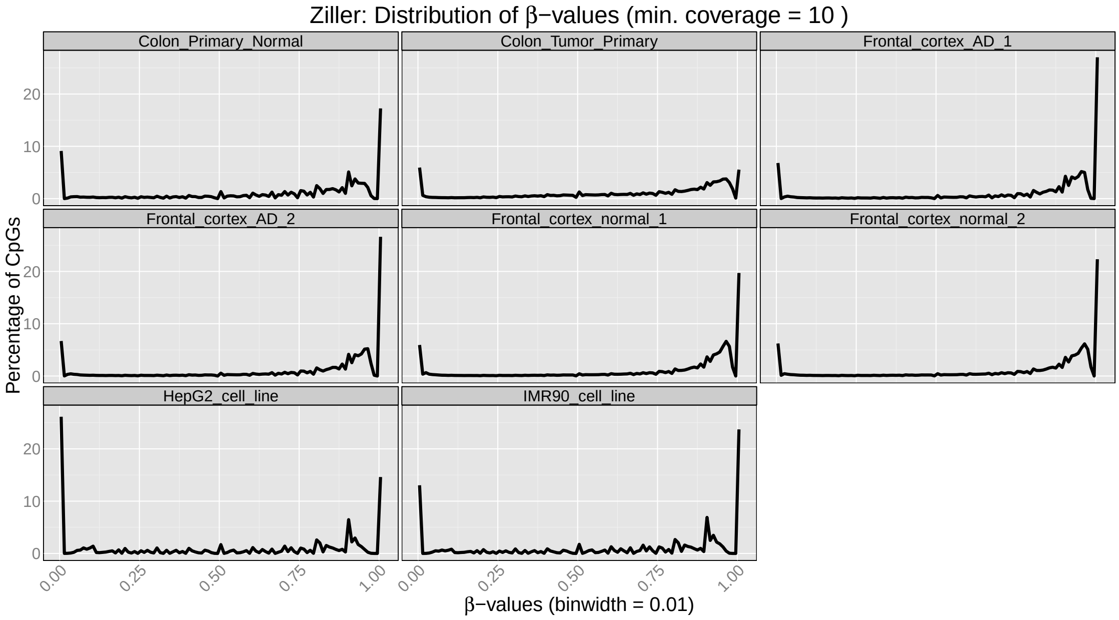


Figure 4.9: Frequency polygon of the genome-wide distribution of CpG β-values for the *Ziller* data. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Observations have been combined across strands and only CpGs with at least 10× sequencing coverage are included.

The bimodality of the genome-wide β-value distributions is driven by the fact that most CpGs in CpG islands are unmethylated whereas those outside of the CpG islands are mostly methylated. Figures 4.10, 4.11, 4.12 and 4.13 show the kernel density plots of the β-value distributions stratified by CpG island status for the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets, respectively. These distributions are normalised so that each density has a maximum value of 1.

These plots show that CpG islands have a more strictly bimodal distribution than do the non-islands. While the majority of CpGs in CpG islands are unmethylated, there are a subset of methylated CpGs in CpG islands in each sample (except for the *E16.5_male_1* sample). The *H1_r1* and *H1_r2* samples, replicates of an embryonic stem cell line, stand out for having CpGs in CpG islands being more methylated than unmethylated. These plots also show that most of the intermediate methylation occurs outside of the CpG islands.

Because these densities are normalised, these plots don't show the proportion of CpGs in CpG islands. Therefore, Figures 4.14, 4.15, 4.16 and 4.17 show the same data but with β-values grouped into 0.01-width bins and plotted against the percentage of total CpGs in each bin. These plots highlight that the majority of unmethylated CpGs occur in CpG islands and that most of the intermediate methylation occurs outside of CpG islands, owing to most CpGs being outside a CpG island.

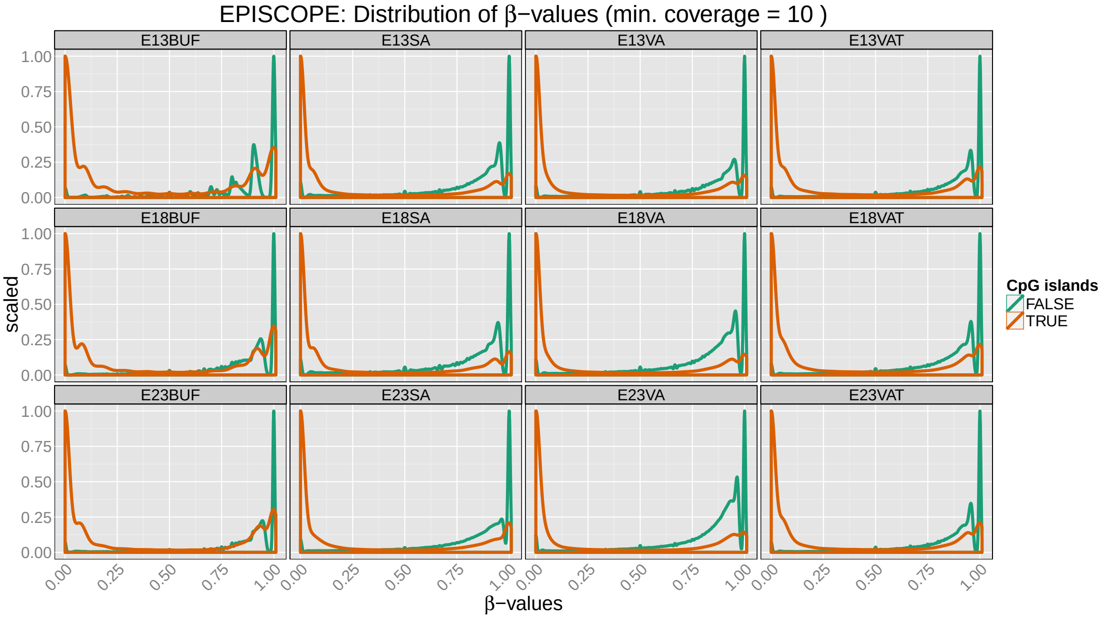


Figure 4.10: Kernel density estimates of the genome-wide distribution of CpG β-values for the *EPISCOPE* data, stratified by whether the CpG is in a CpG island. Only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

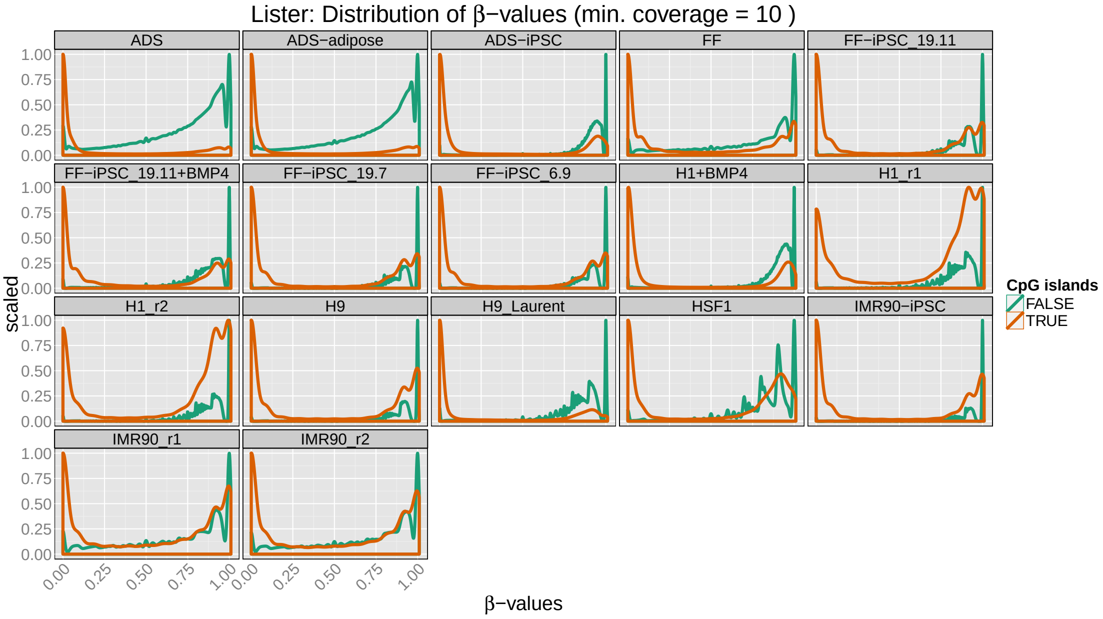


Figure 4.11: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Lister* data, stratified by whether the CpG is in a CpG island. Only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

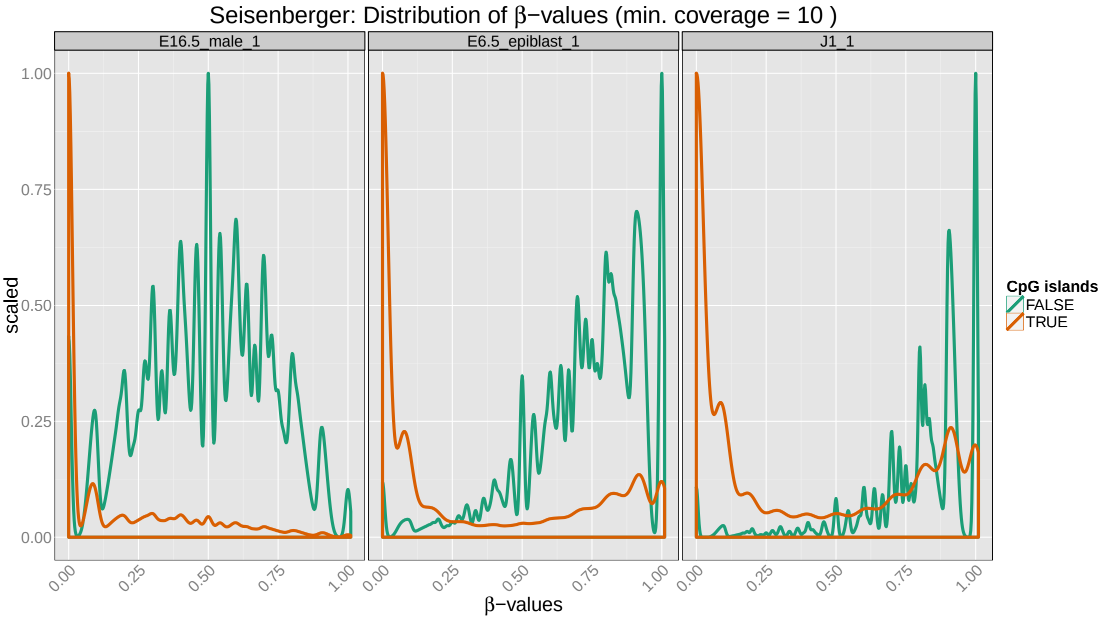


Figure 4.12: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Seisenberger* data, stratified by whether the CpG is in a CpG island. Only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

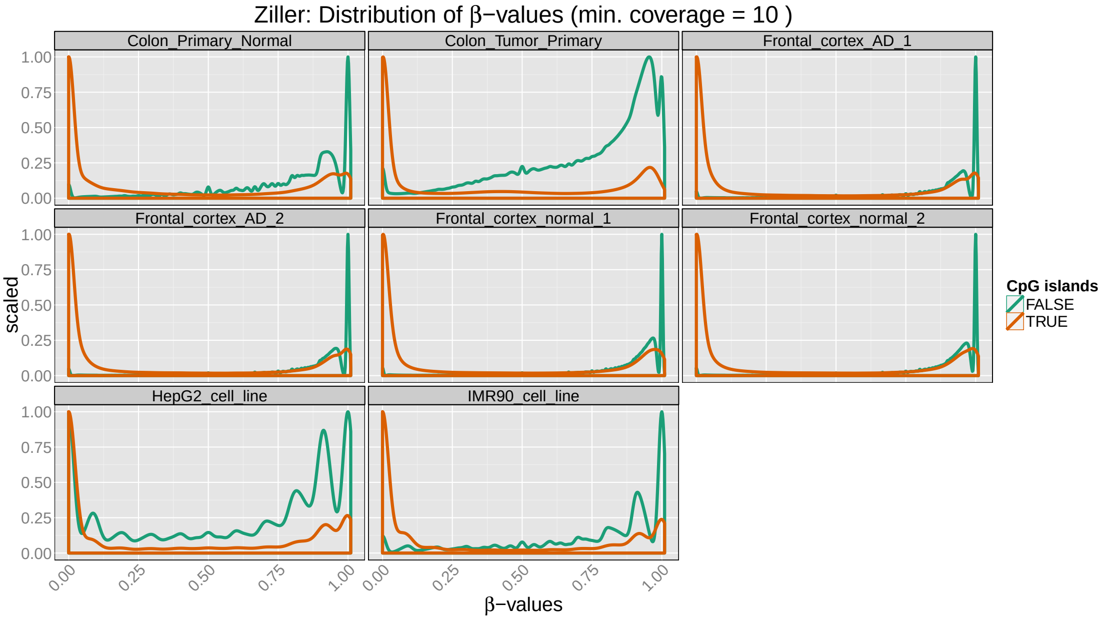


Figure 4.13: Kernel density estimates of the genome-wide distribution of CpG β-values for the *Ziller* data, stratified by whether the CpG is in a CpG island. Only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

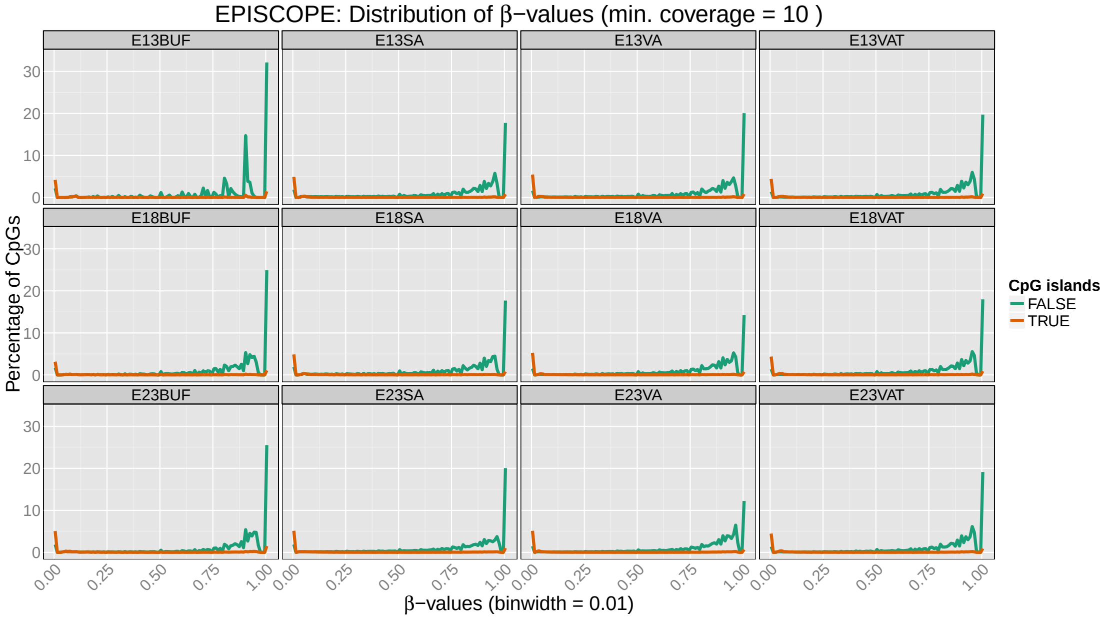


Figure 4.14: Frequency polygon of the genome-wide distribution of CpG β-values for the *EPISCOPE* data, stratified by whether the CpG is in a CpG island. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Only CpGs with at least 10× sequencing coverage are included. Percentages are with respect to all CpGs with at least 10× sequencing coverage, unstratified by CpG island status.

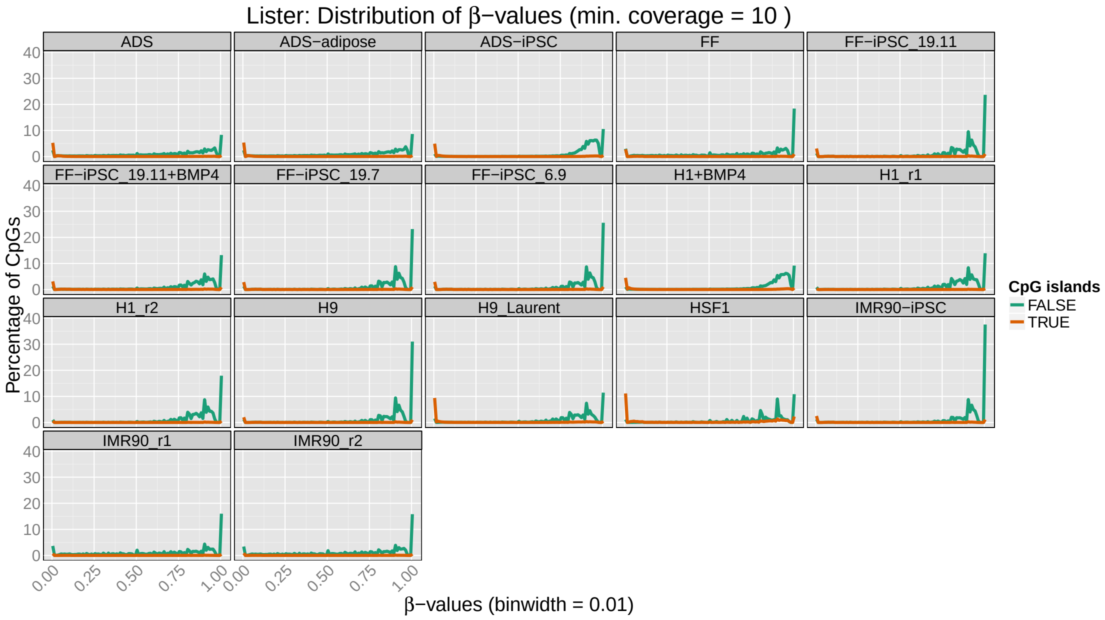


Figure 4.15: Frequency polygon of the genome-wide distribution of CpG β-values for the *Lister* data, stratified by whether the CpG is in a CpG island. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Only CpGs with at least 10× sequencing coverage are included. Percentages are with respect to all CpGs with at least 10× sequencing coverage, unstratified by CpG island status.

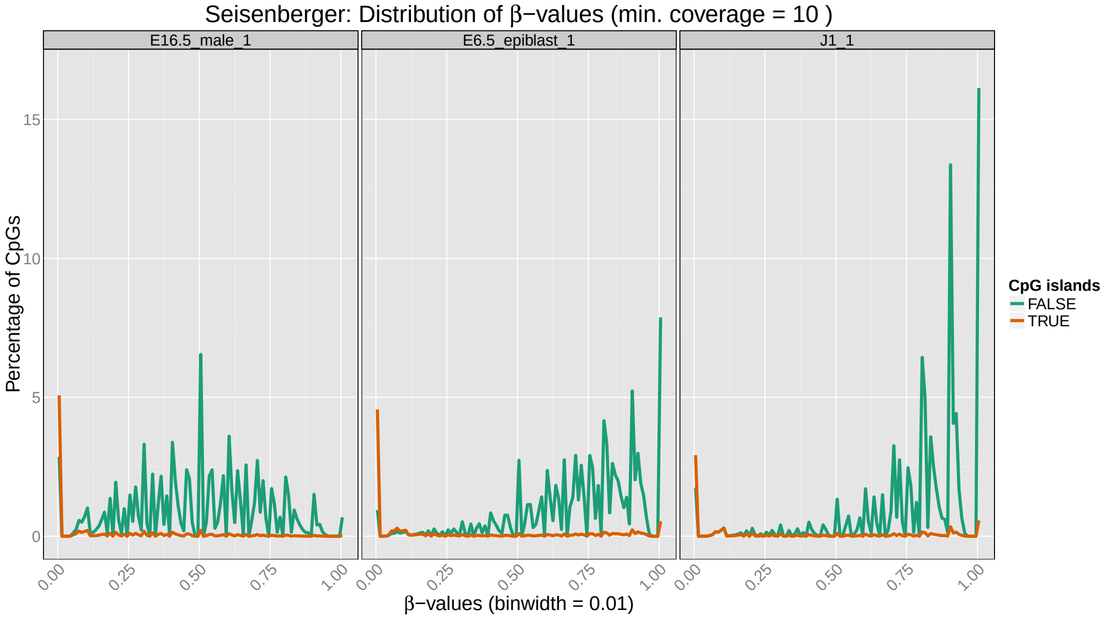


Figure 4.16: Frequency polygon of the genome-wide distribution of CpG β-values for the *Seisenberger* data, stratified by whether the CpG is in a CpG island. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Only CpGs with at least 10× sequencing coverage are included. Percentages are with respect to all CpGs with at least 10× sequencing coverage, unstratified by CpG island status.

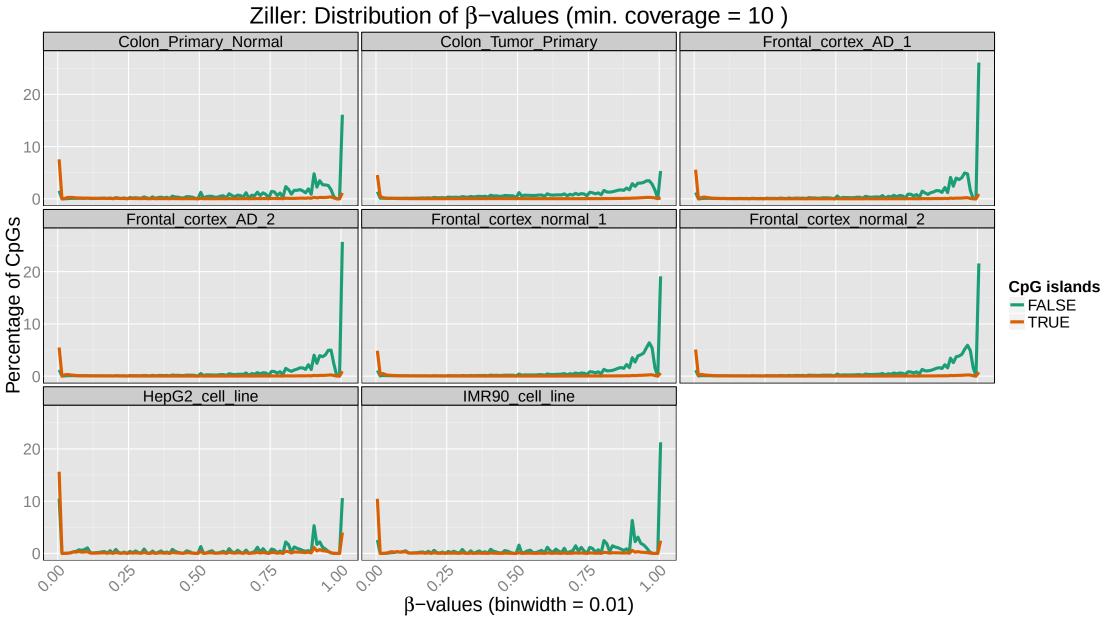


Figure 4.17: Frequency polygon of the genome-wide distribution of CpG β-values for the *Ziller* data, stratified by whether the CpG is in a CpG island. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Only CpGs with at least 10× sequencing coverage are included. Percentages are with respect to all CpGs with at least 10× sequencing coverage, unstratified by CpG island status.

Strand-specific β-values

CpG β-values are often computed by aggregating the m and u counts across the forward and reverse strands. On average, this doubles the sequencing coverage of each CpG but presupposes that the two strands are indeed symmetrically methylated. To investigate the validity of this assumption we can compute the correlation of strand-specific β-values. Figures 4.18, 4.19, 4.20 and 4.21 report the Pearson correlation of these strand-specific β-values for varying sequencing coverage cutoffs.

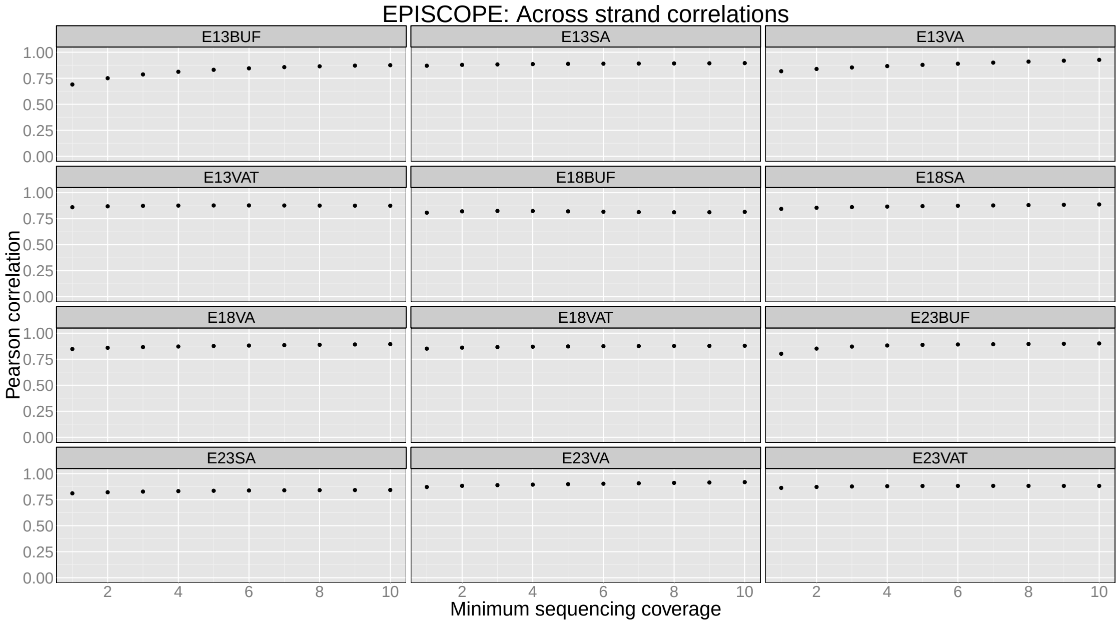


Figure 4.18: Correlations of β-values across strands for the *EPISCOPE* dataset using different minimum sequencing coverage cutoffs.

There is a considerable amount of noise in the estimates of β-values when using low sequencing coverage, as can be seen from the smaller strand-correlations at these lower cutoffs. Once we require a minimum sequencing coverage of 5×, we see that most samples have a very high correlation of β-values across strands, $r = 0.8$ to 0.9, with some notable exceptions.

Some of the embryonic stem cell samples (*H1_r1*, *H1_r2* and *HSF1* from the *Lister* dataset) have less correlated strand-specific β-values, $r = 0.5$ to 0.7. The other embryonic stem cell samples have higher correlations of β-values across strands, although the estimates are quite different between the two replicates of the same cell line (*H9*: $r = 0.77$,

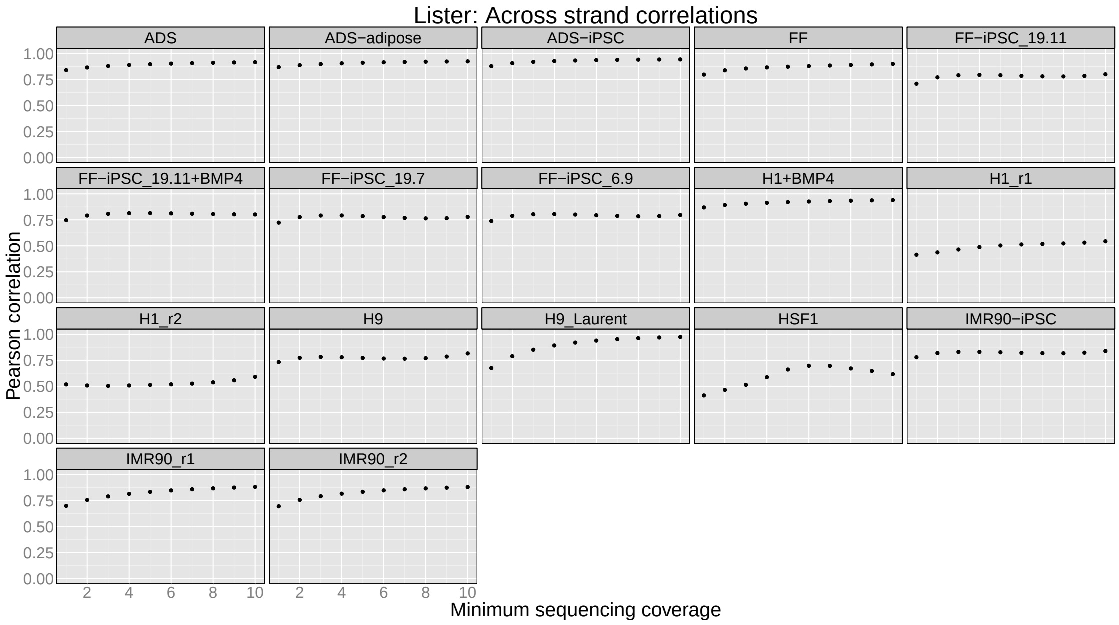


Figure 4.19: Correlations of β-values across strands for the *Lister* dataset using different minimum sequencing coverage cutoffs.

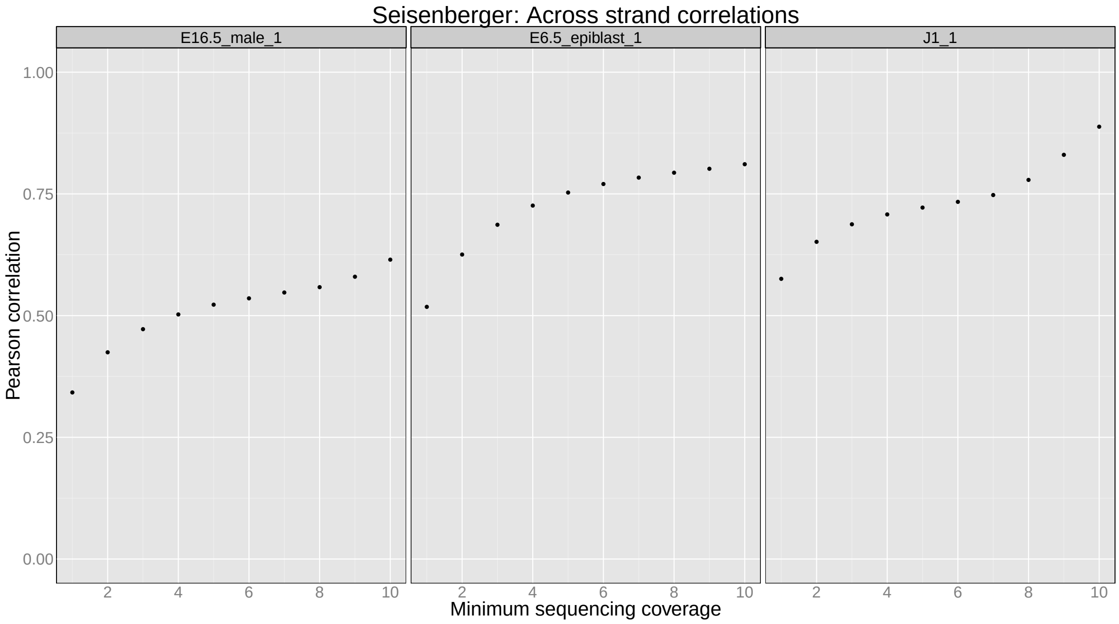


Figure 4.20: Correlations of β-values across strands for the *Seisenberger* dataset using different minimum sequencing coverage cutoffs.

H9_Laurent: $r = 0.92$). This suggests that caution may be warranted in combining CpG methylation levels across strands for embryonic stem cell samples.

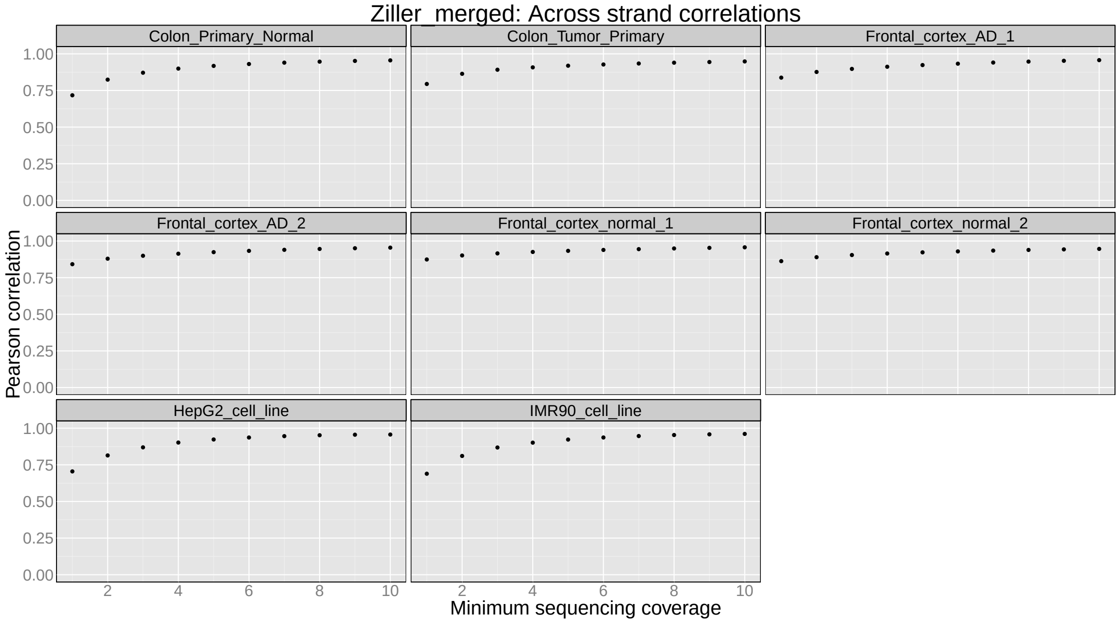


Figure 4.21: Correlations of β-values across strands for the *Ziller* dataset using different minimum sequencing coverage cutoffs.

All three *Seisenberger* samples have noticeably less correlated strand-specific β-values, including the embryonic stem cell, *J1_1*. However, since these data are from pooled DNA, the source of this reduced correlation is difficult to identify.

Overall, with the exception of embryonic stem cells, it seems that most samples have highly correlated strand-specific CpG β-values, which means that these data can safely be combined across strands. However, it remains a good idea to first check this assumption prior to combining data across strands.

4.4.2 Interpretation of β-values

Laird [2003] says in a review paper on DNA methylation that, "about 70% of the CpG dinucleotides in the mammalian genome are methylated". Similar statements are made in many papers about DNA methylation, but how should these be interpreted?

In the context of a whole-genome bisulfite-sequencing experiment, this can be interpreted as the expected β-value of a randomly selected CpG. However, as can be seen from Figures 4.6, 4.7, 4.8 and 4.9, the bimodality of the β-value distributions means that the expected value is not a particularly useful estimate of the methylation level of a particular CpG. To make a useful statement about the methylation level of a particular CpG really requires more information, such as whether it is within a CpG island.

The "70%" statement can also be interpreted as an estimate of the probability that a cytosine randomly sampled from a haploid copy of a mammalian genome is a methylcytosine. Note that this refers to individual cytosines, $Z_{h,i}$, and not the genomic position of the locus, pos_i. Also recall that most assays measure DNA methylation from a pool of cells, not a single haploid copy of the genome. The methylation state at the i^{th} locus may vary across the H_i DNA fragments. Therefore, I do not think it makes sense to describe a locus, i, as being a 'methylcytosine'. However, several important whole-genome bisulfite-sequencing papers, Lister *et al.* [2008, 2009, 2011][5], have used this latter definition, which I believe to be an unnecessary source of confusion.

Lister *et al.* introduced a method "to identify the presence of a methylated cytosine" [Lister *et al.* 2008, Supplementary Material]. In the language of these papers, a "methylcytosine" is a cytosine in the reference genome where "at least s [sic; I believe this should be "a"] subset of the genomes within the sample were methylated" [Lister *et al.* 2009, Supplementary Material]. This amounts to testing the hypothesis $H_0 : \beta_{i,j} = 0$ against the one-sided alternative $H_1 : \beta_{i,j} > 0$. This can be thought of as testing the null hypothesis that the observed number of methylated reads at the i^{th} cytosine were simply due to 'error', where the 'error' is a combination of the estimated sequencing error and the estimated bisulfite-converstion error.

Although the exact procedure is not particularly well described in any of Lister *et al.* [2008, 2009, 2011], nor is any code made available, I believe the method is as follows[6]. For each cytosine they compute the probability of observing more than m_i methylated reads by chance, $P_i = \sum_{k=m_i+1}^{k=d_i} Pr(X = k)$, where $X = Binom(d_i, \epsilon)$ and ϵ is the estimated 'error'. Any site with an FDR-adjusted P-value below a threshold was declared a "methylcytosine"[7]

[5] It is worth noting that this concept was not used in more recent paper from the same group [Lister *et al.* 2013].

[6] The earliest of these papers, Lister *et al.* [2008], includes a short non-mathematical description, while the most detailed description is given in the supplementary material of Lister *et al.* [2009]. Lister *et al.* [2011] simply refers to Lister *et al.* [2009]

[7] Lister *et al.* [2008] used an FDR-adjusted P-value cutoff of 0.05; Lister *et al.* [2009] used an FDR-adjusted P-value cutoff of 0.01. I presume the FDR-adjustment to be based on the Benjamini-Hochberg procedure [Benjamini and Hochberg 1995], although this is not explicitly stated. This procedure was performed separately for each methylation context in Lister *et al.* [2009], but it is not clear if this is the case for Lister *et al.* [2008] (a study of *Arabidopsis thaliani*, which has large amounts of non-CG methylation)

The ϵ are estimated on a per-sample basis, with bisulfite-conversion error rates estimated from the unmethylated chloroplast genome [Lister *et al.* 2008] or from the genome of the lambda phage spike-in control [Lister *et al.* 2009, 2011]. It is not clear how the sequencing error rates were estimated, particularly since the base qualities are not included in the data available from the website.

The proportion of cytosines that are identified by this procedure as "methylcytosines" is a poor estimator of the probability that a cytosine randomly sampled from a haploid genome is methylated, unless the sample is incredibly homogeneous. For example, suppose we had a sample where the true methylation level of every CpG, B_i, was 0.2. Given sufficient sequencing coverage, every CpG in the genome would be declared a "methylcytosine" when in fact for any haploid copy of the genome only 20% of CpGs would be expected to be methylcytosines.

Furthermore, referring to CpGs as "methylcytosines" results in a loss of information since two "methylcytosines" may have very different β-values. For example, in Lister *et al.* [Supplementary Figure 2a of 2009] the authors use a Venn diagram to compare the number of "methylcytosines" called in two biological replicates to summarise the concordance between the two biological replicates. A far better summary of the biological replicability is to plot the β-values from each replicate against one another as a scatter plot, as this includes the magnitude of the β-values and not just whether they are statistically different from zero.

In summary, for bisulfite-sequencing experiments where the DNA for each sample comes from multiple cells I do not think it makes sense, nor is it useful, to refer to individual cytosines, i, as being methylated or unmethylated. Instead, it is better to summarise the methylation level at a CpG by a β-value since this has a natural interpretation as the estimated proportion of haploid genomes in the sample that are methylated at that CpG. Unfortunately, β-values are not without their own issues, as discussed in Section 4.4.3.

or Lister *et al.* [2011] (a study that includes pluripotent human cell lines that have non-negligible levels of non-CG methylation). This affects the false discovery rate correction since the number of tests is far greater if all cytosines are simultaneously corrected compared to a separate correction for each context.

4.4.3 Sources of bias in β-values

The natural interpretation of β_i is the average level of methylation at the i^{th} locus. However, this will be biased if the probability of sequencing a fragment with a methylated site is different from the probability of sequencing a fragment with an unmethylated site. In fact, it has been shown that methylated DNA is overrepresented in bisulfite-sequencing data due, with the problem exacerbated by higher rounds of PCR amplification and dependent on the bisulfite-conversion protocol [Ji *et al.* 2014]. PCR amplification can result in overreprestation of one of the DNA strands in bisulfite-sequencing data [Warnecke *et al.* 1997].

Lab-based solutions to overcome these biases exist for targeted bisulfite-sequencing, but are technically difficult and their cost prohibits their extension to whole-genome bisulfite-sequencing [Ji *et al.* 2014]. Computational correction for these biases have been proposed [Ji *et al.* 2014], but as yet these have not been implemented in any available software.

4.4.4 Transformations of β-values

β-values are the *de facto* standard unit for reporting methylation levels due to their natural interpretation as an estimate of the average level of methylation at the locus. However, they are not necessarily the best unit for statistical inference. This is because a β-value is an estimate of a proportion and there are a well-known statistical challenges when working with proportion data, including:

1. The estimate of the standard error depends on the estimate of the mean (i.e. β), through $se(\beta) = \sqrt{\frac{\beta(1-\beta)}{d}}$. Taking the derivative of this with respect to β, we see that the maximum standard error, $\sqrt{\frac{0.25}{d}}$, occurs at $\beta = 0.5$ and the minimum standard error, 0, occurs at $\beta = 0, 1$.
2. We need to know more than just the β-value to have a sense of how precise an estimate it is. Essentially, we need to also know the sequencing coverage of the methylation loci. Consider two CpGs, one with $m = 1, u = 3$ and the other with $m = 100, u = 300$. Both CpGs have $\beta = 1/4$ but the second CpG is measured with much greater precision. Assuming the binomial model, the first CpG has $se(\beta) = \sqrt{\frac{1/4 \times 3/4}{4}} = 0.22$ whereas the second CpG has $se(\beta) = \sqrt{\frac{1/4 \times 3/4}{400}} = 0.02$.

3. Proportions are bound between 0 and 1, inclusive.

To address (1), proportion data are often transformed via a variance stabilisation transformation. The aim is to make the variance independent of the mean, at least approximately. Popular variance stabilisation transformations include:

- The arcsine transformation, $\arcsin(\sqrt{\frac{m+1}{m+u+1}})$ [Anscombe 1948]. A small value, in this case 1, is added to both m and u to avoid $\beta = 0, 1$.
- The "averaged arcsine" transformation, $\arcsin\sqrt{\frac{m}{m+u+1}} + \arcsin\sqrt{\frac{m+1}{m+u+1}}$ [Freeman and Tukey 1950]. One problem with this transformation is that it does not have a unique inverse [Nunes and Nason 2009].

However, the use of variance stabilising transformations for proportion data has fallen out of favour with the widespread availability of generalised linear model software, in particular for the logistic regression model [Warton and Hui 2011].

One transformation that remains popular, at least in the analysis of DNA methylation microarray data, is the logit-transformation, also known as $\mathcal{M}$-values. An $\mathcal{M}$-value is defined as $logit_2(\beta) = \log_2\left(\frac{\beta}{1-\beta}\right) = \log_2\left(\frac{m+\alpha}{u+\alpha}\right)$, where here m and u are the intensities from the methylated and unmethylated probes, respectively, and α is an offset to avoid a numerator or denominator that is zero. $\mathcal{M}$-values are also known as log-ratios and are widely used in the analysis of RNA expression two-colour microarrays [e.g., Smyth 2005].

Du *et al.* [2010] advocate for the use of $\mathcal{M}$-values for conducting differential methylation analysis from microarray data[8]. The main reason they advocate for the use of $\mathcal{M}$-values is that they are approximately *homoscedastic*, i.e. their variances are approximately constant across the full range of $\mathcal{M}$-values. As already noted, the logit-transformation is not the only possible variance-stabilising transformation, but the familiarity of log-ratios to bioinformaticians and genomics researchers makes it a favourable choice. As with β-values, the $\mathcal{M}$-values derived from bisulfite-sequencing data cannot generally be directly analysed due to the variable sequencing coverage across loci.

[8] Du *et al.* [2010] also recommend that the results of analyses are reported as β-values owing to their "more intuitive biological interpretation".

4.4.5 Spatial correlations of β-values

Many researchers have observed that DNA methylation is spatially correlated along the genome, [e.g., Eckhardt *et al.* 2006, Cokus *et al.* 2008, Li *et al.* 2010, Hansen *et al.* 2011, Hebestreit *et al.* 2013, Wang *et al.* 2011, Pedersen *et al.* 2012, Lacey *et al.* 2013, Sofer *et al.* 2013, Liu *et al.* 2014, Lyko *et al.* 2010, Landan *et al.* 2012, Lister *et al.* 2009]. I call this spatial correlation of methylation levels *co-methylation.*

I examine in detail the spatial correlations of β-values in Chapters 6 and 7.

4.5 Summary

This chapter has defined a mathematical framework for describing data from whole-genome bisulfite-sequencing data. It has addressed some subtleties and complications that arise due to within-sample and between-sample differences in where DNA methylation is measured. By using a common statistical framework we can better understand how different statistical methods relate to one another. Using this framework, we described common estimators of DNA methylation levels and some of their statistical properties. We then examined the empirical distributions of these variables across a diverse set of 40 whole-genome bisulfite-sequencing samples.

From these analyses we have seen that the CpG islands drive the strong bimodal distribution of β-values that is observed in almost all samples. We have also observed that most intermediate methylation occurs outside of the CpG islands. The most distinct methylomes are the hypermethylated embryonic stem cells (*H1_r1* and *H1_r2*) and the hypomethylated cancer cell lines (*HepG2_cell_line*). The Seisenberger samples also stand out, particularly the hypervariable progenitor germ cells (*E16.5_male_1*). The genome-level results for the *Seisenberger* samples are more difficult to interpret, however, since they are from pooled DNA. Nonetheless, some of the increased variability in the *Seisenberger* data will also reflect that these samples are from developmental timepoints during which DNA methylation is very dynamic.

In contrast, the somatic samples, particularly those from tissue samples rather than cell lines, have very 'well-behaved' β-value distributions that are globally similar between samples. The may reflect that the DNA methylome is well-established in these samples and relatively static. The increased level of partial methylation in somatic cell lines compared to somatic tissue samples may be attributable to the development of sub-clones during the culturing of the sample. This is consistent with the high epipolymorphism observed in a study that tracked the dynamics of DNA methylation in an *in vitro* evolutionary cell culture system [Landan *et al.* 2012]. Somatic samples also have very highly correlated CpG β-values across strands, meaning that these data can generally be combined across strands to increase the sequencing coverage of each CpG.

Induced pluripotent stem cell lines also appear to have a quite strictly regulated methylome, with little intermediate methylation. This is likely a consequence of the fact that during the induction of pluripotency, the methylome of the sample is 'reset' [Lister *et al.* 2011, Stricker *et al.* 2013] thereby resulting in a homogeneous population of cells. These samples also have highly correlated strand-specific β-values, meaning that these data can generally be combined across strands to increase the sequencing coverage of each CpG.

All of the above highlights the considerable variability of DNA methylation data, both between samples and within a sample, and the care with which β-values must be interpreted. While an attractively simple measure, analyses based on β-values are based on the 'average' behaviour, where the averaging is over many sources of variation. Analyses based on β-values also do not make full use of the information available in bisulfite-sequencing data, as we shall see in Chapter 5.

Chapter 5

Downstream analyses of whole-genome bisulfite-sequencing data

Overview

This chapter discusses methods for the *downstream analysis* of bisulfite-sequencing data. Downstream analyses proceed the processing of the raw data (Chapter 2) to address the scientific questions of interest. Most downstream analyses are based on methylation counts at 1-tuples, however, there is growing interest in analyses based on methylation patterns at m-tuples. I discuss the statistical questions underlying these downstream analyses, paying particular attention to those based on m-tuples (m > 1) since these have received less attention in the literature.

A barrier to analyses based on m-tuples has been a lack of software. To help eliminate this barrier, I develop `MethylationTuples`, an R package for managing, analysing and visualising methylation patterns at m-tuples. `MethylationTuples` complements `methtuple` (Chapter 2) by providing a framework for the manipulation and analysis of methylation patterns at m-tuples. I describe methods available in `MethylationTuples` for the downstream analysis of whole-genome bisulfite-sequencing data.

5.1 Methods based on 1-tuples

The majority of downstream analysis methods use methylation patterns at 1-tuples, i.e. $\mathbf{m} = (m_1, \ldots, m_i, \ldots, m_{N_{loci}})$ and $\mathbf{u} = (u_1, \ldots, u_i, \ldots, u_{N_{loci}})$[1]. Methods based on 1-tuples have been developed to address a variety of scientific questions including testing for differential methylation (Section 5.1.1), testing for differentially variable methylation (Section 5.1.4) and identifying regulatory regions of the genome (Section 5.1.5).

5.1.1 Differential methylation

By far the most common analysis of bisulfite-sequencing data is to identify differentially methylated cytosines (DMCs) and differentially methylated regions (DMRs). Consequently, there has been a flurry of methods proposed for identifying differential methylation [e.g., Akalin *et al.* 2012b, Chen *et al.* 2014a,b, Dolzhenko and Smith 2014, Gokhman *et al.* 2014, Jaffe *et al.* 2012a, Lacey *et al.* 2013, Lister *et al.* 2009, Rijlaarsdam *et al.* 2014, Sofer *et al.* 2013, Xie *et al.* 2014, Feng *et al.* 2014, Hebestreit *et al.* 2013, Sun *et al.* 2014, Park *et al.* 2014, Hansen *et al.* 2012]. Robinson *et al.* [2014] recently reviewed methods for identifying DMCs and DMRs and so I give but an overview of this important topic.

Experimental design

In any analysis of differential methylation, we want the DMCs and DMRs to be both *biologically* and *statistically* significant; it's no good if all the differences are simply due to technical artefacts or random fluctuations. Key to ensuring biological relevance is good experimental design, such as the use of *replicates* in each experimental group. A distinction is often made in the literature between *technical replicates* and *biological replicates.* Briefly, biological replicates are experimental units that all undergo the same treatment and are used to estimate the within-group variability of the treatment. Technical replicates are repeated measurements of the same experimental unit, perhaps with slight variations in the sample preparation, and are used to estimate the variability of the sample preparation and measurement process.

[1] It is insufficient to use $\boldsymbol{\beta} = \frac{\mathbf{m}}{\mathbf{m}+\mathbf{u}}$ because the conversion to β-values loses information about the precision with which each methylation locus is measured (i.e. the sequencing depth, $\mathbf{d} = \mathbf{m} + \mathbf{u}$).

The boundary between biological and technical replication is not always clear. For example, Lister *et al.* [2009] state that, “For each cell type, two **biological** replicates were performed with cells of different passage number [emphasis added]”. I contend that these are better defined as technical replicates since each replicate came from the same cell line, underwent passaging under near-identical conditions and differ only by the number of cell passages in each media[2].

Initial experiments with whole-genome bisulfite-sequencing rarely had replicates of any kind or, if they did, these were pooled prior to analysis (e.g., the analysis of the H1 and IMR90 cell lines in Lister *et al.* [2009]). Simply pooling replicates and analysing as if they were a single sample ignores all variability between replicates and should not be used.

Technical variability is ideally orthogonal to the biological variability, but this rarely occurs in practice. Indeed, high-throughput sequencing experiments are particularly susceptible to batch effects, and other sources of unwanted variation, that can swamp the biological variation of interest [Leek *et al.* 2010]. This again emphasises the importance of good experimental design, with randomisation, replication and the use of positive and negative controls.

5.1.2 Differentially methylated cytosines

A differentially methylated cytosine (DMC) is one where the true methylation level, B_i, is different between experimental conditions. This is typically framed as a test of the mean levels of methylation at the locus in each group[3]. Suppose we have a two-group experiment and let B_{i,j_k} denote the true methylation level of the i^{th} methylation locus for samples in the k^{th} group ($k = 0, 1$). We wish to test the null hypothesis of $H_0 : B_{i,j_0} = B_{i,j_1}$ against the alternative hypothesis $H_1 : B_{i,j_0} \neq B_{i,j_1}$. As such, identifying DMCs boils down to identifying differences in means, for which there is an enormous body of statistical literature. This problem can be viewed as a ‘stand-alone’ test, such as a t-test, or framed as a regression problem to allow for the inclusion of additional covariates.

[2] In the case of IMR90 cell line, the first replicate, IMR90_r1, underwent 4 cell passages and the second replicate, IMR90_r2, underwent 5 cell passages. In the case of the H1 cell line, the first replicate, H1_r1, underwent 25 passages in the first media and 9 passages in the second media, and the second replicate, H1_r2, underwent 27 passages in the first media and 5 passages in the second media.

[3] This could alternatively be framed as a test of the median methylation level at the locus in each group (or of some other location parameter of the distribution of methylation levels).

Regardless of the statistical test used, all attempts to identify DMCs from whole-genome bisulfite-sequencing data must pay a large multiple-hypothesis testing penalty. Correcting for multiple hypothesis testing is standard practice in the analysis of genomics data, but the number of tests, in this case approximately 25 million, is at least one order of magnitude greater than what is commonly tested in other genomics experiments[4]. Various methods have been used to correct for this multiple testing, as illustrated in Table 5.1.

Table 5.1: Methods proposed for adjusting for the multiple hypothesis testing performed in an analysis of DMCs. Several papers use describe their analysing as performing a "false discovery rate adjustment" or "false discovery rate correction" without explicitly stating what they are doing or citing a reference. One paper uses the Bonferonni correction, leading to a very conservative analysis since this correction aims to control the family-wise error rate.

Method	Used by
Benjamini and Hochberg [1995]	Akalin *et al.* [2012b], Jaffe *et al.* [2012a], Lacey *et al.* [2013], Rijlaarsdam *et al.* [2014], Sofer *et al.* [2013]
Wang *et al.* [2011]	Akalin *et al.* [2012b]
"False discovery rate correction/adjustment"	Dolzhenko and Smith [2014], Gokhman *et al.* [2014], Lister *et al.* [2009], Xie *et al.* [2014]
Storey [2007]	Jaffe *et al.* [2012a]
Benjamini and Yekutieli [2001]	Sofer *et al.* [2013], Hebestreit *et al.* [2013]
"Bonferonni adjustment"	Feng *et al.* [2014]

One thing to note, however, is that the *effective* number of tests is fewer than the actual number of tests. This is because the methylation levels at neighbouring loci are correlated (see Chapters 6 and 7), which means that tests of differential methylation are generally positively correlated, thus reducing the effective number of independent tests. The classical Benjamini-Hochberg procedure also controls the false discovery rate under certain forms of positive dependence [Benjamini and Yekutieli 2001].

While there have been reports of DMCs resulting in a phenotypic difference [e.g., Fürst *et al.* 2012], DMCs are mostly tested as a prelude to the identification of differentially methylated regions (DMRs). Moreover, with approximately 25 million CpGs in the human genome, not to mention the many, many more non-CpG cytosines, it is an optimist who aims for the reliable detection of DMCs from whole-genome bisulfite-sequencing experiments.

[4] For example, there are approximately 20,000 tests in studies of differential gene expression and two million tests in genome-wide association studies.

This will remain true while sample sizes can be counted on one or two hands and the average sequencing depth is 10× to 30×.

Several software packages are now available for identifying DMCs. Most are limited to analysing two-group experiments. Rather than directly analysing the **m** and **u**, these software generally make additional modelling assumptions, such as the beta-binomial model (Section 4.3.2), and/or perform some transformation of the data, such as smoothing of the β-values (Section 4.4.4).

`DSS` [Feng *et al.* 2014], `BiSeq` [Hebestreit *et al.* 2013], `MOABS` [Sun *et al.* 2014], `methylSig` [Park *et al.* 2014] and `RADmeth` [Dolzhenko and Smith 2014] all use a beta-binomial hierarchical model of DNA methylation, although the exact details differ considerably between packages. `DSS` and `MOABS` use empirical Bayes methods to estimate parameters whereas `methylSig`, `BiSeq` and `RADmeth` use maximum likelihood estimation. `BiSeq` and `methlySig` also perform spatial smoothing of the data.

The statistical test used to identify DMCs in these regression models is variously a Wald test (`BiSeq`, `DSS`), a likelihood ratio test (`methlySig`, `RADmeth`) or based on the Bayesian credible interval of the difference in methylation between the two groups (`MOABS`).

Not all software for identifying differential methylation are designed for identifying DMCs. For example, both `bsseq` [Hansen *et al.* 2012] and `Aclust` [Sofer *et al.* 2013] are methods explicitly designed for identifying differentially methylated regions rather than DMCs.

5.1.3 Differentially methylated regions

A differentially methylated region (DMR) is a region of the genome where there are multiple cytosines with evidence of differential methylation. Importantly, not all cytosines in the region need necessarily be genome-wide statistically significant DMCs. Rather, the idea is that a DMR might capture a weaker but broader difference in methylation. For example, it may be more biologically relevant to identify a broad region with a consistent, albeit small, difference in methylation than it is to identify individual cytosines with large differences in methylation.

There are two very different strategies for identifying DMRs:

1. Using regions that are defined *a priori*, which are then tested for differential methylation. Such regions might be CpG islands [e.g., Huang *et al.* 1999, Doi *et al.* 2009][5]; MspI restriction fragments [e.g., Stockwell *et al.* 2014]; a predefined genomic feature, such as a gene promoters or transcription factor binding sites; or general predefined bins [e.g., 100 bp bins used by Park *et al.* 2014].
2. Using data-driven regions, such as those defined from an analysis of DMCs, which are then tested for differential methylation.

The former is much simpler to analyse but is limited in its ability to discover novel DMRs. It is also hampered because the correct unit or scale for differential methylation may not be known for the experiment.

The latter offers the opportunity to identify novel regions that are subject to differential methylation. Included in this is the opportunity to discover the scale over which differential methylation acts. However, valid statistical inference of these regions is far more challenging.

Using *a priori* regions

The idea of testing for differential methylation at *a priori* defined regions is relatively straightforward. Suppose we have a two-group experiment and let $\bar{B}_{r,j_k}$ be the true average level of methylation for the r^{th} region for samples in the k^{th} group ($k = 0, 1$). The null hypothesis is $H_0 : \bar{B}_{r,j_0} = \bar{B}_{r,j_1}$ against the alternative hypothesis $H_1 : \bar{B}_{r,j_0} \neq \bar{B}_{r,j_1}$.

We might estimate $\overline{B_r^k}$ by the group-wise average of the sample-wise weighted average of β-values for all methylation loci in the region, where the weights are proportional to the sequencing coverage. Identifying differential methylation at *a priori* defined regions simply boils down to identifying differences in means, just as is the case for testing for DMCs. Again, this problem can be viewed as a 'stand-alone' test, such as a t-test, or framed as a regression problem to allow for the inclusion of additional covariates.

The above description brushes over some technicalities, such as how to handle CpGs with insufficient sequencing coverage. An alternative approach for testing *a priori* defined regions for differential methylation is offered by `BiSeq` [Hebestreit *et al.* 2013].

[5] Both these examples are from microarray studies, but the same idea can be applied to sequencing studies.

A hierarchical procedure for testing *a priori* defined regions for differential methylation

`BiSeq` [Hebestreit *et al.* 2013] uses a hierarchical procedure to test for differential methylatoin at *a priori* defined regions. To begin, `BiSeq` first defines *CpG clusters* by identifying CpGs that are within a user-specified genomic distance of one another and that have sufficient sequencing coverage across the set of samples[6]. While there is clearly a 'data-driven' component to these cluster definitions, I reserve the use of 'data-driven regions' for those that are based on the methylation levels of loci rather than their genomic co-ordinates.

Once these clusters are defined, the β-values in each cluster are smoothed for each sample using a local binomial likelihood smoother. This procedure will create a smoothed β-value for each CpG, even those with insufficient sequencing coverage. Then, for each CpG, `BiSeq` fits a beta regression model[7] to the smoothed β-values, which is tested for evidence of differential methylation at that cytosine (i.e. a test of whether the cytosine is a DMC).

Based on these P-values, `BiSeq` then use a hierarchical testing procedure to control the false discovery rate at both the cluster-level and locus-level. This method is based on several papers by Yoav Benjamini and colleagues [Benjamini and Hochberg 1997, Benjamini and Yekutieli 2001, Benjamini *et al.* 2006, Benjamini and Heller 2007]. It aims to first control the false discovery rate at the cluster-level and then refines the signal by trimming non-DMCs from those clusters that have been declared as differentially methylated. Finally, these differentially methylated clusters are *post hoc* filtered to ensure they are *consistent*, i.e. that the differences in methylation are in the same direction.

Using data-driven regions

Methods for identifying *data-driven* DMRs are statistically *ad hoc*. The most common approach is to scan the genome for clusters of DMCs and declare these to be DMRs. The initial scan for DMCs will typically use a relaxed statistical significance threshold (i.e. not necessarily genome-wide significant). Notably, many of these methods do not include a

[6] As an alternative to creating these CpG clusters, Hebestreit *et al.* [2013] also suggest using the target regions of the assay, such as MspI fragments in the case of RRBS.

[7] This is different to the beta-binomial regression framework described in Section 4.3.2.

formal statistical test of differential methylation at the region-level [e.g., Lister *et al.* 2009, 2011, Hansen *et al.* 2011, Feng *et al.* 2014].

For example, Hansen *et al.* [2011] start by testing all CpGs for differential methylation and retain all those with a P-value in the lowest 5%. They then declare putative DMRs to be contiguous runs of such CpGs that are within a given distance of one another and with "all differences in the same direction" (i.e. the region is consistent). These putative DMRs may be subject to further filtering, such as requiring that they contain a minimum number of CpGs and span a minimum number of bases, and the merging of nearby putative DMRs into a single putative DMR [Hansen *et al.* 2011].

It is challenging to perform valid statistical inference of differential methylation at these data-driven regions. We must be careful when 'double-dipping' into the data, whereby the same data are being used to define the regions as are being used to test their significance. These regions have been *selected* because loci in these region display a difference and therefore tests of whether the region has a difference are biased towards rejecting the null hypothesis.

The challenges of valid statistical inference at such data-driven regions are not unique to the problem of testing for DMRs. Similar problems arise in the analysis of chromatin immunoprecipitation sequencing (ChIP-seq) experiments [Schwartzman *et al.* 2011a, Lun and Smyth 2014] and in the field of signal processing [Schwartzman *et al.* 2011b].

One way to avoid this issue, and I would argue the best, is to test these regions using a separate dataset, which completely avoids the issue of statistical 'double-dipping'. Of course, this requires that such a dataset is available or that resources exist to create it, which is frustratingly rare.

If the sample size is large enough, then permutation testing may also be appropriate. For example, Hansen *et al.* [2014] permute the group labels of their samples and re-ran the analysis to determine for each of the observed DMRs "how often we see another block of similar length and effect size anywhere in the genome and in any of the permutations". The chief limitation of the permutation strategy is the restricted number of permutations that are possible from small sample sizes, along with the often substantial time and computational resources it takes to analyse the data for each permutation[8].

Related to the method of creating DMRs by forming clusters of DMCs is that of "bump-hunting". Initially developed for application to methylation microarray data [Jaffe *et al.* 2012a], and now available for broader use in the R/Bioconductor package `bumphunter` (`http://bioconductor.org/packages/bumphunter/`), bump-hunting may be used to identify DMRs. The idea is as follows. Firstly, each each CpG is tested for differential methylation. The resulting test statistic is then considered as a function of the position in the genome[9] and processed with an algorithm to identify "bumps" in the signal. Bumps are defined as contiguous regions of the genome where the signal is above some threshold. The algorithm may include an error term to account for the spatial correlation of the signal and the significance of these peaks may be assessed using a permutation strategy.

Another alternative for combining individual loci into data-driven DMRs uses the locus-specific P-values rather than the locus-specific test statistics. These methods can be thought of extensions to Fisher's method for combining P-values [Fisher 1936], that attempt to account for the correlation of tests at nearby methylation loci. `comb-p` [Pedersen *et al.* 2012] uses the Stouffer-Liptak-Kechris [Stouffer 1949, Kechris *et al.* 2010, Zaykin 2011] correction for spatially correlated P-values. `methylKit` [Akalin *et al.* 2012b] uses `SLIM` [Wang *et al.* 2011] to do a similar correction.

Finally, there are a class of methods that turn the problem of identifying data-driven DMRs on its head by constructing the regions without first testing the individual loci for differential methylation. Then, only once these regions are defined, these methods test for differential methylation. This is different to using *a priori* regions, since the regions are still data-defined, but not with respect to differential methylation. The only example of such a method that I am aware of is `Aclust` [Sofer *et al.* 2013]. `Aclust` first clusters CpGs into candidate regions by performing agglomerative nested clustering of the between-sample co-methylation. Briefly, this is the correlation of methylation levels at two loci across the samples[10]. These clusters are then tested for differential methylation using generalised estimating equations.

Regardless of the statistical method used to identify differential methylation, it remains

[8] For example, Hansen *et al.* [2014] only performed nine permutations to estimate significance.

[9] The test statistic may be smoothed to reduce variation at the expense of increasing bias.

[10] See Chapter 6 for further details of between-sample co-methylation.

important to validate these differences. This validation should ideally be performed in a new dataset and perhaps using a different assay in order to mitigate potential biases.

5.1.4 Differentially variable methlyation

It has been hypothesised that increased variability in DNA methylation indicates an epigentic *plasticity* that may be highly relevant to common diseases, in particular, cancer [Feinberg and Irizarry 2010]. Differential variability is distinct from differential methylation. Whereas the analysis of differential methylation is based on statistical tests of differences in means, the analysis of differential variability is based on statistical tests of differences in variances. Analogously, we define variably methylated cytosines (VMCs) and variably methylated regions (VMRs). To emphasise, a locus (resp. region) may be a DMC (resp. DMR) while not a VMC (resp. VMR) and *vice versa*.

Jaffe *et al.* [2012b] first developed formal statistical tests for differential variability of methylation for both the one-group and two-group experiments. These methods were developed for use with data from the CHARM array (see 1.4.1). In a one-group experiment, a variably methylated region is one that has increased variability compared to 'similar' regions elsewhere in the genome. In a two-group experiment, a differentially variable locus is one where the variation in one group is significantly larger than that in the other group.

Statistical tests of variances are well known to be more difficult than tests of means and require larger sample sizes. A more subtle difficulty is in dealing with outliers. An outlier in one group will greatly increase the variation in that group, but this does not necessarily mean that the locus is differentially variably methylated. It might, for example, be due to an error in the assay. It is not difficult to envisage an example where the two groups in fact have very similar variability once the outlier is excluded.

Tests of differential variability that are based on the F-test [e.g., Hansen *et al.* 2011] or Bartlett's test [e.g., Teschendorff and Widschwendter 2012] will be susceptible to calling loci with such outliers as being differentially variable. By contrast, `DiffVar` [Phipson and Oshlack 2014] uses Levene's test [Olkin 1960] to test for differential variability since it is robust to outliers.

5.1.5 Epigenome segmentation

Methylation data in the form of the vectors of **m** and **u** may also be used by methods to segment the genome or epigenome into regulatory regions [Stadler *et al.* 2011]. One type of region that has received particular attention are so-called *partially methylated domains* (PMDs). Partially methylated domains are long stretches of the genome where the average methylation level, B_i, is away from the extremes of 0 and 1, typically in the range 0.2 to 0.7.

PMDs were first identified in the *IMR90* methylome by Lister *et al.* [2009]. Using a simple sliding window algorithm, Lister *et al.* found that approximately 40% of every autosome was a PMD and that the average length of these PMDs was a large 153 kb. They also showed that these PMDs were not simply due to a methylated subpopulation and an unmethylated subpopulation of cells in the sample. This did this by showing that individual reads mapped to these PMDs contained both methylated and unmethylated bases.

A subsequent study found that PMDs are a common feature of somatic cell lines and that they comprise $> 30\%$ of the genome [Lister *et al.* 2011]. Perhaps even more intriguingly, across four somatic cell lines profiled with whole-genome bisulfite-sequencing, Lister *et al.* found a large amount of these genomes (664 Mb) comprised shared PMDs.

PMDs have also been identified within tumour methylomes [Berman *et al.* 2012, Hansen *et al.* 2011]. Hansen *et al.* [2011] found that these PMDs overlap with other important genomic features called large organized chromatin lysine modifications (*LOCKs*) and lamina associated domains (*LADs*). However, PMDs are conspicuous by their absence in pluripotent cell lines, including both embryonic stem cells and induced pluripotent stem cells [Lister *et al.* 2011]. Their absence in the induced pluripotent cell lines may reflect the fact that the methylome is 'reset' upon induction of pluripotency [Lister *et al.* 2011, Stricker *et al.* 2013].

Recently, more sophisticated methods have been proposed to identify these PMDs. `methylSeekR` [Burger *et al.* 2013] is one such method. It uses a hidden Markov model of the β-values, combined with other filters, to segment the genome into unmethylated, lowly methylated and partially methylated regions.

5.2 Methods based on m-tuples (m $>$ 2)

This section reviews the different types of questions that can be addressed by expanding our analysis to use methylation patterns at m-tuples (m $>$ 2) rather than just 1-tuples. This extra information is only available from sequencing-based assays. We also review existing software that implement some of these methods.

5.2.1 Methylation entropy

The natural interpretation of methylation entropy is a measure of 'disorder'. It has been used to quantify how heterogeneous DNA methylation is at a locus [e.g., Xie *et al.* 2011, He *et al.* 2013][11]. These methylation entropies can be analysed to identify heterogeneous regions of the genome or perhaps tested for an association with a phenotype.

On the one hand, if we only observe a single unique methylation pattern, then the m-tuple has the minimum methylation entropy of zero. On the other hand, if we observe all possible 2^m methylation patterns at equal frequency, then the m-tuple has the maximum methylation entropy (typically normalised to one). Depending on the frequency of the observed methylation patterns patterns, we obtain intermediate values of the methylation entropy.

5.2.2 Allele-specific methylation

In a diploid cell, allele-specific methylation occurs when only one of the parental chromosomes is methylated at a particular locus, where the locus may be an individual cytosine or a broader region such as a gene promoter. A particularly interesting form of allele specific methylation occurs at imprinted genes, where one copy of the gene is active in a parent-specific manner. However, it is now apparent that allele-specific methylation is far more prevalent than at just these imprinted regions [Tycko 2010, Shoemaker *et al.* 2010].

The obvious method to detect allele-specific methylation from bisulfite-sequencing requires reads that contain a heterozygous genetic variant, such as a single nucleotide

[11] This is closely related to the idea of identifying epialleles, for which methylation entropy has also played a role [Li *et al.* 2014] and which I discuss in Section 5.2.3. Methylation entropy has also been used to identify differential methylation, however, I do not discuss this further since it uses a different definition that is not based on analysing methylation patterns at m-tuples [Zhang *et al.* 2011, Su *et al.* 2013].

polymorphism, along with at least one methylation locus. The heterozygous variant allows these reads to be separated by the observed allele[12], which can then be used to test for allele-specific methylation. For example, Shoemaker *et al.* [2010] construct a 2 × 2 contingency table, like that shown in Table 5.2, and test for an association between the allele and the methylation state using Fisher's exact test [Fisher 1922].

Table 5.2: 2×2 table used to test for allele-specific methylation. m^1 is the number of reads with the first allele and that are also methylated at the methylation locus, u^1 is the number of reads with the first allele and that are also unmethylated at the methylation locus, etc.

	Allele 1	Allele 2
m	m^1	m^2
u	u^1	u^2

While straightforward, this approach is also obviously limited to the small number of methylation loci that are nearby to a heterozygous genetic variant. Fang *et al.* [2012] and Peng and Ecker [2012] published methods to detect allele-specfic methylation that do not require heterozygous genetic variants nearby to the methylation locus of interest. These methods rely on the probabilistic assignment of reads to alleles (which are treated as missing data). Unfortunately, there is no publicly available software implementing the method proposed by Peng and Ecker [2012] and so I do not discuss it further.

Fang *et al.* [2012] use reads containing multiple methylation loci and looks for regions of the genome where there are two distinct methylation patterns at the read-level that occur at roughly equal proportions, indicating one pattern comes from one allele and the other pattern from the other allele. The likelihood of allele-specific methylation is computed using an expectation-maximisation algorithm, which assigns reads to one of the two possible alleles. Neighbouring regions displaying allele-specific methylation are then joined together. While not mentioned in the paper, the proposed method is now available in the `MethPipe` software (`http://smithlabresearch.org/software/methpipe/`).

5.2.3 Epialleles

A DNA sequence may have multiple epigenetic states. For example, the cytosine in the sequence *TCGA* may be methylated or unmethylated; each of the methylated and

[12] This does not give parent-specificity unless the phase of the genotype is also known.

unmethylated versions of that sequence is an *epiallele*. Rakyan *et al.* [2002] define an *epiallele* as “an allele that can stably exist in more than one epigenetic state, resulting in different phenotypes”. The latter requirement, while obviously more interesting than the alternative, may be unduly restrictive. After all, we refer to alternative forms of a genetic sequence as *alleles* regardless of whether we know of a phenotypic consequence of the variant.

Most of the examples of epialleles with a phenotypic consequence come from the plant kingdom and, even there, the number of such epialleles is small: a review from 2012 put the number at “about a dozen” [Weigel and Colot 2012]. In mammals, the study of epialleles has focused on identifying *metastable* epialleles, which are epialleles that are mitotically heritable [Rakyan *et al.* 2002]. The poster child for the potential importance of epialleles in mammals is the Agouti viable yellow (A^{vy}) allele [Morgan *et al.* 1999]. Genetically identical mice with different versions of the A^{vy} allele are phenotypically distinct. Those mice with an unmethylated version of the allele have a yellow coat, are obese, diabetic, and have an increased susceptibility to tumours; those mice with a methylated version of the allele have a *pseudoagouti*[13] (brown) coat and none of the associated health defects.

In humans, there have been several interesting studies using putative epialleles to infer the clonality and evolution of cancer [Siegmund *et al.* 2009, Li *et al.* 2014], as well as to study the evolution of methylation dynamics and the rate of epipolymorphism of various loci in an immortalised cell line [Landan *et al.* 2012].

Regardless of where you draw the line as to what constitutes an epiallele, it has become clear in the analysis of bisulfite-sequencing data that the occurrence of multiple methylation patterns at an m-tuple is the norm rather than the exception.

Restricting our attention to CpG methylation, a sequence with m CpGs has 2^m potential epialleles. In other words, an epiallele is just a methylation pattern at an m-tuple, with the additional constraint that the underlying DNA sequence also be identical. An epiallele may also be described as an *epimutation* if it is different from the ‘normal’ methylation state.

The *rate of epipolymorphism* of a locus is defined as the probability that two epialleles

[13] These mice are properly described as *pseudoagouti* rather than agouti. They are heterozygous for the wildtype agouti allele (A^{vy}/a) but are phenotypically indistinguishable from true agouti mice, which are homozygous for the wildtype gene (a/a).

randomly sampled from the locus are different from one another [Landan *et al.* 2012][14]. Landan *et al.* define the rate of epipolymorphism of an m-tuple as $1 - \sum_{p=1}^{p=2^m} f_p^2$, where f_p is the estimated frequency of the p^{th} methylation pattern, i.e. the number of times the p^{th} pattern is observed divided by the total number of reads mapped to that m-tuple[15]. Unlike genetic polymorphisms, where the population is typically a set of chromosomes from multiple individuals, the population of epipolymorphisms is often within an individual, even within a tissue within an individual.

The obvious challenge in estimating the frequency of an epiallele is in distinguishing a 'real' epiallele from a spurious one (perhaps caused by incomplete bisulfite-conversion) sequencing error or mapping error. Another difficulty, perhaps unavoidable with current technology, is the effect of PCR amplification bias, which will bias estimates of the relative abundance of each epiallele.

Of the downstream analyses based on methylation patterns at m-tuples, the study of epialleles has received the most attention with respect to methods and software development.

`methclone` [Li *et al.* 2014] is a method to estimate the frequency of epialleles at m-tuples (the rate of epipolymorphism) and to identify "shifts" in these distributions between a pair of samples. `methclone` is based on computing and comparing two forms of methylation entropy, the "foreground" and "background". The foreground combinatorial entropy, S, is based on the observed frequency of epialleles in the two samples. The background combinatorial entropy, $\tilde{S}$, is the expected frequency of epialleles in the two samples if "all patterns of epialleles are uniformly mixed between the two [samples]". The difference in these combinatorial entropies, $\Delta S = S - \tilde{S}$, a kind of observed-to-expected log-ratio, is used to identify shifts in the epiallele distribution between a pair of samples. A $\Delta S = 0$ corresponds to no change and a $\Delta S = -144$ corresponds to maximal difference in entropy. It isn't clear whether the range of ΔS depends on the size of the m-tuples nor is it clear how to choose the threshold at which to declare a significant shift in the distribution of epialleles.

[14] Landan *et al.* [2012] actually call this the "epipolymorphism" of the locus rather than the "rate of epipolymorphism" of the locus. However, I think this is better described as a rate since it refers to the frequency at which we observe epialleles/epipolymorphisms.

[15] Strictly speaking, this is in fact an estimate of the rate of epipolymorphism of the locus under a model that assumes sampling with replacement or, equivalently, an infinite population size. While neither assumption is true, the correction for sampling without replacement from a finite population will not substantially affect the results provided that the sequencing depth is high.

methclone uses the observed frequencies of methylation patterns as being unbiased estimates of the true epiallele frequencies and does not attempt to account for potential sources of bias. In contrast, MPFE (http://bioconductor.org/packages/MPFE, http://f1000.com/posters/browse/summary/1097258), an R/Bioconductor package for "[estimating] the distribution of methylation patterns [i.e. epialleles]" at m-tuples, uses a probabilistic model to account for some of these biases.

MPFE is designed to estimate the frequency of epialleles by maximising a multinomial likelihood that includes error terms for both incomplete bisulfite-conversion and sequencing error. The maximisation of this likelihood is computatationally demanding, as evidenced by the need for a "fast" algorithm that approximates the likelihood. MPFE is designed for amplicon bisulfite-sequencing and may not scale to whole-genome data. The input is a file containing the number of times each methylation pattern was observed at that m-tuple. Unfortunately, MPFE does not provide a way to create this file.

Methods designed to detect allele-specific methylation, specifically those that are based on the observed methylation patterns [e.g., Fang *et al.* 2012, Peng and Ecker 2012], might also be adapted to identify epialleles and their associated frequencies. It is worth emphasising that since all of the methods described in this section are based entirely on the observed methylation patterns, none of these actually check that the underlying DNA sequencing is identical, which, strictly speaking, is a requirement for the m-tuple to be an epipolymorphic locus.

5.2.4 Software for analysing methylation patterns at m-tuples

Generally speaking, there are fewer software options for analysing methylation patterns at m-tuples ($m > 2$) than there are for analysing methylation patterns at 1-tuples. Furthermore, the available options are often difficult to extend since they are typically developed for a specific task and not for general computations with methylation patterns at m-tuples.

I have also experienced considerable difficulty in applying some of these methods owing to poor software implementations. To give two examples, DMEAS [He *et al.* 2013] is only available as a Windows binary or as a Perl script that itself is only available as a PDF file, and I have been unable to install methclone due to compilation errors.

MethPipe is perhaps the best documented and potentially extensible software for analysing methylation patterns at m-tuples. MethPipe is mostly written in C++ and is designed as a suite of tools for a complete 'pipeline' analysis of bisulfite-sequencing data. As such, it does not feature tools that are particularly amenable to interactive or exploratory analyses. In fact, I am unaware of any software that allows easy exploratory analyses of methylation patterns at m-tuples, which in part motivated the development of MethylationTuples.

5.3 MethylationTuples

In order to facilitate the development of downstream analysis methods based on methylation patterns at m-tuples, I saw the need for two pieces of software:

1. Software for extracting methylation patterns at m-tuples,
2. Software for manipulating, analysing and visualising these methylation patterns.

I have made significant progress towards the first goal with methtuple (see Section 2.4) and now introduce MethylationTuples (https://github.com/PeteHaitch/MethylationTuples) to address the second missing link.

5.3.1 Design

MethylationTuples is an R package for managing, analysing and visualising methylation patterns at m-tuples. It is released under an Artistic-2.0 license, consistent with core Bioconductor packages. I chose to write this software in R because it is a very popular language for data analysis, particularly in bioinformatics, and facilitates both batch and interactive usage. R is also my computational mother tongue and an R package is a convenient unit for sharing reusable code. To improve the performance of key functionality, parts of MethylationTuples are written in C++, making use of the Rcpp package [Eddelbuettel *et al.* 2011, Eddelbuettel 2013].

While initially developed to support my research into co-methylation (Chapter 7), the data structures developed in MethylationTuples are well-suited to other analyses based on methylation patterns of m-tuples such as methylation entropy, allele-specific methylation and the identification of epialleles. Of course, `MethylationTuples` can also be used to develop methods based on 1-tuples, such as identifying differential methylation, since 1-tuples are just a particular type of m-tuple.

`MethylationTuples` is written to work within the Bioconductor project [Gentleman *et al.* 2004][16]. Bioconductor makes extensive use of R's S4 object system and encourages developers to re-use existing Bioconductor infrastructure. From a developer's perspective, this avoids the need to re-invent the wheel when tackling common tasks. And from the user's perspective, it helps avoid multiple versions of the wheel, each that might otherwise act slightly differently and that may not be as well-tested.

Bioconductor already has excellent support for working with data defined on genomic ranges via the `IRanges` and `GenomicRanges` packages [Lawrence *et al.* 2013, Lawrence and Morgan 2014]. Genomic tuples, however, such as the co-ordinates of an m-tuple, do not naturally fit into this framework[17]. Therefore, I first wrote a Bioconductor package for working with genomic tuples, rather unimaginatively called `GenomicTuples`, first released as part of Bioconductor version 3.0 (`http://bioconductor.org/packages/GenomicTuples`). In fact, `GenomicTuples` is heavily based on the `GenomicRanges` package, with modifications for tuple-specific operations. This makes it easy to use for users already familiar with the `GenomicRanges` package. For example, there is a tuple-specific method for the `findOverlaps` generic function to identify genomic tuples with equal co-ordinates (i.e. `type = 'equal'`). Since the classes in `GenomicTuples` extend those defined in `GenomicRanges`, these have excellent interoperability with existing Bioconductor infrastructure.

In the `MethylationTuples` package I define the `MethPat` class to store the genomic co-ordinates of m-tuples and the associated counts of each methylation pattern. A `MethPat` object is as a matrix-like object, where rows represent m-tuples and columns represent samples. The `MethPat` class extends the `GenomicRanges::SummarizedExperiment`[18] class

[16] `MethylationTuples` has not yet been submitted to Bioconductor but its development is being published to `https://github.com/PeteHaitch/MethylationTuples`).

[17] The difference between a genomic range and a genomic tuple can be thought of as the difference between an interval and a set. Namely, an interval includes the co-ordinates in between the start and end whereas a set only includes those co-ordinates listed in the set. For example, the genomic interval `chr3:+:[10, 12]` includes the co-ordinates `chr3:10`, `chr3:11` and `chr3:12` on the forward strand, whereas the genomic 2-tuple `chr3:+:{10, 12}` only includes the co-ordinates `chr3:10` and `chr3:12` on the forward strand.

[18] This uses the NAMESPACE notation of R: `GenomicRanges::SummarizedExperiment` can be read as

but makes use of classes defined in the `GenomicTuples` package to store the genomic co-ordinates of the m-tuples. Currently, it is a requirement that all m-tuples in a `MethPat` object have the same size (i.e. same m).

Figure 5.1 is a schematic of a `MethPat` object storing methylation patterns at 3-tuples for n samples. The similarities to the output format of `methtuple` are clear (see Figure 2.10), with the added advantage that a single `MethPat` object can contain data from multiple samples.

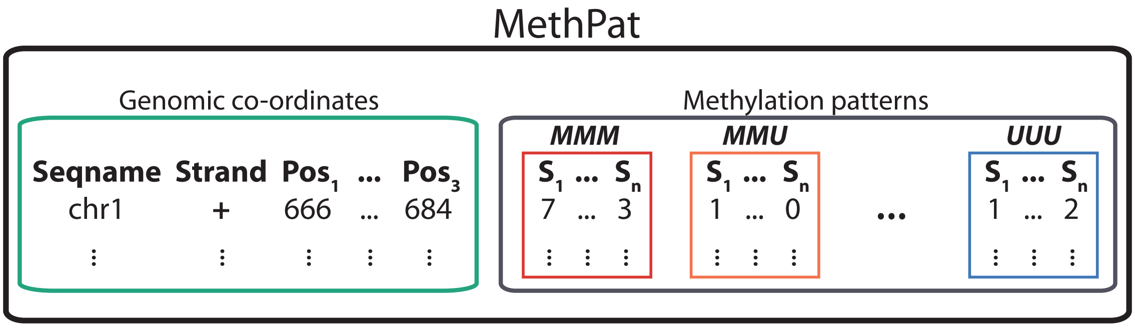


Figure 5.1: Schematic of the `MethPat` class, shown here for 3-tuples. Each row represents a 3-tuple to which the genomic co-ordinates of the tuples (green box) and the counts of the methylation patterns (grey box) are aligned. The counts of each methylation pattern $(MMM, MMU, \ldots, UUU)$ are stored as separate matrices where the columns represent samples $(S_1, \ldots, S_n)$. Some samples may not have any sequencing coverage for a particular m-tuple, in which case the corresponding frequencies are recorded as `NA`.

5.3.2 Methods

A `MethPat` object can be constructed directly using the `MethPat()` constructor function or from the output files of `methtuple` via the `readMethtuple()` function.

The `MethPat` object provides fast subsetting by rows (m-tuples) and columns (samples) via the “`[`” method. It also benefits from fast subsetting based on overlaps of m-tuples with genomic features via the `findOverlaps()`-based methods. Several other useful utility functions for working with `MethPat` objects include:

- `collapseStrand()`: Collapse strand-specific data by aggregating the counts. Only applicable to CpG methylation loci.

“the `SummarizedExperiment` class is part of the `GenomicRanges` package”.

- `combine()`: Combine multiple `MethPat` objects into one.
- `filterOutVariants()`: Remove m-tuples that contain a known variant. Variants must be provided as a `VCF` file.
- `findMTuples()`: Find m-tuples of a given size in a reference genome.
- `getCoverage()`: Compute the sequencing coverage of each m-tuples in each sample.
- `IPD()`: Compute the IPD vector for each m-tuple.
- `methLevel()`: Compute β-values or $\mathcal{M}$-values. Only applicable to 1-tuples.
- `tuples()`: Extract the $pos_1, \ldots, pos_m$ of the m-tuples.

These are in addition to the many useful methods inherited from the `GenomicRanges::SummarizedExperiment` class.

With the `MethPat` class, its associated methods and other utility functions, the `MethylationTuples` package provides a toolbox for manipulating methylation patterns at m-tuples. Aside from providing the necessary infrastructure to analyse methylation patterns at m-tuples, `MethylationTuples` currently includes specific methods to analyse and visualise co-methylation (Chapter 7) with the `cometh()` and `methLevelCor()` methods. I also plan to add methods for estimating epialleles and epipolymorphism. It is my hope that `MethylationTuples` will provide a useful foundation on which others can implement their own methods for analysing methylation patterns at m-tuples.

5.3.3 Compatability with other Bioconductor packages

Since `MethylationTuples` is based on core Bioconductor functionality, it is highly compatible with existing Bioconductor packages. In particular, `MethPat` objects containing 1-tuples are readily coerced for use with differential methylation calling packages, e.g., `bsseq` and `BiSeq`, or to identify partially methylated domains with `MethylSeekR`. I also make extensive use of `MethylationTuples` in my `methsim` software (Chapter 8)

5.3.4 Computational challenges and future directions

The challenges of working with large datasets in R are well-known. These are in large part due to R being designed as an 'in-memory' application and its implementation of 'copy-on-modify' semantics [Wickham 2014]. More generally, with large datasets, there is often a trade-off to be made between storage efficiency and algorithm simplicity; a more efficient way of storing the data may be less convenient to work with and *vice versa*.

I have used the `MethylationTuples` package to analyse various sized m-tuples from datasets containing up to 17 whole-genome bisulfite-sequencing samples (the *Lister* data). I have found the `MethPat` class to be a very convenient representation of the data, however, it has also raised challenges that will apply for larger datasets.

One such challenge is the size of a `MethPat` object in memory. The `MethPat` class currently favours a simpler implementation at the expense of storage efficiency. The main inefficiency with the `MethPat` class is that the matrices storing the counts of each methylation pattern grow increasingly sparse as the size of the tuples increases.

Shown in Table 5.3 is the size of the `MethPat` objects in memory for the *EPISCOPE* whole-genome bisulfite-sequencing data with various sized m-tuples. We see that as the size of the m-tuples increases, the data become sparser: most counts are 0 (meaning that particular methylation pattern was not observed in that particular sample) or `NA` (meaning that that m-tuple was not observed in that particular sample). However, for values of $m < 5$, and particularly for 'dense' data, such as RRBS, this is far less of an issue.

Table 5.3: Size of `MethPat` objects for the *EPISCOPE* data ($N_{samples} = 12$). All m-tuples are stranded. The 'size' of the `MethPat` object, reported in gigabytes (GB), is computed using the `pryr::object_size()` function (`http://cran.r-project.org/web/packages/pryr/index.html`). The 'number of rows' corresponds to the number of m-tuples in the object. The 'number of assays' is 2^{m}, where m is the size of the m-tuples. The final column is a measure of how sparse the data are: a 0 value means that particular methylation pattern was not observed in that particular sample and an `NA` value means that that m-tuple was not observed in that particular sample.

	Size (GB)	Number of rows	Number of assays	Percentage of 0 and `NA` values
1-tuples	5.9	56,348,522	2	28%
2-tuples	20.1	100,586,237	4	80%
2-tuples (`--all-combinations`)	60.0	299,814,999	4	78%
3-tuples	43.3	109,376,348	8	93%
4-tuples	80.5	102,625,758	16	97%

It will be possible to improve the storage efficiency by using a different approach to the internal storage of the data. For example, it may be possible to use sparse matrices to store the counts or to create an index so that counts can be re-ordered and stored as run-length encodings, a very space-efficient storage scheme.

In the short term, my aim for `MethylationTuples` is to release a version to Bioconductor with the core infrastructure for manipulating data of methylation patterns at m-tuples. In the longer term, I would like to extend the set of downstream analyses of m-tuples that are available in the package and to help users extend the package to add their own methods.

5.4 Summary

Most methods for downstream analyses of bisulfite-sequencing data have focused on the problem of identifying differential methylation using methylation calls at 1-tuples. However, there is a growing interest in questions related to the heterogeneity of DNA methylation and these might be better addressed by analyses based on methylation patterns at m-tuples.

A barrier to these type of analyses has been a lack of software for extracting and manipulating these data. It is my hope that the `methtuple` and `MethylationTuples` software will prove useful in facilitating the development of methods for these new types of downstream analyses.

Chapter 6

A critical review of co-methylation

Overview

Co-methylation is the dependence structure of DNA methylation data. This chapter critically reviews different definitions of co-methylation and puts them in the statistical framework of Chapter 4. It highlights various shortcomings of existing methods for quantifying co-methylation, which led me to develop the methods and software described in Chapter 7. We also review how methods to detect differential methylation have sought to account for, and leverage, co-methylation.

6.1 What is co-methylation and why study it?

I define co-methylation as the dependence structure of DNA methylation data. This broad definition encompasses several similar concepts previously described in the literature. These include: "the presence of methylation over a stretch of neighbouring CpG positions" [Schatz *et al.* 2004]; "the relationship between the degree of methylation over distance" [Eckhardt *et al.* 2006]; the "correlation of two [loci] across many samples" [Akulenko and Helms 2013]; and the "*vertical* (i.e. progenitor to descendant) correlation between the same CpG site in neighbouring cell types" [Capra and Kostka 2014].

It is important to understand that there exists both within-sample and between-sample co-methylation.

Within-sample co-methylation is the spatial dependence of DNA methylation along the genome within a single sample. Many researchers have observed that there is a strong spatial dependence of DNA methylation along the genome, that is, methylation loci near to one another in the genome tend to be similarly methylated [e.g., Eckhardt *et al.* 2006, Cokus *et al.* 2008, Lister *et al.* 2009, Lacey and Ehrlich 2009, Li *et al.* 2010, Lyko *et al.* 2010, Landan *et al.* 2012]. These studies also found that the strength of this dependence itself depends on the distance between the loci and on the genomic context.

The fundamental level of within-sample co-methylation is the dependence structure of DNA methylation events occurring on the same DNA fragment, which I call *within-fragment co-methylation*. At a higher level is the *correlation of aggregate methylation levels*, such as the spatial correlation of β-values within a sample.

Between-sample co-methylation is the relationship between methylation levels at a pair of loci across a set of samples[1]. For example, Akulenko and Helms [2013] reported 187 pairs of genes whose methylation level was highly correlated (Pearson $|r| \geqslant 0.75$) across more than 300 breast cancer samples. It is worth noting that measures of within-sample co-methylation can themselves be compared between samples.

There have been several attempts to analyse co-methylation, most with the aim of better understanding the biological processes that lay down and regulate DNA methylation. An increased understanding of co-methylation could also lead to cheaper and more efficient assays of DNA methylation. For example, we might identify methylation loci whose methylation status is highly predictive of the methylation level of a surrounding region and design an assay to interrogate those highly predictive loci. This idea is analogous to the use of tag-SNPs on genotyping microarrays.

There are also reasons for studying co-methylation that are less to do with improving our biological understanding and more to do with improving statistical techniques for analysing DNA methylation data. If co-methylation can be better estimated and understood, then statistical techniques can be developed that explicitly account for and leverage these

[1] One measure of between-sample correlation of aggregate methylation levels that does not fall under my definition of co-methylation is the following: Consider a pair of samples, j and j', and compute the correlation between the set of β-values for sample j against sample j', i.e. $cor(\{\beta_{i,j}\}_{i=1}^{i=N_{loci}}, \{\beta_{i,j'}\}_{i=1}^{i=N_{loci}})$. This measure is often reported as evidence for the “concordance” of methylation levels between replicates, but is not what I consider a form of co-methylation.

dependencies to create more powerful and efficient analysis methods.

A common difficulty in translating biological observations into mathematical or statistical language are vague, ambiguous or otherwise insufficiently detailed descriptions of the analyses performed. This is frequently coupled with a lack of software to implement the proposed method. These deficiencies can bring into question the results and certainly hinder the development of new analysis methods and software. Many of the papers reviewed in this chapter suffer from these deficiencies and in these cases I have made my best attempt at deciphering the methods, often via correspondence with the authors.

6.2 Correlations of aggregate methylation levels

To date, most analyses of co-methylation have been based on analysing correlations of aggregate methylation levels, such as β-values, within individual samples [Eckhardt *et al.* 2006, Cokus *et al.* 2008, Li *et al.* 2010, Lyko *et al.* 2010, Lacey *et al.* 2013]. These analyses have an implicit assumption of stationarity, that is, that the correlation between two loci depends only on the distance between them and not the actual position of the loci in the genome.

6.2.1 Eckhardt *et al.* [2006]

Eckhardt *et al.* [2006] is one of the first publications to apply bisulfite-sequencing to a large number of regions and samples. Of particular relevance to this chapter, Eckhardt *et al.* [2006] is often cited as evidence of co-methylation. For example, Hebestreit *et al.* [2013] cite Eckhardt *et al.* [2006] as evidence that "methylation levels are strongly spatially correlated" and Hansen *et al.* [2011] cite Eckhardt *et al.* [2006] as evidence that "proximal CpGs [have] similar methylation levels".

Eckhardt *et al.* [2006] performed PCR-amplicon bisulfite-sequencing of $2,524$ amplicons on human chromosomes 6, 20 and 22. PCR-amplicon bisulfite-sequencing is a labour-intensive assay and so it was a considerable effort to produce these data. Nonetheless, it is important to bear in mind that these data are nowhere near as comprehensive as data generated using modern whole-genome bisulfite-sequencing assays. The samples are a variety of human tissue types from multiple donors of different ages, both male and female.

As part of their analysis, the authors explore co-methylation, which they define as "the relationship between the degree of methylation over distance". They state that "we were able to establish a significant correlation for comethylation over short distances ($\leqslant 1,000$ bp), it deteriorated rapidly for distances $> 2,000$ bp" and reference the plot reprinted in Figure 6.1 in support of this claim. In Figure 6.1, the authors plot the "percentage identical methylation" against "distance in bp", which ranges from 0 to $20,000$ bp. This is not a correlation in the usual sense of the word. Nowhere in the original paper is it clearly described what the "percentage identical methylation" is nor how it is computed[2].

The average amplicon length is reported as 411 bp with a standard deviation of 71 bp. Therefore, it is impossible for the "percentage identical methylation" to have been obtained from methylation events co-occuring on the same amplicon, at least for those pairs separated by more than approximately 600 bp.

My best guess as to the definition of "percentage identical methylation" is that it is the proportion of pairs of identical methylation calls at different CpGs from (generally) different amplicons, that is, $Pr(Z_{h,i} = Z_{h',i'})$. I will refer to my best guess of "percentage identical methylation" as PIM^{*}, and now turn to some of its properties.

Let $0 \leqslant p_h \leqslant 1$ be the genome-wide average CpG methylation level for the sample from which $Z_{h,i}$ is sampled, that is, $Pr(Z_{h,i} = 1) = p_h$. For two distinct CpGs sequenced on two (potentially) distinct amplicons, $Z_{h,i}$ and $Z_{h',i'}$, we want to compute the probability that the methylation states are identical, that is, $Pr(Z_{h,i} = Z_{h',i'})$. Not only may the two CpGs be from distinct amplicons, but in the original analysis the amplicons themselves may be from distinct samples with different genome-wide average methylation levels. This most general setting is described by $Z_{h,i} \overset{d}{=} \text{Bernoulli}(p_h)$ and $Z_{h',i'} \overset{d}{=} \text{Bernoulli}(p_{h'})$, with $cor(Z_{h,i}, Z_{h',i'}) = c$. In this setting we can compute the theoretical PIM^{*} as follows:

[2] An email discussion with the senior author, Professor Stephan Beck (University College London), was unable to clarify the definition or method.

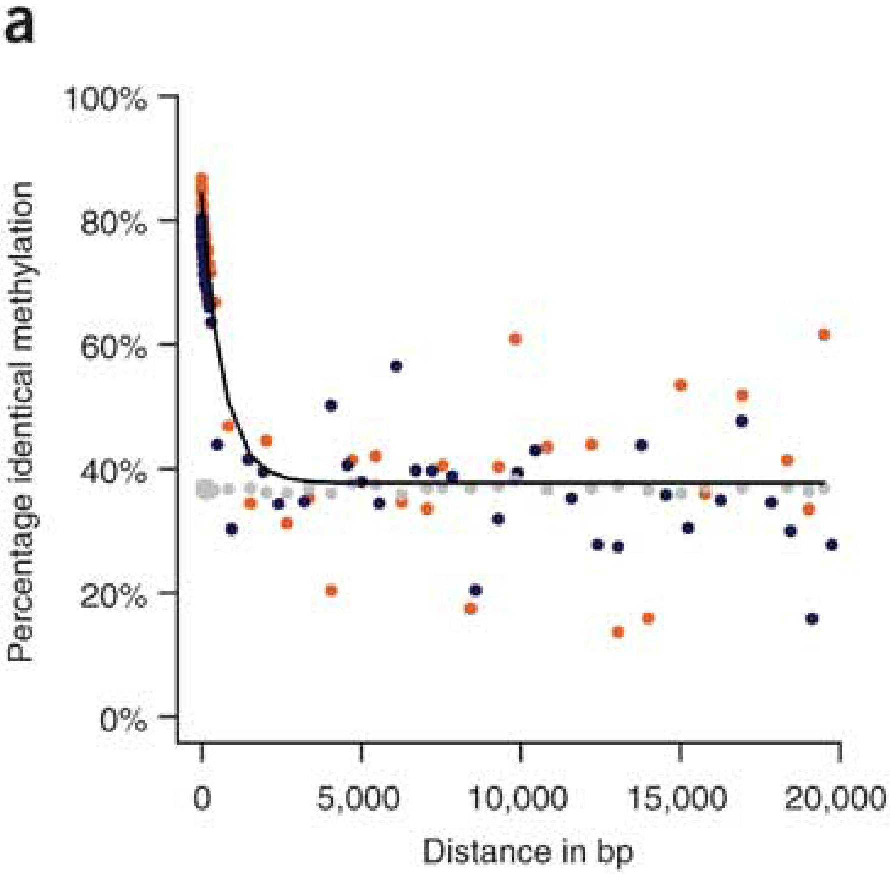


Figure 6.1: Correlation of DNA methylation with spatial distance. (a) Correlation between comethylation and spatial distance. Orange dots represent CpG methylation values aggregated and averaged over $25,000$ individual measurements. Gray dots represent CpG methylation values based on resampling of random CpG positions. Blue dots indicate CpG methylation values based on resampling of amplicon positions. At distances $> 1,000$ bp, we did not detect any correlation between CpG methylation and spatial distance. Adapted by permission from Macmillan Publishers Ltd: [Nature Genetics] (Eckhardt *et al.* [2006]), copyright (2006).

$$\begin{aligned}
PIM^* &= Pr(Z_{h,i} = Z_{h',i'}) \\
&= Pr(Z_{h,i} = 0, Z_{h',i'} = 0) + Pr(Z_{h,i} = 1, Z_{h',i'} = 1) \\
&= Pr(Z_{h',i'} = 0|Z_{h,i} = 0)Pr(Z_{h,i} = 0) + Pr(Z_{h',i'} = 1|Z_{h,i} = 1)Pr(Z_{h,i} = 1) \\
&= \big[(1 - p_{h'}) + cp_h\big](1 - p_h) + \big[p_{h'} + c(1 - p_h)\big]p_h \\
&= (1 - p_h)(1 - p_{h'}) + p_h p_{h'} + 2cp_h(1 - p_h)
\end{aligned}$$

The proof of this result is given in Appendix A.1.

In general, $PIM^* \in [0, 1]$ since it is a probability, which is already unlike a correlation coefficient that takes values in $[-1, 1]$. There are a few important special cases that are worth discussing:

1. $PIM^*_{inid} = (1-p_h)(1-p_{h'})+p_h p_{h'}$: Independent and non-identically distributed (*inid*), i.e. $Z_{h,i} \overset{d}{=} Bernoulli(p_h)$ and $Z_{h',i'} \overset{d}{=} Bernoulli(p_{h'})$ and with $cor(Z_{h,i}, Z_{h',i'}) = c = 0$.
2. $PIM^*_{did} = (1 - p)^2 + p^2 + 2cp(1 - p)$: Dependent and identically distributed (*did*), i.e. $Z_{h,i}, Z_{h',i'} \overset{d}{=} Bernoulli(p)$ and $cor(Z_{h,i}, Z_{h',i'}) = c \neq 0$.
3. $PIM^*_{iid} = (1 - p)^2 + p^2$: Independent and identically distributed (*iid*), i.e. $Z_{h,i}$, $Z_{h',i'} \overset{d}{=} Bernoulli(p)$ and $cor(Z_{h,i}, Z_{h',i'}) = c = 0$.

As for PIM^*, $PIM^*_{inid} \in [0, 1]$. In contrast, it is easy to show by differentiation that $PIM^*_{iid} \geqslant 0.5$ and that $PIM^*_{did} \geqslant 0.5(c + 1)$, which is > 0.5 (resp. < 0.5) for $c > 0$ (resp. $c < 0$).

To return to the plot shown in Figure 3a; the "percentage identical methylation" asymptotes at 40%, or 0.4, which is less than the theoretical lower-bound of PIM^*_{iid}. If we believed that all samples had the same average CpG methylation levels, p_h, then we would conclude that the methylation states of two CpGs separated by a large distance are in fact **anti-correlated** ($c < 0$) for CpGs separated by approximately $1,000$ bp. However, since we know that there is sample-to-sample variability in the average level of CpG methylation, the more likely explanation for this result is that the p_h vary between samples and therefore $PIM^* < 0.5$ are possible, even supposing $c > 0$.

Of course, the above discussion pre-supposes that what is labelled "percentage identical methylation" in Figure 6.1 is mathematically equivalent to PIM^*. I believe this to be true but due to the aforementioned inadequacies of the method description it is not possible to verify this derivation.

Quite apart from what exactly is plotted in Figure 6.1, the analysis is based on data from only $2,524$ amplicons. These limited data cannot give a genome-wide picture of co-methylation. Furthermore, the data come from a variety of tissues and donors, which makes it difficult to know how applicable these results are for a particular tissue. Caution is warranted in extrapolating these results to general statements about the correlation of aggregate methylation levels.

6.2.2 Cokus *et al.* [2008]

Cokus *et al.* [2008] published the BS-seq protocol for performing whole-genome bisulfite-sequencing. They demonstrated this technique by generating a genome-wide map of DNA methylation in wildtype *Arabidopsis thaliana*, a small flowering plant whose methylome is widely studied. They performed an "autocorrelation" analysis of the β-values for CG, CHG and CHH methylation[3]. The autocorrelation computes, for a set of distances ($0 - 5,000$ bp), the (Pearson) correlation coefficient of methylation levels for all pairs of methylation loci separated by each distance. To emphasise, I believe this analysis uses all pairs of methylation loci, that is, regardless of how many intervening methylation loci there are for each pair.

Cokus *et al.* [2008] found that methylation levels are highly correlated, particularly for CG methylation levels, and that the correlation decays as a function of genomic distance between methylation loci. From this they concluded that the "significant correlation between methylated cytosines for distances up to 5,000 nucleotides or more [were] probably a reflection of regional foci of methylation throughout the genome and of large blocks of pericentromeric heterochromatin". They also identified correlations between different methylation contexts, e.g., CG vs. CHG, which "suggest[s] complex interactions between the different types of methylation".

The autocorrelation plots showed two periodicities, one of approximately 167 bp (CpG, CHG and CHH) and the other of 10 bp (CHH only). Both periodicities are visible in plots of the raw autocorrelations and are confirmed by a Fourier analysis of the autocorrelation signal.

The 167 bp period "is similar to, but slightly shorter than, estimates of the average spacing of nucleosomes in plant chromatin" and Cokus et al. hypothesise that this periodicity is due to histones dictating access to DNA by the DNA methyltransferases.

[3] Many plant species, such as *A. thaliana*, have much higher levels of non-CG methylation than do mammals.

The 10 bp period is equal to the length of one helical DNA turn. From this result Cokus et al. hypothesise that the DOMAINS REARRANGED METHYLASE 2 (*DRM2*), "the main enzyme controlling asymmetric methylation in *Arabidopsis*", contains two active sites that methylate sites 8 to 10 bp apart[4], like its mammalian homologue, DNA methyltransferase 3 (*DNMT3*). The 10 bp CHH-only period in the autocorrelation analysis of the β-values is also observed in a *within-fragment analysis* (discussed in Section 6.3)

6.2.3 Lister *et al.* [2009]

Lister *et al.* [2009] report an 8 to 10 bp period in the occurrence of "methylcytosines". Recall that Lister *et al.* define a "methylcytosine" as a cytosine that has a non-zero β-value, as determined by a binomial test (see Section 4.4.2 for details and a criticism of this concept). With its reference to "methylcytosines", this may first appear to be a within-fragment analysis of co-methylation, but it is in fact based on β-values. However, unlike other methods described in this section, the analysis is not based on a correlation of these β-values, but rather on patterns in the spacing of "methylcytosines" along the genome. This leads to problems.

Lister *et al.* [2009] tabulate the number of neighbouring[5] pairs of "methylcytosines" as a function of the IPD. These counts are plotted as a bar chart, with IPD along the x-axis and count on the y-axis, along with a cubic spline fit to the counts. This is done separately for each combination of methylation context (CG, CHG and CHH) and genomic context (exonic, intronic or random) for a total of nine plots (reproduced in Figure 6.2).

An 8 to 10 bp period in the fitted cubic spline is observed in most, but not all, of these nine plots. This periodicity is very weak in several contexts and is absent for exonic CpGs. The weakness of the signal is partially attributable to the small number of "methylcytosines" in the CHG and CHH contexts in human samples. More concerning is that the reported 8 to 10 bp periods are weaker than a 3 bp period for "methylcytosines" at exonic CpGs. Unlike the 8 to 10 bp period, the 3 bp period can clearly be observed in the raw counts,

[4] This is also referred to as "co-methylation" by Jurkowska *et al.* [2011], here meaning the methylation of multiple loci in a single event. Jurkowska *et al.* [2011] review a plausible biochemical model for how *DNMT3a* could co-methylate two CpGs separated by 8 to 10 bp; this model includes both the scenario where both CpGs are on the same DNA strand and the scenario where the CpGs are on opposite strands.

[5] I think that "neighbouring" is what is meant by "non-redundant pair-wise distances" in the caption of Figure 6.2, but I am not certain.

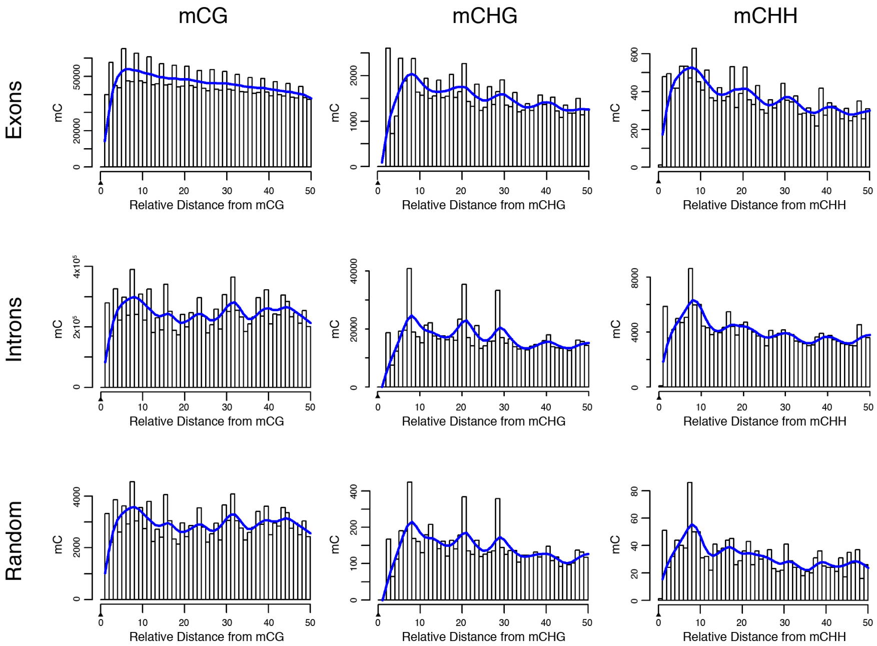


Figure 6.2: Spacing of Adjacent Methylcytosines in Different Contexts. Prevalence of mCHG/mCHH sites (y-axis) as a function of the number of bases between adjacent mCHG/mCHH sites (x-axis) based on all non redundant pair-wise distances up to 50 nt in exons, introns and random sequences. The blue line represents smoothing by cubic splines. Adapted by permission from from Macmillan Publishers Ltd: [Nature Genetics] (Lister *et al.* [2009]), copyright (2009)

although the cubic spline smoother seemingly ignores it. Lister *et al.* do not comment on the 3 bp period.

Based on these nine plots, Lister *et al.* [2009] draw parallels to the 10 bp period in the autocorrelation of CHH β-values observed by Cokus *et al.* [2008] and speculate that "*DNMT3A* may be responsible for catalysing the methylation at non-CG sites". However, I suspect that the results in these nine plots are in fact driven by periodicities in where cytosines are located in the genome rather than any periodicities in the actual methylation of these cytosines. That is not to say that there aren't periodicities in the methylation of cytosines, but that is not what is being measured in this analysis. To explain, these plots use the raw counts of "methylcytosines", which are not adjusted for the frequency of pairs of cytosines with a given IPD. If pairs of cytosines are observed more often at certain IPDs then even if methylation was totally independent of IPD we would expect more pairs of "methylcytosines" at these same IPDs. Therefore, at a minimum, it is necessary to normalise the number of “methylcytosines” by the number of cytosines.

In fact, there are well known periodicities in the distribution of bases in genomes. For example, in eukaryotic genomes there is a 10 bp period of G+C dinucleotides that acts out of phase with a 10 bp period of A+T dinucleotides for nucleosome-bound DNA [Segal *et al.* 2006] and a “well-known period-3 oscillation in coding regions” of the genome [Li 1997]; the former likely explains the 8 to 10 bp period and the latter the 3 bp period for exonic CpGs.

6.2.4 Chodavarapu *et al.* [2010]

Chodavarapu *et al.* [2010] use MNase-seq to study nucleosome positioning and the relationship to DNA methylation, including co-methylation, in *Arabidopsis* and an embryonic human stem cell line (*HSF1*). The authors plot a “weighted average methylation”, which I presume to be a weighted version of the β-values, against the distance from predicted nucleosome start sites, which are obtained via MNase-seq. The nucleosome start sites anchor the methylation observations to a common grid and this plot is used to show that nucleosome bound DNA has a greater level of methylation than DNA that is not bound to nucleosomes. A 10 bp period is readily observed in these plots, consistent with the discovery of 10 bp period in the autocorrelation of β-values published by Cokus *et al.* [2008]. Again, the observed periodicity is confirmed by analysing a Fourier transform of this correlation signal.

In addition to the genome-wide analyses, the same method is used to examine promoters, genes, repeats, euchromatic regions and centromeric regions of the *Arabidopsis* sample, and promoters, genes, repeats and CpG islands of the *HSF1* sample, which confirmed the 10 bp periodicity. In fact, these results revealed that the 10 bp period is common to all methylation types, including symmetric CpG methylation, at least for nucleosome-bound DNA. This led Chodavarapu *et al.* to discard their earlier hypothesis[6] as to the source of the 10 bp period in favour of one that “nucleosomes are to some extent dictating access to the DNA and therefore setting the register of methylation **for all DNA methyltransferases** [emphasis added]” and not just for the DRM2 methyltransferase.

[6] Many of the same authors contributed to Cokus *et al.* [2008] and Chodavarapu *et al.* [2010].

6.2.5 Li *et al.* [2010]

Li *et al.* [2010] compute “the correlation of methylation level of any two nearby CpGs and the relationship between spatial distance (from one CpG to another) and strength of this correlation”. Li *et al.* report that “co-methylation deteriorates over distance and becomes nearly undetectable at distances $> 1,000$ bp”, and conclude that their results are consistent with those in Eckhardt *et al.* [2006]. In addition to this genome-wide analysis, Li *et al.* [2010] perform the same analysis for CpGs within 19 different genomic contexts such as upstream of a gene, untranslated regions, coding DNA sequences and various repetitive elements. Somewhat surprisingly, they do not specifically compare CpG islands to non CpG islands, although some of these differences will be captured by the analysis of regions upstream of a gene, which include many CpG island promoters.

Unfortunately, just as in Eckhardt *et al.* [2006], Li *et al.* [2010] never define exactly what is being computed in their analysis of co-methylation and no software is available. All the co-methylation results are presented as supplementary figures, which lack figure legends and are supported by only very brief captions. So, once again, I have been forced to make my best guess as to what these figures show and how it was computed.

Figure 6.3 plots the “correlation” against “distance of CpG” (0 to 1000), which I believe are estimates of the correlation of aggregate methylation levels over the range 0 to 1000 bp. Based on the “$n = 18,936,995$” quoted in the header of panel (a) of this figure, which I interpret as the number of pairs used in this plot of “correlation” against “distance of CpG” ($0 - 1000$), I believe that what is shown are the correlations of pairs of CpG β-values where the pairs are separated by a common distance and have no intervening CpGs (i.e. $NIL = 0$ pairs). If all pairs were used, n would be much larger since there are some 25 millions CpGs in the haploid human genome. To emphasise, the correlation computed by Li *et al.* [2010] is not the same as the autocorrelation computed in Cokus *et al.* [2008] since Li *et al.* [2010] only use pairs of adjacent CpGs whereas Cokus *et al.* [2008] use all pairs of CpGs, regardless of the number of intervening CpGs.

The plot of genome-wide correlations (panel (a) in Figure 6.3) contains a period of approximately 170 bp, very similar to the 167 bp period identified in *Arabidopsis* by Cokus *et al.* [2008]. This is further evidence that nucleosome spacing effects the strength of

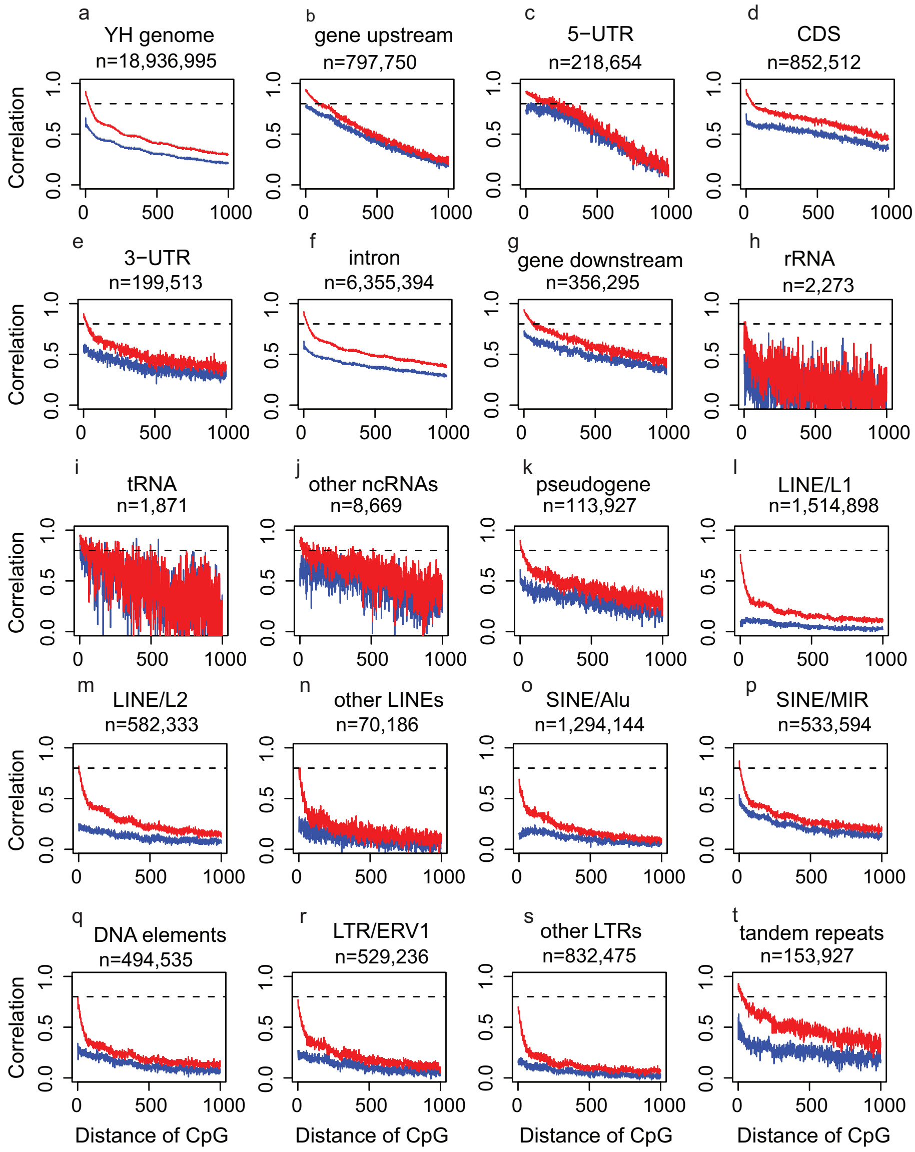


Figure 6.3: Co-methylation patterns for different genomic features. A reproduction of Supplementary Figure 6 from Li *et al.* [2010] under the Creative Common Attribution (CCBY) licence.

co-methylation. However, Li *et al.* [2010] found no evidence of higher frequency periods, such as the 10 bp period reported by Cokus *et al.* [2008] and Chodavarapu *et al.* [2010].

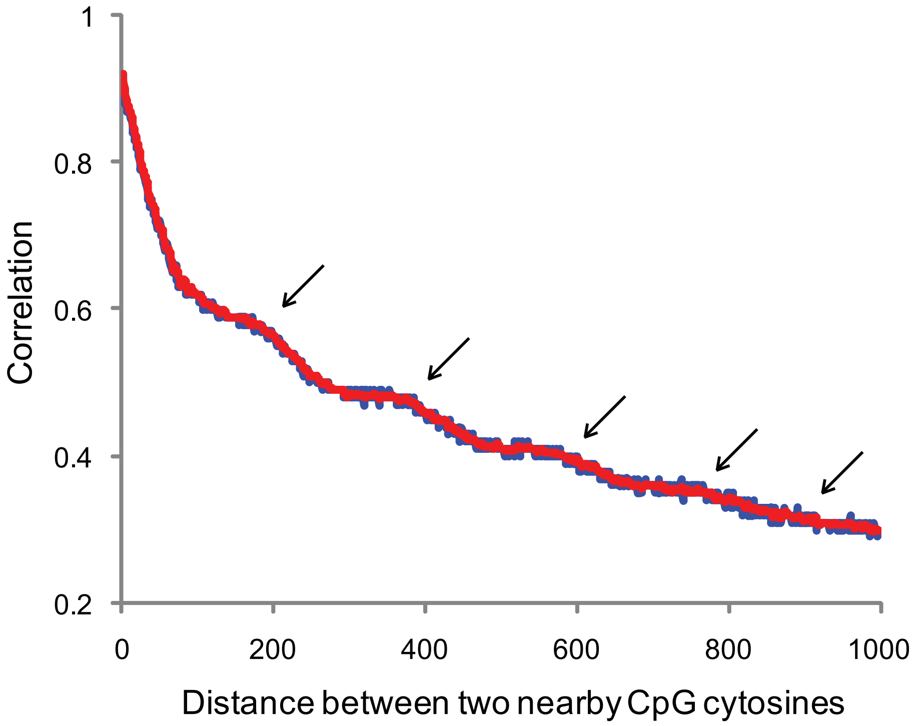


Figure 6.4: Transformation of methylation correlation at nearby CpG cytosines. A reproduction of Supplementary Figure 7 from Li *et al.* [2010] under the Creative Common Attribution (CCBY) licence.

6.2.6 Lyko *et al.* [2010]

In their study of differential methylation in queen and worker honey bees (*Apis mellifera*), Lyko *et al.* [2010] perform an autocorrelation analysis of CpG methylation. They report that the "correlation of methylation status of neighbouring CpGs increases sharply between 1 bp and 20 bp, then drops rapidly between 40 bp and 100 bp, and then slowly fades away". This analysis, however, is potentially flawed.

Figure 6.5 reproduces the results for the co-methylation analysis. Note that the maximum value of the "CpG autocorrelation" is less than 0.015 (panel (A) of Figure 6.5). I suspect that the autocorrelation has been incorrectly computed. Specifically, what I believe has been computed is the autocorrelation of all positions in the genome, not just CpGs, where non-CpGs have been artificially assigned a β-value of zero. It does not make sense to include non-CpG positions, particularly non-cytosines that can never be methylated, in an analysis of CpG co-methylation.

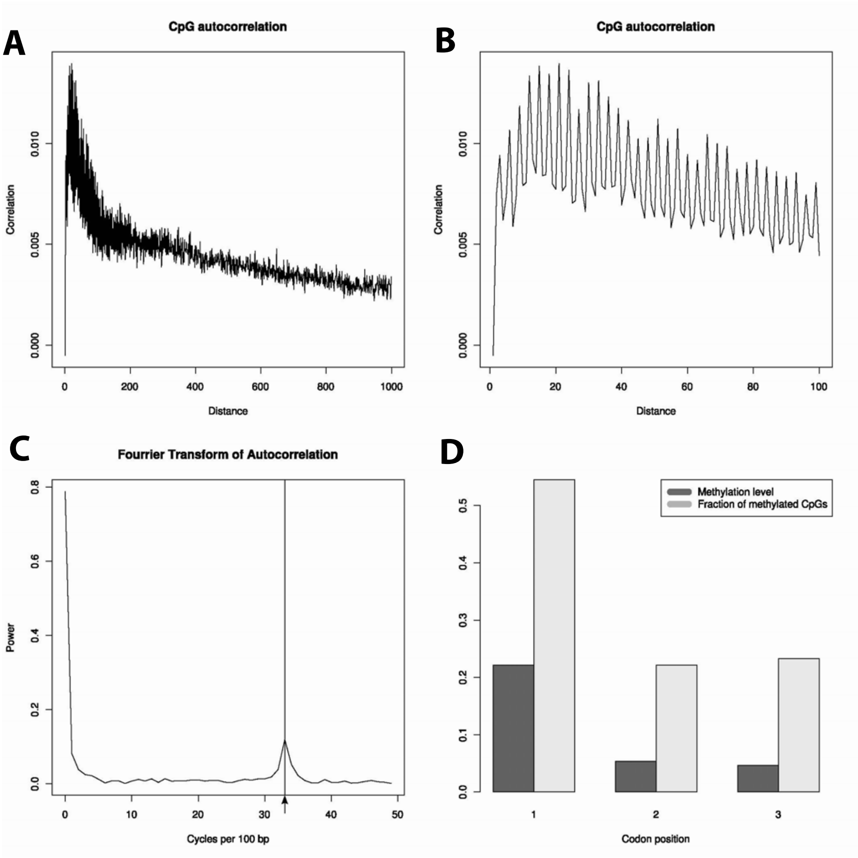


Figure 6.5: Periodicity of methylation patterns. (A) Autocorrelation of CpG methylation status over 1 kb. (B) Autocorrelation over 100 bp. Figures A and B show that the correlation of methylation status of neighboring CpGs increases sharply between 1 bp and 20 bp, then drops rapidly between 40 bp and 100 bp, and then slowly fades away. CpGs within a neighborhood of 2 bp to 100 bp are thus more likely to share the same methylation status than more distant CpGs. (C) Fourrier [sic] transform of autocorrelation showing a clear periodicity peak at 33 cycles per 100 bp (every 3 bp). (D) Distribution of codon position of mCs, and distribution of methylation level depending on the position. These two panels indicate that the distance between methylated CpGs is often a multiple of three and that the methylated cytosine corresponds most frequently to the first nucleotide of an arginine codon. A reproduction of Supplementary Figure 7 from Lyko *et al.* [2010] under the Creative Common Attribution (CCBY) licence.

I do not have the data used in Lyko *et al.* [2010], but the following example supports my belief. This example uses the chromosome 22 data from the *FF* sample in the *Lister* dataset (see Section 3.2 for full details of this sample).

Let $\boldsymbol{\beta}^*$ be the augmented vector of β-values, where:

$$\beta^*_{i'} = \begin{cases} 0 & \text{if } i' \text{ is not a CpG} \\ \beta_i & \text{if } i' \text{ is a CpG} \end{cases}$$

Figure 6.6 is a plot of the autocorrelation of the augmented vector, $\boldsymbol{\beta}^*$ for the chromosome 22 data from the *FF* sample. We see that it is remarkably similar to that of panel (A) in Figure 6.5; the correlation first increases then decreases and the maximum autocorrelation is very small, close to zero.

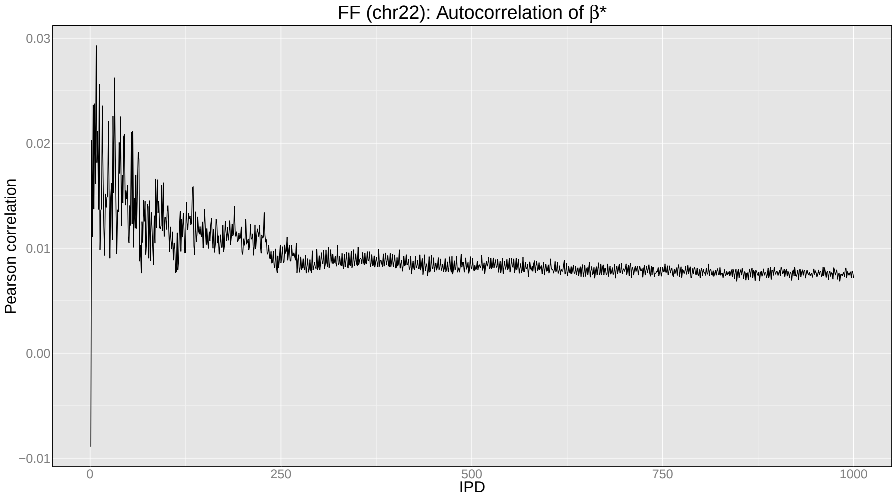


Figure 6.6: Autocorrelation of $\boldsymbol{\beta}^*$ for the chromosome 21 data of the *FF* sample.

On the one hand, this suggests that this may be the cause of the negligible autocorrelations in Figure S7 of Lyko *et al.* [2010]. On the other hand, supposing that the maximum autocorrelation of the β-values is truly less than 0.015, then this suggests that co-methylation itself is negligible in honeybees, in contrast to results from other organisms.

The effort required to process the Lyko dataset, and the insufficient detail in the paper to facilitate the reproducibility of the analysis, mean that I decided not to invest further effort into exploring potential interpretations of the co-methylation analysis of Lyko *et al.* [2010].

6.2.7 Lacey *et al.* [2013]

Lacey *et al.* [2013] explore co-methylation in their paper on modelling and analysing reduced representation bisulfite-sequencing data (RRBS). To do so, Lacey *et al.* fit a

Gaussian variogram model to the CpG β-values from chromosome 11 of a normal myotube cell line (*MTCTL2*). The empirical and fitted variograms asymptote for CpGs separated by approximately 3000 bp (Figure 1d of Lacey *et al.* [2013]; I am unable to reproduce the figure here due to rights restrictions) and so they conclude that the methylation level of CpGs "show a strong correlation for sites in close proximity, decaying to near independence at distances beyond 3000 bp".

Caution is warranted in extrapolating these results to statements about genome-wide co-methylation. Firstly, this analysis uses only approximately $60,000$ CpGs[7], far fewer than the 25 million CpGs in the haploid human genome. Secondly, these data are from a single sample from a single tissue, which cannot capture the sample-to-sample nor tissue-to-tissue variability in co-methylation. Finally, RRBS targets CpG dense regions of the genome with low levels of methylation and so these data do not capture co-methylation outside of these regions, where the majority of the CpGs are found and which are highly methylated.

6.3 Within-fragment co-methylation

As discussed in Section 2.4.2, bisulfite-sequencing allows for the study of methylation patterns at m-tuples from individual DNA fragments. Since each read comes from a single haplotype from a single cell[8], these data can be used to study co-methylation at the level of individual DNA molecules. I argue that this is the more biologically meaningful level at which to study co-methylation, since this is the same scale as the physical process that lays down and maintains DNA methylation.

I call this *within-fragment co-methylation*, where 'fragment' refers to a fragment of DNA. More precisely, within-fragment co-methylation is the dependence of DNA methylation at methylation loci that occur on the same DNA fragment. It can be thought of as the dependence structure of the binary stochastic process $\mathbf{Z_h} = (Z_{h,1}, Z_{h,2}, \ldots, Z_{h,N_{loci}})$. There is less previous research studying within-fragment co-methylation than there is studying

[7] The data for MTCTL2 are not publicly available so I estimated the number of sequenced CpGs on chromosome 11 from three comparable myoblast RRBS libraries (samples `wgEncodeHaibMethylRrbsHsmmtubefshdDukeSitesRep1`, `wgEncodeHaibMethylRrbsHsmmtubefshdDukeSitesRep2`, and `wgEncodeHaibMethylRrbsHsmmtubefshdDukeSitesRep3` available from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/`).

[8] Here I ignore the possibility of chimeric reads that may be artificially produced by DNA fragments ligating to one another during the library preparation or sequencing process.

correlations of aggregate methylation levels. Chapter 7 extensively explores within-fragment co-methylation and how different variables affect this dependence, such as the distance between methylation loci and the genomic context of the methylation loci.

If we could sequence entire chromosomes by single reads, then we could analyse co-methylation between any set of methylation loci. However, the most commonly used technology for high-throughput bisulfite-sequencing, which is based on Illumina's sequencing technology, can only generate reads that span approximately 200 to 250 bp. We are therefore limited to studying within-fragment co-methylation between methylation loci separated by at most 200 to 250 bp[9]. Long-read sequencing technologies (e.g., Pacific Biosciences, Roche 454 and Oxford Nanopore) would overcome this limitation but the lower throughput of these technologies, and the reliability with which they can identify methylcytosines, create new complications. An alternative solution, although completely hypothetical, is to use a technique that generates synthetic long reads, similar to those generated by Illumina's TruSeq Synthetic Long-Read technique[10].

To summarise, current technologies mean that we are limited to studying within-fragment co-methylation at loci that are within regions of approximately 200 to 250 bp.

6.3.1 Cokus *et al.* [2008]

As discussed in Section 6.2.2, Cokus *et al.* [2008] identify a 10 bp period in the autocorrelation of β-values for CHH methylation from a whole-genome bisulfite-sequencing experiment of *Arabidopsis thaliana*. They also report that they find this same period "when individual reads are examined directly".

The sequencing technology at the time of this publication produced very short reads, on average only 31 bp of usable sequence. For the within-fragment co-methylation results, Cokus *et al.* first identify all reads containing multiple CHH loci. Then, for a pair of CHH loci in the same read, they estimate the probability that the second CHH in the pair is

[9] NB assembling these reads into longer contigs will not solve this problem since the assembly does not guarantee that the reads assembled into longer fragments actually originated from the same cell or haplotype.

[10] My understanding is that Illumina's TruSeq Synthetic Long-Read technique cannot be directly applied to bisulfite-sequencing for a number of technical reasons, including the fragility of large DNA fragments treated with sodium bisulfite.

methylated given that the first CHH in the pair is methylated. This is estimated by simply counting the number of times this occurs and dividing by the number of pairs of CHH loci that occur within the same read[11]. They estimate this probability separately for each distance between CHH loci and separately for each of the five *Arabidopsis* chromosomes. Cokus *et al.* then plot the average across the five chromosomes as a function of the number of bp between the CHH loci. The plot includes pointwise 95% confidence intervals, computed under an assumption of normality, and a running average of these averages. This plot, reproduced in Figure 6.7, shows evidence of an approximately 10 bp period, supporting the interpretation of the results observed in the autocorrelation of the β-values.

I consider these results as based on a restricted definition of within-fragment co-methylation. To explain, these results tell us nothing about the probability that both CHH loci are *unmethylated* nor the joint probability distribution of methylation for two CHH loci on the same fragment; they tell us only about the probability that both CHH loci are methylated.

Specifically, Cokus *et al.* [2008] only estimate the marginal probability that a CHH is methylated, $Pr(Z_{h,i} = 1)$, and the probability that two CHH loci on the same fragment are both methylated, $Pr(Z_{h,i'} = 1|Z_{h,i} = 1)$, where $i < i'$. From these two quantities we can only estimate two of the four joint probabilities, $Pr(Z_{h,i} = x, Z_{h,i} = y), x, y \in \{0, 1\}$, namely, $Pr(Z_{h,i} = 1, Z_{h,i'} = 1)$ and $Pr(Z_{h,i} = 1, Z_{h,i'} = 0)$. It does not allow estimation of $Pr(Z_{h,i} = 0, Z_{h,i'} = 0)$, nor of $Pr(Z_{h,i} = 0, Z_{h,i'} = 1)$, since these require estimates of $Pr(Z_{h,i'} = 0|Z_{h,i} = 0)$. While it is trivial to additionally estimate $Pr(Z_{h,i'} = 0|Z_{h,i} = 0)$, I argue in Section 7.3 that it is more useful and simpler to estimate the odds ratio, $\psi = \frac{Pr(Z_{h,i}=1,Z_{h,i'}=1)Pr(Z_{h,i}=0,Z_{h,i'}=0)}{Pr(Z_{h,i}=1,Z_{h,i'}=0)Pr(Z_{h,i}=0,Z_{h,i'}=1)}$, rather than the conditional probabilities, $Pr(Z_{h,i'} = 1|Z_{h,i} = 1)$ and $Pr(Z_{h,i'} = 0|Z_{h,i} = 0)$.

Furthermore, by aggregating all pairs of CHH loci separated by a common distance and then estimating these probabilities, there is an implicit assumption that all such pairs on the same chromosome have the same conditional probability, $Pr(Z_{h,i'} = 1|Z_{h,i} = 1)$. There is good reason to suspect that this is not true and that there exists significant

[11] A read may contain multiple pairs of CHH loci and it is not clear whether Cokus *et al.* consider all pairs of CHH loci or only the first pair of CHH loci in a read. The number of pairs of CHH loci that occur within the same read is not the same as the number of reads containing multiple CHH loci. Either method should give similar results since there will be few 31 bp reads containing multiple pairs of CHH loci.

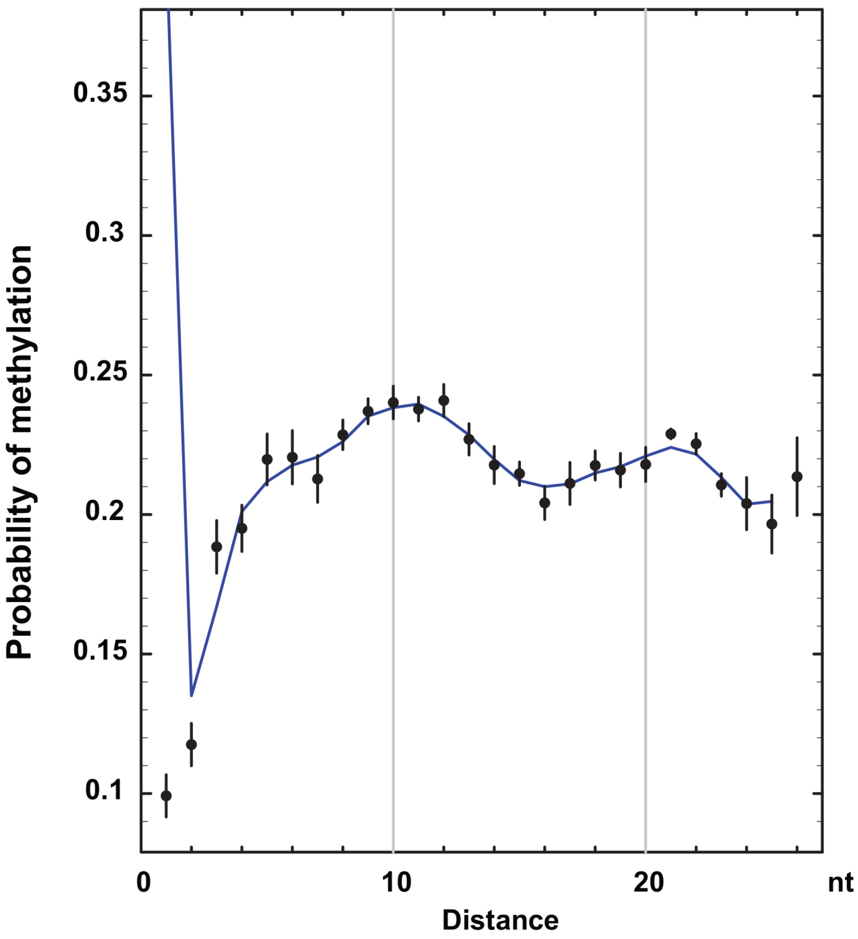


Figure 6.7: Within-read probability of additional methylation of CHH sites within a given distance from a methylated CHH site. Data were derived from individual BS-Seq reads. The x-axis indicates the distance between the two cytosines. The y-axis indicates the probability of methylation of CHH sites within the given distance from another methylated CHH site. Each point is the mean value from averaging the probability from each of the five Arabidopsis chromosomes, and the blue line is a running average of these mean values. Error bars represent 95% confidence intervals via critical values of Student *t* distributions. Adapted by permission from from Macmillan Publishers Ltd: [Nature Genetics] (Cokus *et al.* [2008]), copyright (2008).

pair-to-pair variability. For example, we know for CpGs that those within CpG islands have a vastly different probability of being methylated than do those outside of islands; it is reasonable to suspect that this is also true for CHH methylation in different regions of the *Arabidopsis* genome and hence for the co-methylation of these same loci. This aggregation may also introduction artificial associations into the 2×2 contingency table from which these probabilities are estimated [Good and Mittal 1987].

6.3.2 Ball *et al.* [2009]

Ball *et al.* [2009] use bisulfite padlock probes with 36 bp single-end reads to profile CpG methylation at approximately $7,000$ CpGs in an Epstein-Barr virus transformed B-lymphocyte cell line. As part of their study, Ball *et al.* sought to investigate whether the co-methylation reported by Eckhardt *et al.* [2006] occurs at the single-molecule level.

To do so, Ball *et al.* took all reads containing multiple CpGs and mapping to positions with intermediate β-values ($0.2 < \beta < 0.8$) with at least $100\times$ sequencing coverage. From these reads they computed the Pearson correlation of within-read methylation states. Ball *et al.* found that the distribution of these correlations is centred around 0.5 with very few values below zero, evidence that methylation levels at CpGs on the same DNA fragment "are generally positively correlated".

The obvious limitation to this analysis is the small number of CpG pairs interrogated by the assay. The short reads also only allow for the exploration of within-fragment co-methylation at very close CpGs and Ball *et al.* do not explore the effect of distance between CpGs (IPD) on co-methylation.

6.3.3 Lacey and Ehrlich [2009]

Lacey and Ehrlich [2009] develop stochastic models of DNA methylation replication during mitosis (cell division) for a single double-stranded DNA molecule in humans. The most complex model that Lacey and Ehrlich consider, called the *neighbouring sites model*, uses a one-step Markov random field. Under this model, the methylation state of one locus, $Z_{h,i}$, depends on the methylation states of the adjacent loci, $Z_{h,i-1}$ and $Z_{h,i+1}$. The physical distances between these neighbours, the IPDs, are not considered.

The neighbouring sites model requires an analysis of within-fragment co-methylation in order to estimate the transition probabilities. They estimate these parameters from two small regions of 10 human ovarian carcinoma samples and a pool of somatic controls, each region approximately 200 bp in length and containing 13 and 14 CpGs, respectively. These regions were sequenced using hairpin-bisulfite PCR, an assay that measures DNA methylation on both strands of a double-stranded DNA molecule[12]. For each region, 12 to

[12] This is unlike traditional bisulfite-sequencing, which only measures methylation along a single DNA

16 clones were sequenced per sample[13]

For each of these regions, Lacey and Ehrlich found that the methylation level of a CpG strongly depends on that of its neighbours. In other words, they found significant within-fragment co-methylation. They also found that the overall level of methylation of the CpGs in a region influenced the strength of the within-fragment co-methylation.

In summary, this analysis uses high resolution measurements from a moderate number of clones from a moderate number of samples for a very small fraction of the human genome. While the methods and models are quite interesting, the data are limited to such a small region that it is difficult to know whether these results give an accurate genome-wide picture of within-fragment co-methylation.

6.3.4 Landan *et al.* [2012]

Landan *et al.* [2012] is notable for its use of the within-read information available from bisulfite-sequencing data. The authors explore the within-sample heterogeneity of methylation patterns at 4-tuples and 6-tuples of CpGs from ultra-deep bisulfite-sequencing (the median sequencing coverage of each CpG in the target is greater than $10,000\times$) of 45 cancer-related CpG islands and study how methylation dynamics at this regions evolves over time.

As part of their analysis, Landan *et al.* construct "methylation linkage diagrams" based on a Pearson correlation estimate of within-fragment co-methylation. For each CpG 2-tuple in their targeted regions, they create a 2×2 contingency table. The general form of the table is shown below in Table 6.1. They then compute the Pearson correlation of each table[14], $r = \frac{n_{mm}n_{uu} - n_{mu}n_{um}}{\sqrt{n_{u+}n_{m+}n_{+u}n_{+m}}}$. Importantly, r is based on within-read measurements from a single sample and so these r are estimates of within-read, within-sample co-methylation.

So-called "methylation linkage diagrams" are created by plotting the Pearson correlation of each 2-tuple, $r_{(i,i')}$ as a heatmap (*à la* heatmaps of linkage disequilibrium) to identify blocks of CpGs with significant co-methylation. It should be noted that the concepts of

strand.

[13] Essentially, $12 - 16$ reads were generated per sample.

[14] The formula I give for r is also known as the ϕ coefficient of a 2×2 table. The ϕ coefficient of a 2×2 table of binary variables, such as Table 6.1, is mathematically equivalent to the Pearson correlation coefficient of the same data.

Table 6.1: Summary of read counts at a CpG 2-tuple. For example, n_{um} is the number of reads methylated at the first CpG and unmethylated at the second CpG. We use the 'plus' notation to denote the sum over the index it replaces, e.g., $n_{u+} = n_{uu} + n_{um}$.

		Second CpG		
		Unmethylated	Methylated	Total
First CpG	Unmethylated	n_{uu}	n_{um}	n_{u+}
	Methylated	n_{mu}	n_{mm}	n_{m+}
	Total	n_{+u}	n_{+m}	$n = n_{++}$

'subject' and 'population' are rather different for "methylation linkage" than they are for genetic linkage (see Table 6.2).

Table 6.2: Notions of 'subject' and 'population' in traditional genetic linkage and the "methylation linkage" of Landan *et al.* [2012]. NB 'person' could be any organism, e.g, 'mouse', 'dog', etc.

	Subject	Population
Genetic linkage	Person	Population
Methylation linkage	(Haploid) genome of cell	Person

From the methylation linkage analysis of these 45 regions, Landan *et al.* conclude that:

> "[the] correlation between the methylation states of pairs of CpGs was generally very low. This lack of correlation suggests that methylation dynamics are typically independent for different CpGs, making the methylation state of one CpG (whether high or low) uninformative on the methylation state of nearby CpGs."

As we will see in Chapter 7, this result is contradicted by my analyses of within-fragment co-methylation using whole-genome data.

6.4 Between-sample co-methylation

As is now abundantly clear, there are multiple definitions of within-sample co-methylation. An almost orthogonal concept is between-sample co-methylation, the correlation of methylation levels for a single pair of loci between a set of samples.

Since it is a loci-specific, population-level correlation, between-sample co-methylation is somewhat analogous to the idea of linkage disequilibrium between genetic loci. Briefly, for two loci, (i, i'), the correlation of aggregate methylation levels is computed across samples using, for example, the Pearson correlation coefficient, $r = cor(\boldsymbol{\beta_i}, \boldsymbol{\beta_{i'}})$, where $\boldsymbol{\beta_i} = (\beta_{i,1}, \ldots, \beta_{i,N_{samples}})$.

The concept of between-sample co-methylation leads to the attractive idea of identifying "methylation tag-loci", whose methylation level is a good proxy for the methylation level across a larger region. These tag-loci could act like tag-SNPs used in genome-wide association studies, which would allow researchers to assay a small subset of loci in the methylome and capture much of the methylation variability. This idea has also been explored by Barrera and Peinado [2012] with respect to the use of *HpaII* sites as proxies for the methylation level of CpG islands.

Unfortunately, unlike linkage disequilibrium, we cannot assume that between-sample co-methylation is identical between different tissues from the same individuals since DNA methylation, unlike DNA sequence, varies between tissues. Therefore, at best, we might hope to generate tissue-specific maps of between-sample co-methylation.

In order to leverage between-sample co-methylation in this way, we also need to know the size of these co-methylation blocks, that is, the length of the regions over which between-sample co-methylation is sufficiently strong. The longer these blocks are, the more sparsely we can afford to sample individual methylomes.

Bell *et al.* [2011] identify evidence of between-sample co-methylation using the Illumina 27k microarray to measure DNA methylation in lymphoblastoid cell lines from 77 HapMap Yoruba individuals. Of note, they found that between-sample co-methylation decays as a function of the distance between CpGs (IPD) and is stronger within CpG islands than outside of CpG islands. However, it must be recognised that the Illumina 27k microarray measures DNA methylation at less than 1% of all CpGs as is biased towards CpG islands.

Another example of between-sample co-methylation is given by Liu *et al.* [2014]. The authors compute the correlation of pairs of β-values between samples in three studies ($n = 247, n = 91$ and $n = 305$, respectively) to identify clusters of correlated CpGs and relate methylation changes at these clusters to nearby SNP genotypes. The resolution of their data is better than Bell *et al.* [2011] since they use the Illumina 450k microarray, however it still only captures less than 2% of all CpGs in the genome. Liu *et al.* conclude that "DNA methylation is correlated over regions with a median length of 274 bp from our data set". They also note that, "methylation codependency [co-methylation] itself varies across the genome", as is to be expected.

A simple way to estimate between-sample co-methylation is to create a matrix of β-values, $\boldsymbol{\beta}$, where the (i, j) entry is $\beta_{i,j}$. Then, compute $R = cor(\boldsymbol{\beta})$, whose (k, l) entry is $cor(\boldsymbol{\beta}_{\cdot,k}, \boldsymbol{\beta}_{\cdot,l})$, where $\boldsymbol{\beta}_{\cdot,k}$ is the column vector of β-values for the k^{th} methylation loci. This is what the R package, `coMET` [Martin *et al.* 2015], does to estimate the between-sample co-methylation of Illumina HumanMethylation450 array data. As far as I am aware, `coMET` is the only software specifically designed to estimate between-sample co-methylation and integrate this with other genomic annotations.

6.5 Leveraging co-methylation in downstream analyses

The variety of definitions, rigour, and reproducibility of the published analyses of co-methylation, not to mention the lack of software implementing the proposed methods, has made it difficult to properly account for co-methylatoin in downstream analyses. Ideally, co-methylation would not only be *accounted for*, but *leveraged*, in downstream analyses of DNA methylation data. In fact, some authors have already sought to do this.

For example, the R/Bioconductor packages `bsseq` [Hansen *et al.* 2011, 2012] and `BiSeq` [Hebestreit *et al.* 2013] both cite Eckhardt *et al.* [2006] as providing evidence of co-methylation, which can be leveraged by spatial smoothing of the β-values prior to testing for differential methylation. Specifically, by assuming that CpGs close to one another have similar methylation levels, the smoothed β-values are, on average, more accurate than the raw β-values. Both `bsseq` and `BiSeq` smooth β-values on a per-sample basis.

Smoothing is particularly powerful for experiments with low sequencing coverage since the error in estimating the raw β-values is inversely proportional to the sequencing coverage and smoothing reduces this error. An important parameter when smoothing is the choice of window size over which to smooth, also known as the *bandwidth* of the smoother. The default window size in `bsseq` is defined as one that contains at least 70 CpGs and is at least

2000 bp wide. BiSeq uses a much smaller default window size of 80 bp, a decision that is driven by its focus on RRBS data which has a higher CpG density than whole-genome bisulfite-sequencing data. Ideally, a data-driven bandwidth would be used, one driven by the strength of co-methylation in the region, but this is difficult to implement in practice. A simpler but potentially useful hybrid would be to define different window sizes for genomic features with generally different co-methylation structures, e.g., CpG islands vs. the rest of the genome (see Chapter 7).

In addition to smoothing the β-values, BiSeq also estimates and tries to account for the spatial correlation of test statistics when testing for differential methylation at neighbouring loci. This is not the same as looking for correlations between the β-values, but is rather based on looking for correlations amongst the *differences* in β-values, since these differences are what the test statistics of differential methylation are based upon. Specifically, BiSeq fits a semivariogram to the test statistics of differentially methylated CpGs. The correlations estimated using the semivariogram are in turn used in estimates of the standard deviations when identifying differentially methylated regions.

More generally, methods such as SLIM [Wang *et al.* 2011] and comb-p [Pedersen *et al.* 2012] attempt to account for the correlation amongst test statistics in genomics data. Rather than analysing the distribution of the test statistics themselves or the underlying data, they do this by analysing the distribution of the resulting P-values. This makes these methods quite general, although application-specific methods may be more powerful. comb-p has been used in the analysis of DNA methylation data from the Illumina microarray platforms [Pedersen *et al.* 2012].

These methods leveraging correlations amongst the test statistics or their P-values are not directly based on co-methylation estimates. However, what drives the correlations amongst the test statistics is a complex mixture of within-sample and between-sample co-methylation.

Aclust [Sofer *et al.* 2013] is notable for its use of between-sample co-methylation to identify differentially methylated regions. Unlike other methods for detecting DMRs, which test for DMCs and then cluster these DMCs into DMRs, Aclust clusters loci prior to testing for differential methylation. That is, Aclust tests for DMRs and not DMCs. This has the advantage of reducing the dimensionality of the data. Clusters are formed using a modified form of agglomerative nested clustering, where the modification restricts clusters to comprise neighbouring loci. The distance metric used in the clustering is based on an estimate of the between-sample co-methylation for each pair of adjacent loci.

Capra and Kostka [2014] also make use of between-sample co-methylation, albeit a specialised form of it, to study the dynamics of DNA methylation in differentiating cell lineages. For each locus there is a correlation between its methylation state in the precursor cell and its dependent cell types, which they call "vertical" correlations (in contrast to the "horizontal" correlations of within-sample co-methylation). In this specialised setting, where 'between-samples' means "between samples in a lineage", they show that a method based on "vertical" correlations is better at imputing the methylation level of CpGs than a method based on "horizontal" within-sample co-methylation.

6.6 Summary

While it has long been observed that DNA methylation at nearby loci is highly correlated, pinning down exactly what this means, and its implications, is surprisingly difficult. As I have shown, there are problems and limitations with most, if not all, previous analyses. These problems range from a lack of methodological detail and limited data, through to serious flaws in the methodology or its implementation. Moreover, despite numerous papers including an analysis of co-methylation, few of these have includes software to reproduce their method, further complicating the assessment of these methods.

It is clearly inappropriate to treat methylation at nearby loci as independent and methods have been proposed to account for or leverage this in various ways. These methods would be better informed by a more rigorous and reproducible analysis of co-methylation, which I give in Chapter 7.

Chapter 7

Co-methylation

Overview

This chapter describes two measures of within-sample co-methylation, *within-fragment co-methylation* and *correlation of β-values*. In order to analyse within-fragment co-methylation, we first study methods for estimating meaningful summaries of dependence in sparse $2 \times 2 \times K$ contingency tables.

Using estimates of both within-fragment co-methylation and correlation of β-values, we explore how co-methylation is affected by the distance between methylation loci and the genomic context of the methylation loci. We apply these methods to 40 whole-genome bisulfite-sequencing samples to study how co-methylation varies between different samples.

This chapter focuses on co-methylation of CpGs, although the methods described are also applicable to non-CpG methylation loci.

7.1 Correlations of β-values

Chapter 6 documented several previous analyses based on 'correlations' of aggregate methylation levels, such as β-values. What was lacking from these were clear descriptions of the proposed methods and software implementations. In this chapter I try to rectify the former and with the `MethylationTuples` I make available a software implementation of these methods.

7.1.1 Methods

Here I describe two simple methods based on correlations of β-values. Both of these are implemented in the `methLevelCor()` function that is part of the `MethylationTuples` software.

The aim of this analysis is to address the question of 'how are aggregate methylation levels correlated as a function of the distance between loci?'. There are two steps in this analysis:

1. Define the pairs of methylation loci.
2. For each intra-pair distance (IPD), compute the correlation of β-values for all pairs of methylation loci separated by that IPD.

The second step may be further stratified by the genomic context of the pair of CpGs, such as whether they are in a CpG island, or other variables of interest.

The same analysis could be performed using $\mathcal{M}$-values rather than β-values, and this is available as an option in the `methLevelCor()` function.

Constructing the pairs

I consider two different strategies for creating pairs of CpGs. Both strategies are based on CpGs in the reference genome after having filtered out those reference-specific CpGs. That is, I begin with $\mathcal{I}^{ref}$ and, based on the `Bis-SNP` output, I filter out loci that are not CpGs in the sample's genome. Rather than actually remove these sites, I simply set the corresponding β-value to `NA`. I also set the β-values of 'missing' CpGs, those without sufficient sequencing coverage to call a β-value, to `NA` so that these are still included in the set of CpGs.

The first strategy creates all pairs of adjacent CpGs on the same chromosome and strand, which are then stratified by IPD and any secondary variables, such as genomic context. For a chromosome with N_{loci} CpGs there are $(N_{loci} - 1)$ pairs. I call this the 'adjacent pairs' or $NIL = 0$ strategy, since all pairs have no intervening loci (NIL = number of intervening loci)[1].

[1] Recall that this is with respect to the reference genome and not the sample genome.

The second strategy uses all pairs of CpGs on the same chromosome and strand, which are then stratified by IPD and any secondary variables, such as genomic context. In practice, the second strategy only uses pairs from a set of IPDs, e.g., $IPD = 2$ to 2000 bp, otherwise the number of pairs becomes unwieldy. This is what is called an 'autocorrelation' analysis of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{N_{loci}})$ by Cokus *et al.* [2008], although it is somewhat different to the traditional notion of an autocorrelation analysis. I refer to this as the $NIL \geqslant 0$ strategy. Using the $NIL \geqslant 0$ strategy, two pairs of CpGs with an identical IPD may have a very different number of intervening CpGs.

The interpretation of correlations computed using the $NIL \geqslant 0$ is complicated because the set of pairs are more heterogeneous than those under the $NIL = 0$ strategy. For this reason I prefer the $NIL = 0$ strategy, however, it is useful to consider and contrast the two strategies, which is what I have done in what follows.

Stratifying by a genomic feature

Two pairs of CpGs with an identical IPD may come from two regions, even when using the $NIL = 0$ strategy. These regions may have very different methylation dynamics. In other words, the vector of β-values for a given sample is highly non-stationary; the correlation between two loci depends on more than just the IPD. An obvious variable to investigate is how strong an influence the genomic context has on these correlations of β-values. To do this, we can stratify our analysis by a genomic feature. While this does not eliminate this heterogeneity, it may help identify factors that drive some of the variation.

For any genomic feature, a pair of loci may be inside, outside or spanning the boundary of the feature. Figure 7.1 illustrates the 'feature status' of some pairs of loci. Note that a pair where each locus is in a distinct feature from the same class is declared to be inside the feature, i.e. elements need not be in the same feature but rather in the same type of feature. The genomic feature should partition the genome so that each locus is either inside or outside of the feature. Since the number of 'spanning pairs' is substantially fewer than those inside or outside of the feature, and the boundaries of the feature are oftentimes fuzzy, I have excluded these 'spanning pairs' in the results shown below.

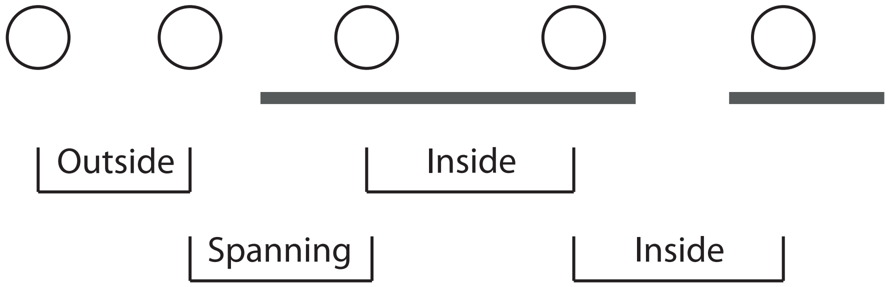


Figure 7.1: Schematic illustrating the 'feature status' of four pairs of methylation loci. The circles represent loci and the grey bars a genomic feature, such as CpG islands. Each pair's feature status is determined by whether both loci are outside of the feature (outside), one locus is inside and one outside of the feature (spanning), or both inside the feature (inside). Note that the rightmost pair is 'inside' even though the two loci that comprise the pair are in different elements of the genomic feature.

Choice of correlation coefficient

Three commonly used correlation coefficients are Pearson's r [Pearson 1895], Spearman's ρ [Spearman 1904] and Kendall's τ [Kendall 1938]. Briefly, Pearson's product-moment correlation is designed to detect linear relationships between two variables. However, it is not robust to outliers, which motivates the use of Spearman's and Kendall's rank-based correlation coefficients. Spearman's ρ is simply the Pearson product-moment correlation of the ranks of the data while Kendall's τ is based on the relative frequency of concordantly ranked pairs.

Owing to the size of whole-genome bisulfite-sequencing data, I have chosen to only compute Pearson's r and Spearman's ρ, which are much faster to compute than Kendall's τ. The `methLevelCor()` function in the `MethylationTuples` package will compute a confidence interval, 95% by default, for the Pearson correlation. This feature is not yet available when using Spearman's or Kendall's correlation coefficient.

Interpretation of correlation

Before computing correlations and trying to interpret these, it is a good idea to look at the data from which these correlations are computed. Figure 7.2 shows kernel density smoothed scatterplots for pairs of β-values separated by $IPD = 2, 20, 200$ or 2000 bp with $NIL = 0$ or $NIL \geqslant 0$ from the *ADS* sample from the *Lister* dataset. It highlights two major problems.

Firstly, the bimodality of the distribution of β-values means that the points concentrate in the corners of the plot: $(0,0),(0,1),(1,0)$ and $(1,1)$. A correlation coefficient is not well suited to summarising these distributions of points. Secondly, when using the $NIL = 0$ strategy, as the IPD increases, the number of points in each scatterplot decreases and so the correlations are very unstable.

In summary, the correlations of pairs of β-values should be cautiously interpreted, bearing in mind that these are somewhat crude summaries that hide many details of the underlying scatterplots.

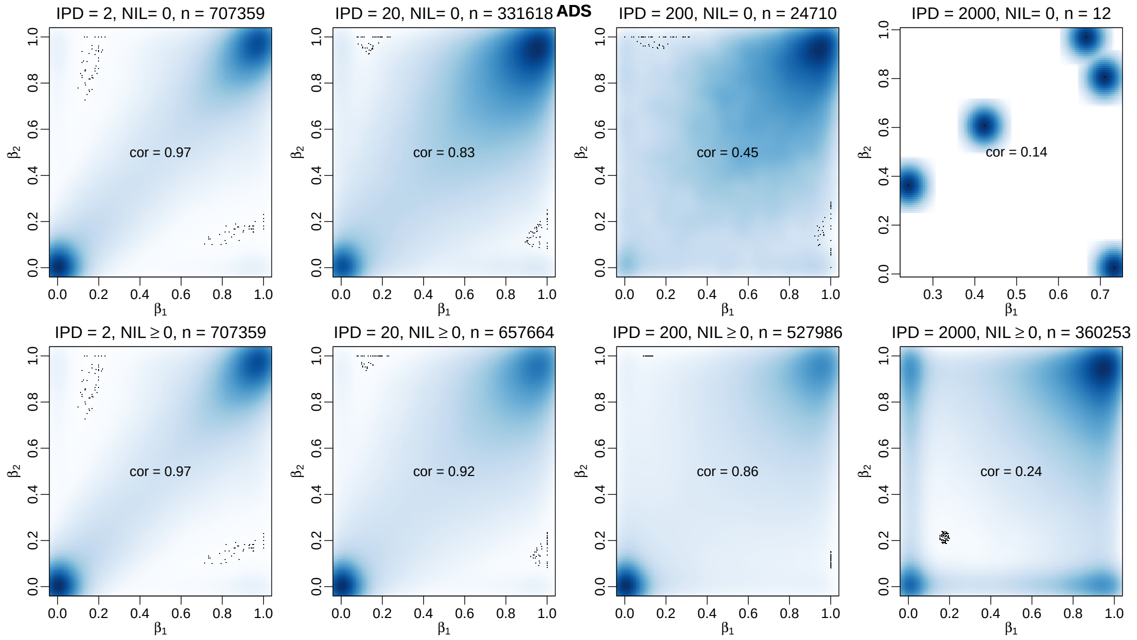


Figure 7.2: Kernel density smoothed scatterplots for pairs of CpG β-values, (β_1, β_2), from the ADS sample. β-values are computed using strand-collapsed counts. n is the number of pairs in each scatterplot. The Spearman correlation of each scatterplot is also reported. Individual plots created using the `smoothScatter()` function in R.

7.1.2 Results

For all 40 samples in the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets, I compute the Pearson and Spearman correlations using both the $NIL = 0$ and $NIL \geqslant 0$ strategies. I compute these genome-wide and after stratifying CpG pairs by whether they are in a CpG island. These analyses are performed separately for each strand and also for strand-collapsed β-values. The plots of these analyses show the raw estimates as semi-transparent points and are overlaid with a loess fit to these points.

The effect of strand

We have seen that the CpG β-values are very highly correlated across strands for most samples when the minimum sequencing coverage is at least 5× ($r = 0.8$ to 0.9, Section 4.4.1). While there are some notable exceptions (the embryo-derived and embryonic stem cell samples), I have opted to collapse by strand in most of the analyses in Section 7.1.2 in order to simplify the plots.

Figures 7.3, 7.4, 7.5 and 7.6 are the only plots to stratify by strand and show the Spearman correlations of β-values for $NIL = 0$ pairs of CpGs. These correlations are very similar across strands, even for those embyro-derived and embryonic stem cell samples.

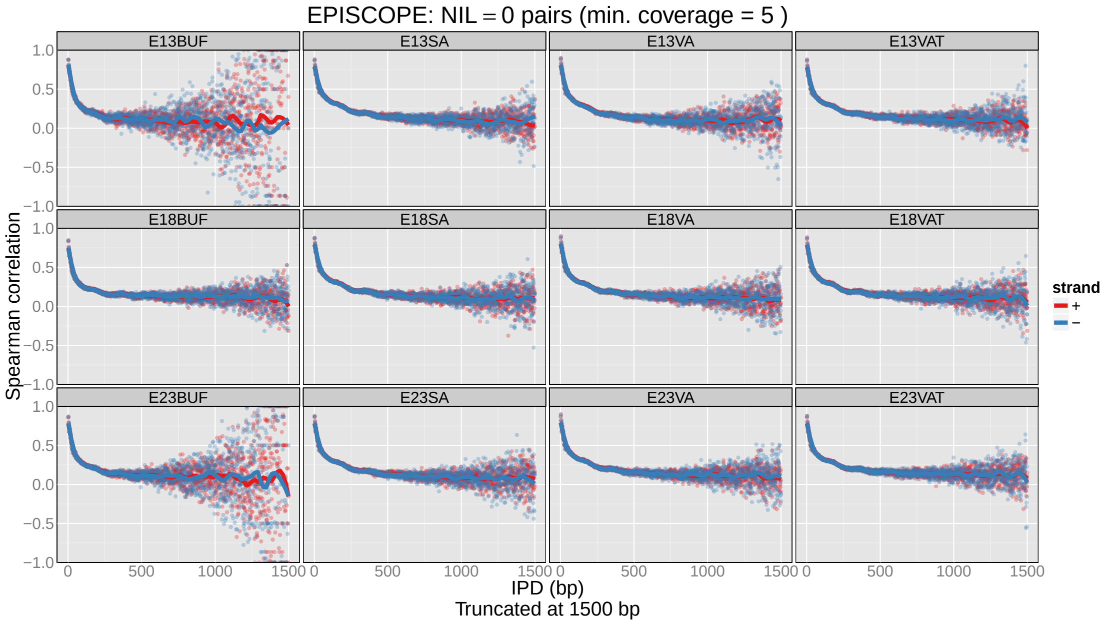


Figure 7.3: Correlations of β-values for strand-specific pairs of CpGs with $NIL = 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

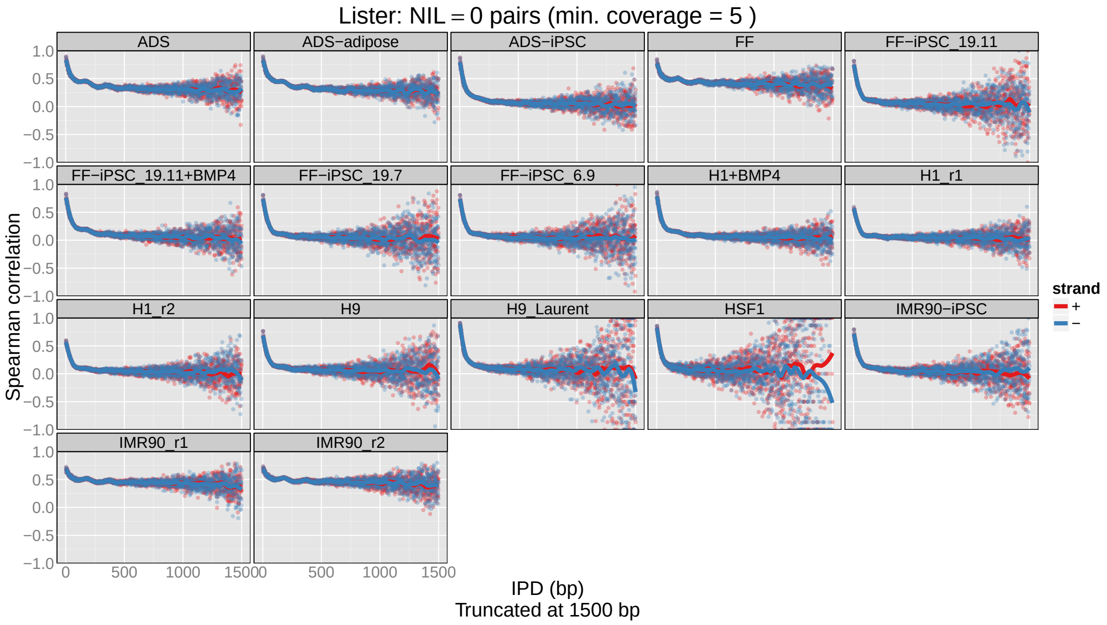


Figure 7.4: Correlations of β-values for strand-specific pairs of CpGs with $NIL = 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

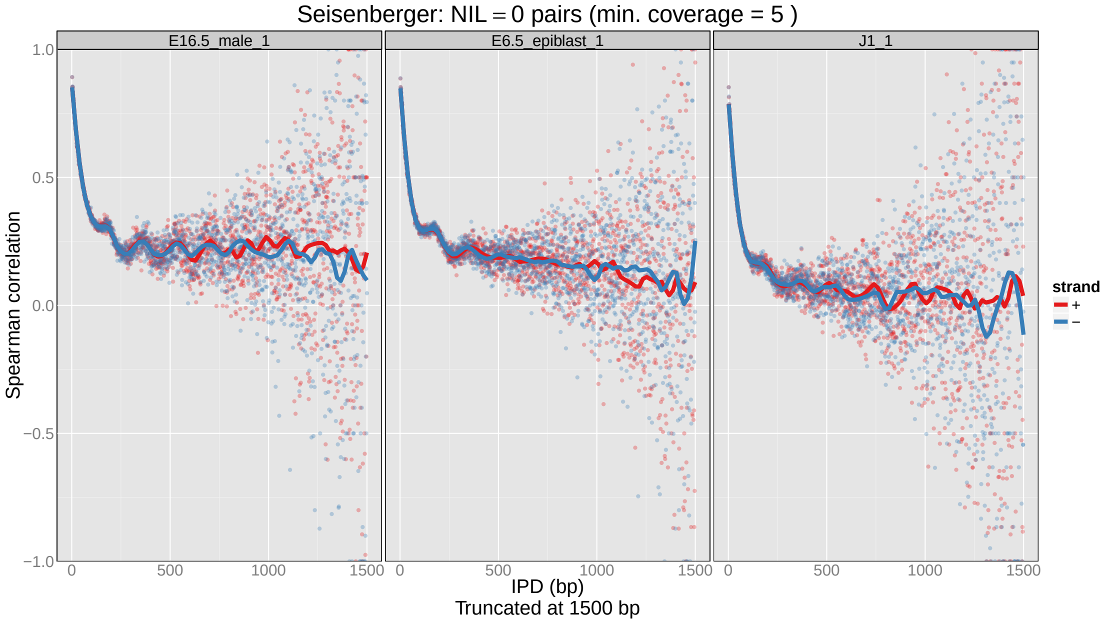


Figure 7.5: Correlations of β-values for strand-specific pairs of CpGs with $NIL = 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

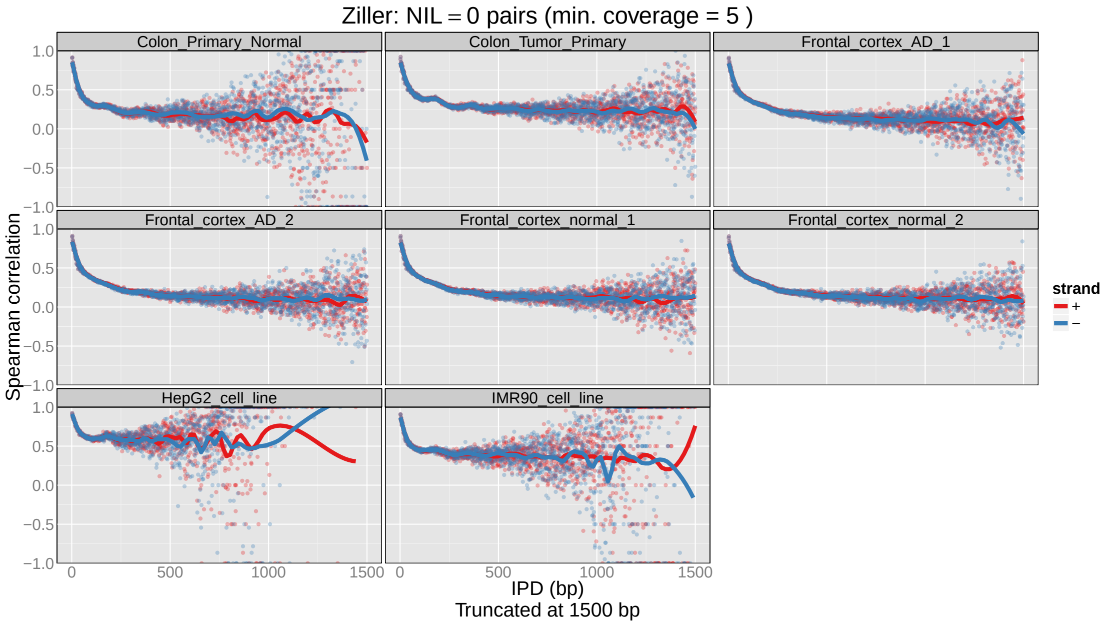


Figure 7.6: Correlations of β-values for strand-specific pairs of CpGs with $NIL = 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

Choice of correlation coefficient

Figures 7.7, 7.8, 7.9 and 7.10 compare the Pearson correlation coefficient to the Spearman correlation coefficient. The two correlation coefficient give qualitatively similar results, however, the Spearman correlation is consistently smaller in magnitude than the Pearson correlation.

I have elected to use Spearman's correlation coefficient in what follows because it is more robust to outliers than Pearson's correlation [Kraemer 2006]. However, this does not remove the general limitations of using correlation coefficients (discussed in Section 7.1.1).

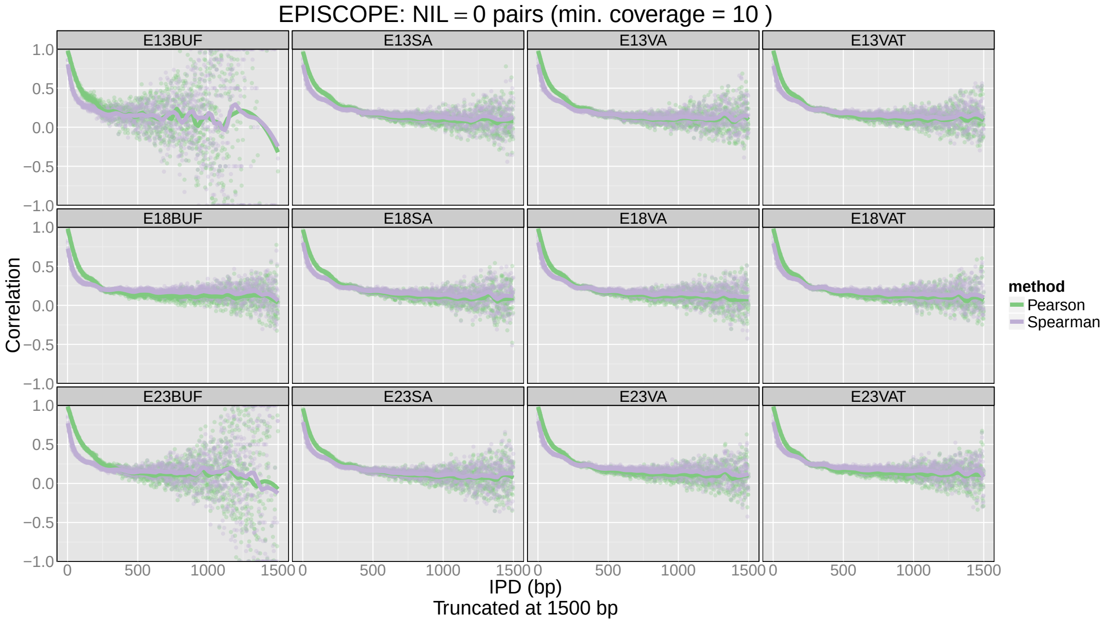


Figure 7.7: Pearson correlation versus Spearman correlations of β-values for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

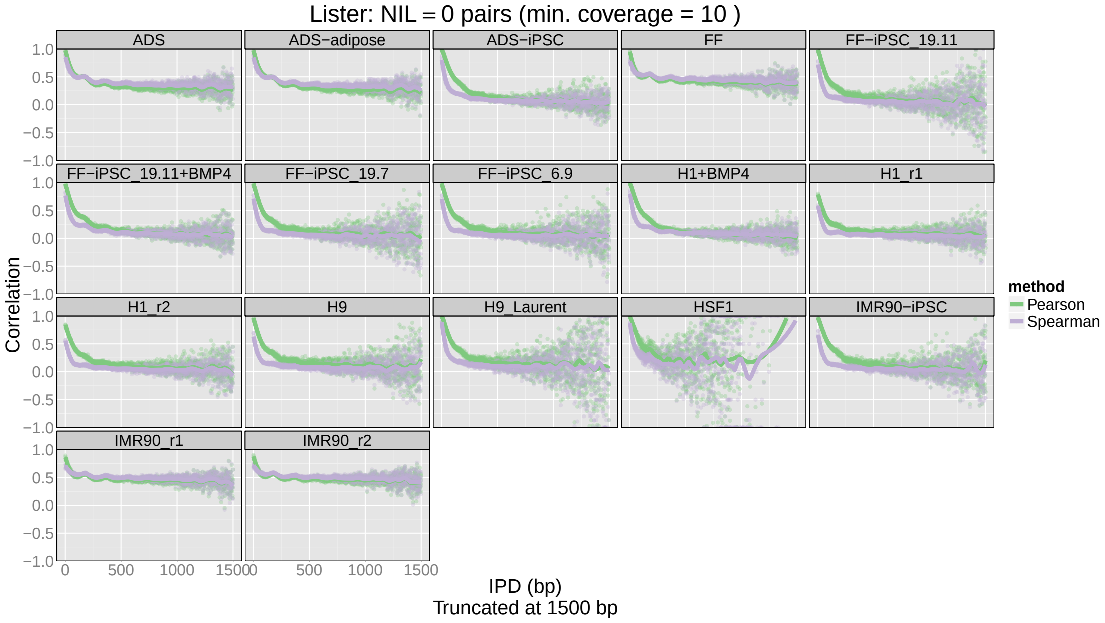


Figure 7.8: Pearson correlation versus Spearman correlations of β-values for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

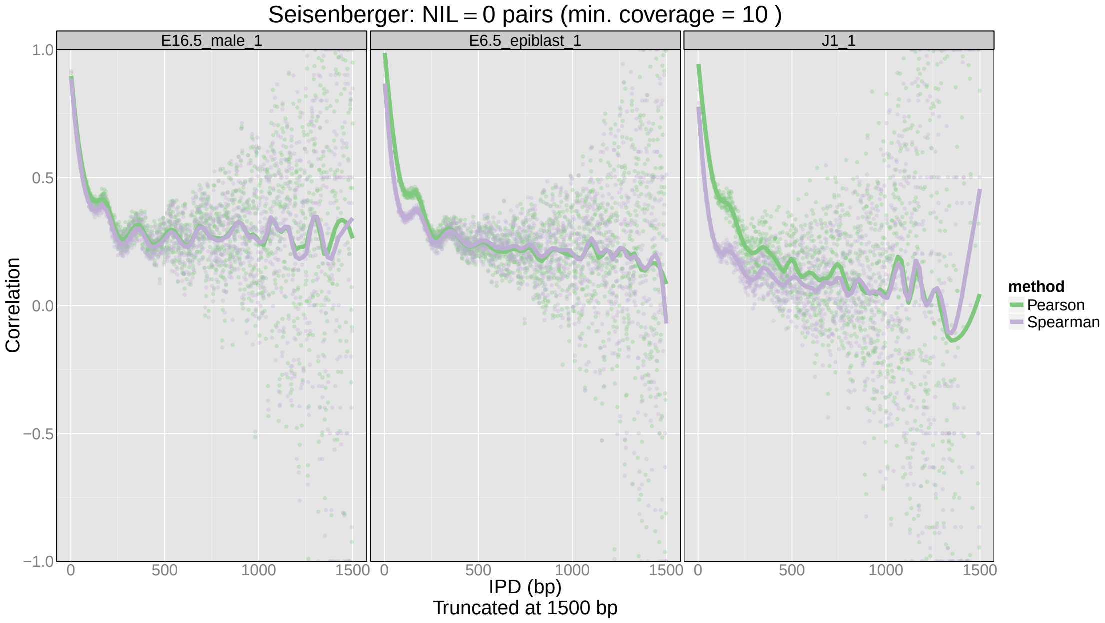


Figure 7.9: Pearson correlation versus Spearman correlations of β-values for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

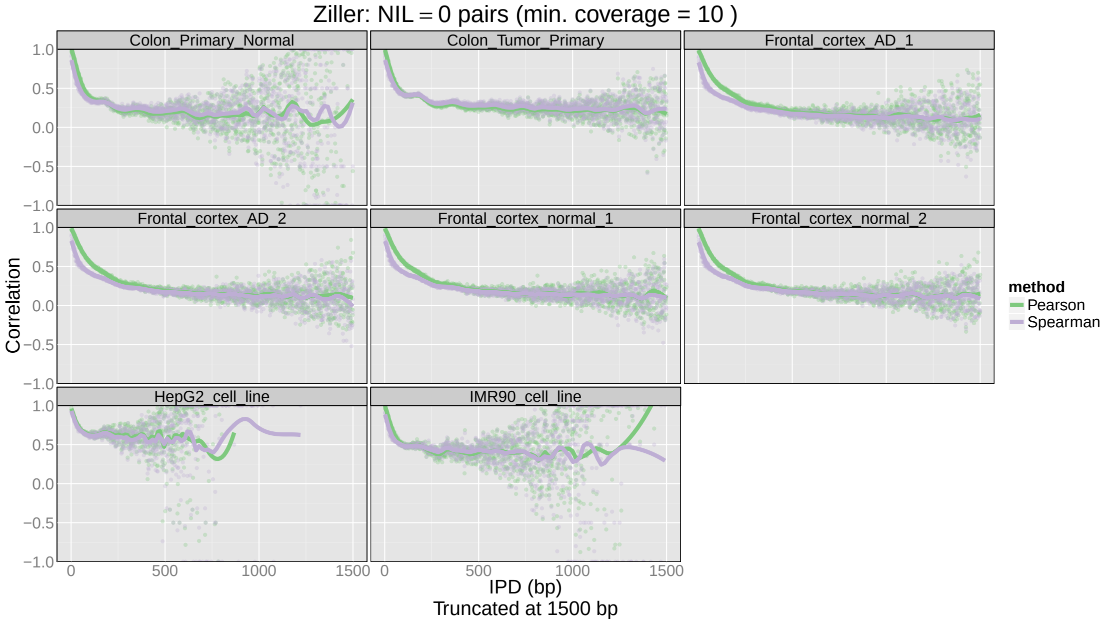


Figure 7.10: Pearson correlation versus Spearman correlations of β-values for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

7.1.3 CpG pairs with $NIL = 0$

Figures 7.11, 7.12, 7.13 and 7.14 show the Spearman correlations of strand-collapsed β-values for $NIL = 0$ pairs of CpGs. Two things immediately stand out. Firstly, for all samples the trend is that the strength of the correlation decays as a function of IPD; pairs of CpGs separated by larger distances are on average less correlated than pairs of CpGs separated by shorter distances. Secondly, the estimated correlations violently jump around for larger IPDs. The latter is simply due to the aforementioned lack of $NIL = 0$ pairs at larger IPDs with which to estimate these correlations. Therefore, it is the trend, rather than the individual values, that should be analysed in these plots.

There are some consistent patterns in comparing these trends across samples. The correlations very quickly drop well below 0.5, within approximately 50 bp for all the *EPISCOPE* and *Seisenberger* samples and most of the *Lister* and *Ziller* samples. Samples that are exceptions to this trend (*ADS*, *ADS-adipose*, *ADS-iPSC*, *FF*, *IMR90_r1*, *IMR90_r2*, *IMR90_cell_line* and *HepG2_cell_line*) have a slower decay in these correlations and remain more highly correlated over $IPD = 2, \ldots, 1500$. These same samples are also notable because they have significant amounts of intermediate methylation (Section 4.4.1). I believe this is what drives the stronger and more slowly decaying correlations. The scatterplots of β-values for these samples will have more β-values in the middle of the scatterplot, and fewer in the corners, hence the stronger correlations.

Most samples have a correlation close to 1 for neighbouring CpGs ($IPD = 2$). The notable exceptions are the embryonic stem cell replicates, *H1_r1* and *H1_r2* (Lister dataset), which begin with a maximum correlation of only around 0.5. This is not observed in the other embryonic stem cells (*H9*, *H9_Laurent*, *HSF1* and *J1_1*), which makes it difficult to know whether β-values of embryonic stem cells are truly less correlated than other cell types.

The trend of the correlations for all these samples plateaus out close to zero, slightly higher for those samples with significant intermediate methylation. It should not be concluded from this, however, that the methylation of CpGs separated by more than a few hundred base pairs are uncorrelated. Recall that these results are only for $NIL = 0$ pairs of CpGs, which are increasingly rare for larger IPDs. These results tell us about the ‘first-order’ correlations of β-values, which decay very quickly but whose importance also decreases quite quickly owing to the vast majority of CpGs being within 100 bp of another CpG (e.g., Figure 1.4 shows the distribution of IPDs for humans).

One final aspect of these genome-wide plots bears mentioning. There are hints of an approximately 150 to 200 bp periodicity in these plots. This is a similar scale to the previously reported periodicities in correlations of β-values (see Chapter 6). While this periodicity can readily be detected by eye, it is somewhat more difficult to verify using classical methods from spectral analysis. The chief reason is that the data are very noisy beyond the second or third putative cycle due to the limited data at these larger IPDs.

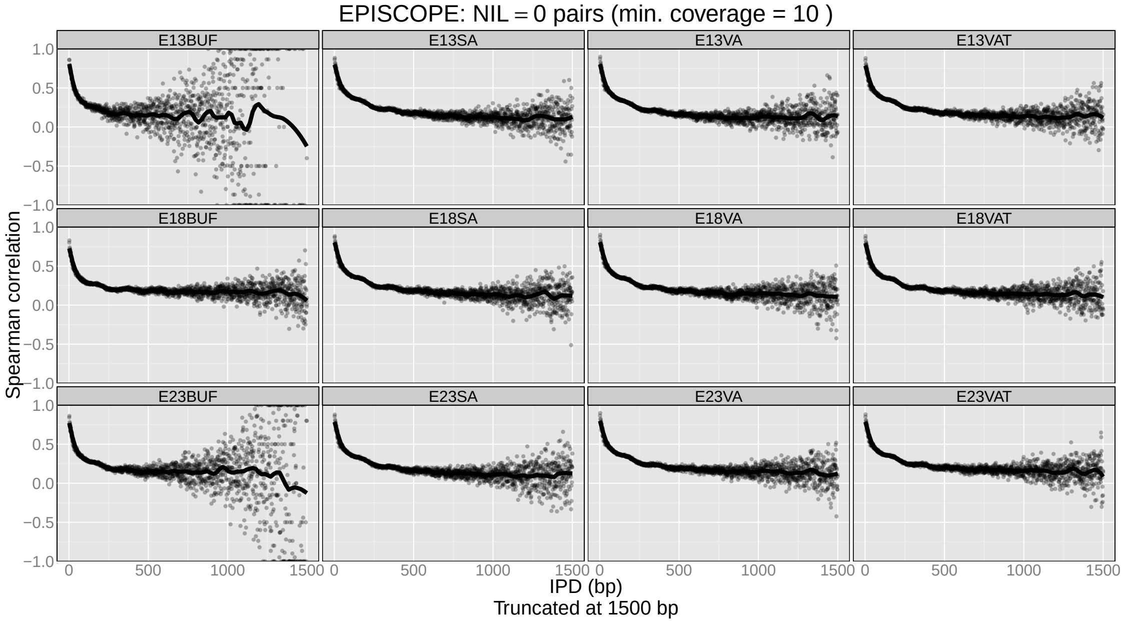


Figure 7.11: Spearman correlations of β-values as a function of IPD for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

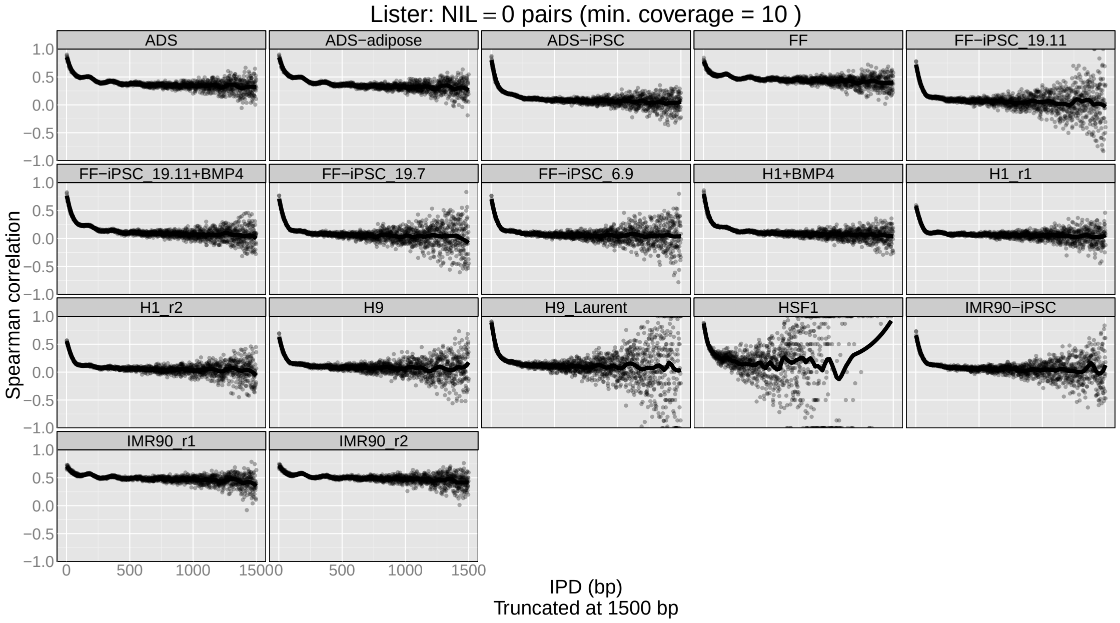


Figure 7.12: Spearman correlations of β-values as a function of IPD for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

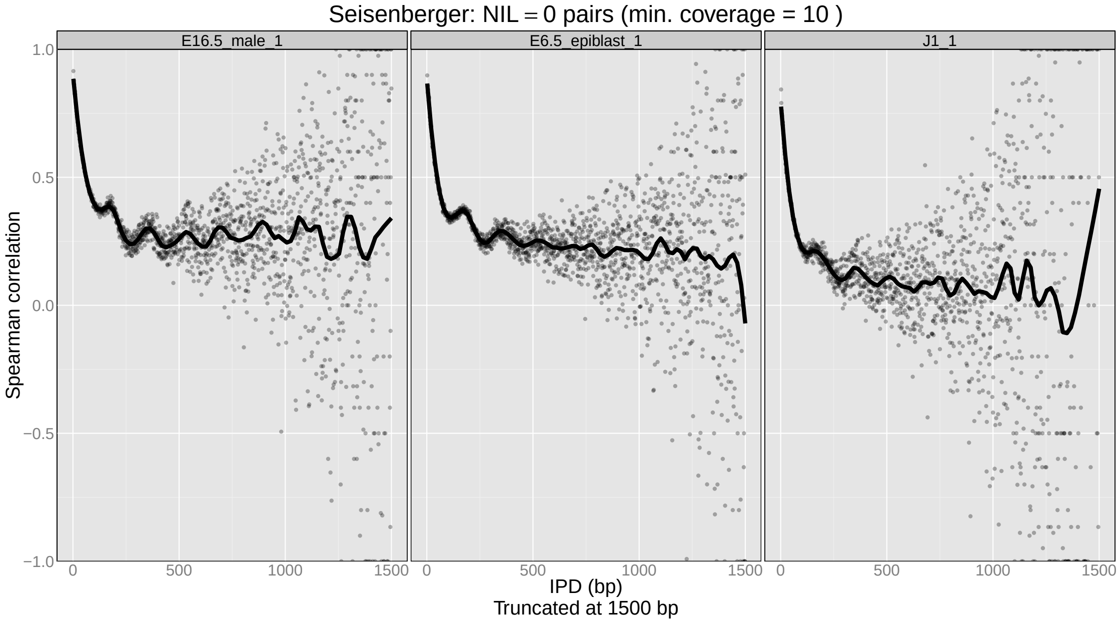


Figure 7.13: Spearman correlations of β-values as a function of IPD for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

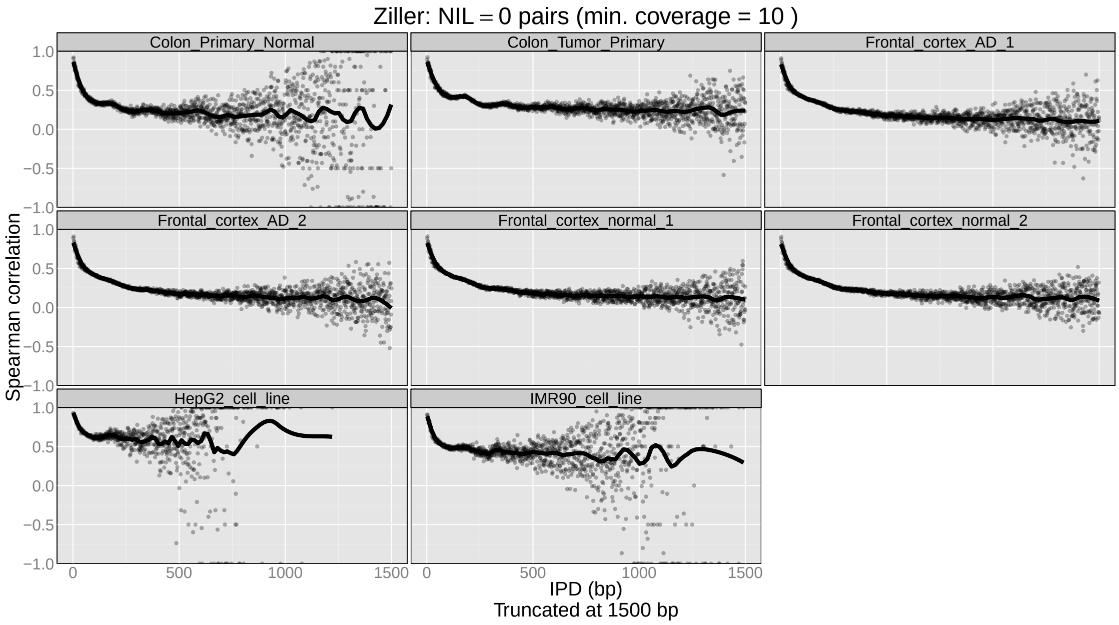


Figure 7.14: Spearman correlation of β-values as a function of IPD for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

Figures 7.11, 7.12, 7.13 and 7.14 are genome-wide plots; all pairs of CpGs with a given IPD are treated identically. However, there are good reasons to expect that the correlations of β-values, like the β-values themselves, might be influenced by their genomic context. To investigate this, we can stratify pairs of CpGs pairs by a genomic feature and repeat the analysis. The obvious feature to stratify on is CpG islands.

Figures 7.15, 7.16, 7.17 and 7.18 are plots of Spearman correlations of β-values for pairs of CpGs stratified by whether the pair is inside or outside of a CpG island[2]. Again, the substantial noise in these plots is due to a lack of $NIL = 0$ pairs with large IPDs, particularly inside CpG islands.

The message from these plots is clear, β-values of $NIL = 0$ pairs within CpG islands are much more correlated than those pairs outside of islands. Stratifying by CpG island status also reveals that the correlation of the methylation levels of CpGs inside CpG islands is less influenced by IPD. In fact, for some samples it appears that the most correlated $NIL = 0$ pairs inside CpG islands are not those with the minimum $IPD = 2$, but those with a slightly larger IPD. From these plots we can now see that the apparent sharp decline in correlations over the first 50 bp from the genome-wide data is really driven by a shift from pairs that are frequently inside CpG islands to pairs that are mostly outside of islands.

[2] Pairs spanning the boundary of a CpG island are ignored. There are few such pairs when $NIL = 0$. While there are considerably more such pairs when $NIL \geqslant 0$, I have ignored these in favour of simplicity.

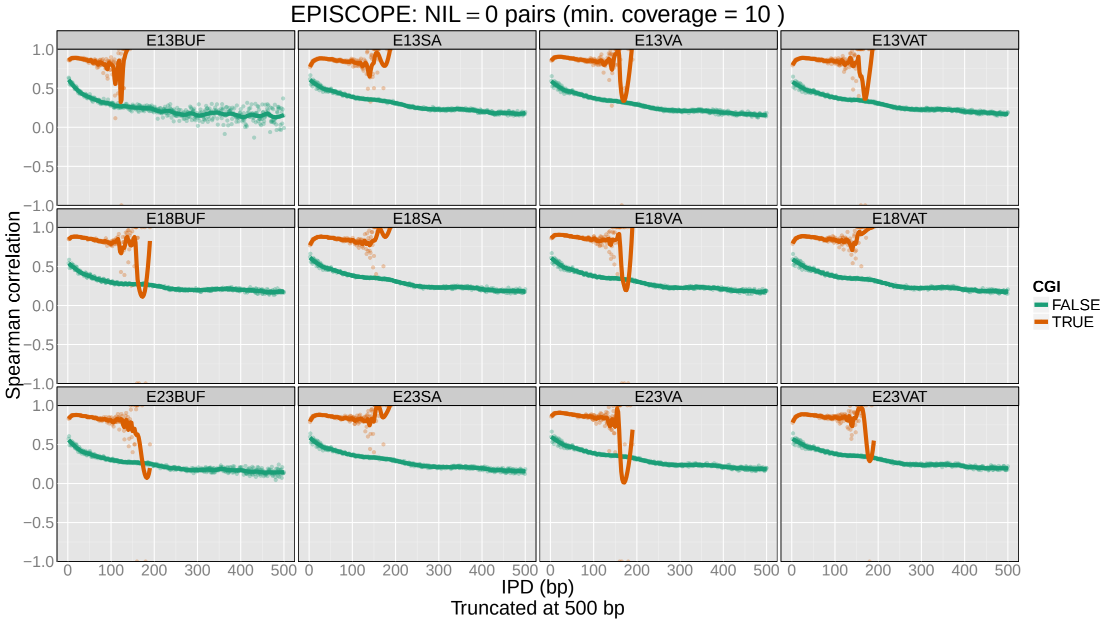


Figure 7.15: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points ($\texttt{span} = 0.1$).

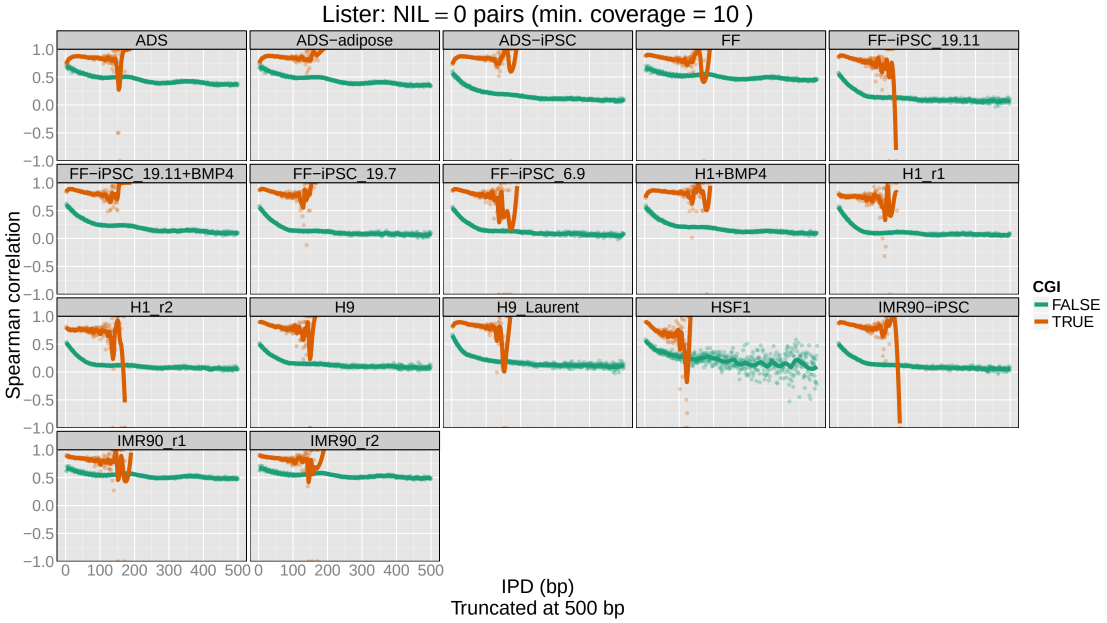


Figure 7.16: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points ($\texttt{span} = 0.1$).

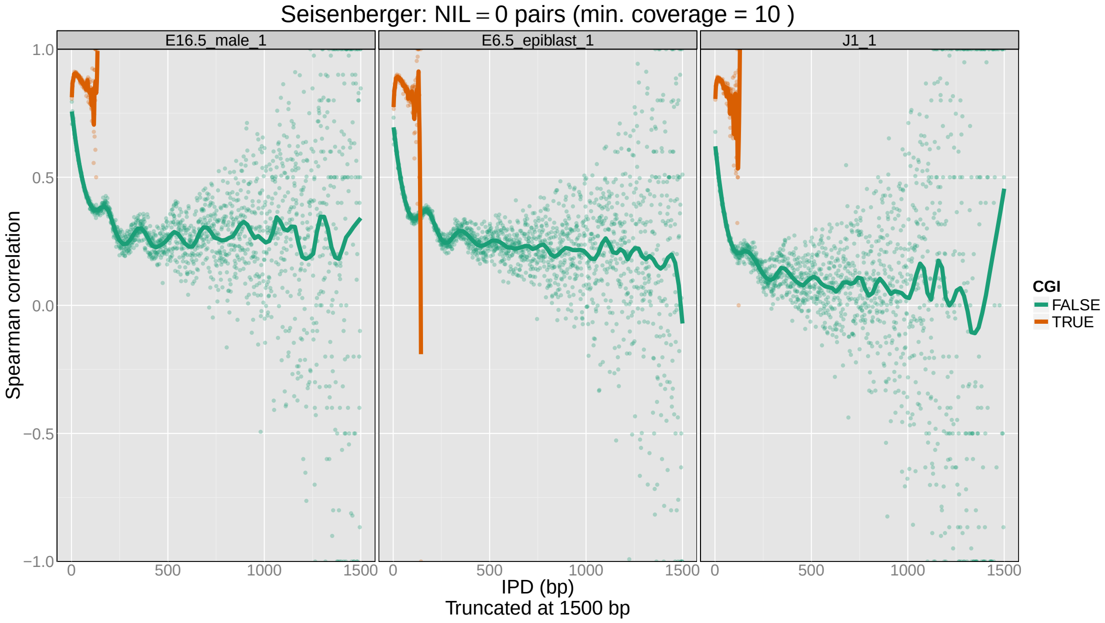


Figure 7.17: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

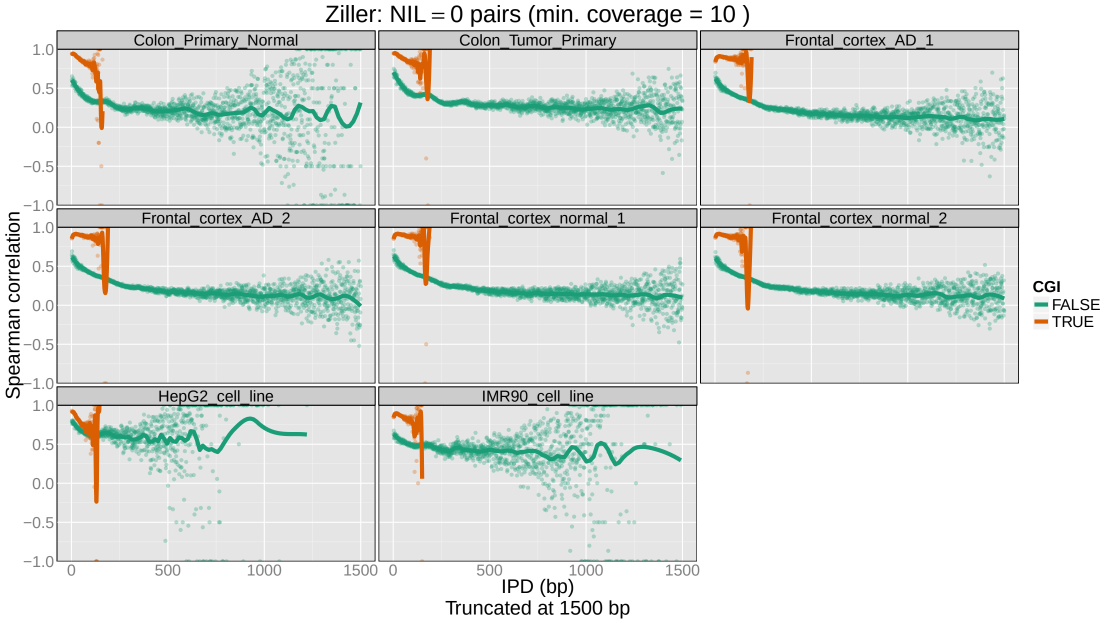


Figure 7.18: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL = 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

7.1.4 CpG pairs with $NIL \geqslant 0$

Figures 7.19, 7.20, 7.21 and 7.22 are the plots of CpG β-value correlations against IPD for $NIL \geqslant 0$ pairs. In contrast to the $NIL = 0$ results, the correlations here are relatively stable, at least for $IPD = 2, \ldots, 2000$[3]. This is because there are many more $NIL \geqslant 0$ pairs than there are $NIL = 0$ pairs.

When we aggregate over as many factors as we are in these genome-wide $NIL \geqslant 0$ plots, we find that there is little difference in the correlation curves between samples. All samples have a very high correlation, close to 1, for small IPDs that steadily decays to a lower correlation of 0.2 to 0.4 by $IPD = 2000$.

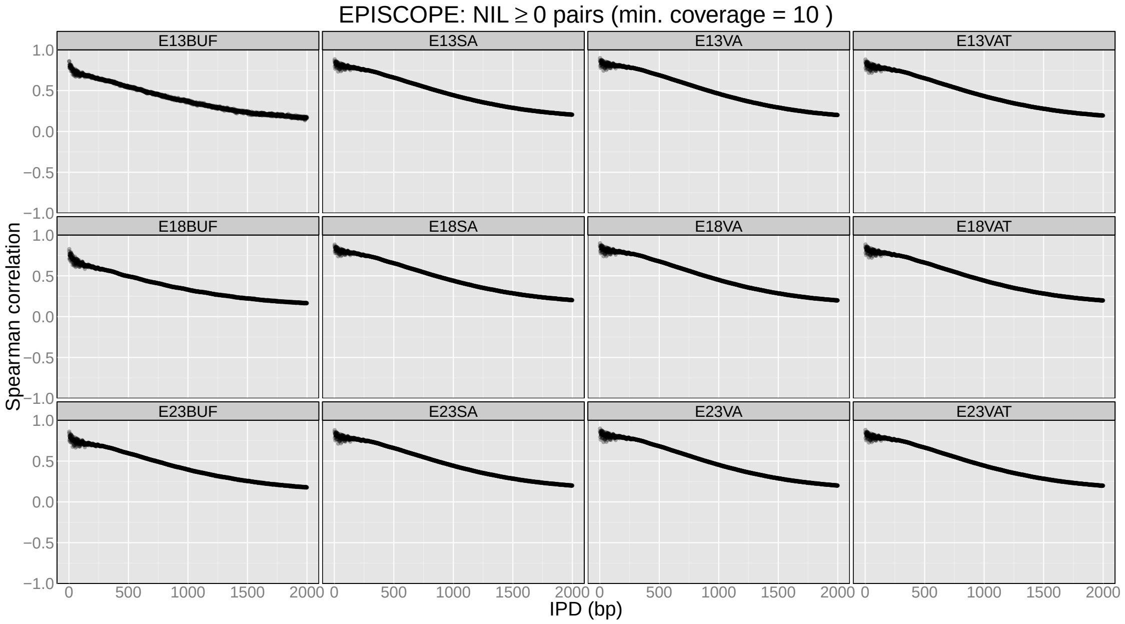


Figure 7.19: Spearman correlation of β-values for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

[3] These correlations could be computed for $IPD > 2000$. However, constructing all $NIL \geqslant 0$ pairs requires much more time and memory than constructing all $NIL = 0$ pairs because there are many more such pairs. For this reason I have focused on $IPD = 2, \ldots, 2000$ for the $NIL \geqslant 0$ pairs.

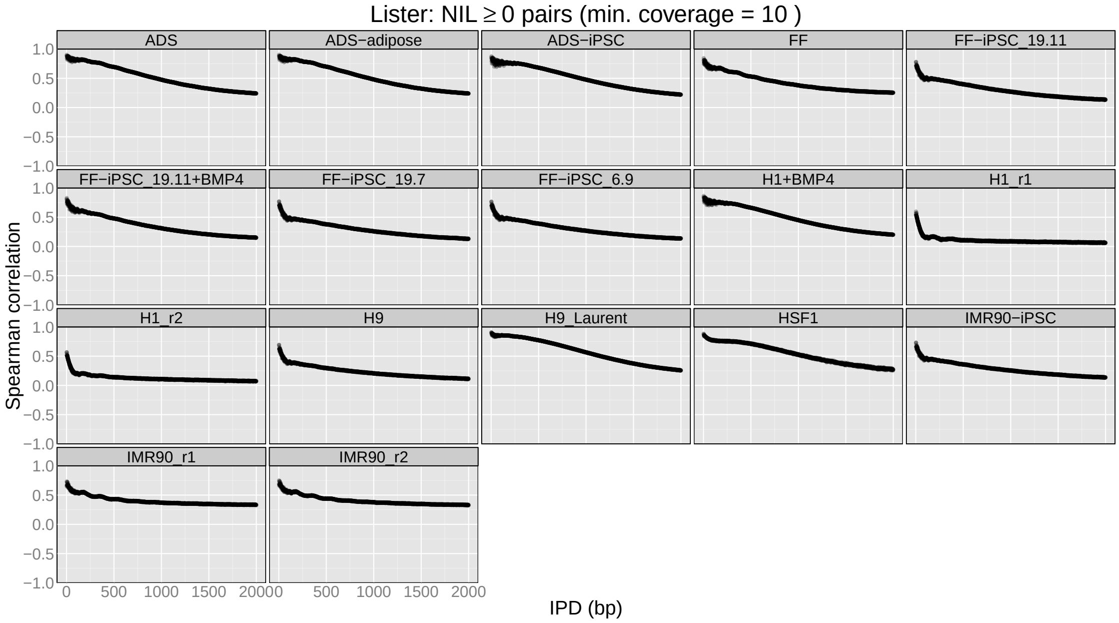


Figure 7.20: Spearman correlation of β-values for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

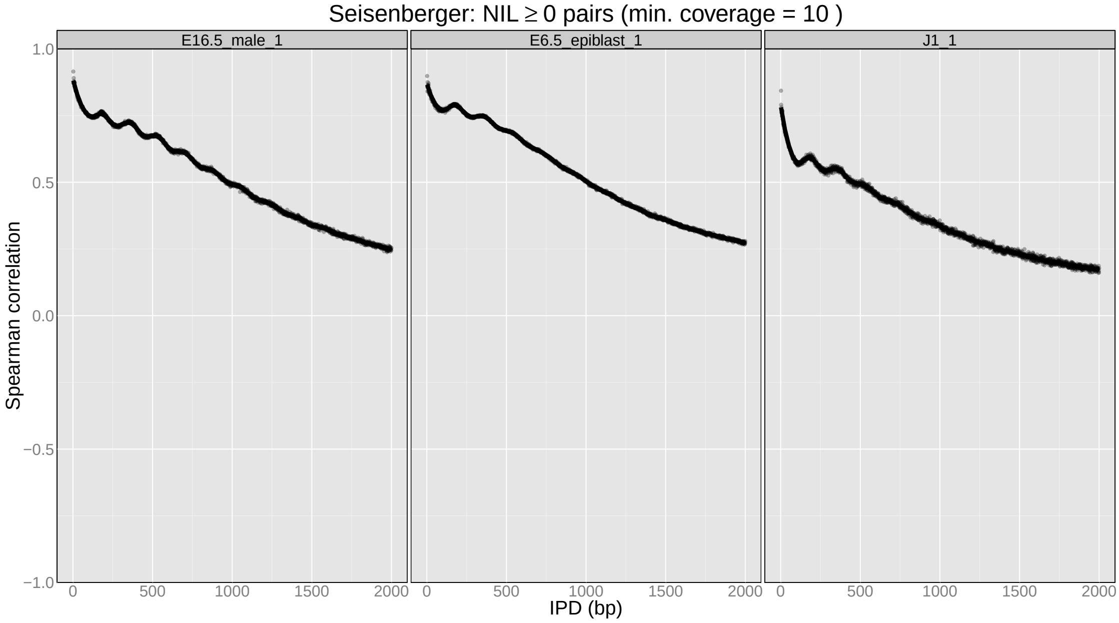


Figure 7.21: Spearman correlation of β-values for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

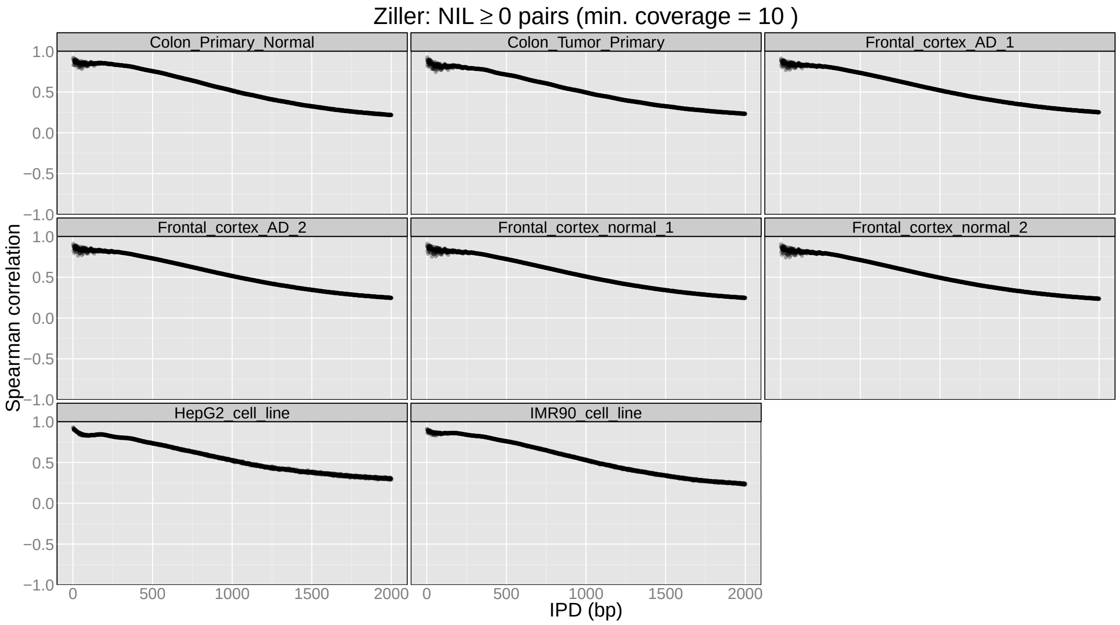


Figure 7.22: Spearman correlation of β-values for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

Comparing CpG islands to non-islands, we again see that the methylation levels of pairs of CpGs within islands are consistently more correlated than those outside of islands[4] (Figures 7.23, 7.24, 7.25 and 7.26).

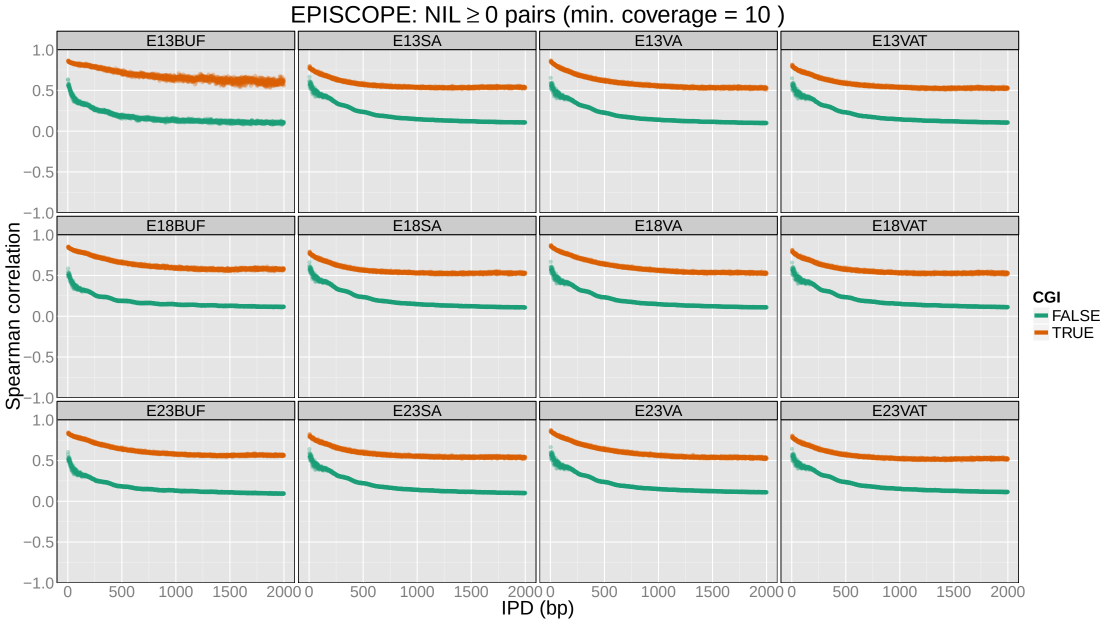


Figure 7.23: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *EPISCOPE* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` = 0.1).

The 150 to 200 bp periodicity is also evident from the $NIL \geqslant 0$ data and is particularly clear for the CpG pairs outside of CpG islands. This may be due to different regulation of DNA methylation outside of CpG islands or simply because CpG islands have such consistently high correlations that these periodicities are not evident.

[4] The minor exception is the *HepG2_cell_line* where the CpG island and non-island lines intersect by $IPD = 1500$, but we have already seen that this sample has an unusual methylome compared to the other samples, particularly with respect to CpG islands (see Figure 4.17).

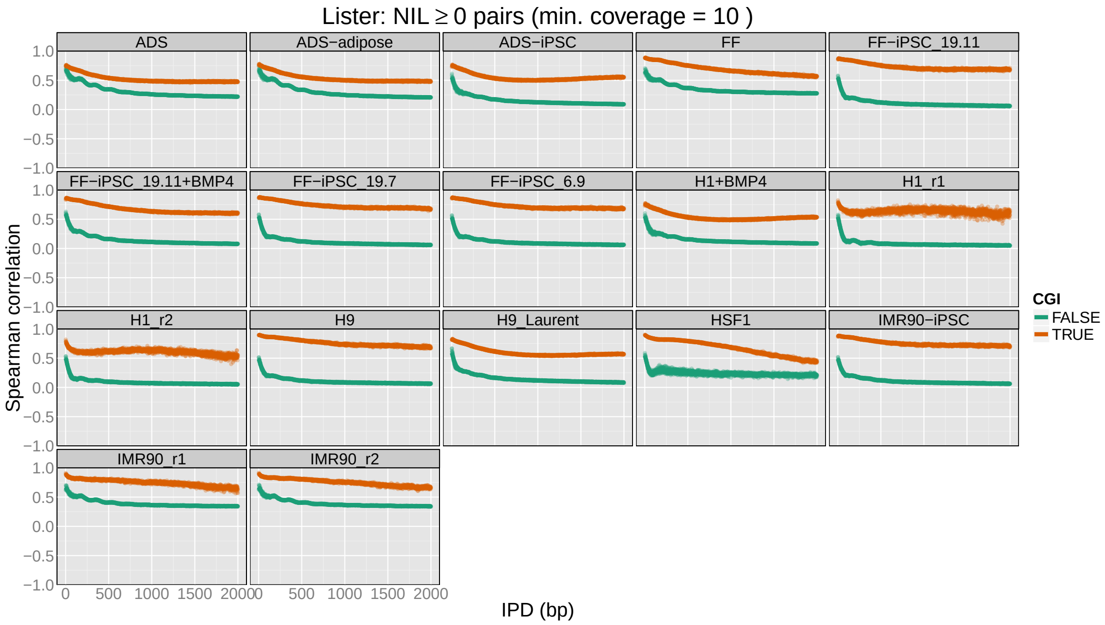


Figure 7.24: Spearman correlations of β-values as a function of *IPD* stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Lister* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

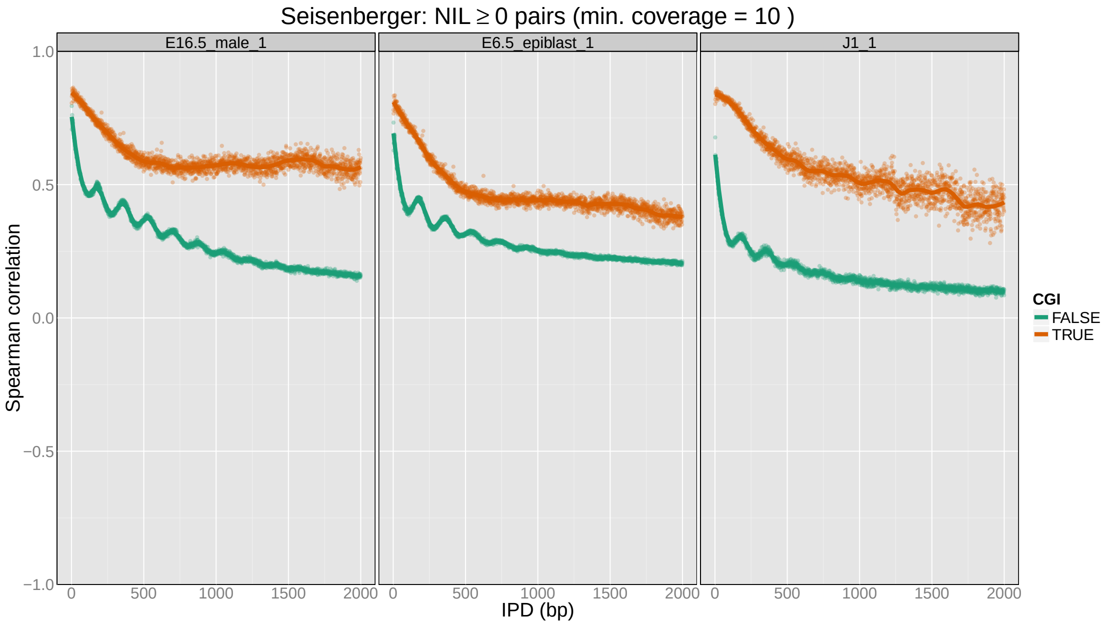


Figure 7.25: Spearman correlations of β-values as a function of *IPD* stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Seisenberger* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

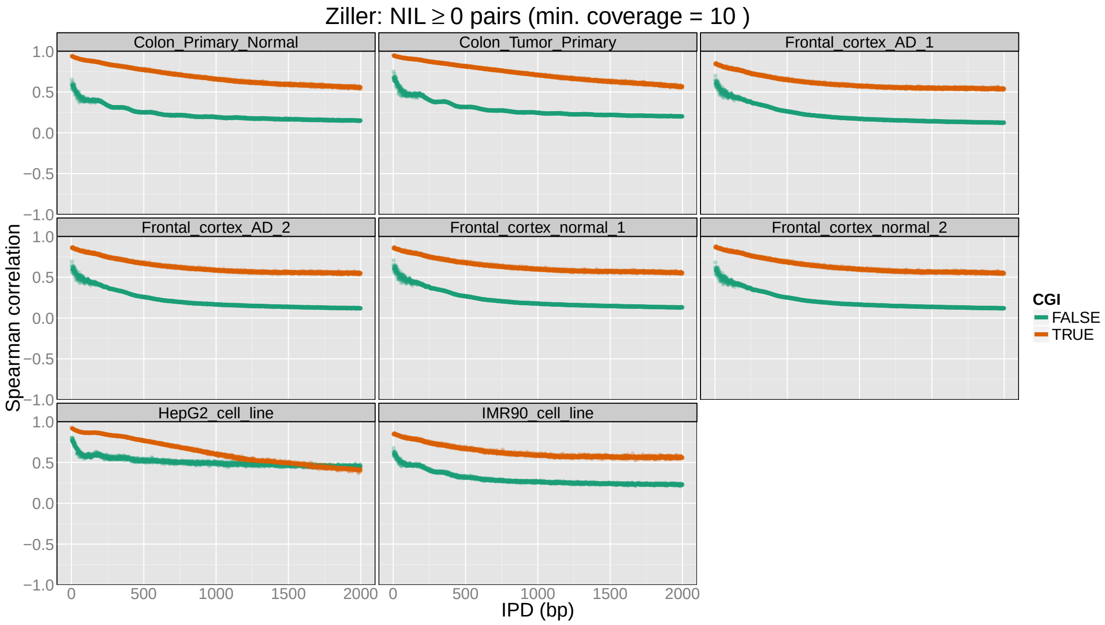


Figure 7.26: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL \geqslant 0$ for samples from the *Ziller* dataset. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points ($\texttt{span} = 0.1$).

7.1.5 Limitations

A limitation to these analyses is that they are based on correlations of measurements that are aggregates from many thousands of DNA molecules. We are not measuring co-methylation on the scale at which the biological process of DNA methylation acts, namely at the level of individual DNA molecules. As noted by Ball *et al.* [2009], “the clonal feature of Illumina sequencing [allows us] to investigate whether co-methylation occurs at the single-molecule level”. We explore this topic in Section 7.3, but before doing so we first must first discuss the challenge of estimating meaningful measures of dependence in sparse $2 \times 2 \times K$ contingency tables. While perhaps not yet apparent, this will be essential for the analysis of within-fragment co-methylation.

7.2 Estimating dependence in sparse $2 \times 2 \times K$ contingency tables

We will begin with an overview of the simplest case, estimating dependence in a 2×2 contingency table (i.e. $K = 1$). We will then progressively introduce the complications of sparsity, $K > 1$, and heterogeneity across the K tables. We conclude with a simulation study highlighting the performance of several different estimators of dependence in sparse $2 \times 2 \times K$ contingency tables and determine which is most appropriate for a study of within-fragment co-methylation.

7.2.1 2×2 contingency tables

Consider two binary random variables, X and Y. Let π_{ij} denote the probability that a subject has response i for variable X and response j for variable Y $(i, j, = 1, 2)$. We wish to estimate the *odds ratio*, which is defined as $\psi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$. The odds ratio is a convenient summary of dependence in a 2×2 table. Provided that all $\pi_{ij} > 0$, the odds ratio, ψ, can be interpreted as follows:

- $\psi = 1$: X is independent of Y.
- $\psi > 1$: Subjects are more likely to have the same level for both X and Y, e.g.,

subjects with $(X, Y) = (1, 1)$ or $(X, Y) = (2, 2)$ are more likely than subjects with $(X, Y) = (1, 2)$ or $(X, Y) = (2, 1)$.

- $\psi < 1$: Subjects are more likely to have the the opposite level for X than they do for Y, e.g., subjects with $(X, Y) = (1, 2)$ or $(X, Y) = (2, 1)$ are more likely than subjects with $(X, Y) = (1, 1)$ or $(X, Y) = (2, 2)$.

The odds ratio is a simple summary of the dependence in a 2 × 2 table. An attractive feature is that it is independent of the marginal probabilities. The odds ratio is not perfect, however. For one, the possible values of ψ are highly skewed, with $0 < \psi < 1$ and $1 < \psi < \infty$ corresponding to the two distinct dependence scenarios. Rather than estimating ψ directly, we will instead focus our attention on the *log odds ratio*, $\theta = \log \psi$. The log odds ratio is symmetric about 0, which makes statistical inference somewhat simpler. I will routinely switch between discussing the estimation of ψ and θ, noting here that we can always convert one to the other by appropriate exponentiation or taking of logarithms.

Suppose we observe (X, Y) for n samples. Denote by n_{ij} the number of subjects with response i for variable X and response j for variable Y. We use the 'plus' notation to denote the sum over the index it replaces, e.g., $n_{1+} = n_{11} + n_{12}$. The general form of this 2 × 2 contingency table is shown in Table 7.1.

Table 7.1: Notation for a 2 × 2 contingency table.

		Y		
		1	2	Total
X	1	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
	Total	n_{+1}	n_{+m}	$n = n_{++}$

The simplest estimator of θ is the unconditional maximum likelihood estimator, $\widehat{\theta}_U = \log \frac{n_{11}n_{22}}{n_{12}n_{21}}$. The asymptotic standard error of $\widehat{\theta}_U$ is given by $\hat{\sigma}(\widehat{\theta}_U) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$. The asymptotic distribution of $\widehat{\theta}_U$ as $n \to \infty$ is $Gaussian(\theta, \hat{\sigma}(\theta)^2)$.

A problem with the estimator $\widehat{\theta}_U$ is that it is 0 or ∞ if any $n_{ij} = 0$ and undefined if either the row or column sums are zero, events that have positive probabilities. We can avoid such problems by adding $\frac{1}{2}$ to each of the n_{ij} and defining a modified estimator $\widehat{\theta}_{0.5} = \log \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}$ with corresponding asymptotic standard error

$\hat{\sigma}(\widehat{\theta}_{0.5}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$. Haldane [1956] and Anscombe [1956] showed that $\widehat{\theta}_U$ and $\widehat{\theta}_{0.5}$ have the same asymptotic distribution around θ as $n \rightarrow \infty$ but that $\widehat{\theta}_{0.5}$ has reduced first-order bias. The modified estimator, $\widehat{\theta}_{0.5}$, is therefore generally recommended.

Another alternative to $\widehat{\theta}_U$ is the *conditional* maximum likelihood estimator, $\widehat{\theta_C}$. This is computed by first conditioning on n_{1+}, n_{+1}. Some algebra then leads to the conditional distribution of n_{11}, $f(n_{11}; n_{1+}, n_{+,1,\psi})$, which is a hypergeometric distribution [Fisher 1935]. The maximum likelihood estimator of ψ, $\widehat{\psi}_C$, is solved using iterative methods. The conditional estimator, $\widehat{\theta_C} = \log \widehat{\psi}_C$, works better than the unconditional estimator, $\widehat{\theta}_U$, when the sample size, n, is small [Agresti 2007, pp. 157-158].

Estimation of the odds ratio is made more difficult when the contingency table is *sparse*. We say that a contingency table is sparse when some of the n_{ij} are small. Sparsity can occur even when the sample size, n, is large. Sparsity becomes more of a problem as we move from a 2×2 table to a $2 \times 2 \times K$ table.

7.2.2 $2 \times 2 \times K$ contingency tables

Suppose we now measure the same two binary variables X and Y in K different strata. We summarise these data in a $2 \times 2 \times K$ contingency table, $\{n_{ijk}\}$, where the k^{th} 'slice' of the table corresponds to the 2×2 table for the k^{th} stratum. We assume for now that the odds ratio, $\psi = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}$, remains constant across the K tables but we allow the marginal probabilities, $\pi_{+1k} = 1 - \pi_{+2k}, \pi_{1+k} = 1 - \pi_{2+k}$ to vary.

We can again compute an unconditional maximum likelihood estimator, a '0.5-adjusted' unconditional maximum likelihood estimator, and a conditional maximum likelihood estimator of the odds ratio ψ [see Breslow 1981].

Another estimator of the common odds ratio, ψ, is the Mantel-Haenszel estimator [Mantel and Haenszel 1959], $\widehat{\psi}_{MH} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{n_{12k} n_{21k} / n_{++k}}$. Robins *et al.* [1986] derived a simple robust estimator of the variance for $\widehat{\theta}_{MH} = \log \widehat{\psi}_{MH}$, $\hat{\sigma}^2(\widehat{\theta}_{MH})$. This variance estimator did not appear until some 25 years after the initial publication of the Mantel-Haenszel estimator[5]. This delay was in part due to the existance of two asymptotic forms of $2 \times 2 \times K$ contingency tables.

[5] Other estimators had been proposed but were supplanted by the work of Robins *et al.* [1986].

The first asymptotic environment, $A1$, assumes that K remains small while the the sample sizes in each strata, n_{++k}, grows large. The second asymptotic environment, $A2$, assumes that K grows while n_{++k} remains small. The variance estimator proposed by Robins *et al.* [1986] is consistent under both $A1$ and $A2$. For our study of within-fragment co-methylation, $A2$ is the more relevant asymptotic environment.

Breslow [1981] showed that under $A2$ that the unconditional maximum likelihood estimator does not converge to the true odds ratio. In contrast, both the Mantel-Haenszel estimator and the conditional maximum likelihood estimator do converge to the true odds ratio, with the Mantel-Haenszel estimator having the advantage of a simple closed form expression [Breslow 1996].

7.2.3 Homogeneity of odds ratios assumption

So far we have assumed a common odds ratio for the K 2×2 tables. This is also known as the *homogeneity of odds ratios* assumption. As noted by Liang and Self [1985], Mantel and Haenszel [1959] did not explicitly assume homogeneity of the odds ratio in their original work. Nonetheless, it does raise the question of how to test this assumption and the effect on these estimators when this assumption is not true.

Under asymptotic environment $A2$, where the 2×2 tables are sparse and $K \to \infty$, Liang and Self [1985] developed three conditional tests of the homogeneity of odds ratios assumption and compared these with two unconditional tests that were designed under $A1$. Unsurprisingly, Liang and Self found that the three tests designed for $A2$ are more appropriate than either of the tests designed for $A1$ when the data are simulated under $A2$. The disadvantage of these conditional tests is the extra computation required. It has been noted, however, that statistical tests of homogeneity versus heterogeneity typically have low power and are only able to detect large departures form homogeneity [e.g., Hauck 1989]. These tests are therefore not always of great practical use.

7.2.4 Simulation study

We carry out a simulation study to explore the effects of sparsity and heterogeneity of odds ratios on estimates of odds ratios in $2 \times 2 \times K$ contingency tables. The simulation requires that we specify several parameters. In order to relate these back to DNA methylation, I describe each parameter using the statistical framework proposed in Chapter 4:

- the number of tables, K. The number of CpG pairs (2-tuples).
- the marginal probabilities, $\pi_{2+} = Pr(Z_i)$ and $\pi_{+2} = Pr(Z_{i'})$. Some simulations use a common set of marginal probabilities across all tables but others simulate the marginal probabilities from a specified distribution.
- the odds ratio of each table, ψ_k. Most simulations use a common odds ratio across the K tables, $\psi_k = \psi$.
- the sample size of each table, $n_{++k} = d_{(i,i')}$. The number of reads containing both the i^{th} and i'^{th} methylation locus (sequencing depth). Most simulations use a truncated negative binomial distribution to model the sequencing depth. The truncated negative binomial distribution is parameterised by the mean (μ), the dispersion parameter (size) and the truncation level (trunc). Unless otherwise noted, simulations use $\mu = 30$, size $= 10$ and trunc $= 9$, corresponding to an average sequencing depth of 30× but where only pairs with at least 10× sequencing depth are used in downstream analyses[6].

For each set of parameters, we generate K 2×2 contingency tables having the desired marginal probabilities and odds ratio. We then compute the following statistics:

1. The K 0.5-adjusted unconditional log odds ratio estimates, $\widehat{\theta}_{0.5}$, from each 2×2 table.
2. The unconditional log odds ratio estimate from the 2×2 table formed by collapsing over k, $\widehat{\theta}_U$.
3. The Mantel-Haenszel log odds ratio estimate from the $2 \times 2 \times K$ table, $\widehat{\theta}_{MH}$.

The results of each simulation are summarised by a plot of $\widehat{\theta}_{0.5}$ (histogram), the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange), and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured

[6] In the simulation study, a larger value of 'trunc' results in higher quality data since we can simply keep simulating until we have K tables with sufficient sequencing depth. In analysing real data, however, increasing this cutoff may be detrimental since we end up with fewer tables to analyse.

blue). The density of θ_k, the value of the true log odds ratio in the k^{th} slice of the $2 \times 2 \times K$ table, is shown by a black curve[7].

Results

To begin, we explore the effect of the marginal probabilities, π_{2+} and π_{+2}, when $\theta = 0$ ($\psi = 1$) identically across the K tables (corresponding to independence of X and Y) for $K = 100$ (Simulation 1a, Figure 7.27), and $K = 1000$ (Simulation 1b, Figure 7.28). In this setting we expect both $\widehat{\theta}_U$ and $\widehat{\theta}_{MH}$ to converge to the true value, $\theta = 0$, as $K \to \infty$. The rate of convergence, however, also depends on the the marginal probabilities, π_{2+} and π_{+2}. We see that when the marginal probabilities are uniformly near the boundaries ($\pi_{2+} = \pi_{+2} = 0.01$ and $\pi_{2+} = \pi_{+2} = 0.99$), that the approximation is poor when $K = 100$ (Simulation 1a, Figure 7.27) but is reasonable when $K = 1000$ (Simulation 1b, Figure 7.28). Regardless of whether $K = 100$ or $K = 1000$, the histograms of $\widehat{\theta}_{0.5}$ are not centred around the true value, $\theta = 0$, unless the marginal probabilities are well away from the boundaries ($0.25 < \pi_{2+} = \pi_{+2} < 0.75$).

The bias can be eliminated only by having ultra-deep sequencing coverage, such as in Simulation 1c where the average sequencing coverage is greater $10,000\times$. Even then, however, the variation of $\widehat{\theta}_{0.5}$ across the $K = 1000$ tables remains non-negligible when the marginal probabilities, π_{2+} and π_{+2} are near the boundaries.

[7] In many simulations θ is constant across the K tables and therefore the density is degenerate.

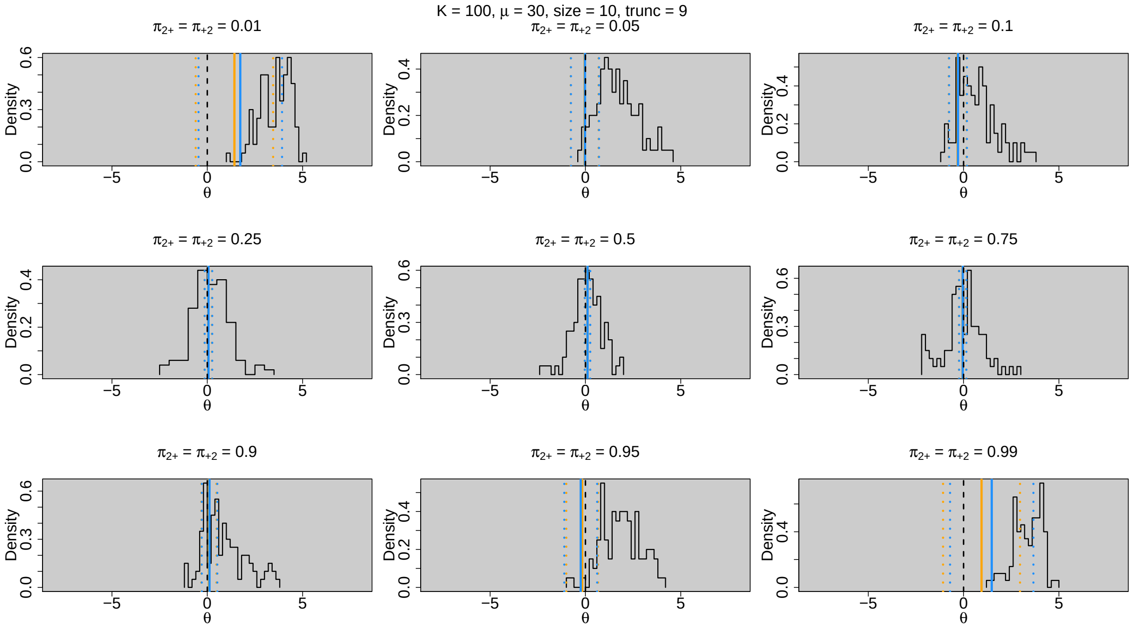


Figure 7.27: Results of simulation 1a ($K = 100$). Estimating the log odds ratio, $\theta = 0$ (shown by the black line), from a $2 \times 2 \times 100$ contingency table when the marginal probabilities, π_{2+} and π_{+2}, are equal. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

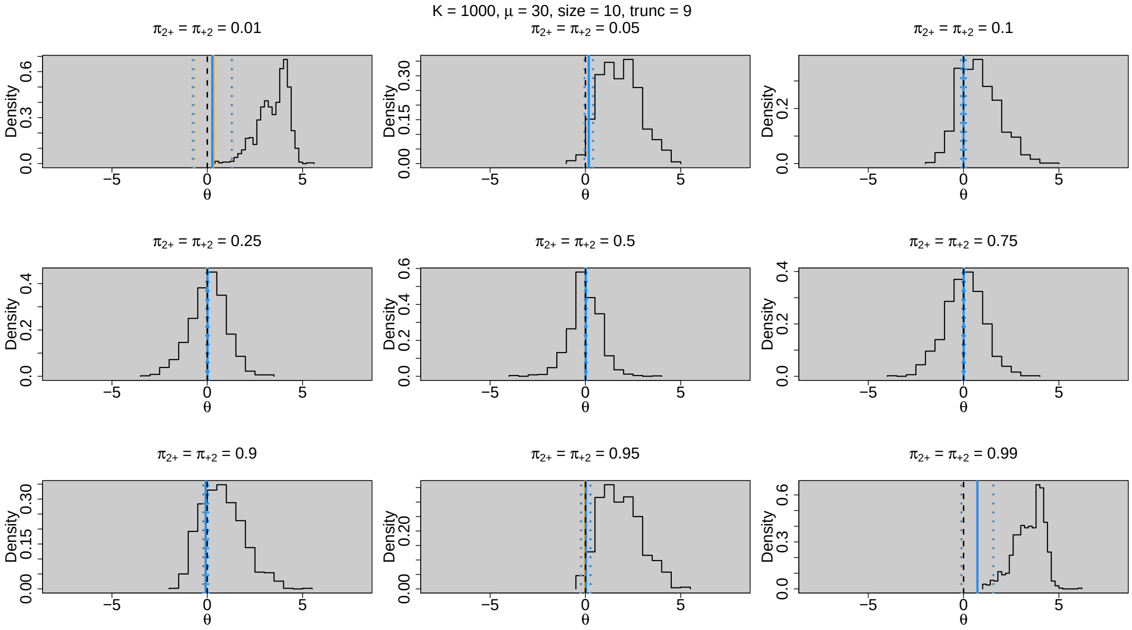


Figure 7.28: Results of simulation 1b ($K = 1000$). Estimating the log odds ratio, $\theta = 0$ (shown by the black line), from a $2 \times 2 \times 1000$ contingency table when the marginal probabilities, π_{2+} and π_{+2}, are equal. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

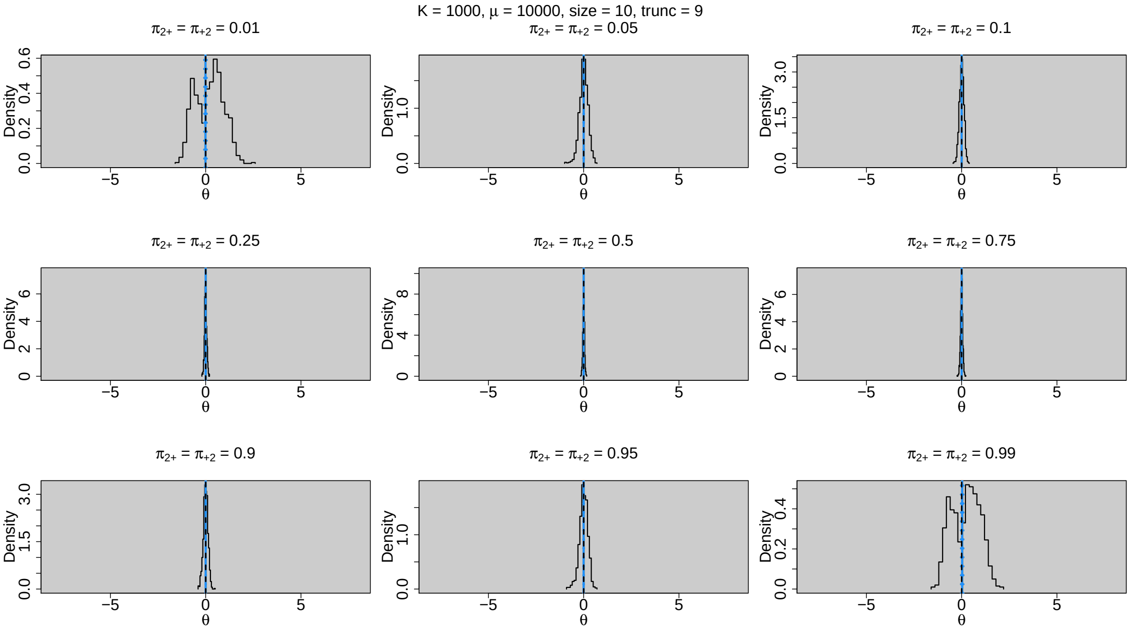


Figure 7.29: Results of simulation 1c ($K = 1000$). Estimating the log odds ratio, $\theta = 0$ (shown by the black line), from a $2 \times 2 \times 1000$ contingency table when the marginal probabilities, π_{2+} and π_{+2}, are equal. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is $10,000$.

Next, we explore what happens when we allow the marginal probabilities, π_{2+} and π_{+2}, to vary across the $K = 100$ tables for a fixed (possibly non-zero) value of θ. To do so, we simulate $\pi_{2+}, \pi_{+2} \overset{d}{=} Uniform(0, 1)$ for each 2×2 table. As is expected from the theory, the results of Simulation 2 show that the Mantel-Haenszel estimator does an excellent job of estimating the true value of θ (Figure 7.32). By contrast, the estimate formed by collapsing over the K tables, $\widehat{\theta}_U$, is well away from the true value and the distribution of $\widehat{\theta}_{0.5}$ is skewed and widely dispersed.

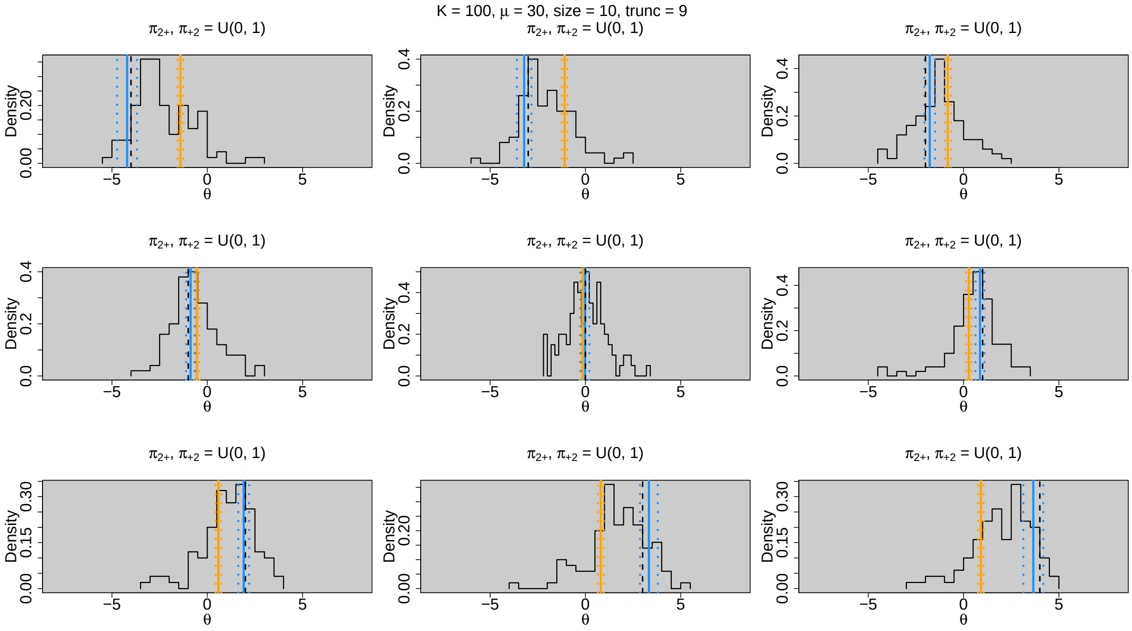


Figure 7.30: Results of simulation 2 ($K = 100$). Estimating the log odds ratio, θ (shown by the black line, $-4, -3, \ldots, 3, 4$ across the panels), from a $2 \times 2 \times 100$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $Uniform(0, 1)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

In a sense, the $\pi_{2+}, \pi_{+2} \overset{d}{=} Uniform(0, 1)$ assumption in Simulation 2 is too easy. As we have seen in Chapter 4, the marginal probabilities of DNA methylation are not uniformly distributed but, rather, are bimodal and mostly near the boundaries. Furthermore, each methylation locus in a pair will likely have a similar marginal probability of methylation. To simulate this more relevant scenario, Simulation 3 uses $\pi_{2+} = \pi_{+2} \overset{d}{=} Beta(0.3, 0.2)$, which has a similar shape to the genome-wide distribution of β-values observed in most mammalian genomes (Figure 7.31).

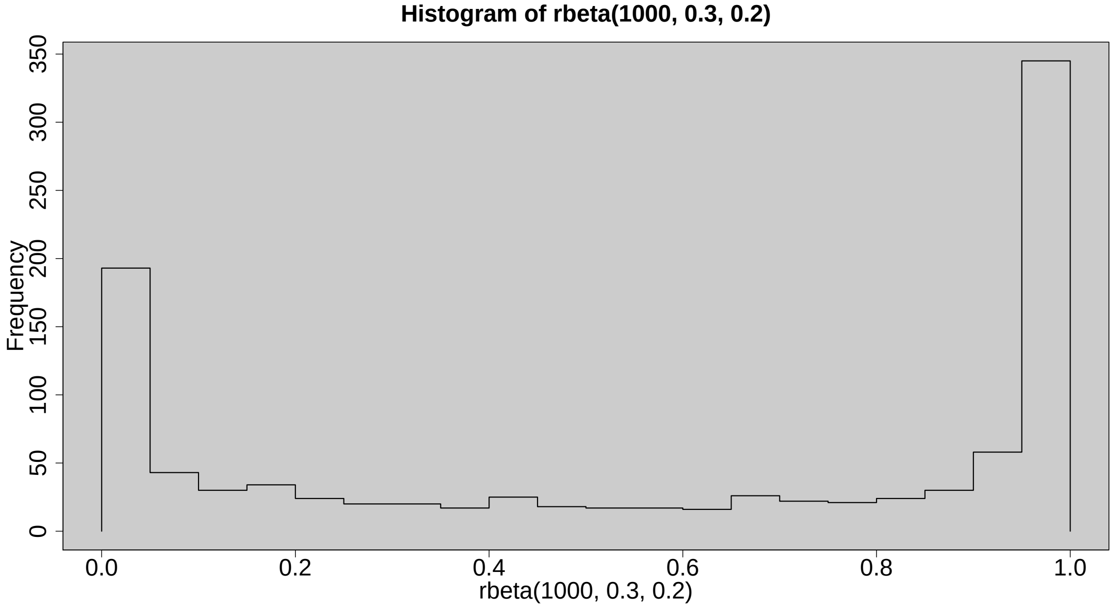


Figure 7.31: Histogram of 1000 points simulated from a $Beta(0.3, 0.2)$ distribution.

The results of this simulation (Simulation 3, Figure 7.32) are basically an exaggerated version of Simulation 2. The Mantel-Haenszel estimator, $\widehat{\theta}_{MH}$, remains an excellent estimator; the unconditional log odds ratio estimator of the collapsed table does an awful job, regularly returning $\widehat{\theta}_U > 0$ when in fact $\theta < 0$; and the distribution of $\widehat{\theta}_{0.5}$ is even more widely dispersed and typically not centred around θ.

We now turn our attention to what happens when the odds ratio is heterogeneous across the K tables. Although results are shown for $\widehat{\theta}_{0.5}$ and $\widehat{\theta}_U$, we do not discuss these estimators further since we have seen that these perform poorly in even the simplest simulation scenarios.

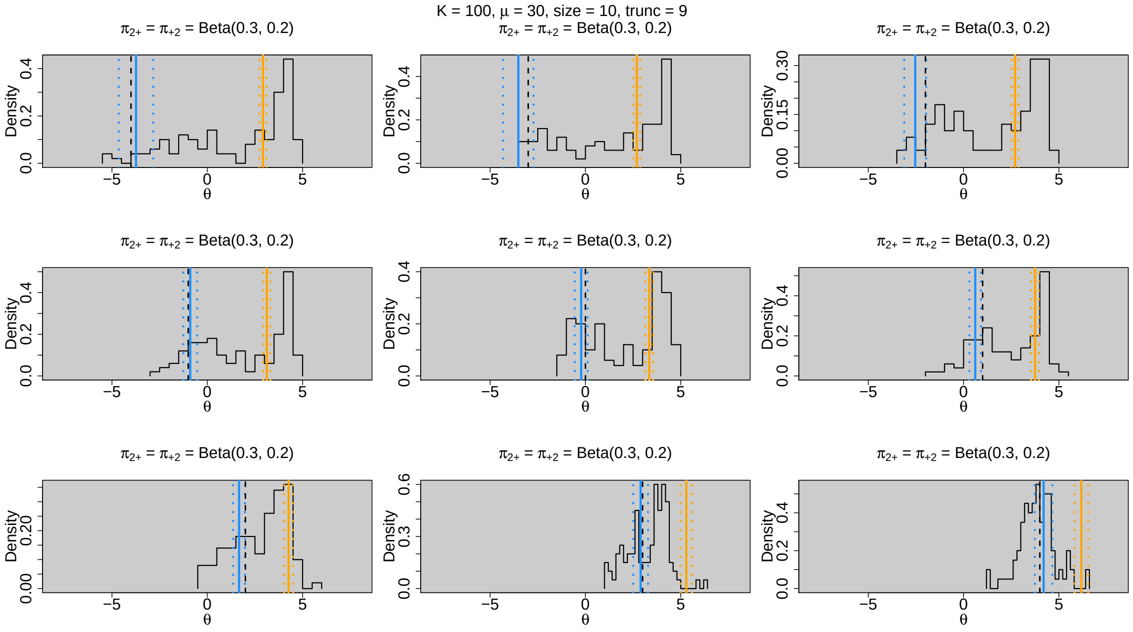


Figure 7.32: Results of simulation 3 ($K = 100$). Estimating the log odds ratio, θ (shown by the black line, $-4, -3, \ldots, 3, 4$ across the panels), from a $2 \times 2 \times 100$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $Beta(0.3, 0.2)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

In these remaining simulations, all parameters are as for Simulation 3 except that the true log odds ratio of each table, θ_k, comes from the following distributions:

- Simulation 4a and Simulation 4b: $\theta_k \overset{d}{=} Gaussian(\theta_0, 1)$, where $\theta_0 = -4, \ldots, 4$ is identical across the K tables.
- Simulation 5: $\theta_k \overset{d}{=} Skewed\text{-}Gaussian(\theta_0, 1, -10)$, where $\theta_0 = -4, \ldots, 4$ is identical across the K tables.
- Simulation 6: $\theta_k \overset{d}{=} 0.5 \times Gaussian(-\theta_0, 1) + 0.5 \times Gaussian(\theta_0, 1)$, i.e. a mixture distribution.

Simulations 4a, 4b and 5 (Figures 7.33, 7.34 and 7.35) demonstrate that the Mantel-Haenszel estimator, $\widehat{\theta}_{MH}$, is a reasonable estimator of the centre of the distribution of the true log odds ratios, θ_k. It performs better in Simulation 4a and 4b, where the distribution of the θ_k is symmetric, as opposed to Simulation 5, where the distribution of the θ_k is skewed. Unsurprisingly, $\widehat{\theta}_{MH}$ does a better job when $K = 1000$ (Simulation 4b, Figure 7.34) than when $K = 100$ (Simulation 4a, Figure 7.33).

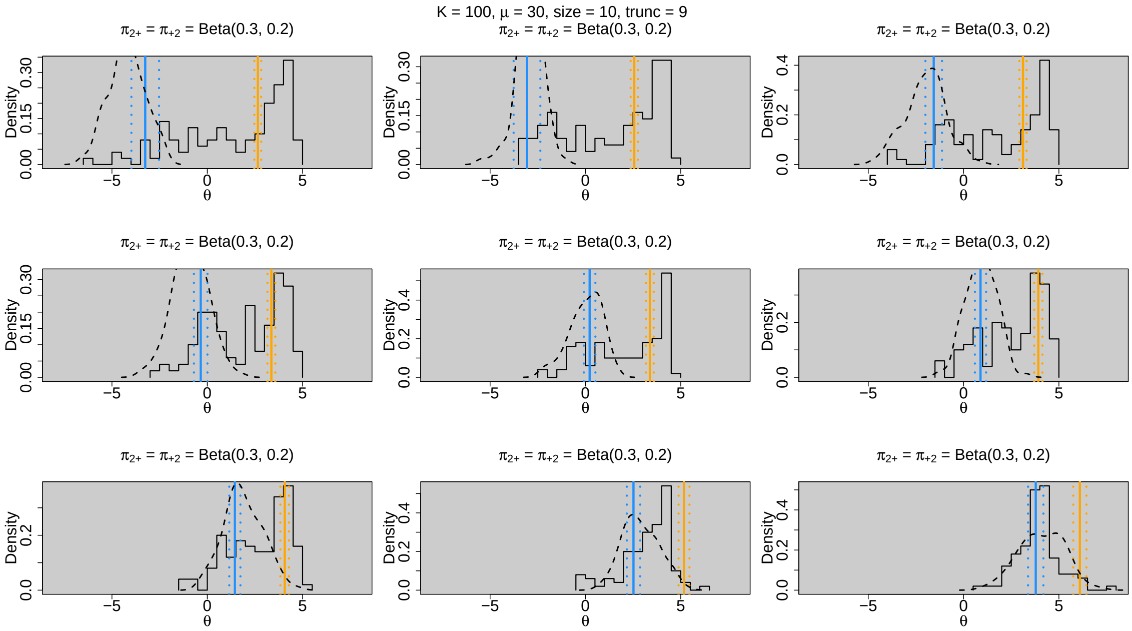


Figure 7.33: Results of simulation 4a ($K = 100$). Estimating a heterogeneous log odds ratio, $\theta_k = Gaussian(\theta_0, 1)$ (density shown by the black curve), from a $2 \times 2 \times 100$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $Beta(0.3, 0.2)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

To take this heterogeneity even further, Simulation 6 (Figure 7.36) shows that if the distribution of θ is multimodel, then $\widehat{\theta}_{MH}$ will estimate the 'average' effect. Of course, the utility of any point estimate is severely reduced when the true effect is multimodal.

One final note. When the true log odds ratios, θ_k, is heterogeneous, then the asymptotic variance of the estimator Mantel-Haenszel estiamtor, $\hat{\sigma}(\widehat{\theta}_{MH})^2$, is **not** an estimate of the variance of the true odds ratios, $var(\theta_k)$. We can see in Simulations 4a, 4b, 5 and 6 (Figures 7.33, 7.34, 7.35 and 7.36) that the asymptotic 95% confidence interval of $\widehat{\theta}_{MH}$ only covers a small amount of the variation in the θ_k.

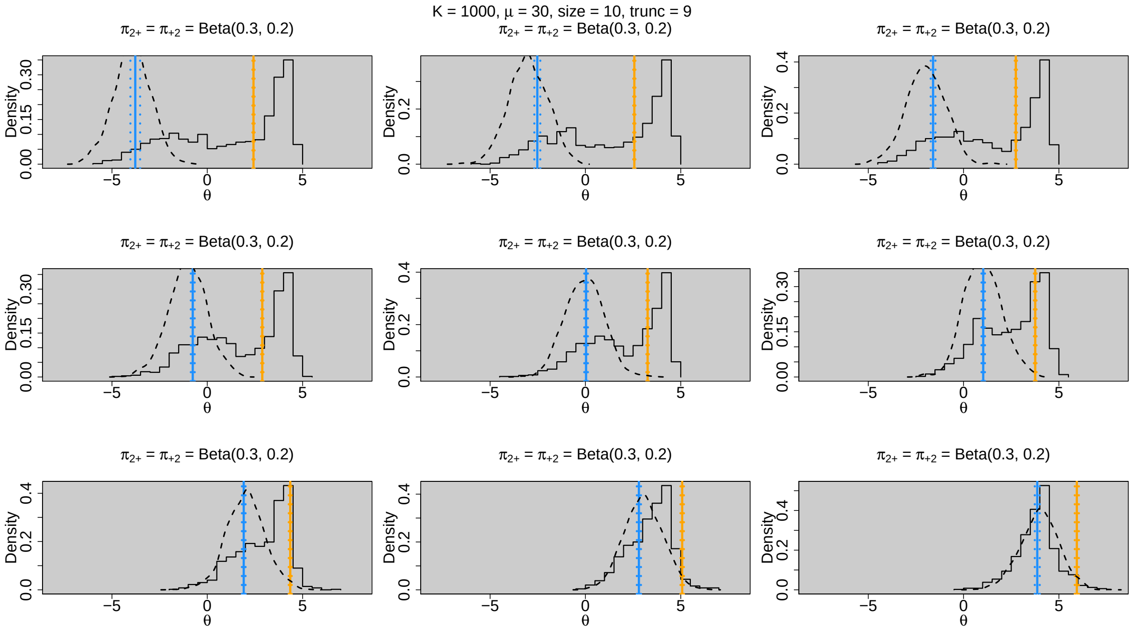


Figure 7.34: Results of simulation 4b ($K = 1000$). Estimating a heterogeneous log odds ratio, $\theta_k = Gaussian(\theta_0, 1)$ (density shown by the black curve), from a $2 \times 2 \times 1000$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $Beta(0.3, 0.2)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

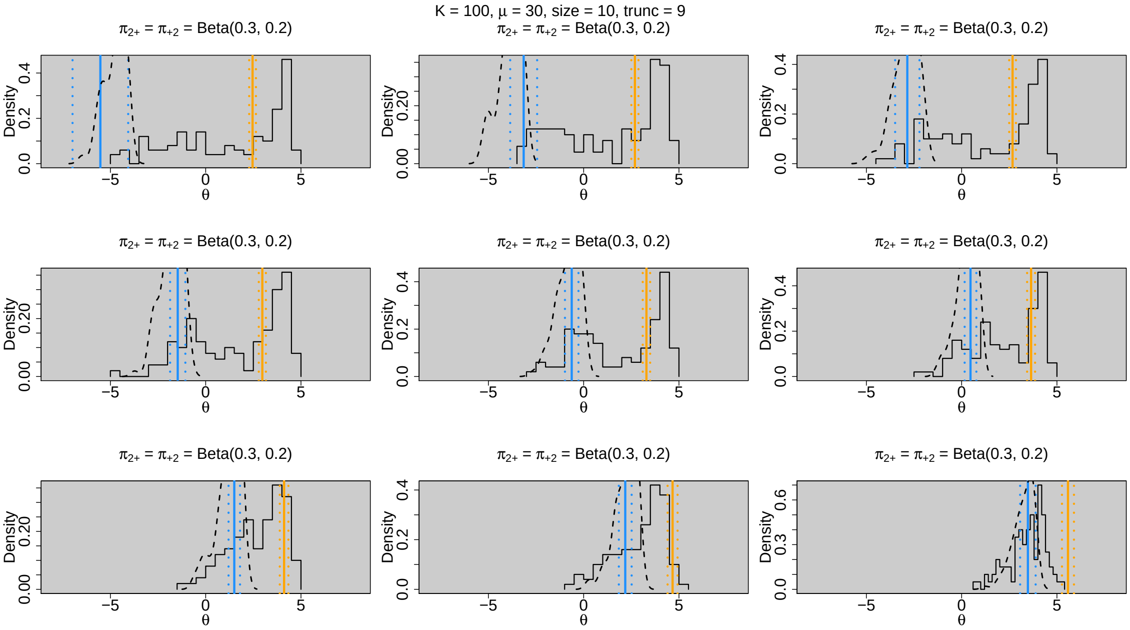


Figure 7.35: Results of simulation 5 ($K = 100$). Estimating a heterogeneous log odds ratio, $\theta_k = \mathit{Skewed\text{-}Gaussian}(\theta_0, 1, -10)$ (density shown by the black curve), from a $2 \times 2 \times 100$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $\mathit{Beta}(0.3, 0.2)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

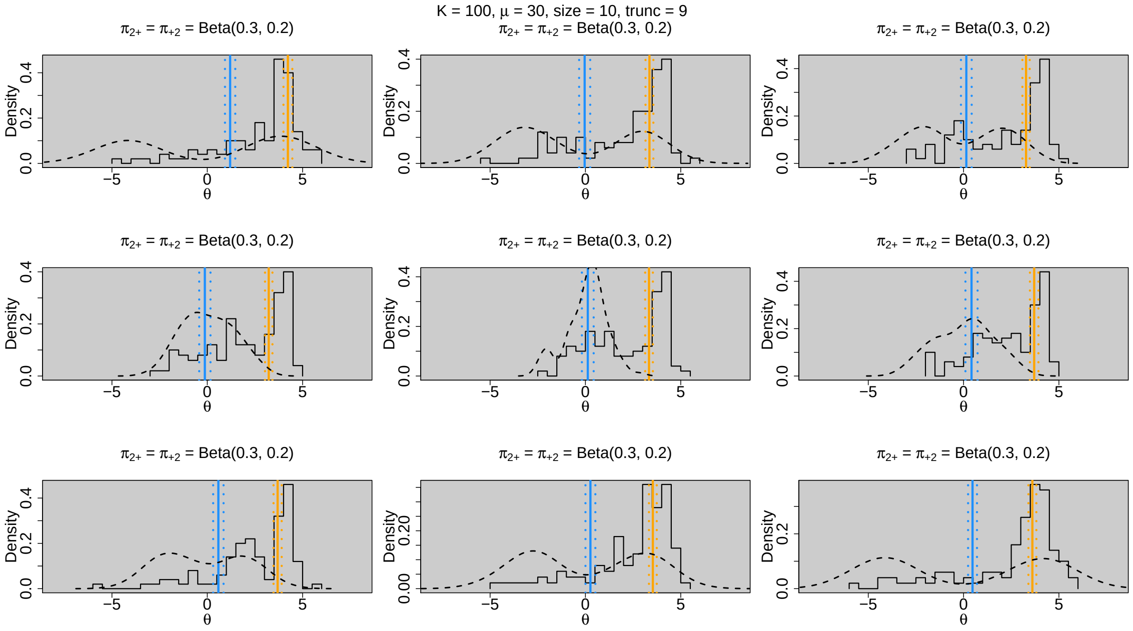


Figure 7.36: Results of simulation 5 ($K = 100$). Estimating a heterogeneous log odds ratio, $\theta_k = 0.5 \times Gaussian(-\theta_0, 1) + 0.5 \times Gaussian(\theta_0, 1)$ (density shown by the black curve), from a $2 \times 2 \times 100$ contingency table where the marginal probabilities, π_{2+} and π_{+2}, are distributed as $Beta(0.3, 0.2)$ random variables. The results of three estimators of θ are shown: $\widehat{\theta}_{0.5}$ (histogram); the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_U$ (coloured orange); and the point estimate (inner line) and 95% confidence interval (outer lines) of $\widehat{\theta}_{MH}$ (coloured blue). The average sample size (sequencing-coverage) of each 2×2 table is 30.

7.2.5 Summary

We summarise the results of the simulation study with respect to our aim of estimating within-fragment co-methylation from whole-genome bisulfite-sequencing data.

This simulation study demonstrates that we cannot estimate within-fragment co-methylation at individual pairs of methylation loci using the simple odds ratio estimator, $\theta_{0.5}$, unless ultra-deep[8] sequencing is used ($d > 10000\times$). The distributions of $\widehat{\theta}_{0.5}$ at typical sequencing depths ($d \approx 30$) are highly biased and dispersed, even under very simple simulation scenarios. Therefore, some degree of aggregation is required in order to study within-fragment co-methylation.

Suppose that we aggregate K CpG pairs to form a $2 \times 2 \times K$ contingency table. When little can be assumed about this $2 \times 2 \times K$ contingency table, the Mantel-Haenszel estimator is the most appropriate of those considered. Firstly, it is highly robust to different marginal probabilities across the K tables. Secondly, when the odds ratios are heterogeneous across the K tables, the Mantel-Haenszel estimator will estimate the 'average' effect.

Of course, the utility of this 'average' will depend on the distribution of θ, which is unknown in practice. However, we might explore its variability by computing the Mantel-Haenszel estimate under a range of aggregation strategies (i.e. across different k).

7.3 Within-fragment co-methylation

This section describes a simple method to study within-fragment co-methylation by analysing methylation patterns at pairs of methylation loci (2-tuples). For simplicity, we will use pairs of CpGs, but the method is applicable to any methylation type. This method is to be implemented in the `cometh()` function that is part of the `MethylationTuples` software[9].

The aim of this analysis is to address the questions:

1. How dependent are methylation states at nearby methylation loci on the same DNA fragment?

[8] Ultra-deep sequencing has its own problems, such as an increased effect of PCR-bias.

[9] At the time of writing, this is available as a stand alone `mantelhaen()` function, but it will soon be one of several strategies of estimating within-fragment co-methylation using the `cometh()` function.

2. What factors influence this dependence?

7.3.1 Methods

There are three important decisions to make in our analysis:

1. How to define the CpG pairs?
2. How to aggregate the CpG pairs?
3. Which estimator to use?

Answers to the latter two are informed by Section 7.2, in particular the simulation study.

How to define the CpG pairs

The construction of CpG pairs is very similar to constructing the pairs of β-values, described in Section 7.1.1. However, we are now limited to analysing co-methylation of loci that can be captured within a single read. With current technology this means we are limited to studying within-fragment co-methylation for pairs with IPD less than approximately 250. As in Section 7.1, we will consider pairs with $NIL = 0$ and $NIL \geqslant 0$.

How to aggregate the CpG pairs?

The counts of methylation patterns at each CpG pair can be summarised by a 2×2 contingency table. The sequencing depth of each pair is typically low. Furthermore, for each locus in the pair, the marginal probability that it is methylated is very often close to zero or one. Therefore, in light of the results in Section 7.2.4, we must aggregate these 2×2 tables in order to perform any useful inference.

The ideal level of aggregation combines those pairs with homogeneous odds ratios. Of course, this information is unknown to us. Instead, we will aggregate by chromosome to explore the variation of co-methylation across chromosomes and compare these to genome-level estimates. The resulting estimates must be interpreted as the ‘average’ degree of within-fragment co-methylation across the aggregation levels.

Which estimator to use?

We will use the Mantel-Haenszel odds ratio estimator since it is robust to variable marginal probabilities. This allows us to combine data for pairs from different regions of the genome, e.g., regions that are hypomethylated, partially methylated and hypermethylated.

7.3.2 Results

We analyse all 40 samples in the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets. For each sample we compute $\widehat{\theta}_{MH}$ using different levels of aggregation (listed here in decreasing order):

1. IPD-only: A $2 \times 2 \times K$ table for each IPD.
2. IPD-CGI: A $2 \times 2 \times K$ table for each IPD and 'CpG island status' (CGI-status) combination. The CGI-status of each pair is whether it is *inside* a CpG island or *outside* a CpG island[10].
3. IPD-chromosome: A $2 \times 2 \times K$ table for each IPD and chromosome combination.

No minimum sequencing coverage cutoff is required since the Mantel-Haenszel estimator appropriately downweights pairs with low sequencing coverage.

We will first compare the results of chromosome-level estimates to genome-level estimates[11].

Chromosome-level $NIL = 0$ analyses

Figure 7.37, 7.38, 7.39 and 7.40 show the results using CpG pairs with $NIL = 0$ for the *EPISCOPE*, *Lister*, *Seisenberger* and *Ziller* datasets, respectively.

Unsurprisingly, there is variation in the chromosome-level estimates for a given IPD. Overall, however, we see that the genome-level and chromosome-level estimates follow a

[10] Pairs spanning the boundary of a CpG island are ignored. There are few such pairs when $NIL = 0$. While there are considerably more such pairs when $NIL \geqslant 0$, I have ignored these in favour of simplicity.

[11] Some female samples have apparent Y chromosome data, e.g., all the *EPISCOPE* samples. This is indicative of mapping errors and such data should be ignored.

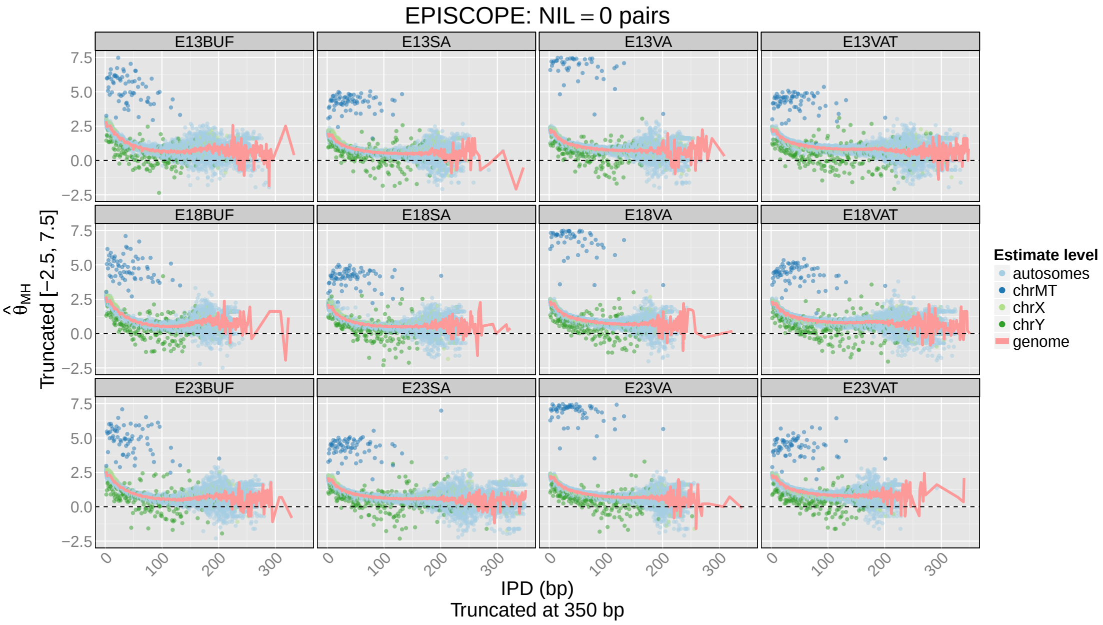


Figure 7.37: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *EPISCOPE* dataset using $NIL = 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads.

similar trend for all samples[12]: co-methylation is strongest for smaller IPDs, decaying fairly rapidly before levelling out by $IPD = 20$ to 50. As the IPD increases, we see more variation in the chromosome-level and genome-level estimates, but this is due to the smaller sample sizes we have for pairs with $NIL = 0$ at larger IPDs.

Notably, although the genome-level estimates decay as a function of IPD, $\widehat{\theta}_{MH}$ is almost entirely positive over the observable range of IPDs, suggesting that within-fragment co-methylation extends for at least a few hundred bp. Furthermore, in samples with sufficiently long paired-end reads, there is a hint of an upturn in the genome-wide $\widehat{\theta}_{MH}$ at approximately $IPD = 180$. This is approximately the distance between two CpGs on adjacent nucleosomes (see Section 1.1.2), and may be evidence of the three-dimensional structure of the genome affecting co-methylation.

[12] The obvious outlier on each of these plots is the mitochondrial DNA. The mitochondria are typically almost completely unmethylated, i.e. the marginal probability of methylation at a CpG on the mitochondria is very close to zero, and so the concept of co-methylation is perhaps not particularly meaningful here.

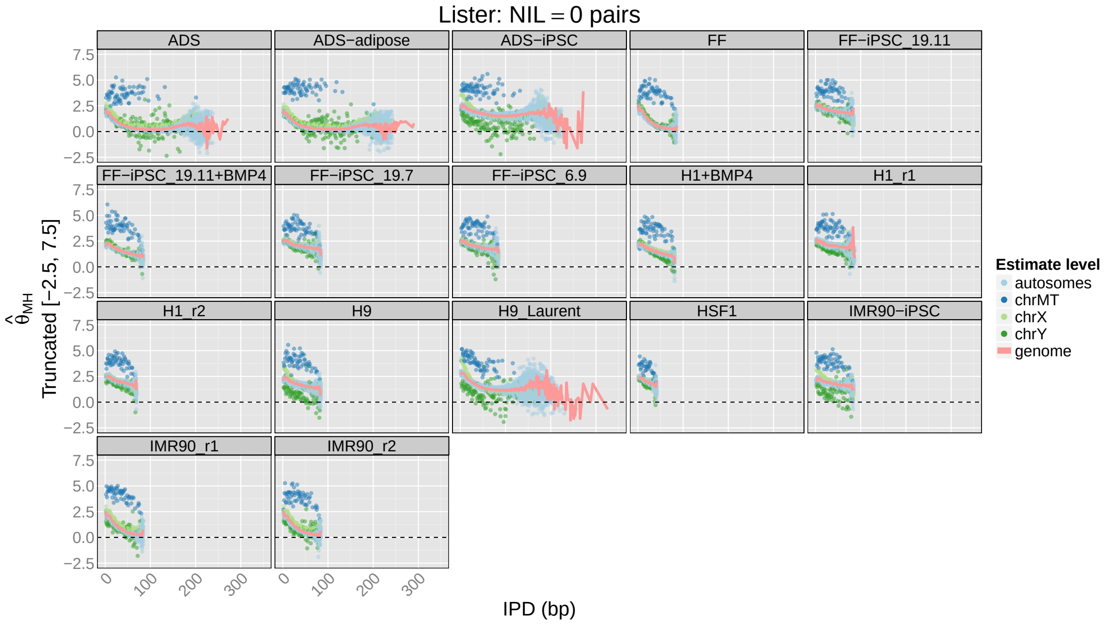


Figure 7.38: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Lister* dataset using $NIL = 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. Only the *ADS*, *ADS-adipose*, *ADS-iPSC* and *H9_Laurent* samples were sequenced with paired-end reads.

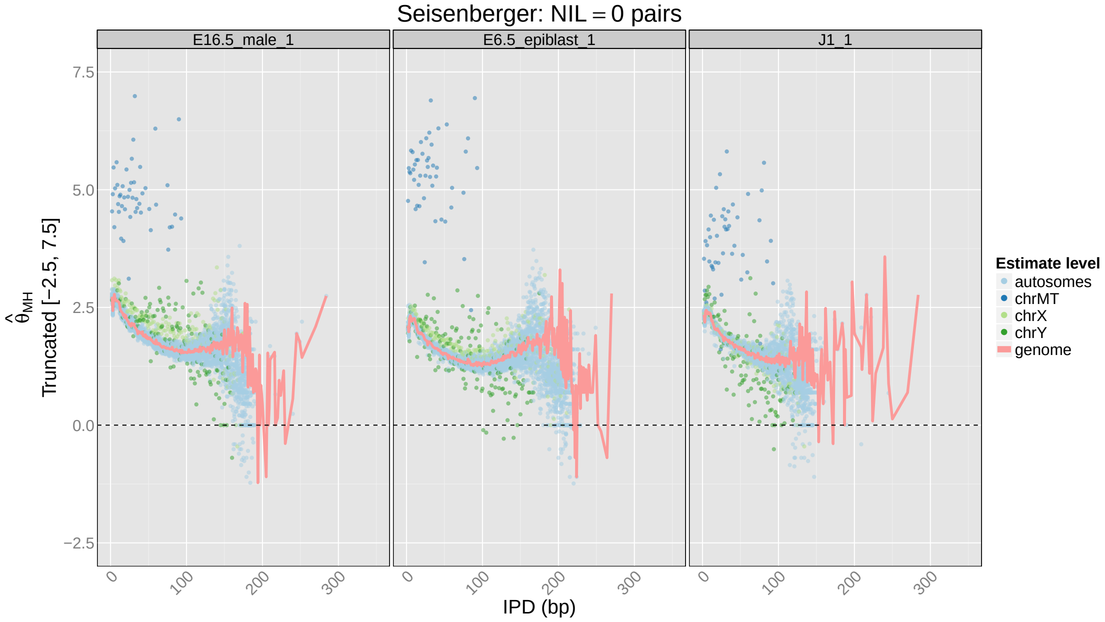


Figure 7.39: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *Seisenberger* dataset using $NIL = 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads.

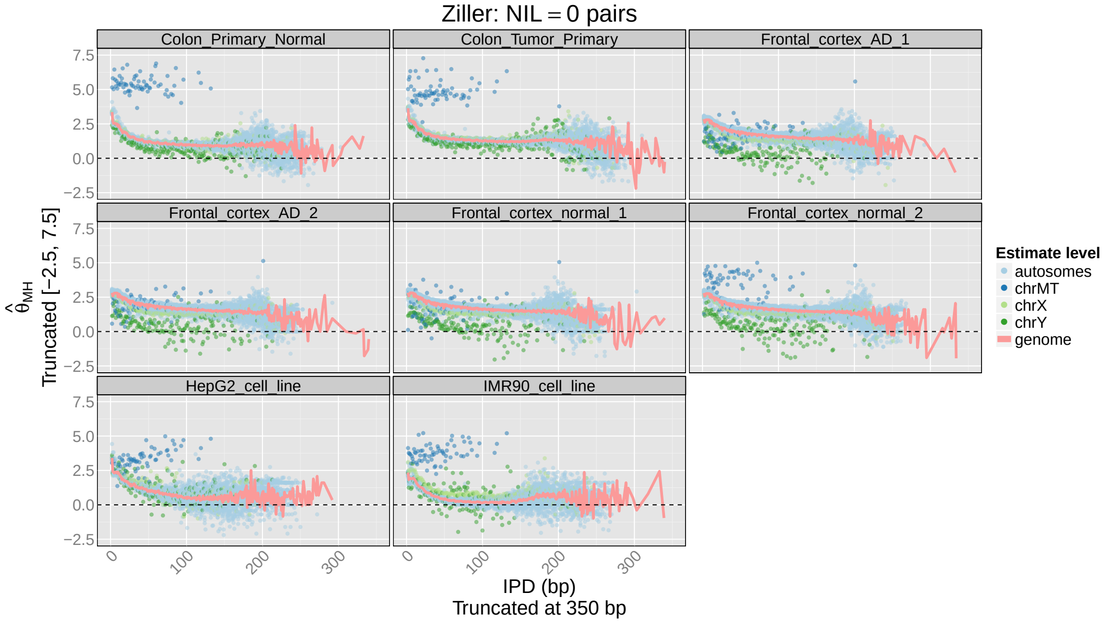


Figure 7.40: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Ziller* dataset using $NIL = 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads.

Chromosome-level $NIL \geqslant 0$ analyses

The equivalent plots for the $NIL \geqslant 0$ pairs paint a similar picture, as shown in Figures 7.41, 7.42, 7.43 and 7.44. In addition to the mitochondria, the estimates for X chromosome are frequently distinguishable from the rest of the chromosome-level estimates; e.g., the *ADS*-based samples (*ADS*, *ADS-adipose* and *ADS-iPSC*; *ADS* is a female cell line), the *H9*-based samples (*H9* and *H9_Laurent*; sex of cell line not reported in Lister *et al.* [2011]), the *IMR90*-based samples (*IMR90_r1*, *IMR90_r2*, *IMR90_cell_line*, *IMR90-iPSC*; *IMR90* is a female cell line).

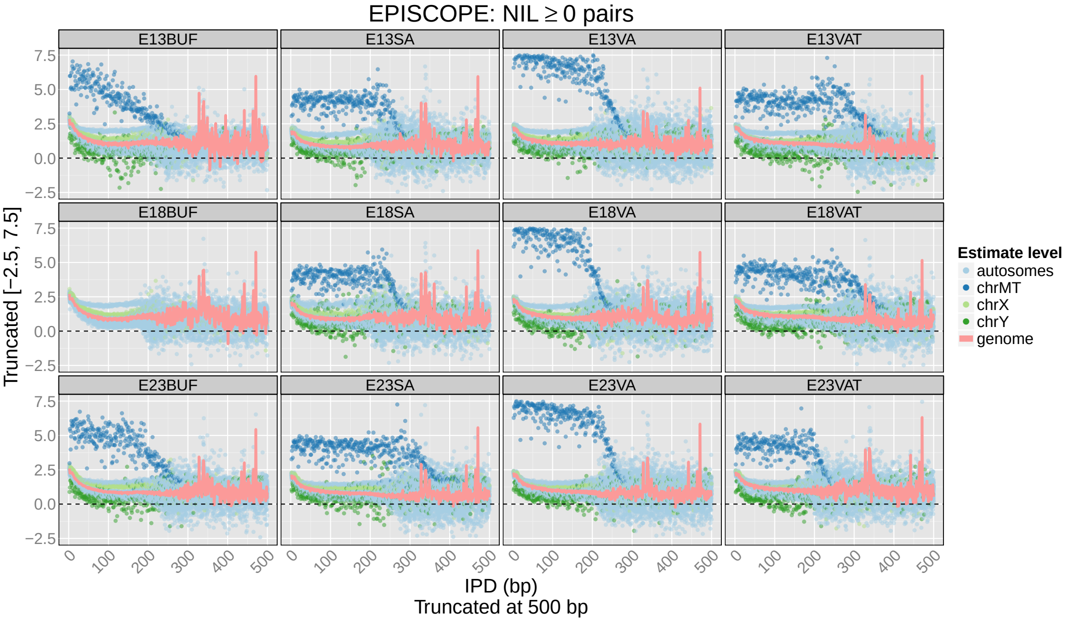


Figure 7.41: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *EPISCOPE* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads. There is no Y chromosome or mtDNA data for *E18BUF* due to a coding error that meant these chromosomes weren't processed by `methtuple` for `2ac` tuples. This omission does not affect the other results.

By analysing CpG pairs with $NIL \geqslant 0$ we are increasing the heterogeneity of the K 2×2 tables, which consequently makes more difficult the interpretation of these results. However, it does give the advantage of allowing the analysis of pairs with larger *IPD*s. Nonetheless, the short DNA fragments and read lengths of our data make it difficult to perform reliable chromosome-level inference in even the highest-quality samples for

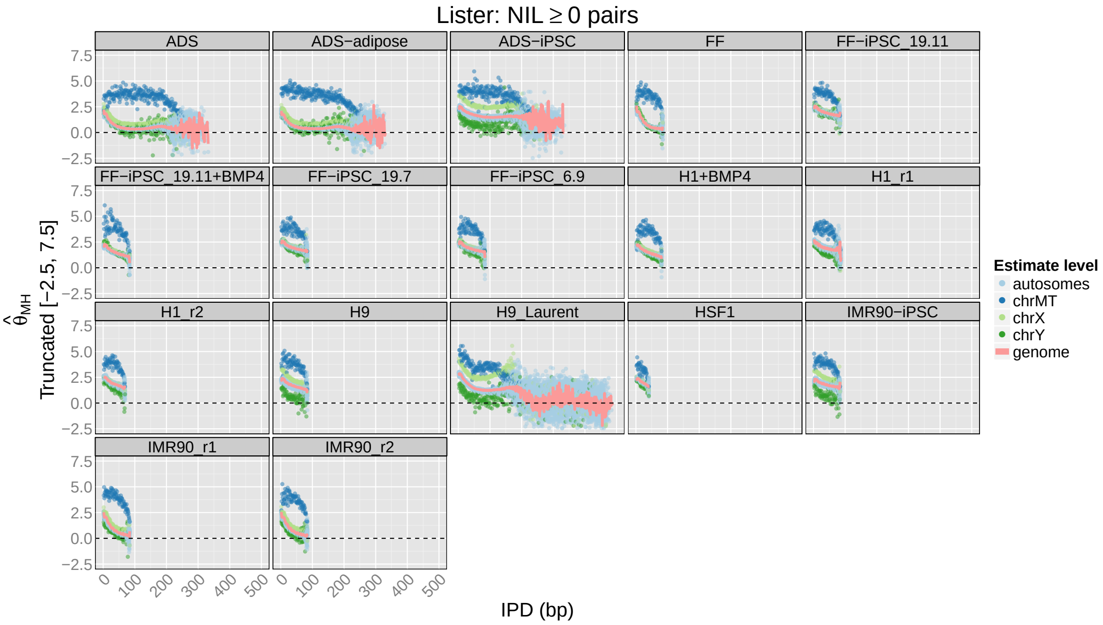


Figure 7.42: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Lister* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. Only the *ADS*, *ADS-adipose*, *ADS-iPSC* and *H9_Laurent* samples were sequenced with paired-end reads.

IPDs greater than 250 to 300 bp. Retreating to a genome-level analysis suggests that within-fragment co-methylation is active over at least a few hundred basepairs in distance.

In many of the $NIL \geqslant 0$ plots there are samples with autosomal $\widehat{\theta}_{MH}$ that are noticeably separated from the main cloud of autosomal estimates. For example, this can be readily observed in the *Colon_Primary_Normal*, *Colon_Primary_Tumour* samples from the *Ziller* dataset (Figure 7.44), but it is also apparent in other samples from the *Ziller* dataset and the *EPISCOPE* dataset (Figure 7.41). These plots do not distinguish estimates from different autosomes. Further analysis, however, reveals that these outliers almost all come from a single chromosome, chromosome 21.

Figures 7.45 (*EPISCOPE*), 7.46 (*Lister*) and 7.47 (*Ziller*) show the exact same plots as before, but with the chromosome 21 data highlighted by a black line. The chromosome 21 data stand out remarkably from the rest of the autosomal data for all samples from the *EPISCOPE* dataset and several samples from the *Ziller* dataset. Just as remarkable, however, is that this phenomenon is not observed in the *Lister* dataset nor for any of the

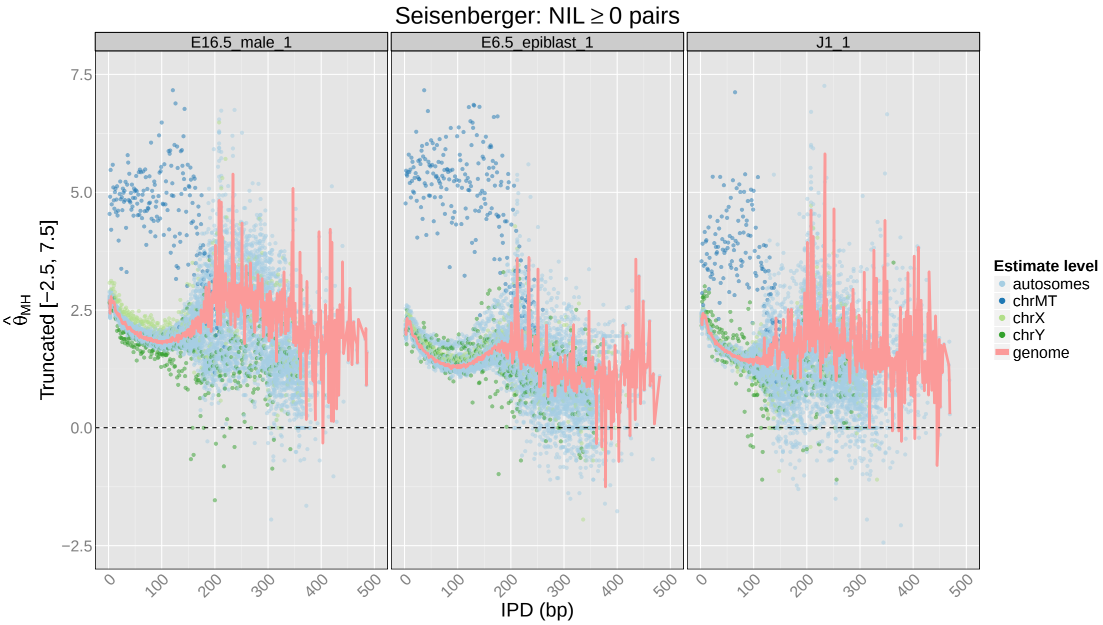


Figure 7.43: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *Seisenberger* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads.

frontal cortex samples form the *Ziller* dataset[13]. At this time I have no explanation for this result. I cannot rule out it being a technical artefact, although preliminary analyses suggest this is not the case (data not shown), and I do not have a hypothesis for a potential biological cause of this phenomenon.

[13] This phenomenon cannot occur in the *Seisenberger* dataset since these samples are mice, and mice only have 19 autosomes.

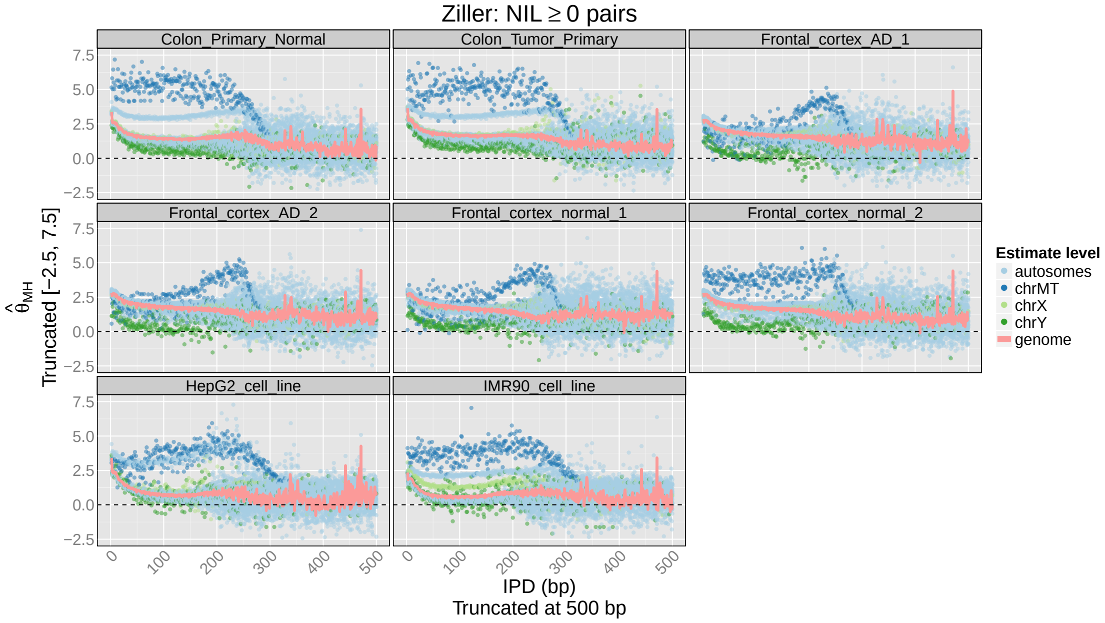


Figure 7.44: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Ziller* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. All samples were sequenced with paired-end reads.

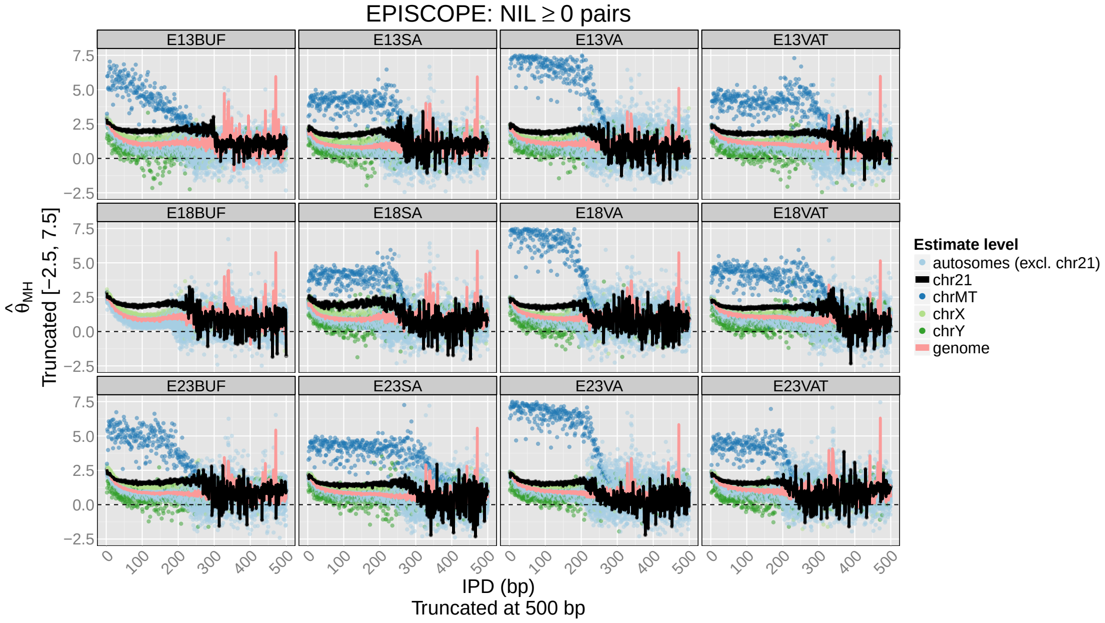


Figure 7.45: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *EPISCOPE* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. The chromosome 21 data are shown by the black line. All samples were sequenced with paired-end reads. There is no Y chromosome or mtDNA data for *E18BUF* due to a coding error that meant these chromosomes weren't processed by `methtuple` for `2ac` tuples. This omission does not affect the other results.

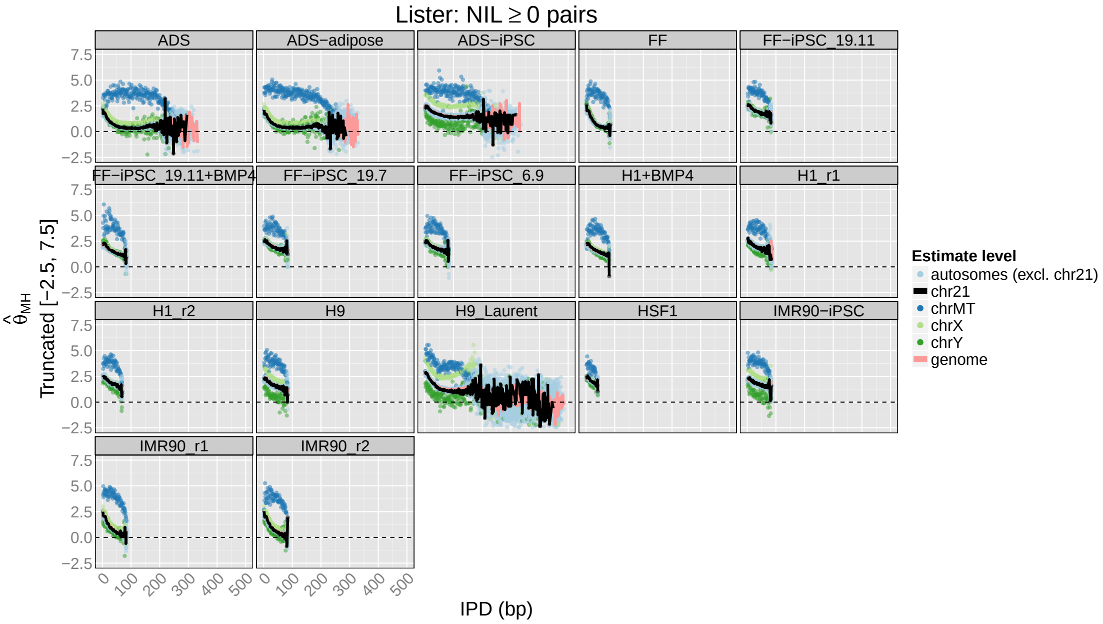


Figure 7.46: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Lister* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. The chromosome 21 data are shown by the black line. Only the *ADS*, *ADS-adipose*, *ADS-iPSC* and *H9_Laurent* samples were sequenced with paired-end reads.

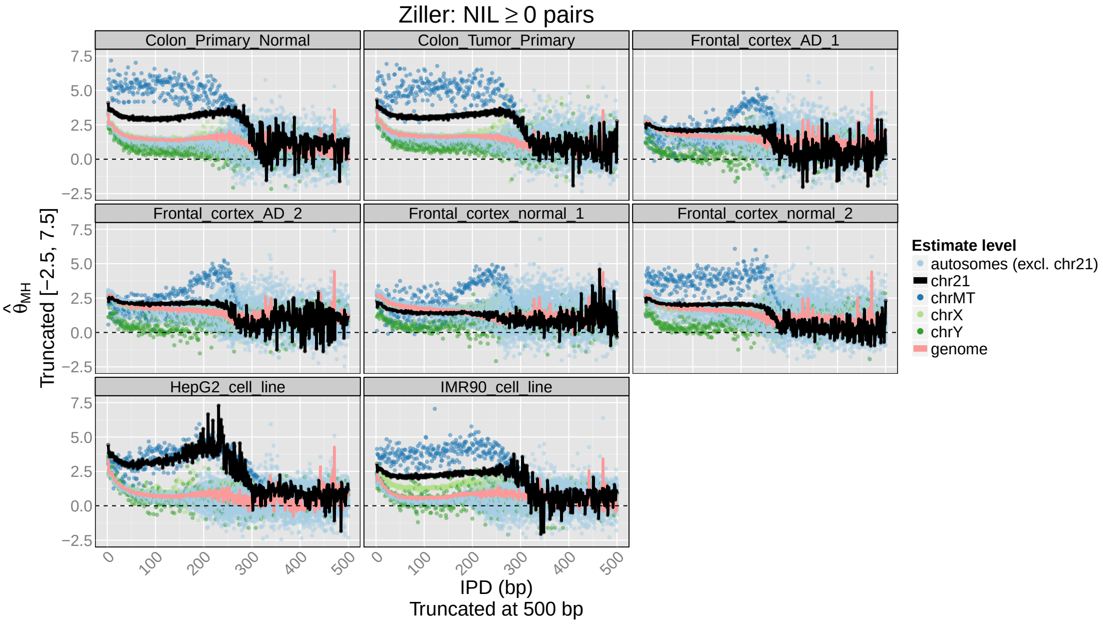


Figure 7.47: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Ziller* dataset using $NIL \geqslant 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points coloured by whether the chromosome is an autosome, X chromosome, Y chromosome or mtDNA. The chromosome 21 data are shown by the black line. All samples were sequenced with paired-end reads.

The effect of CpG islands

Since we must aggregate pairs in order to perform any meaningful analysis of within-fragment co-methylation, there will obviously be unaccounted-for heterogeneity in our estimates. In general, it will be difficult to identify sources of this heterogeneity. One obvious candidate, however, are CpG islands and so we compare estimates of within-fragment co-methylation inside CpG islands and outside of CpG islands. In order to simplify the figures for these analyses, the chromosome-level estimates are not shown. We have seen that the genome-wide estimates capture the trend of the chromosome-level estimates, at least for the autosomes, and so these are sufficient for our purpose.

We observe that there is very little difference between genome-wide estimates of θ inside and outside of CpG islands for pairs with $NIL = 0$ (Figures 7.48, 7.49, 7.50). This is perhaps a little surprising; CpG islands are typically more homogeneous in their average methylation levels and so we would reasonably expect their within-fragment methylation states are more dependent than elsewhere in the genome. However, the $NIL = 0$ results do not tell the full story.

When we analyse pairs with $NIL = 0$, we are only estimating the ‘first-order’ within-fragment co-methylation. But suppose that the methylation state of the current locus ($Z_{h,i}$) depends not only on the state at the previous locus ($Z_{h,i-1}$) but also on the states at the two previous states ($Z_{h,i-2}$ and $Z_{h,i-3}$). Or it might depend on the states at both upstream and downstream loci (i.e. $Z_{h,i+i'}$ and $Z_{i-i''}$, $i', i'' > 0$). The $NIL = 0$ analyses reduce heterogeneity by focusing on adjacent loci but cannot (by themselves) tell us whether within-fragment co-methylation is truly ‘first-order’.

To look for evidence of higher-order within-fragment co-methylation we turn to the analysis of pairs with $NIL \geqslant 0$ (Figures 7.52, 7.53, 7.54 and 7.55). We now see a fairly clear separation of the genome-wide estimates of θ, with CpG islands being more co-methylated than the rest of the genome. Moreover, the estimates of θ within CpG islands are approximately constant with respect to IPD. Also notable is that the pairs of CpGs outside of CpG islands have fairly similar estimates of θ in both the $NIL = 0$ and $NIL \geqslant 0$ analyses.

Taken together, these results suggests that within-fragment co-methylation in CpG

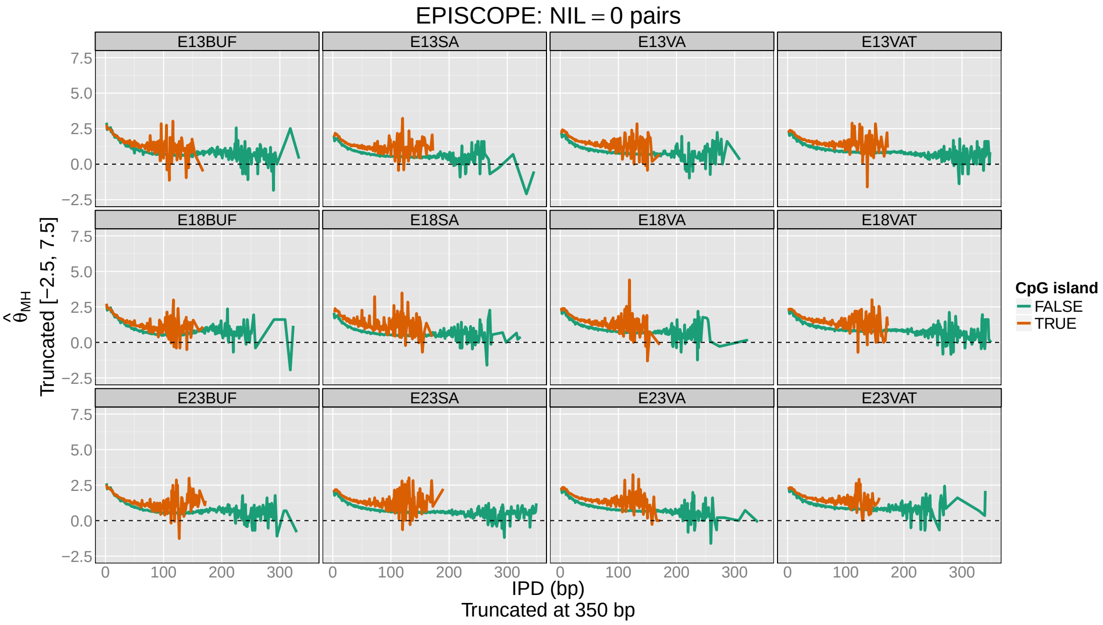


Figure 7.48: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the $EPISCOPE$ dataset using $NIL = 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

islands depends on multiple loci and that the genomic distance between these loci is of little direct importance[14]. In contrast, within-fragment co-methylation outside of CpG islands depends mostly on the adjacent locus and the distance to that locus (IPD)[15]. We might try to estimate the order of within-fragment co-methylation in CpG islands by analysing m-tuples with m > 2. This quickly becomes complicated, however, since there are $2^{\mathrm{m}-1}$ odds ratios to estimate for m-tuples. We have therefore not yet pursued this refinement.

[14] The overall density of methylation loci is likely important, which is of course related to IPDs.

[15] The ratio $\frac{NIL>0\mathrm{pairs}}{NIL\geqslant 0\mathrm{pairs}}$ is lower outside of CpG islands than inside CpG islands owing to the decreased density of CpGs outside of CpG islands. We therefore can't rule out that the similarity of $NIL = 0$ and $NIL \geqslant 0$ estimates outside of CpG islands is simply due to a lack of data.

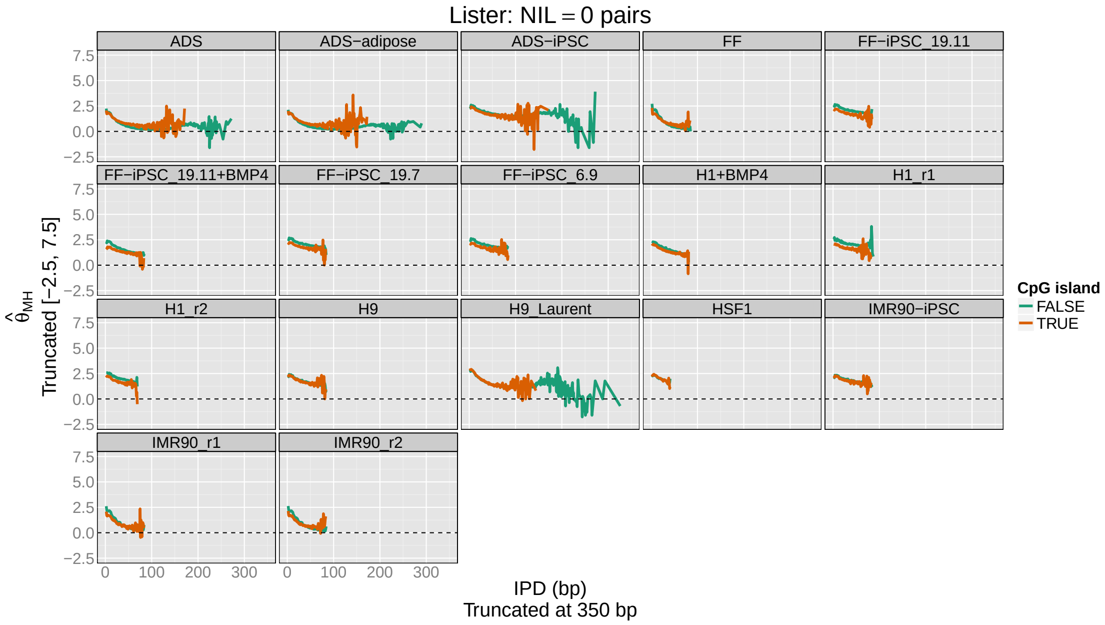


Figure 7.49: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Lister* dataset using $NIL = 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). Only the *ADS*, *ADS-adipose*, *ADS-iPSC* and *H9_Laurent* samples were sequenced with paired-end reads.

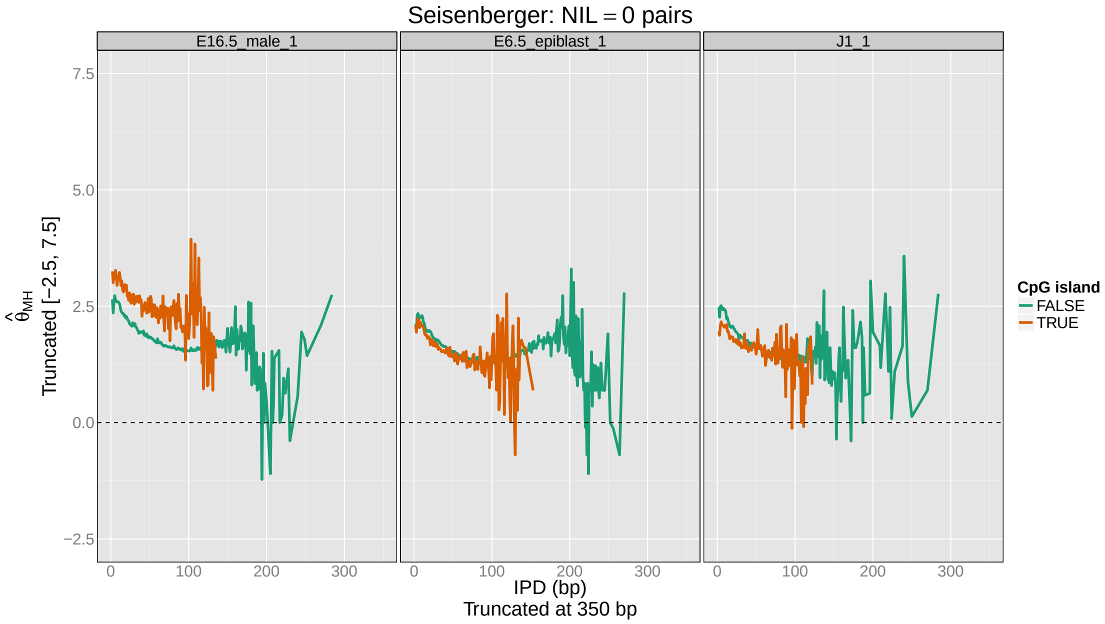


Figure 7.50: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Seisenberger* dataset using $NIL = 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

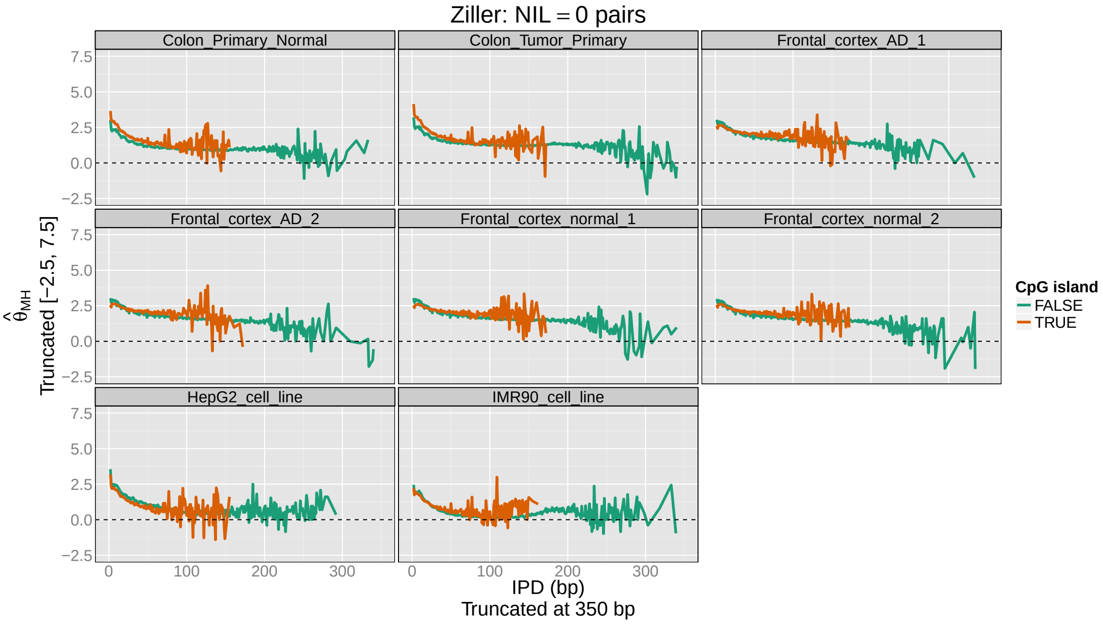


Figure 7.51: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *Ziller* dataset using $NIL = 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

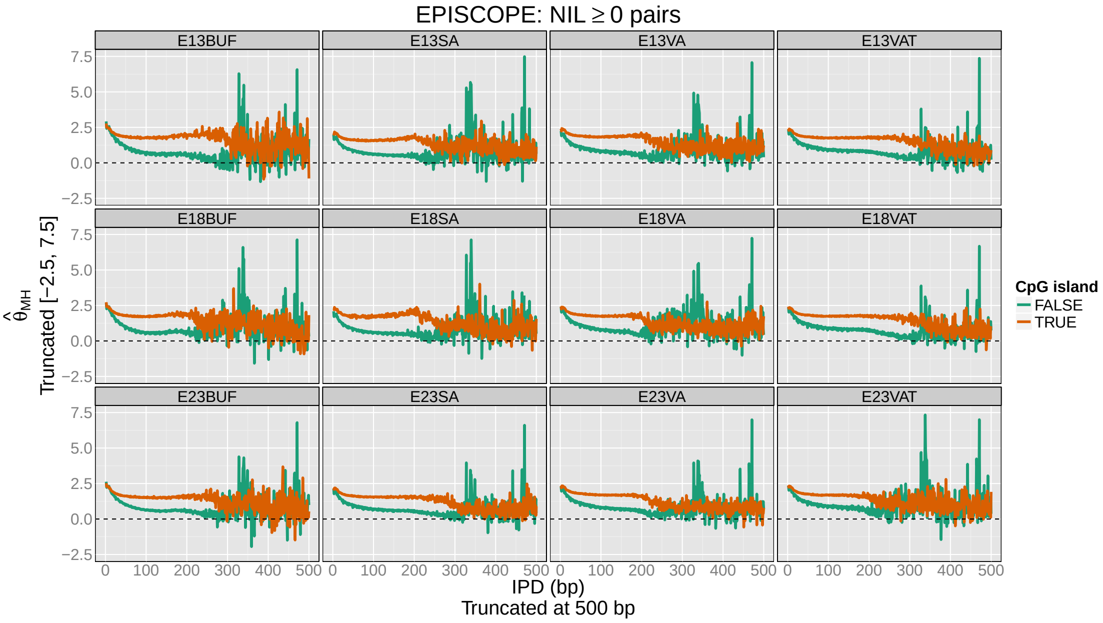


Figure 7.52: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of *IPD* for the *EPISCOPE* dataset using $NIL \geqslant 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

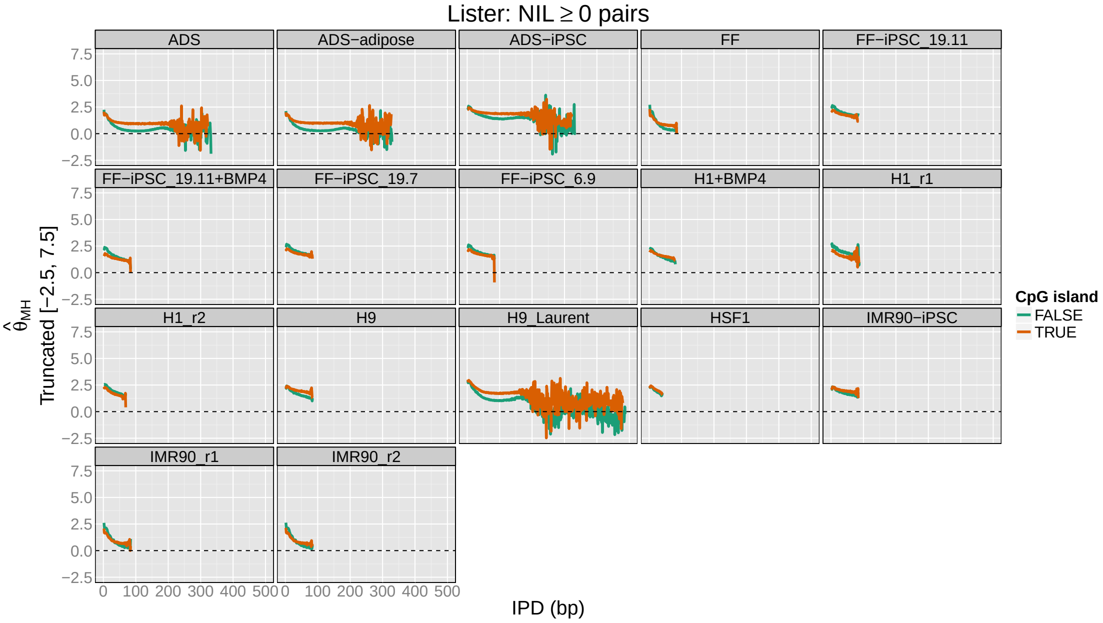


Figure 7.53: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Lister* dataset using $NIL \geqslant 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). Only the *ADS*, *ADS-adipose*, *ADS-iPSC* and *H9_Laurent* samples were sequenced with paired-end reads.

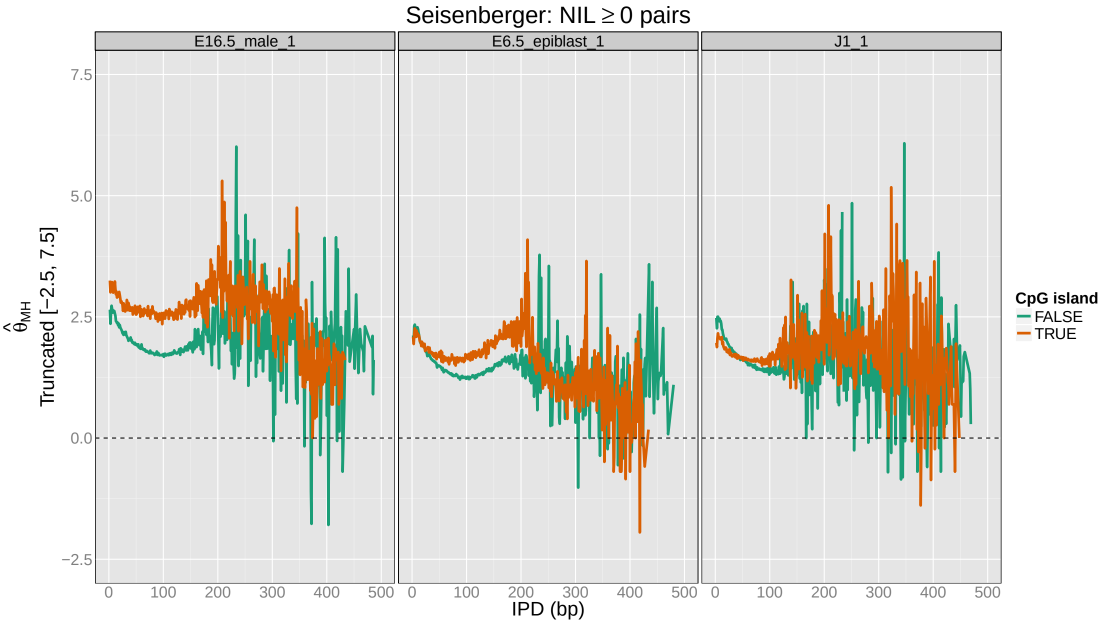


Figure 7.54: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Seisenberger* dataset using $NIL \geqslant 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

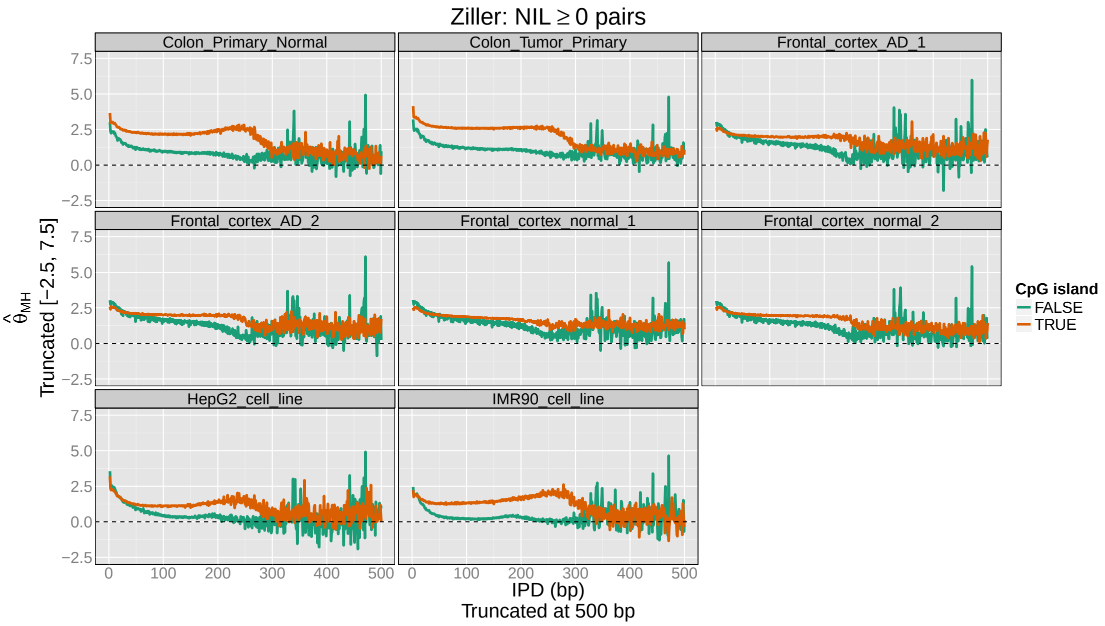


Figure 7.55: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the *Ziller* dataset using $NIL \geqslant 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI). All samples were sequenced with paired-end reads.

Variation of within-fragment co-methylation between samples

The analyses so far have highlighted the broad similarities of within-fragment co-methylation between a diverse set of samples. Reasuringly, samples done in technical duplicate (*IMR90_r1* and *IMR90_r2*; *H1_r1* and *H1_r2*) have very similar genome-level estimates of θ. Furthermore, biological samples assayed by different laboratories (*H9* and *H9_Laurent*; *IMR90_r1*, *IMR90_r2* and *IMR90_cell_line*) also have fairly similar genome-level estimates of θ[16].

There are of course differences, and we now turn our attention to these. The results of this section should be interpreted with some caution: few samples have replicates and so we will have little evidence to conclude what is driving any observed variation nor do we have sufficient power to definitively rule out large differences in co-methylation. Nonetheless, I will offer some hypotheses and support for these as appropriate.

The *EPISCOPE* dataset is the best dataset for studying between sample variation of within-fragment co-methylation. It has a fully crossed design of donors and tissues. The genome-level estimates of θ are similar across all 12 samples. Visually, the variation between tissues is greater than that between donors (the plots are more similar by column than by row).

Turning to the *Ziller* dataset, there is little difference between the genome-level estimates of θ for *Colon_Normal_Primary* and *Colon_Tumour_Primary* nor between the various frontal cortex samples. This does not exclude the possibility that there exist more local differences in co-methylation but there is no evidence that within-fragment co-methylation is globally affected by the disease status in each case.

The most interesting differences between samples occur in the *Lister* dataset. Recall that the *Lister* dataset contains four 'mini datasets': *Lister-ADS*, *Lister-FF*, *Lister-IMR90* and *Lister-H1* (see section 3.2). In each mini datasets, the pluripotent cell lines (induced pluripotent stem cells or embryonic stem cells) have larger genome-level estimates of θ than do their precursor forms. This difference is less pronounced, even absent, in the samples differentiated *in vitro* with BMP4 from the induced pluripotent stem cell lines (iPSCs).

[16] Differences in the $NIL \geqslant 0$ results between the *IMR90_r1*, *IMR90_r2* and *IMR90_cell_line* samples are due to the *IMR90_cell_line* being sequenced with paired-end reads and therefore having many more $NIL > 0$ pairs to analyse.

This suggests that pluripotent cells have stronger within-fragment co-methylation than do somatic or differentiated cells. Consistent with this is that the *H9*, *H9_Laurent* and *J1* samples (embryonic stem cell lines), the *E6.5_epiblast_1* sample (epiblasts derived from day 6.5 embryos) and the *E16.5_male_1* (male progenitor germ cells derived from day 16.5 embryos) all have relatively high genome-wide estimates of θ.

One explanation of this result, at least for the iPSCs, is that it is due to 'resetting' of the methylome during embryonic development induction of pluripotency [Lister *et al.* 2011, Stricker *et al.* 2013]. This explanation does not immediately carry over the the embryonic stem cell lines, although there is extensive reprogramming of DNA methylation during embryogenesis, which may be relevant [Seisenberger *et al.* 2013].

A second observation on the results for the *Lister* dataset is that genome-level estimates of θ decay particularly rapidly in the somatic cell lines (*ADS*, *ADS-adipose*, *FF*, *IMR90_r1* and *IMR90_r2*). This is recapitulated in the *IMR90_cell_line* sample from the *Ziller* dataset.

The reduced within-fragment co-methylation in the somatic cell lines may be due to culturing of the cells. As a population of cells develops from a single cell, there are errors in copying the DNA methylation patterns from mother to daughter cell. These errors disrupt the co-methylation of pairs. If the iPSC and iPSC-derived cell lines have had less (developmental) time between the laying down of the original DNA methylation pattern and the point at which they were sequenced, then there is a lower probability that the co-methylation has been disrupted by replication errors. There are also reports that DNA methylation profiles differ widely between cell lines and their precursor primary tissues Nestor *et al.* [2015], which very likely also affect co-methylation.

7.3.3 Limitations

A major challenge in analysing within-fragment co-methylation is the susceptibility of estimates of θ to M-bias, particularly when the IPD is large and measurements are taken towards the ends of reads. As noted in Section 2.2.2, it is difficult to properly account for M-bias in pre-trimmed data such as the *Lister* dataset. Since we can do little about data that has already been pre-trimmed, we are forced to be cautious in our interpretation of within-fragment co-methylation for pairs with larger IPDs from these datasets. This is particularly relevant to the apparent increase in within-fragment co-methylation for $NIL \geqslant 0$ CpG pairs with $IPD \approx 180$. This result is tantalisingly similar to the periodicity we see in the correlations of β-values and the approximate spacing of nucleosomes in human DNA (see Section 1.1.2). However, it is less apparent in other paired-end samples, and this is only likely to be resolved by longer reads that can span multiple nucleosomes.

In analysing these data at the genome-level, and even the chromosome-level, we are aggregating measurements from heterogeneous regions of the genome. Our simulation results of Section 7.2.4 tell us that the Mantel-Haenszel estimator estimates the 'average' effect across these potentially heterogeneous regions. We might reduce this heterogeneity by aggregating over smaller regions, although choosing the resolution is non-trivial, and so we content ourselves with studying an estimate of the average effect.

7.4 Summary

Correlations of β-values and estimates of within-fragment co-methylation are complementary methods to better understand the complex dependence structure of DNA methylatin data. Which is more relevant depends on the question at hand. Within-fragment co-methylation seems closer to the biology because it measures the dependence of DNA methylation on the scale that the DNA methyltransferases act. However, the most common downstream analyses are based on β-values and so these correlations may be more relevant to analyses of DNA methylation data. Ultimately, the two measures are intertwined since the correlations of β-values will in part be driven by co-methylation on individual DNA fragments. This idea is explored in greater detail in Chapter 8.

The presented analyses only measure pairwise dependencies. More specifically, the $NIL = 0$ results measure the 'first-order' dependence whereas the $NIL \geqslant 0$ measure a mixture of different orders of dependences but all are done in a pairwise manner. I find the $NIL = 0$ results easier interpret but they clearly do not tell the whole story.

By comparing the $NIL = 0$ and $NIL \geqslant 0$ results, we see that the strength of the dependency itself depends on not just the distance between the two CpGs, but the number of intervening CpGs. This is most noticeable for within-fragment co-methylation. This indicates that these dependencies are more than ‘first-order’. Moreover, the differences in dependencies between CpG islands and non-islands indicates that the ‘order’ of these dependencies itself varies across the genome. All this means that the DNA methylation measurements for a single sample form a complicated and highly non-stationary process with variable order dependencies.

In theory, we could explore these higher-order dependencies of within-fragment co-methylation by analysing m-tuples with $\mathrm{m} > 2$. However, the complexity of the challenge increases exponentially as 2^{m}. Moreoever, the IPD vector now has $(m - 1)$ dimensions, which makes it difficult to visualise the strength of co-methylation as a function of IPD.

Another complexity that these analyses have brushed over is the fact that the genome is not a one-dimensional object but is rather a complex three-dimensional structure (whose shape also varies in time). Therefore, the relevant distance between two loci is not necessarily the number of base pairs between them but may instead be the Euclidean distance in three-dimensional space. Such complexities cannot be resolved with bisulfite-sequencing data alone but will require measures of the three-dimensional structure of the genome, such as those provided by chromatin conformation capture technologies [e.g., Dekker *et al.* 2013].

In summary, the results of this chapter demonstrate that DNA methylation has a complex dependence structure. Co-methylation exists along individual DNA fragments and also manifests as correlations of aggregate measures of methylation. We have estimated the effect of this co-methylation using two complementary approaches: correlations of β-values and within-fragment co-methylation. Consistent with previous work, our analysis identifies the intra-pair distance between CpGs and the genomic context of these CpGs, in particular, CpG islands, as being important drivers of co-methylation.

Unique to our results are genome-level and chromosome-level estimates of within-fragment co-methylation that reveal a possible role for nucleosome positioning in determining co-methylation. Future work integrating these results with data from assays of nucleosome occupancy may clarify the nature of this relationship. Furthermore, stratifying these analyses by other genomic features, such as genic versus intergenic regions, way help elucidate what other factors are at play.

While it is the highest resolution assay for studying DNA methylation, bisulfite-sequencing can still only give a rather limited picture of co-methylation, particularly of within-fragment co-methylation. Data from assays with longer reads will be immensely useful in to advancing our understanding of DNA methylation dynamics. In particular, reads that can span multiple nucleosomes will allow for co-methylation to be better tied in with, and expand upon, existing results on the relationship between nucleosomes and DNA methylation. Until such data are available, we can try to learn more about plausible models of DNA co-methylation through computational methods such as the simulation method described in the next chapter.

Chapter 8

A simulation model of DNA methylation data

Overview

This chapter describes the `methsim` software to simulate DNA methylation data. `methsim` incorporates a model of within-fragment co-methylation and simulates individual sequencing reads, which sets it apart from existing software. We explore different parameter choices and highlight promising directions of research, as well as areas requiring further improvement in ongoing work.

8.1 Introduction

It can be difficult to design an experiment that can be used for validating or benchmarking different analysis strategies. In fact, the huge variety of experimental factors, and their possible values, can make such a task infeasible. But perhaps the bigger hurdle is that these are not attractive experiments to perform; why spend your time and money on an experiment where you 'know the answer' when you could be spending that same time and money on investigating some new biology? In these scenarios and others, simulation studies play a vital role in applied statistics, where they can be used in the development, validation and benchmarking of different analysis methodologies.

The key advantage of simulated data is that we know the truth *a priori*. Moreover, we can manipulate the truth via parameters in the simulation model and can examine how an analysis method performs under a variety of scenarios. And this manipulation is cheap, a mere matter of changing parameters and re-running a piece of software, which means that we can investigate a broad range of plausible scenarios.

Simulation studies can provide many insights into the performance of a method. At the most basic level, if a method fails or performs poorly when applied to 'easy' simulated data, then it is very unlikely to work well when applied to more complex experimental data. We can also identify which scenarios are 'easy' (ones where most methods are able to identify the truth) and scenarios under which certain methods perform better than others. While simulated data can never fully capture the richness of real data, we can learn a lot about a method by studying how it performs when applied to simulated data.

Simulation methods may also be used to learn about the plausibility of hypothesised models of a phenomenon, be they mechanistic or stochastic. By simulating data from the proposed model, and comparing it to the real data, we can identify hypotheses that are incompatible with reality. This can also help identify shortcomings in the model so that it may be refined in an iterative manner.

There are a few key criteria when designing a simulation method:

1. Realism: The simulated data must be 'similar' to the real data. While this is obvious, it is also often hard to clearly define or agree upon what constitutes 'similar enough'.
2. Cost: It should be fast and cheap to simulate data. The most common use of simulation models in applied statistics is to repeatedly generate datasets under a range of parameter settings. This requires that each simulation is fast and computationally cheap, otherwise it will be prohibitive to explore the full space of scenarios. An exception to this rule may occur when the simulation model is used to test a proposed mechanistic or stochastic model of a phenomenon, such as in studies of molecular dynamics. Even then, however, the cost of a simulation should be less than the cost of performing the equivalent experiments, otherwise the simulation is generally not worth the effort.
3. Usability: There **must** be a software implementation. Simulation models exist to be

simulated from; a simulation model without an implementation is next to useless. The implementation should give the user easy access to the key parameters and have sensible default settings. The output of the software should be in a standard format or readily convertible to a simple, manipulable format.

There is a lack of gold-standard datasets in the field of DNA methylation for the benchmarking and validation of analysis methods. Therefore, the development of realistic, cheap, and user-friendly software to simulate DNA methylation data will be of benefit.

8.2 Literature review

The simulation methods described here can be thought to lie on a spectrum with model-based methods at one end and re-sampling based methods at the other end[1]. It is often simpler to simulate from a model-based simulation, particularly if the model is a well-studied parametric distribution. This simplicity often comes at the cost, however, of increased assumptions, whose validity may be questionable. Simulations based on sampling of real data may reduce the number of assumptions required. However, care must be taken in selecting the units to be sampled so that the sampling process is efficient and so that the sampled data don't grossly distort within-sample and between-sample dependencies.

Any procedure for simulating DNA methylation data should obviously be tailored to its purpose. For example, if the study is comparing alignment software for bisulfite-sequencing data then the simulation software should produce `FASTQ` files with realistic base quality score profiles and sequencing errors. On the other hand, it may be sufficient to simulate some aggregate data, such as β-values, when the simulated data are to be used for comparing methods to identify differential methylation.

8.2.1 Methods for simulating bisulfite-sequencing reads

All of the currently available methods for simulating bisulfite-sequencing reads are designed for the comparison of alignment strategies and are model-based. These are not generally suitable for comparisons of downstream analysis methods.

[1] Of course, the parameters in any model-based simulation should be based on real data, although this may not use a formal estimation procedure such as maximum likelihood.

Sherman is software to simulate bisulfite-sequencing reads as FASTQ files, including various 'contaminants', such as SNPs, basecall errors and sequence artefacts (http://www.bioinformatics.babraham.ac.uk/projects/sherman/). The simulated reads are designed for comparing the performance of different alignment strategies. Sherman has many parameters, of which the ones relevant to our discussion of simulating realistic DNA methylation data are -CG and -CH, the bisulfite conversion rates for CG and CH methylation loci, respectively. These are set by the user with values between 0 and 100 (%). Reads are simulated by sampling from the user-specified reference genome. When a read contains a CG (resp. CH) locus, it is randomly assigned as being converted to a TG (resp. TH) with probability -CG / 100 (resp. -CH / 100).

While appropriate for comparing alignment strategies, Sherman produces data that is not suitable for use in comparing downstream analysis methods. All CG (resp. CH) loci have an average β-value of -CG (resp. -CH) regardless of the genomic context, which we know to be incompatible with real data. Furthermore, the methylation state of each methylation locus is independent, which is clearly inconsistent with the strong co-methylation observed in real data.

DNemulator (http://www.cbrc.jp/dnemulator/README.html) uses a slightly more sophisticated simulation strategy to simulate FASTQ files for use in comparing alignment strategies for bisulfite-sequencing data. DNemulator does this with three separate routines, fasta-methly-sim, fasta-polymorph and fasta-bisulf-sim:

1. fasta-methyl-sim converts cytosines in the reference genome (FASTA file) to a character indicating the methylation level of that locus: C represents 0% methylated, c represents 10% methylated, d represents 20% methylated, v represents 50% methylated and t represents 100% methylated. Each of these conversions has a different probability in the CG and CH contexts.
2. fasta-polymorph simulates a polymorphic, diploid genome based on the modified reference sequence created by fasta-methyl-sim.
3. fasta-bisulf-sim simulates reads by sampling from the simulated genome created by fasta-polymorph. Read are simulated with bisulfite-conversion error and sequencing error.

The reads simulated by DNemulator will result in β-values that have more context-dependence than those resulting from reads generated by Sherman. However, methylation events are still generated independently of one another, which means there is no co-methylation in the simulated data. Therefore, reads simulated by DNemulator, while suitable for comparing alignment strategies, are not suitable for comparing downstream analysis methods.

Other software for simulating individual bisulfite-sequencing reads are FastqToBS (http://users.dimi.uniud.it/~nicola.prezza/projects.html), which uses a similar strategy as Sherman, and BSsim (http://122.228.158.106/BSSim/, and used in Xie *et al.* [2014]), which has a similar strategy to DNemulator.

8.2.2 Methods for simulating aggregate methylation levels

The most widely studied downstream analysis problem is that of identifying differential methylation, which is done by comparing summary measures of methylation, such as β-values, between two or more groups. It is therefore generally sufficient to directly simulate these aggregated measures, rather than simulating reads, for these type of simulation studes.

Most papers that propose a new method for downstream analysis of bisulfite-sequencing data include a simulation study. Generally, such a simulation method exists to support claims about the performance of the proposed method and is not a major feature of the paper. Consequently, the simulation model is often only briefly described and a software implementation is rarely made available. In fact, until recently, of the methods reviewed in this section, only that of Lacey *et al.* [2013] had a software implementation available. As I was writing this chapter, the WGBSSuite software was published [Rackham *et al.* 2015]. WGBSSuite is the only software specifically published for the purpose of simulating whole-genome bisulfite-sequencing data for comparing methods for identifying differential methylation. WGBSSuite is available for download as a collection of R scripts[2].

Fortunately, these simulation methods follow a common framework, even if the details

[2] Unfortunately, WGBSSuite is not available as an R package, which is the "fundamental unit of reproducible R code" (http://r-pkgs.had.co.nz/) which would greatly simplify the installation and use of the software. Furthermore, no license file is included in the download, meaning that it is unclear how the user is permitted to use WGBSSuite and whether they may modify or redistribute the code.

differ:

0. [Optional] Simulate the locations of methylation loci.
1. Simulate the unobserved group-specific true methylation levels.
2. Simulate the observed sample-specific sequencing depths.
3. Simulate the observed sample-specific methylation levels, e.g., the β-values.

A popular choice of parametric model in this framework is the beta-binomial model [e.g., Feng *et al.* 2014, Lacey *et al.* 2013, Xu *et al.* 2013, Chen *et al.* 2014b, Dolzhenko and Smith 2014]. This model, and others, are now discussed.

Simulating methylation loci

Lacey *et al.* [2013] and Rackham *et al.* [2015] are notable in that they choose to simulate the locations of CpGs rather than simply using their locations in a reference genome. Both papers use hidden Markov models to simulate genomes with regions of high and low CpG density.

I do not think this a useful or necessary step, and it may even be counterproductive. While in truth the set of methylation loci do vary between samples due to genetic variation, it is a reasonable approximation to consider the positions of these loci as fixed. If sequence variation is required then it is easily accommodated by sampling from the set of methylation loci in the reference genome. Furthermore, the aim of the simulation model is to realistically simulate methylation *levels*, not the *locations* of these loci.

Almost all downstream analyses are reference-based (see Section 2.3), so it is desirable to know how these methods perform with respect to the relevant reference genome, not a simulated genome in which the location of methylation loci vary from simulation to simulation.

Simulating $B_{i,j}$

Recall that $B_{i,j}$ is the underlying 'true' methylation level at locus i in sample j. In the context of simulating data for comparing differential methylation calling methods, we want these to be group-specific. That is, we want to specify B_{i,j_k}, where j_k indicates that sample

j comes from group k. For non-differentially methylated loci all k groups have identical $B_{i,j_k} = B_{i,j_0}$; differentially methylated loci are simulated by setting $B_{i,j_k} \neq B_{i,j_{k'}}$ for some $k \neq k'$.

Under the beta-binomial model, the true methylation level, $B_{i,j}$, is assumed to follow a $Beta(\mu_{i,j_k}, \phi_{i,j_k})$ distribution, where μ_{i,j_k} is the mean and ϕ_{i,j_k} is the dispersion of the beta distribution[3]. Both the means, μ_{i,j_k}, and the dispersions, ϕ_{i,j_k}, are group-specific and are allowed to vary across methylation loci (i.e. across i). The dispersion parameter models the within-group variability of the $B_{i,j}$, i.e. the within-group *biological variability* of DNA methylation.

As noted by Feng *et al.* [2014], the beta distribution is a very flexible distribution with support on $[0, 1]$ and has "long been a natural choice to model binomial proportions", particularly as a conjugate prior, as it is used in the empirical Bayes model of Feng *et al.* [2014]. The beta-binomial model can also be viewed from a non-Bayesian perspective as a compound distribution or as an overdispersed binomial distribution.

Other distributions may be used instead of the beta distribution for modelling $B_{i,j}$. For example, Xie *et al.* [2014] consider both a single Gaussian distribution and a mixture of Gaussian distributions, while Xu *et al.* [2013] consider both a truncated Gaussian and a mixture of truncated Gaussian distributions.

An alternative to specifying a parametric distribution for the $B_{i,j}$ is to sample these from real data, e.g., by sampling some observed $\beta_{i,j}$ for a particular dataset and treating them as if they were observations on $B_{i,j}$. `WGBSSuite` [Rackham *et al.* 2015] uses a modified form of this approach. In `WGBSSuite`, a hidden Markov model is used to classify every CpG as having an underlying state ("de-methylated", "1st transition", "2nd transition" or "methylated"). Each of these four states has a region-specific average methylation level that is based on the distribution of β-values for a chosen dataset. For example, the average methylation level for all "de-methylated" regions is defined as $B_{de-methylated} = median(\{\beta_{i,j} : \beta_{i,j} \in [0, 0.5)\}])$. Then, each $B_{i,j}$ is a perturbed version of this region-level average methylation, $B_{region-type}$, obtained by adding on a zero-mean Gaussian random variable, i.e. $B_{i,j} = B_{region-type} + \epsilon_{i,j}$,

[3] Note that this is different to the standard parameterisation of the Beta distribution, which is described by two shape parameters, α and β. The relationship between the two parameterisations is $\mu = \frac{\alpha}{\alpha+\beta}$ and $\phi = \frac{1}{\alpha+\beta+1}$ [Feng *et al.* 2014].

where $\epsilon_{i,j} \overset{d}{=} Normal(0, s^2)$. Care needs to be taken that $0 \leqslant B_{i,j} \leqslant 1$.

Simulating sequence depth

The sequencing depth at each methylation loci, $d_{i,j}$, may be sampled from real data [Feng *et al.* 2014, Chen *et al.* 2014b, Dolzhenko and Smith 2014], or simulated from a parametric distribution such as the Poisson [Rackham *et al.* 2015], a rounded Gaussian distribution [Xu *et al.* 2013], or a rounded mixture of Gamma distributions [Lacey *et al.* 2013].

The most sophisticated approach to simulation of sequencing depth in bisulfite-sequencing experiments is given by Lacey *et al.* [2013]. In addition to using a mixture of distributions to capture both the low-coverage and high-coverage modes observed in RRBS sequencing coverage, Lacey *et al.* [2013] model the correlation of sequencing depth across samples for a given region. They do this by using a Gaussian copula to make the set of sequencing depths a jointly dependent set of random variables. While this is undoubtably sophisticated, the effect of correlated versus uncorrelated sequencing depths in a simulation model is not explored in the paper and so the cost-benefit trade-off is unclear. Moreover, it is simpler, and likely more computationally efficient, to include such correlations by a sensible sampling of sequencing depths from real data.

Simulating the observed methylation levels

The final step is to simulate the read counts, $M_{i,j}$ and $U_{i,j}$. These are based on the true underlying methylation level, $B_{i,j}$, and the sequencing depth, $d_{i,j}$. In the beta-binomial model, this is done by binomial sampling where $M_{i,j} \overset{d}{=} Binomial(d_{i,j}, B_{i,j})$. In addition to binomial read sampling, the `WGBSSuite` software also implements (truncated) negative binomial read sampling [Rackham *et al.* 2015], which essentially introduces overdispersion in the read counts.

Simulating differentially methylated regions

As we have seen, the simulation of a differentially methylated locus is straightforward; for the i^{th} locus, simply vary B_{i,j_k} across the k groups. The simulation of a differentially methylated region is more complex.

In principal we can simulate a differentially methylated region by simply simulating runs of differentially methylated loci. However, it must be noted that this requires careful choice of parameters. Such parameters include the length of the DMR, the minimal number of loci it must contain, the maximal intra-pair distances of loci within the DMR and how many of the loci in the DMR must themselves be differentially methylated, e.g., should Figure 8.1 be considered one DMR or two DMRs? Such decisions ultimately have to be made by the user of the simulation software based on the types of events she is interested in analysing.

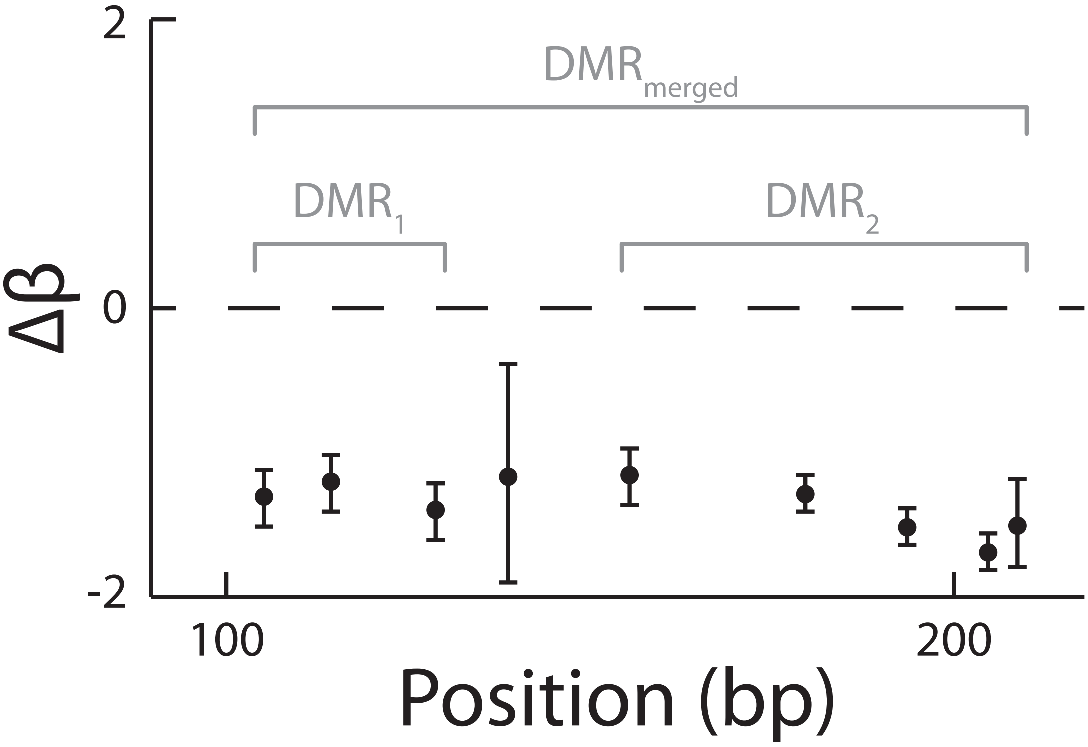


Figure 8.1: A hypothetical region in a study of differential methylation in a two-group experiment. Plotted is the difference in β-value between the two groups ($\Delta\beta$) with associated standard error against the position along the genome (*Position (bp)*). The first three methylation loci are DMCs, as are the last five methylation loci. However, the fourth locus is not a DMC. Should this region be considered as two distinct DMRs or as a single DMR?

8.2.3 Simulating co-methylation

A key feature ignored by the majority of simulation methods, with the notable exceptions of Lacey *et al.* [2013] and Rackham *et al.* [2015], is co-methylation. The methylation states of neighbouring loci are highly dependent and, consequently, the $B_{i,j}$ of nearby loci are highly dependent. To be clear, the majority of studies using simulated DNA methylation data do not model an important feature of DNA methylation data, co-methylation.

Most simulation methods do not simulate individual reads and so cannot simulate within-fragment co-methylation. Instead, they capture correlations of methylation levels by inducing dependence in the $B_{i,j}$. For example, under the beta-binomial model these correlations could be induced by forcing the $\mu_{i,j}$ to be spatially dependent. This idea takes its inspiration from Jaffe *et al.* [2012a] who simulate spatially correlated DNA methylation microarray data by imposing an autocorrelation structure via a lag-1 autoregressive model of the simulated β-values.

Lacey *et al.* [2013] take a different approach, but one that still results in correlated B_{i,j_k} across i, within group k. To begin, they compute β-values from chromosome 11 for a single normal myotube cell line that was sequenced with RRBS. They then fit a Gaussian variogram to these β-values, which shows "a strong correlation for sites in close proximity, decaying to near independence at distances beyong 3000 bp". To simulate spatially correlated $B_{i,j}$ they use an iterative process:

1. Simulate B_{i,j_k} from a Beta distribution with parameters estimated from the chromosome 11 MTCTL2 data. These are estimated under an assumption of independence.
2. Induce correlation amongst the B_{i,j_k} (across i) by a transformation of the B_{i,j_k}.

The second step uses a method published by Zaykin *et al.* [2002]. The transformed values, B^*_{i,j_k}, are created by the transformation $\mathbf{B^*_{j_k}} = 1 - \Phi\Big\{C\Phi^{-1}(1 - B_{j_k})\Big\}$, where $\mathbf{B_{j_k}}$ is the vector of B_{i,j_k}, C is a factor of the correlation matrix $\Sigma = CC^{'}$, where Σ is estimated from the fitted Gaussian variogram, and where $\Phi(\cdot)$ denotes the standard normal distribution function.

`WGBSSuite` induces spatial correlation amongst the B_{i,j_k} in a less direct manner. Recall that each methylation locus is assigned one of the four underlying states ("de-methylated", "1st transition", "2nd transition", "methylated") via a hidden Markov model. The transition matrix of this hidden Markov model ensures that neighbouring loci, $i, (i+1)$, are more likely to be assigned the same state. Furthermore, since all loci within each of the four states are assigned the same (perturbed) underlying methylation level, $B_{i,j_k} = B_{region-type} + \epsilon_{i,j_k}$, neighbouring loci have similar methylation levels. Note that the IPD only plays a direct role in the initial segmentation of the genome, not in the assigning of the $B_{i,j}$.

Other model-based simulations

A separate class of model-based simulation methods are those that simulate the β-values directly, i.e. without simulating sequencing depth [Jaffe *et al.* 2012a, Chen *et al.* 2014a,b]. These models are designed for simulating microarray data and not sequencing data, since they do not include the variability due to variation in sequencing depth. Of these methods, only Jaffe *et al.* [2012a] simulate correlation amongst the B_{i,j_k}.

8.2.4 Methods based on re-sampling real data

A simulation may also be based entirely on re-sampling of real data[4]. This type of simulation is attractive because, through careful sampling, it can capture behaviour that is otherwise very difficult, if not impossible, to capture in a parametric model. At the same time, however, if the sampling units are poorly chosen or the sampling strategy is incorrect, then it may ignore these same features or, worse still, introduce artefacts into the simulated data.

Re-sampling methods are most easily implemented at the level of β-values. Re-sampling reads is more difficult, except in the special case of down-sampling whereby the positions of reads are held constant but only a sub-sample of them are used in downstream analysis. Down-sampling is only really of interest to examine the effects of sequencing coverage on downstream analyses.

Sofer *et al.* [2013] use a re-sampling based simulation method in the development of their `Aclust` software (designed for identifying differential methylation from microarray data). The idea is adapted from Gaile *et al.* [2007], which is a simulation method for array

[4] This may also be referred to as creating a synthetic dataset.

comparative genomic hybridization experiments. Sofer *et al.* [2013] sample "blocks" of CpGs, a region of the genome where all CpGs are within 10 kb of the next, to "generate (spatial) correlation-preserved methylation data". By sampling blocks rather than individual CpGs, this sampling scheme preserves the correlation structure between CpGs occurring in the same block. Two blocks are unlikely to be correlated since there is little evidence that CpGs separated by more than 10 kb have spatially correlated methylation levels.

Sofer *et al.* [2013] sample from a dataset of 539 Illumina 450k microarrays and select a small number of "target" CpGs, which are CpGs whose methylation level is highly variable across the 539 samples. If a block does not contain a target then it is sampled uniformly at random from the 539 samples. If a block does contain a target, however, then the sampling is weighted so that the "cases" are preferentially sampled from blocks with a high level of methylation at the target CpG and the "controls" are preferentially sampled from blocks with a low level of methylation at the target CpG[5]. This is essentially weighted re-sampling of the real data to induce differential methylation.

Due to the correlation structure of the β-values (the co-methylation) it is likely, although not guaranteed, that other CpGs in the blocks containing targets also display differential methylation.

8.2.5 Summary

My initial interest in simulation methods was to explore models of co-methylation, particularly at the level of individual DNA fragments. This requires two key capabilities:

1. Simulation of individual reads.
2. Simulation of co-methylation.

None of the published simulation methods satisfied both these requirements. This motivated the development of `methsim`.

`methsim` is specifically designed to model the co-methylation structure of bisulfite-sequencing data by simulating individual DNA fragments rather than directly simulating summary methylation measurements. In order to model the co-methylation structure,

[5] The use of "cases" and "controls" is arbitrary, as is the choice of highly methylated for "cases" and lowly methylated for "controls" at target CpGs.

I decided that methsim should simulate at the level of individual DNA fragments. An added bonus of this approach is that methsim can (in theory) generate data at multiple resolutions: β-values, methylation patterns at m-tuples, or entire sequencing reads. While my initial motivation in developing methsim was to explore models of co-methylation, I realised that it could also be used in the development and comparison of methods for the downstream analysis of DNA methylation data.

8.3 Methods

Simulating data with methsim involves 3 steps:

1. Simulating the true methylome of each sample.
2. Simulating reads, including sequencing error and bisulfite-conversion error, by sampling from the true methylome.
3. Constructing the output, be it reads, methylation patterns at m-tuples, or β-values.

methsim requires an input dataset from which to estimate key parameters. For the input dataset, methsim requires the methylation patterns at various sized m-tuples, which can be produced by the methtuple software. methsim can currently only simulate CpG methylation and assumes that the methylation states of all CpGs are strand symmetric.

In what follows we will simulate data based on the *ADS* methylome from the *Lister* dataset. We focus on simulating autosomal data, since the sex chromosomes and mitochondrial DNA have very different methylation dynamics.

8.3.1 Implementation

methsim is written in R and builds on the MethylationTuples package described in Chapter 5, as well as several R packages available on Bioconductor and CRAN (see the DESCRIPTION file for a complete list.). It is currently a very experimental package and is therefore not yet published on Biocondutor, but its development can be followed at https://github.com/PeteHaitch/methsim.

methsim makes extensive use of the S4 object system in R. The most important classes defined in `methsim` are `MethylomeParam`, `SimulatedMethylome`, `WGBSParam` and `SimulatedBS`. The most important methods are the `simulate()` methods defined for the `MethylomeParam` and `WGBSParam` classes.

The `MethylomeParam` object contains the empirical distributions of key statistics from which the parameters for simulating the 'true' methylome are sampled. Therefore, the `MethylomeParam` object should be based on data from a relevant sample. To help a new user get started, `methsim` includes `MethylomeParam` objects for the *ADS*, *ADS-adipose* and *ADS-iPSC* samples from the *Lister* dataset. Alternatively, the user may process the `BAM` file for their sample with `methtuple` and then use the helper functions in `methsim` to construct a `MethylomeParam` object based on their sample of interest. The user runs the `simulate()` method on the `MethylomeParam` object to simulate a true methylome; this returns a `SimulatedMethylome` object.

Once we have a true methylome, we can simulate data from it. `methsim` currently supports the simulation of whole-genome bisulfite-sequencing reads via the `WGBSParam` class and associated `simulate()` method. Other assays, such as RRBS or microarrays could in principle be supported. A `WGBSParam` object will contain a `SimulatedMethylome` object, along with parameters such as the read-length, sequencing coverage and error rate of the data to be simulated. When applied to a `WGBSParam` object, the `simulate()` method returns a `SimulatedBS` object or a `MethylationTuples::MethPat` object[6]. A `SimulatedBS` object contains all simulated reads and is generally a large object (on the order of 10 GB). By contrast, the `MethylationTuples::MethPat` object, which summarises the simulated reads for m-tuples of a particular size[7], is much smaller (on the order of 500 MB) but does not contain the full information of the simulation since read-level data are lost. `methsim` also includes a helper function, `asMethPat()`, to coerce a `SimulatedBS` object to a `MethylationTuples::MethPat` object.

[6] This uses the NAMESPACE notation of R: `MethylationTuples::MethPat` can be read as "the `MethPat` class is part of the `MethylationTuples` package". See Section 5.3 for details of this class.

[7] Returning a `MethylationTuples::MethPat` object may be appropriate, for example, if all that is required for downstream analyses are methylation counts at 1-tuples.

8.3.2 Simulating a single methylome

As we have seen, DNA methylation is highly heterogeneous along the genome. Nonetheless, there are clear regions of 'similarity', such as the unmethylated CpG islands and long partially methylated domains. In these locally-similar regions we might hope to model DNA methylation by a simple parametric model. `methsim`, like other simulation methods, is based on the idea of segmenting the genome into 'regions of similarity', fitting a simple model to each region and then 'stitching' the results together to form the true methylome.

The idea of segmenting a globally heterogenous stochastic process into a series of locally homogeneous processes is not new. A hidden Markov model is an example of such a process; while the entire process may highly heterogeneous, conditional on the hidden states the process may be homogeneous. Hidden Markov models, and other models assuming local similarity in spite of global heterogeneity, have been used with great success in bioinformatics.

`methsim` takes the following approach to simulating the true underlying methylome for each sample:

1. Segment the methylome into regions of similarity.
2. Sample parameters for the i^{th} methylation locus based on the segmentation and the empirical distributions of key statistics.

I now describe each step in greater detail.

Segmenting the methylome

`methsim` uses the the R/Bioconductor package, `MethylSeekR` [Burger *et al.* 2013], to segment the input methylome into regions of similarity. `MethylSeekR` was developed to discover regulatory motifs from bisulfite-sequencing data by segmenting the methylome into unmethylated regions (*UMR*s), lowly-methylated regions (*LMR*s) and partially methylated regions[8] (*PMR*s). It does this using a two-stage algorithm applied to the β-values from a sample:

[8] Partially methylated regions are also commonly known as partially methylated domains, but we will refer to them as 'regions' for consistency with UMRs and LMRs.

1. Identify partially methylated regions. A summary statistic, α, which is based on the β-values in a sliding window of 100 CpGs, is used to identify PMRs. Briefly, a two-state hidden Markov model is fit to the α values to identify PMRs and non-PMRs.
2. Identify UMRs and LMRs. The PMRs are masked from the genome and simple heuristics are used to identify UMRs and LMRs based on the average β-values in a window and the number of CpGs in the window.

`methsim` post-processes the segmentation provided by `MethylSeekR` to partition the methylome into regions of similarity, namely *UMR*, *LMR*, *PMRS* and *other*[9]. Roughly speaking, *other* regions are 'mostly methylated regions' (see Figures 8.2, 8.3, 8.4, 8.5), although Burger *et al.* [2013] do not describe these regions as such.

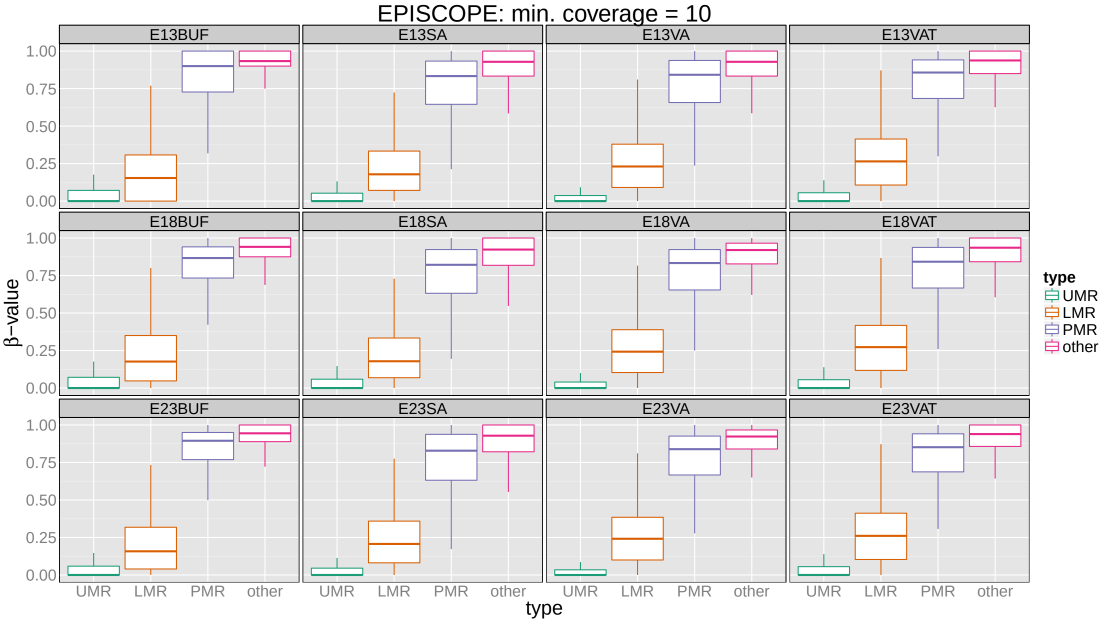


Figure 8.2: Boxplots of CpG β-values in each region-type following segmentation by `MethylSeekR` for the *EPISCOPE* dataset.

While a tailored algorithm may improve this segmentation process, the result produced by `MethylSeekR` is a reasonable approximation to segmenting the methylome into the required 'regions of similarity'.

[9] The output returned by `MethylSeekR` does not strictly partition the methylome since it is neither disjoint nor exhaustive, hence the need for post-processing.

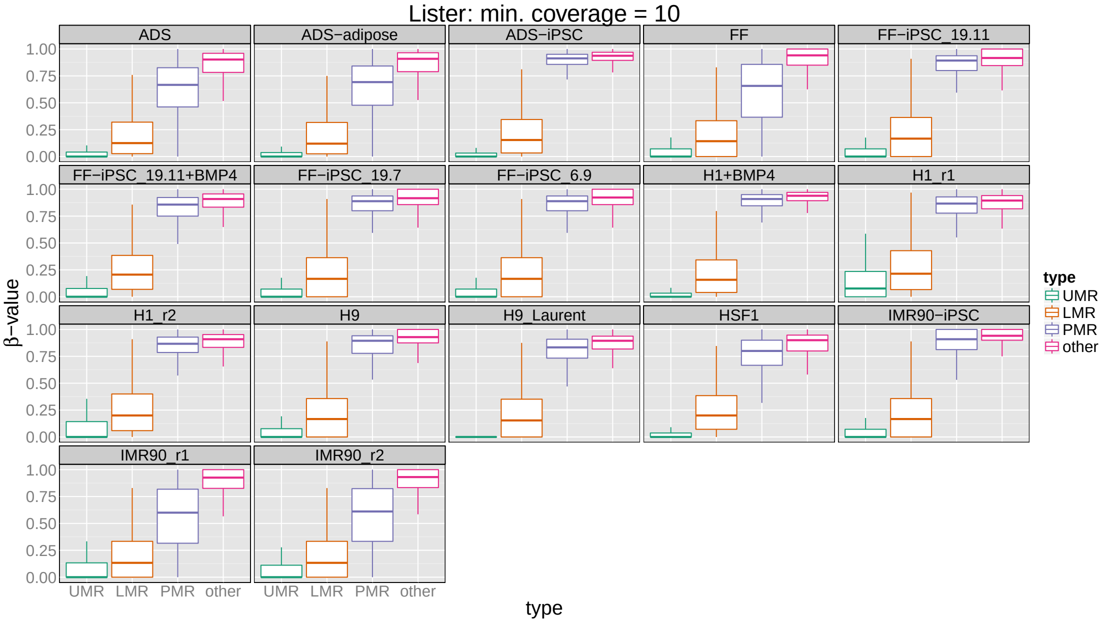


Figure 8.3: Boxplots of CpG β-values in each region-type following segmentation by `MethylSeekR` for the *Lister* dataset.

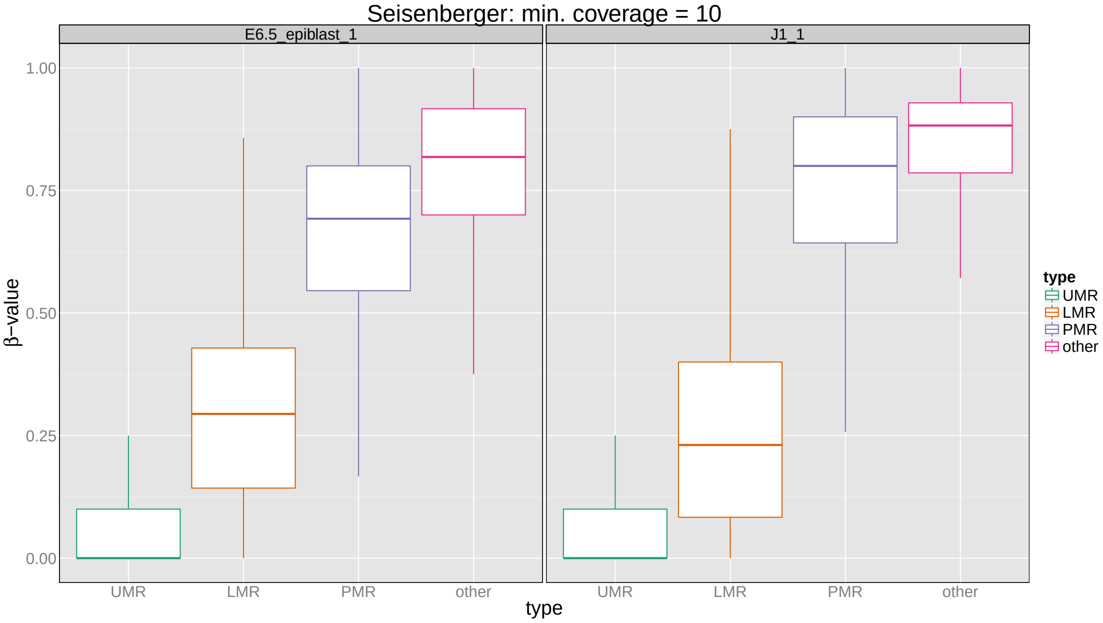


Figure 8.4: Boxplots of CpG β-values in each region-type following segmentation by `MethylSeekR` for the *Seisenberger* dataset. `MethylSeekR` was unable to partition the *E16.5_male_1* methylome due to its unusual β-value distribution, hence no data is shown for this sample.

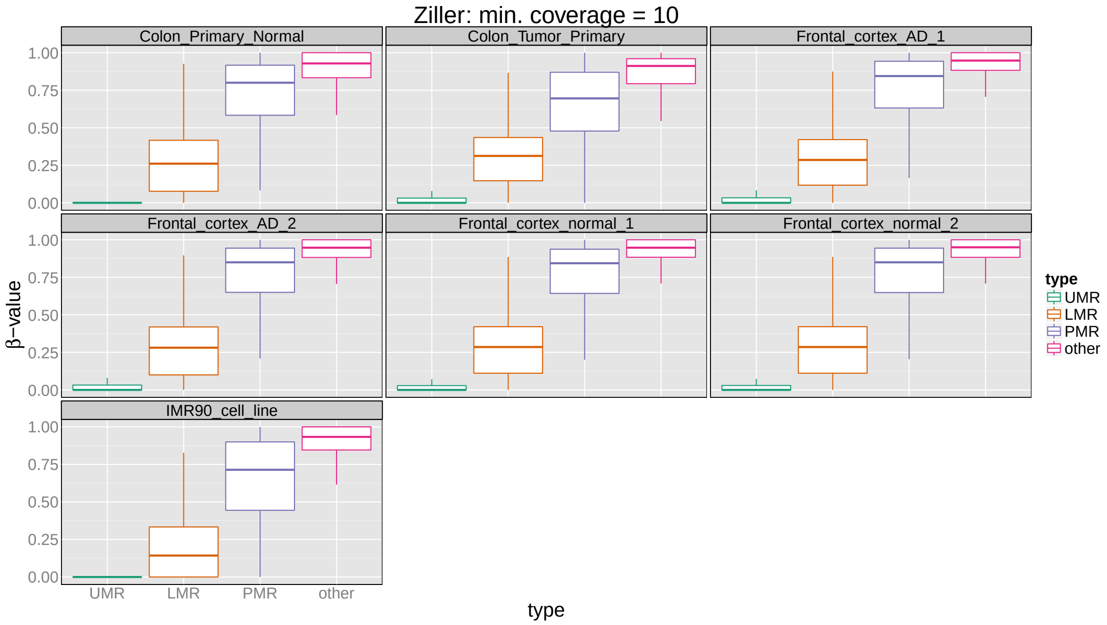


Figure 8.5: Boxplots of CpG β-values in each region-type following segmentation by `MethylSeekR` for the *Ziller* dataset.

Parameterising the model

`methsim` uses a two-state (1 = methylated, 0 = unmethylated), first-order Markov chain to model $\mathbf{Z} = \{\mathbf{Z_h}\}_{h=1}^{h=n_h}$. This allows the incorporation of within-fragment co-methylation into the simulation. The choice of a first-order Markov process is one of computational simplicity and is a reasonable approximation for most of the genome given the available data (see Chapter 7).

In Chapter 7 we saw that the strength of within-fragment co-methylation varies as a function of the intra-pair distance and by the genomic context. For this reason I allow the transition probabilities, $\mathbf{p}$, to vary with i, that is, I allow the Markov chain to be spatially inhomogeneous. In particular, I allow the transition probabilities to depend on the intra-pair distance between the i^{th} and $(i+1)^{th}$ locus and on the region type, r_i (*UMR*, *LMR*, *PMR* or *other*), for each pair of loci[10].

The above-described model is not particularly amenable to analytical calculations due to the spatial inhomogeneity. It is, however, relatively simple to simulate realisations from

[10] Actually, it only depends on the region of the i^{th} locus. Most pairs of loci, $(i, i+1)$, will lie in the same region. For pairs that span the boundary of two different regions I have arbitrarily chosen to use the region of the i^{th} locus.

this model, requiring only a single loop over the set of loci on each chromosome. For a chromosome containing n methylation loci, there are $n-1$ transition matrices to estimate or otherwise assign.

Rather than directly modelling the transition probabilities, $\mathbf{p}$, `methsim` is parameterised by a vector of marginal probabilities, $\mathbf{B} = \{Pr(Z_i = 1)\}$, and a vector of odds ratios, $\boldsymbol{\psi} = \frac{Pr(Z_{i+1}=1|Z_i=1) \times Pr(Z_{i+1}=0|Z_i=0)}{Pr(Z_{i+1}=1|Z_i=0) \times Pr(Z_{i+1}=0|Z_i=1)} = \frac{Pr(Z_i=1, Z_{i+1}=1) \times Pr(Z_i=0, Z_{i+1}=0)}{Pr(Z_i=1, Z_{i+1}=0) \times Pr(Z_i=0, Z_{i+1}=1)}$. We can compute the transition probabilities, $\mathbf{p}$, from $\mathbf{B}$ and $\boldsymbol{\psi}$.

For each 2-tuple, $(i, i+1)$, `methsim` first constructs the joint probability matrix, $P_{i,i+1}$ from the marginal probabilities, B_i and B_{i+1}, and the odds ratio, ψ_i. The general form of $P_{i,i+1}$ is shown below:

$$P_{i,i+1} = \begin{array}{c c} & \begin{array}{cc} 1 - B_{i+1} & B_{i+1} \end{array} \\ \begin{array}{c} 1 - B_i \\ B_i \end{array} & \begin{pmatrix} Pr(Z_i = 0, Z_{i+1} = 0) & Pr(Z_i = 0, Z_{i+1} = 1) \\ Pr(Z_i = 1, Z_{i+1} = 0) & Pr(Z_i = 1, Z_{i+1} = 1) \end{pmatrix} \end{array}$$

`methsim` computes $P_{i,i+1}$ using the iterative proportional fitting algorithm. Iterative proportional fitting is a general method "for constructing tables of numbers satisfying certain constraints" [Speed 2005]. In the case of `methsim`, we use iterative proportional fitting to construct the *unique* 2×2 array (the joint probability matrix) with specified margins (the marginal probabilities of each methylation state at each locus) and the specified cross-ratio (the odds-ratio). An example of this procedure is illustrative.

Let $B_i = 0.7$, $B_{i+1} = 0.6$ and $\psi_i = 2$. To begin the iterative proportional fitting algorithm, form the matrix $P_{i,i+1}^{(0)} = \left(\begin{smallmatrix} \psi_i & 1 \\ 1 & 1 \end{smallmatrix}\right)$. This matrix has the desired cross-ratio, ψ_i, but not the desired row and column margins. At each iteration of the algorithm, iterative proportional fitting adjusts the rows and columns such that the cross-ratio remains ψ_i while the row and column margins converge towards the desired values. The algorithm continues in this manner, forming a series of 2×2 tables that converge pointwise (and uniquely) to a 2×2 table $P_{i,i+1}^{(*)} \equiv P_{i,i+1}$. For our example, $P_{i,i+1}^{(*)} = \left(\begin{smallmatrix} 0.155 & 0.145 \\ 0.245 & 0.455 \end{smallmatrix}\right)$, to three decimal places of precision. It is easily verified that $P^{(*)}$ has the desired cross-ratio and marginal sums.

Once $P_{i,i+1}$ is computed, the desired transition probability is obtained by dividing the appropriate element of $P_{i,i+1}$ by the appropriate marginal probability. To continue our example, suppose $Z_i = 0$, then the probability that Z_{i+1} is also zero is given by $Pr(Z_{i+1} = 0|Z_i = 0) = \frac{Pr(Z_{i=0}, Z_{i+1}=0)}{Pr(Z_i=0)} = \frac{0.155}{0.3} = 0.517$ (to three decimal places). In fact, `methsim` only stores $Pr(Z_{i+1} = 1|Z_i = 0)$ and $Pr(Z_{i+1} = 1|Z_i = 1)$ since $Pr(Z_{i+1} = 0|Z_i = 0) = 1 - Pr(Z_{i+1} = 1|Z_i = 0)$ and $Pr(Z_{i+1} = 0|Z_i = 1) = 1 - Pr(Z_{i+1} = 1|Z_i = 1)$.

The above-described model implicitly assumes that all n_h haplotypes have the same marginal probabilities and co-methylation structure. We may extend this model to define $\mathbf{Z}$ as a *mixture* of a small number of first-order Markov chains defined as in the above. Such a mixture model may be appropriate when simulating a sample that is a combination of cell types.

To simulate values of $\mathbf{B}$ and $\boldsymbol{\psi}$, `methsim` uses the empirical distributions of their estimates, $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\theta}} = \log(\widehat{\boldsymbol{\psi}})$, in the input methylome. More specifically, $\mathbf{B}$ is based on $\boldsymbol{\beta_{i|r_i}}$ (the empirical distribution of β_i conditional on the region type of the i^{th} locus, r_i) and $\boldsymbol{\theta}$ is based on $\widehat{\theta}_{MH}$. We have already seen many examples of the distribution of $\widehat{\theta}_{MH}$ in Chapter 7. The empirical distributions $\boldsymbol{\beta_{i|r_i}}$ for the *ADS* sample are shown in Figure 8.6.

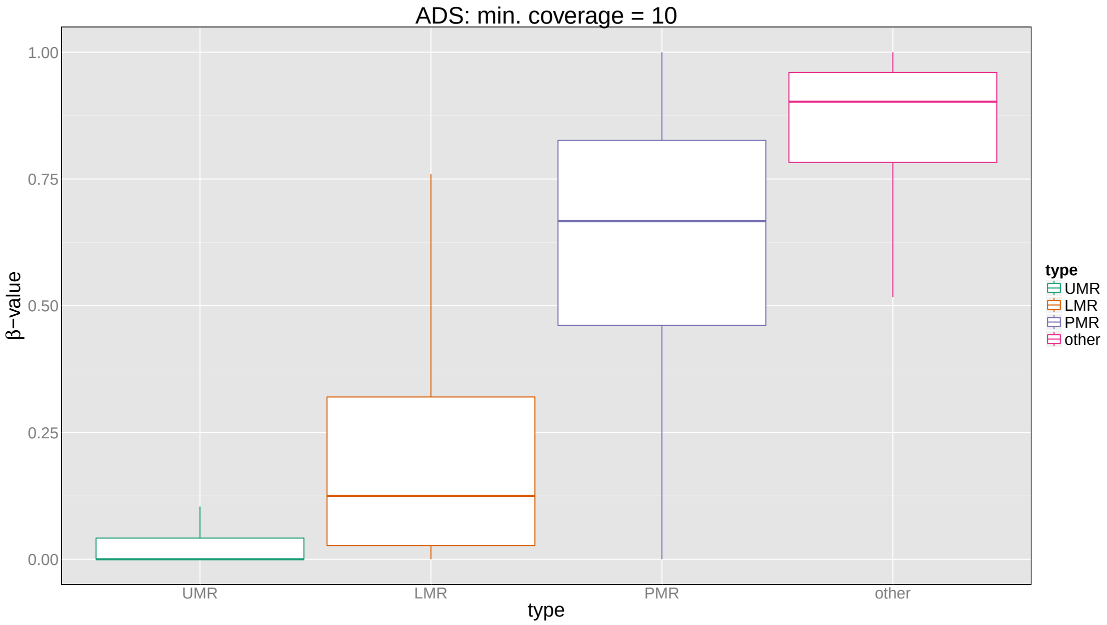


Figure 8.6: $\boldsymbol{\beta_{i|r_i}}$, the distribution of β-values for CpGs in each region type, for the *ADS* sample. Only CpGs with at least 10× sequencing coverage are used. UMR = unmethylated region; LMR = lowly methylated region; PMR = partially methylated region; other = any other region.

There are many ways to simulate values of $\mathbf{B}$ and $\boldsymbol{\theta} = \log \boldsymbol{\psi}$ based on these empirical distributions. I have so far only had time to explore a few simple methods and this remains ongoing work. We will discuss the results from three methods for simulating $\mathbf{B}$ (*m1*, *m2*, and *m3*) and two ways for simulating $\boldsymbol{\theta}$ (A and B):

m1 All loci in each region have the same B, which is sampled from $\boldsymbol{\beta_{i|r_i}}$.

m2 All loci in each region have the same **average** B, which is sampled from $\boldsymbol{\beta_{i|r_i}}$; this average is then perturbed for each locus by a $Gaussian(0, \sigma_B^2)$ random variable. This is similar to what is done by `WGBSsuite`. Here I have used $\sigma_B^2 = B \times (1 - B)$. The resulting $\mathbf{B}$ is truncated so that all values lie between 0.01 and 0.99 (necessary to avoid issues in the iterative proportional fitting algorithm at the boundaries).

m3 $\mathbf{B}$ simulated as in *m2* but with the perturbations in each region simulated from a first-order autoregressive process with coefficient equal to 0.5.

A $\boldsymbol{\theta}$ identically zero. This is to simulate independence of within-fragment methylation states and as a control to check that `methsim` is working as intended for simulating within-fragment co-methylation.

B θ sampled from $Gaussian(\mu_{IPD}, \sigma_{IPD}^2)$, where μ_{IPD} and σ_{IPD} are plug-in estimates computed from the distribution of chromosome-level $\widehat{\theta}_{MH}$ (autosomes only).

Furthermore, we will consider simulations where:

I The sample is 'pure' and all haplotypes have the same parameters.

II The sample is a mixture of subpopulations with different parameters. We will consider a sample with four subpopulations with relative frequencies equal to $(0.6, 0.25, 0.1, 0.05)$.

We do not explore all combinations of these factors in the results presented here but look at five informative combinations: *m1AI*, *m1BI*, *m2BI*, *m3BI* and *m3BII*. We compare these five models to the real *ADS* data. The limitations of these models may already be clear, and we will discuss these further in Section 8.4.

The simulation of $\boldsymbol{\theta} = \log \boldsymbol{\psi}$ requires further explanation. The empirical distributions of $\widehat{\boldsymbol{\theta}}$ will only include values for small IPDs since these are estimated from individual sequencing reads (see Chapter 7 for details). When simulating a methylome, we will encounter 2-tuples with much larger IPDs and the question is how to deal with these.

Again, there are several options. For example, we might assume a parametric form for the distribution of $\boldsymbol{\theta}$ as a function of IPD, such as a exponential decay towards $\theta = 0$ (corresponding to independence). Given the quality and read-lengths of the ADS data, I have elected to use the empirical $\widehat{\theta}_{MH}$ for pairs with $IPD \leqslant 180$ and set $\mu_{IPD} = 0$ for those with $IPD > 180$. σ_{IPD} is given by the median absolute deviation of σ_{IPD} for pairs with $IPD \leqslant 180$. While obviously a gross simplification, the vast majority of CpG pairs with $NIL = 0$ have an $IPD \leqslant 180$ and so will be unaffected by this 'independence' assumption.

8.3.3 Simulating reads

Once we have our 'true' methylome, we want to simulate an assay of this sample. In theory this could be any type of methylation assay, but here we focus on simulating whole-genome bisulfite-sequencing data.

`methsim` uses a simple Poisson-based method for simulating bisulfite-sequencing. The user specifies the desired read length[11], the average sequencing coverage, the error rate (ϵ, which includes both sequencing error and bisulfite-conversion error), and provides the simulated 'true' methylome. We will use 200 bp, single-end reads (equivalently, 100 bp paired-end reads that are always end-to-end) with an average coverage of 23× (mimicking the actual ADS data) in what follows.

The number of reads required is computed by $n_{reads} = \frac{\text{average sequencing coverage}}{\text{read length}} \times$ size of genome. The number of reads per chromosome is assigned proportional to the chromosome length. Then, the start of each read is sampled uniformly across the respective chromosome. We only retain those reads that overlap a methylation locus.

Suppose we have a simulated read, z, that overlaps the 3-tuple $(i, i+1, i+2)$. The methylation state along the read is simulated as follows:

[11] Currently only single-end data are supported. This isn't a big issue. Most paired-end bisulfite-sequencing datasets have overlapping mates, and so paired-end data can be approximated by simply doubling the read length. Also, all reads must currently have the same length, i.e. no simulation of read trimming. I don't consider this feature a priority, but it could easily be implemented by generating read lengths from a given probability distribution.

1. Sample z_i from a Bernoulli(B_i) distribution[12].
2. Sample z_{i+1} from a Bernoulli(p_i) distribution, where p_i is the appropriate transition probability given the simulated z_i.
3. Sample z_{i+2} from a Bernoulli(p_{i+1}) distribution, where p_{i+1} is the appropriate transition probability given the simulated z_{i+1}.
4. Each of element of the simulated z is independently flipped with probability ϵ to simulate sequencing error.

The result of this process is a `SimulatedBS` object storing the ID of each read that overlaps a methylation locus, along with the corresponding genomic co-ordinates and the methylation states of the read.

8.3.4 Constructing the output

The `SimulatedBS` object contains all the data from the simulation. However, it is not always the most convenient or efficient format with which to work. For example, many downstream analyses only make use of β-values, so we might want to summarise our simulated data this way. `methsim` provides the `asMethPat()` function to convert the `SimulatedBS` object to a `MethylationTuples::MethPat` object containing methylation patterns at m-tuples of a given size[13]; all the functionality of the `MethylationTuples` package is then available to the user, such as computing β-values with the `methLevel()` method or the correlations of β-values using the `methLevelCor()` function.

8.3.5 Simulating multiple samples

So far we have described how to simulate a single whole-genome bisulfite-sequencing sample. If we want to simulate six independent samples then we could simply run this procedure six times, with a different `MethylomeParam` object for each realisation. However, we will generally be interested in simulating experiments where there is some relationship between the samples. For example, suppose we want to simulate a two-group experiment with three

[12] If $\mathbf{Z}$ is a mixture of Markov chains then we first simulate which component the read comes from by sampling from a multinomial(w) distribution, where w is the vector of weights of each component. All subsequent steps will be simulated according to which component is sampled at this step.

[13] Alternatively, the user may use the `simplify` argument of the `simulate()` method to return an already 'simplified' `MethylationTuples::MethPat` object rather than the `SimulatedBS` object.

samples per group. Furthermore, suppose we want to include differentially methylated regions between the two groups. How do we simulate such an experiment?

To simulate these type of experiments, we need to be able to control the variation between the resulting `SimulatedBS` objects. In this example, we probably want the three samples in each group to be more similar to each other than to the three samples in the other group. Essentially, we need to be able to control the within-group and between-group variation.

Biological variation is controlled through the `MethylomeParam` object and the method with which we simulate $\mathbf{B}$ and $\boldsymbol{\psi}$ through the `simulate()` method. Technical variation is controlled the `WGBSParam` object. If we want two samples to be very different from one another, then we would use two different `MethylomeParam` objects at the beginning of the process. If we want to simulate two technical replicates, however, then we would use the same `WGBSParam` objects; the two samples have identical 'true' methylomes and they only diverge at the 'sequencing' step of the simulation.

We might also consider more subtle ways of introducing variation. For example, we might use the same `MethylomeParam` object but vary how we simulate $\mathbf{B}$ and $\boldsymbol{\psi}$ between calls to `simulate()`. This avenue remains to be explored.

8.3.6 Simulating differential methylation

An obvious application of `methsim` is to simulate data to be used in a study comparing methods for identifying differential methylation.

One way of doing this would be to take a `SimulatedMethylome` object and create a modified copy where $\mathbf{B}$ has been perturbed by specified amounts at a set of specified loci. We could then simulate bisulfite-sequencing data from each `SimulatedMethylome` and study which analysis methods can identify these simulated DMCs and DMRs.

8.3.7 Performance

While `methsim` is an R package, much of it is written in C++, using the Rcpp package [Eddelbuettel *et al.* 2011, Eddelbuettel 2013], which greatly speeds up the running time of key procedures. Furthermore, several steps of the simulation can be run in parallel for each chromosome. Parallelism is implemented via the `BiocParallel` Bioconductor package (`http://bioconductor.org/packages/BiocParallel`).

The `SimulatedMethylome` and `SimulatedBS` objects are large in memory. This is to some extent unavoidable; we are simulating whole-genome bisulfite-sequencing sequencing experiments that produce an enormous amount of data.

Using up to 8 CPU cores in parallel, it takes less than 9 minutes to simulate a single 30× sequencing coverage whole-genome bisulfite-sequencing assay from a 'true' methylome. These results are indicative of simulating high-coverage whole-genome bisulfite-sequencing for human-sized genomes. Simulations using a smaller genome or lower sequencing coverage will have faster running times and lower memory usage.

8.4 Results

A central feature of `methsim` is that the simulation parameters are sampled from an input sample. At a bare minimum, the first test of `methsim` is that simulated data are similar to the real data on which they are based. To assess this, we compare several summary measures between the simulated data and the real data. The simplest of these are summaries of the distributions of β-values and within-fragment co-methylation, which are explicitly sampled in `methsim`.

Firstly, looking at the distribution of β-values, we see that all of *m1AI*, *m1BI*, *m2BI*, *m3BI* and *m3BII* do a reasonable job of capturing the bimodality of β-values (Figure 8.7) and their relationship to CpG islands (Figure 8.8). The effect of both the independent perturbations (*m2*) and the correlated perturbations (*m3*) appear to be dominated by variation due to variation in sequencing coverage. The lack of within-fragment co-methylation in *m1AI* does not affect the genome-wide distribution of methylation levels. The mixture model, *m3BII*, has noticeably more intermediate β-values owing to the increased heterogeneity at each CpG across the sub-populations.

Turning our attention to within-fragment co-methylation, Figure 8.9 shows the genome-level and chromosome-level Mantel-Haenszel estimates of within-fragment co-methylation. Figure 8.10 shows the genome-level estimates of within-fragment co-methylation stratified by CpG islands. These results are only available for CpG pairs with $NIL = 0$ because the

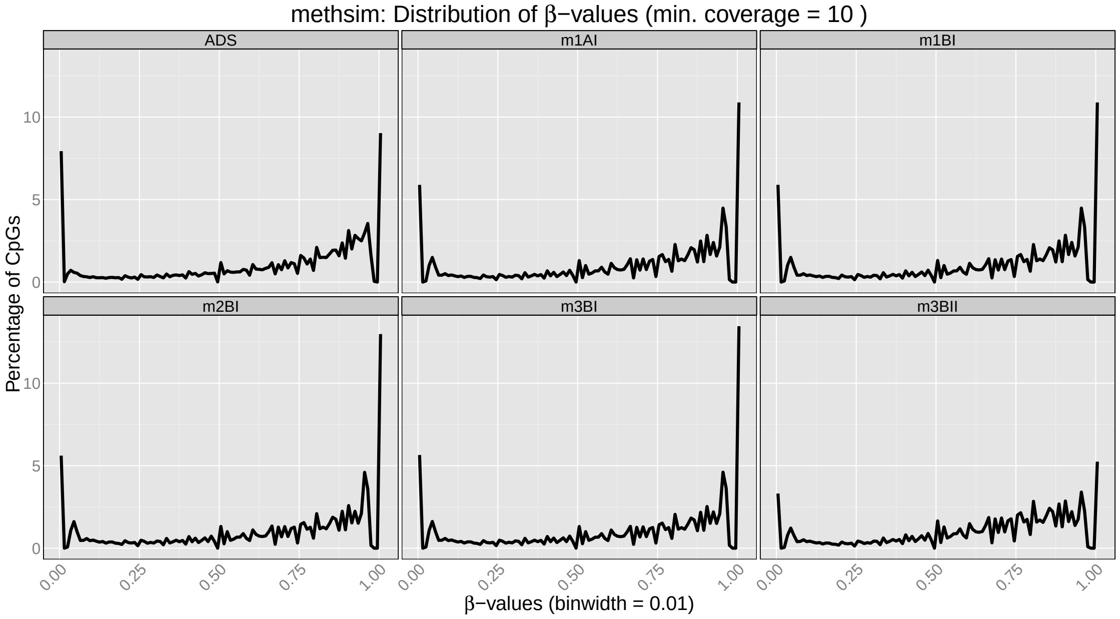


Figure 8.7: Frequency polygon of the genome-wide distribution of CpG β-values for `methsim` simulated data and the *ADS* input sample. β-values are grouped into 0.01-width bins and the percentage of CpGs in each bin is plotted on the y-axis. Observations have been combined across strands and only CpGs with at least $10\times$ sequencing coverage are included.

`asMethPat()` function does not create $NIL > 0$ pairs.

Reassuringly, we see no evidence of within-fragment co-methylation in *m1AI*, which simulates independence of within-fragment methylation states. The trend of the genome-level estimates, by design, very closely match that of the *ADS* input sample. We see, however, that the chromosome-level estimates of *m1BI*, *m2BI*, *m3BI* and *m3BII* are perhaps a little too homogeneous when compared to the *ADS* input sample. We also see in the results for *m1BI*, *m2BI*, *m3BI* and *m3BII* the effect of simulating from a $Gaussian(0, \sigma)$ for $IPD > 180$, as outlined in Section 8.3.2. The *m3BII* model again stands out, this time with a noticeably higher level of within-fragment co-methylation. Each sub-population has its own $\boldsymbol{\theta}$ and, although these are sampled from the same parametric model, this suggests that such heterogeneity in the sample will artificially inflate estimates of within-fragment co-methylation, even those made at the genome-level. While this requires further investigation, it raises the possibility that increased genome-level estimates of within-fragment co-methylation may in fact be measuring increased heterogeneity of the sample.

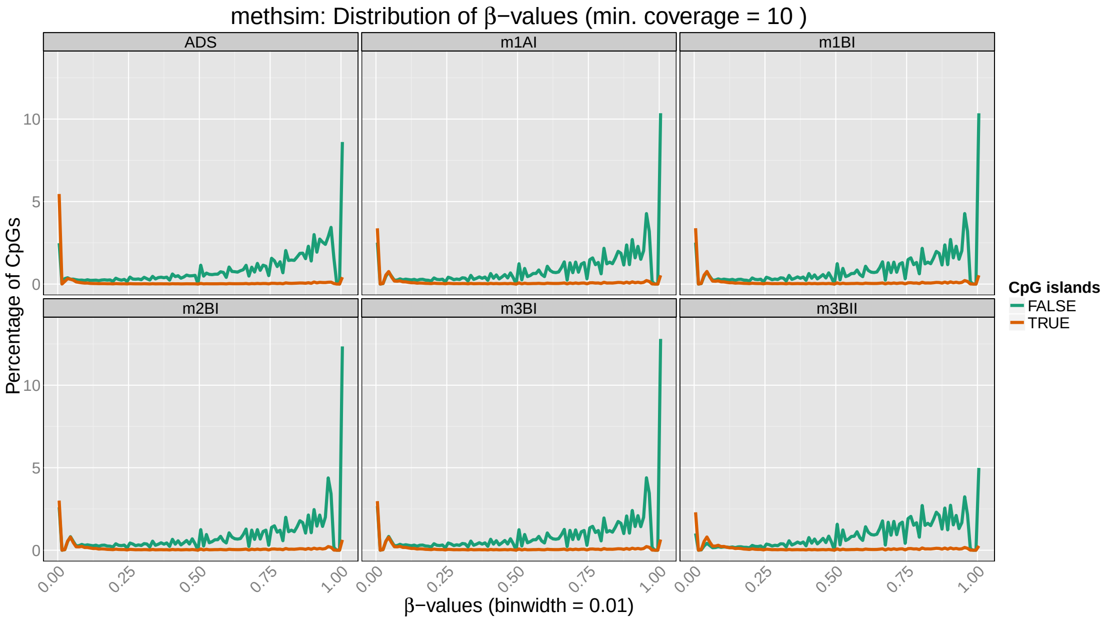


Figure 8.8: Frequency polygon of the genome-wide distribution of CpG β-values for the `methsim` simulated data and the *ADS* input sample, stratified by whether the CpG is in a CpG island. Only CpGs with at least 10× sequencing coverage are included. 'Spikes' in the density estimate are due to the discreteness of β-values.

These caveats aside, all five models do quite a good job of capturing the average level of CpG methylation and the strength of within-fragment co-methylation. Unfortunately, the results are not so promising for the correlations of β-values.

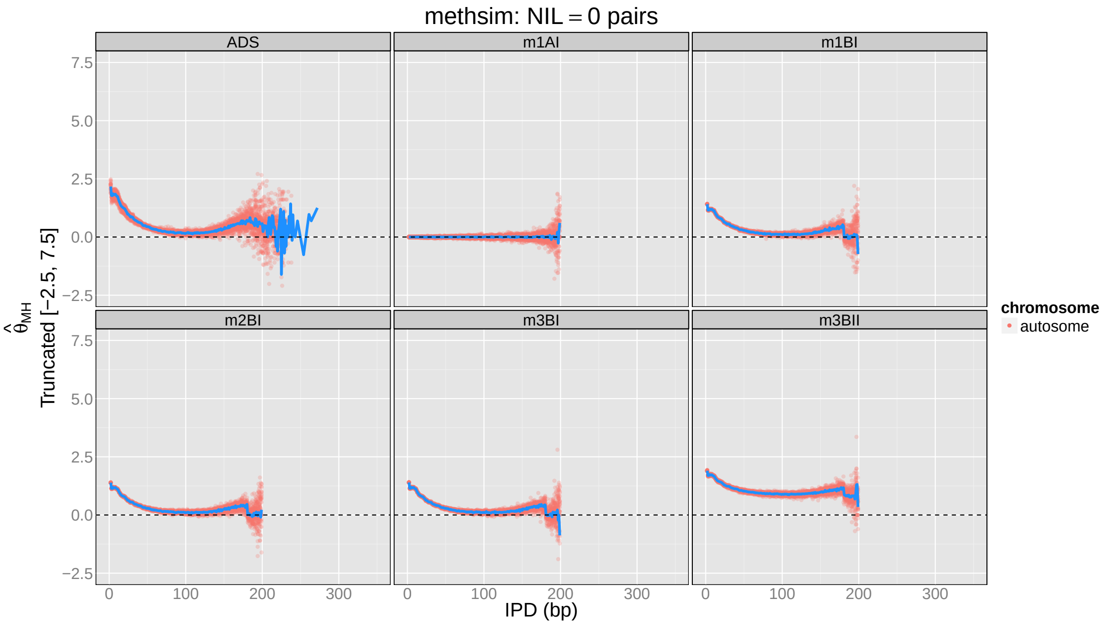


Figure 8.9: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the methsim simulated data and the ADS input sample using $NIL = 0$ pairs. The genome-level estimates are shown by the blue line while the chromosome-level estimates are plotted as points (only autosomal data are simulated, hence only autosomal data are shown).

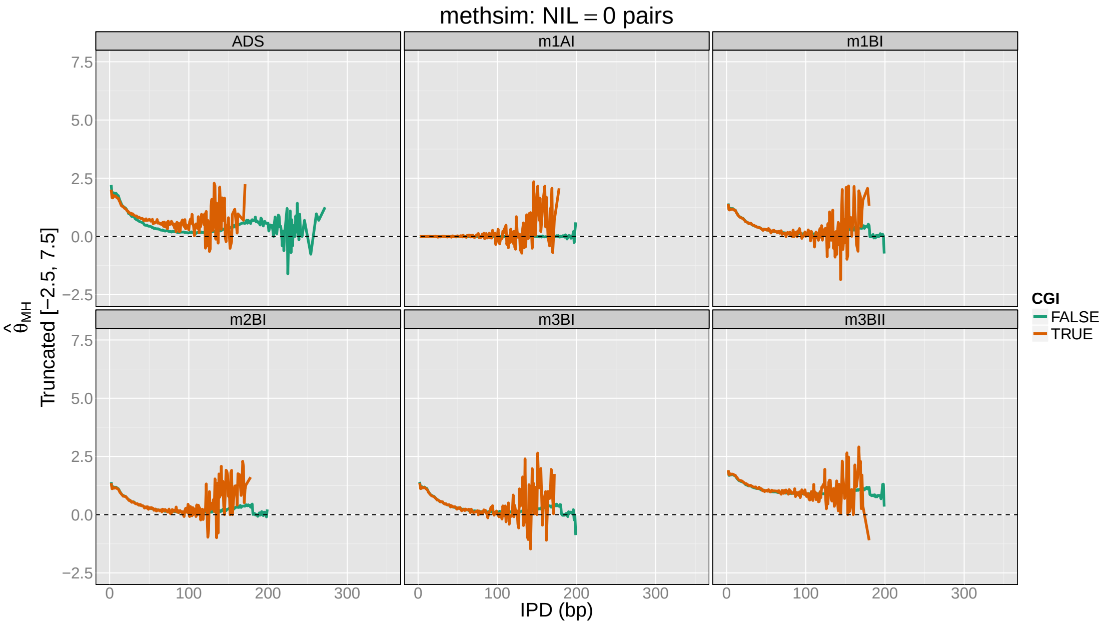


Figure 8.10: The Mantel-Haenszel estimate of the log odds ratio, $\widehat{\theta}_{MH}$, is plotted as a function of IPD for the methsim simulated data and the ADS input sample using $NIL = 0$ pairs. Only the genome-level estimates are shown, stratified by whether the pair is inside a CpG island (CGI).

In contrast to the other two summaries, the intention with `methsim` was to not explicitly model the spatial correlation of β-values. Rather, the idea was to see how the marginal level of methylation and the within-fragment co-methylation affected these correlations[14]. However, this overlooks that we are already imposing some structure, and therefore correlations, on the $\mathbf{B}$ by specifying them in a region-specific manner.

Figure 8.11 shows the dramatic consequences of this modelling decision. When compared to the correlations of β-values from the *ADS* input sample, we see that in all five models the correlations of β-values are far too strong. Stratifying these correlations by CpG islands, we see that the source of this problem is in regions outside of CpG islands (Figure 8.12). This gives some hope that this deficiency may fixed by a more careful modelling of β-values outside CpG islands, and remains a source of ongoing work.

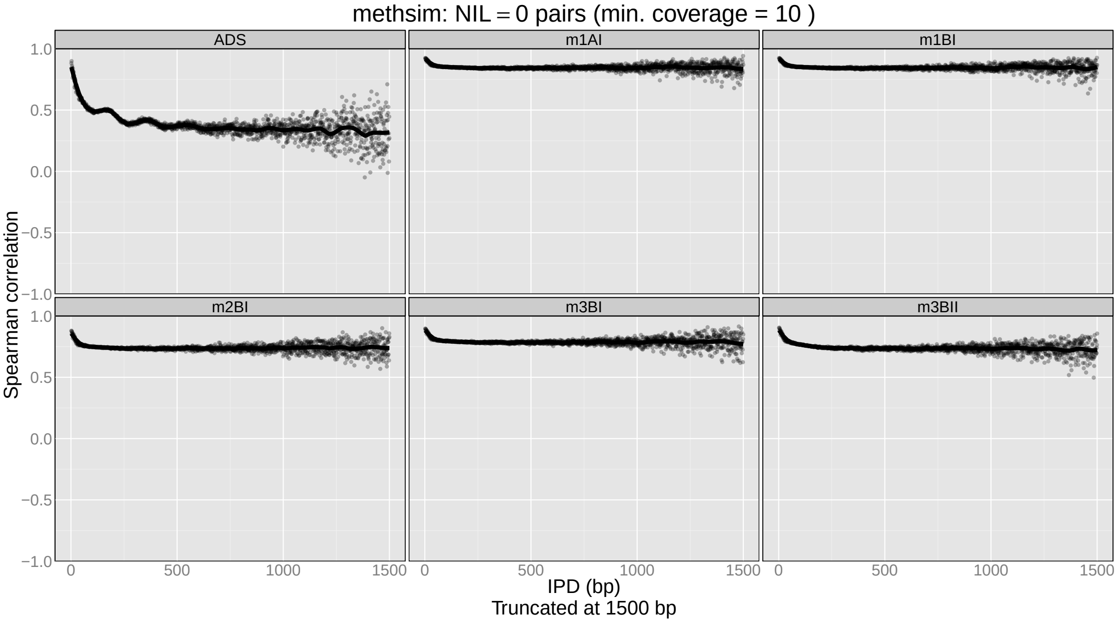


Figure 8.11: Spearman correlations of β-values as a function of IPD for strand-collapsed pairs of CpGs with $NIL = 0$ for `methsim` simulated data and the *ADS* input sample. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

[14] While *m3BI* and *m3BII* do include some spatial dependence of the $\mathbf{B}$ via the autoregressive perturbation, these perturbations are independent of IPD and are really intended to analyse the effect of the independence condition in *m2BI*.

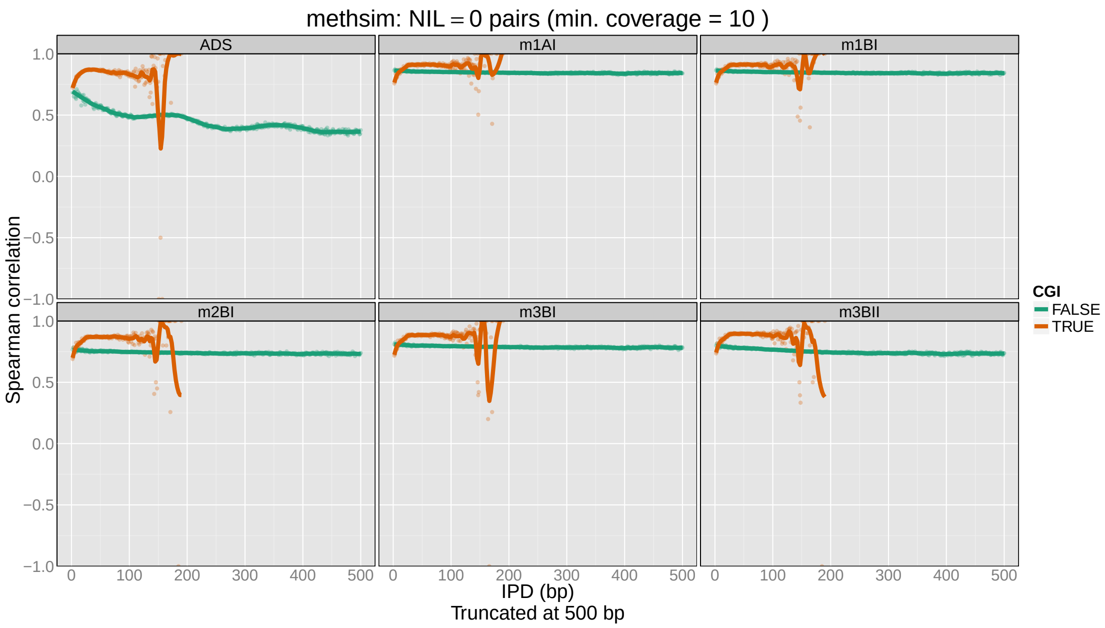


Figure 8.12: Spearman correlations of β-values as a function of IPD stratified by CpG island status for strand-collapsed pairs of CpGs with $NIL = 0$ for `methsim` simulated data and the ADS input sample. The raw estimates of the correlations are shown as semi-transparent points and are overlaid with a loess fit to these points (`span` $= 0.1$).

8.5 Summary

Simulating DNA methylation data is complicated by its heterogeneity. This heterogeneity exists is in multiple 'directions': along the genome, between cells in a sample, and between subjects. As more data become available, we can get a better handle on the causes of the variation but, for now, it remains challenging.

`methsim` provides a framework that I have used to experiment with models for simulating DNA methylation data. The models I have currently explored capture some key aspects of the data including, for the first time, within-fragment co-methylation. Unfortunately, while these results are promising on some fronts, there is clearly much work to be done on others.

Most notably, and frustratingly, I do not yet have an adequate model for the correlations of the true methylation levels, **B**. The model of Lacey *et al.* [2013] is promising, although may be computationally infeasible for whole-genome data given the timings reported in the original publication (13 seconds for $5,000$ CpGs increasing to 80 seconds for $10,000$

CpGs). The model used by `WGBSsuite` Rackham *et al.* [2015] is similar to m2, so I suspect it will produce similar results, but this requires further investigation. I continue to explore how alternative models of **B** might be integrated into `methsim` to improve this aspect of the simulation.

Chapter 9

Concluding remarks

This thesis has examined the statistical and computational challenges raised by high-throughput assays of DNA methylation. The first chapter introduced DNA methylation and then extensively reviewed different high-throughput assays for measuring it, focusing on the current gold standard assay of whole-genome bisulfite-sequencing.

The second chapter detailed the bioinformatics analysis of whole-genome bisulfite-sequencing data. Beginning with the `FASTQ` files created by the DNA sequencers, we reviewed the quality control procedures, read mapping, and post-processing steps that are essential to creating a `BAM` file containing high-quality mapped reads. We then described methods for calling methylation events from bisulfite-sequencing data. The chapter concluded by introducing m-tuples and the `methtuple` software that I wrote. Uniquely, `methtuple` can call methylation patterns at m-tuples from whole-genome bisulfite-sequencing data.

The third chapter described the 40 whole-genome bisulfite-sequencing samples that are analysed throughout the remainder of the thesis.

Chapter 4 laid out a statistical framework for analysing bisulfite-sequencing data. This process illustrated that the commonly used summary of methylation levels, the β-value, aggregates over many potential sources of variation. In light of this, we discussed the interpretation of β-values and examined the distributions of β-values in our 40 samples.

Chapter 5 examined methods for analysing bisulfite-sequencing data to address key biological questions. Most downstream analyses are based on β-values, such as the identification of differential methylation, and we reviewed these. There is, however, a growing interest in analyses based on methylation patterns at m-tuples since these can address novel biological questions. A limitation to the development of these methods has been a lack of software. The chapter concluded with the development of the `MethylationTuples` R package, which, together with `methtuple`, is designed to facilitate analyses based on methylation patterns at m-tuples.

The sixth and seventh chapters are a comprehensive review and analysis of co-methylation, the spatial dependence of DNA methylation. Chapter 6 re-examined previous analyses of co-methylation and found that these have been limited by inadequate data and deficiencies in the statistical methods applied. We addressed some of these limitations with an extensive analysis of co-methylation in Chapter 7. We proposed a novel analysis of co-methylation using methylation patterns from individual DNA fragments, which we call within-fragment co-methylation. A simulation study demonstrated that, in general, we cannot estimate this within-fragment co-methylation for individual pairs of CpGs. By aggregating pairs of CpG, however, the simulation study showed that we can estimate a summary of this spatial dependence of DNA methylation along individual DNA fragments. We applied this method to 40 whole-genome bisulfite-sequencing samples to identify genomic features that influence co-methylation and examined how it varies between different tissues.

To conclude, Chapter 8 detailed our efforts to develop the `methsim` software to simulate whole-genome bisulfite-sequencing data. Unlike existing software, `methsim` seeks to model co-methylation so as to create more realistic simulated data. Such simulated data will be useful in developing and benchmarking methods for the analysis of DNA methylation data. While aspects of `methsim` are promising, the overly-strong correlations of β-values is an obvious inadequacy in the current model. We continue to work to improve this and to develop the software more generally.

DNA methylation data have a complex dependence structure, as demonstrated by the analyses presented in this thesis. Any analysis must bear in mind the multiple sources of variation in the data and the amount of aggregation required in order to extract meaningful results.

The large size of the data generated by these experiments can make it challenging to develop analysis methods. This is particularly true when coupled with a lack of software for extracting, manipulating and summarising the relevant features of the data. Much of the effort in this thesis has gone into developing software to extract and manage this information (`methtuple` and `MethylationTuples`) so that new tools can be built on a higher foundation (`methsim`). It is my hope that, with continued development, these tools may facilitate other analyses to unravel the complexity of DNA methylation and its many biological roles.

Appendix A

Appendix

A.1 The probability that two dependent Bernoulli random variables are identical

Lindqvist [1978] wrote a brief note on Bernoulli trials with dependence. Building on earlier work by Klotz [1973], Lindqvist parameterises the Bernoulli process $X_1, X_2, \ldots$ on $\{0, 1\}$ by the parameters $p = Pr(X_i = 1)$ and $c = cor(X_{i-1}, X_i)$ and shows that the transition matrix is given by

$$\Pi = \begin{pmatrix} (1-p) + cp & p(1-c) \\ (1-p)(1-c) & p + c(1-p) \end{pmatrix}$$

provided that $max(1 - \frac{1}{p}, 1 - \frac{1}{1-p}) \leqslant c \leqslant 1$.

From this we can compute the joint distribution,

$$\begin{aligned} &Pr(X_1 = x_1, \ldots, X_n = x_n) \\ =&Pr(X_1 = x_1)Pr(X_2 = x_2|X_2 = x_2) \cdots Pr(X_n = x_n|X_{n-1} = x_{n-1}) \end{aligned}$$

In particular, in the case $n = 2$ we can compute the probability that two dependent and identically distributed Bernoulli random variables are equal.

To extend the above result to the probability that two dependent and **non-identically** distributed Bernoulli random variables are equal, simply requires that we derive the appropriate transition matrix. Switching notation to that used in Chapter 6, let $Z_{h,i} \overset{d}{=} Bernoulli(p_h)$ and $Z_{h',i'} \overset{d}{=} Bernoulli(p_{h'})$. The transition matrix, $\Pi = \big(\scriptstyle Pr(Z_{h',i'}=z_{h',i'}|Z_{h,i}=z_{h,i})\big)$, is given by

$$\Pi = \begin{pmatrix} (1-p_{h'}) + cp_h & p_{h'} - cp_h \\ (1-p_{h'}) - c(1-p_h) & p_{h'} + c(1-p_h) \end{pmatrix}$$

We can then compute the desired probability

$$\begin{aligned} Pr(Z_{h,i} = Z_{h',i'}) &= Pr(Z_{h,i} = 0, Z_{h',i'} = 0) + Pr(Z_{h,i} = 1, Z_{h',i'} = 1) \\ &= Pr(Z_{h',i'} = 0|Z_{h,i} = 0)Pr(Z_{h,i} = 0) \\ &+ Pr(Z_{h',i'} = 1|Z_{h,i} = 1)Pr(Z_{h,i} = 1) \\ &= \big[(1-p_{h'}) + cp_h\big](1-p_h) + \big[p_{h'} + c(1-p_h)\big]p_h \\ &= (1-p_h)(1-p_{h'}) + cp_h\big[1-p_h\big] + p_hp_{h'} + cp_h(1-p_h) \\ &= (1-p_h)(1-p_{h'}) + p_hp_{h'} + 2cp_h(1-p_h) \end{aligned}$$

A.2 Computing details

All computational work was performed on one of the Bioinformatics Division's HP Blade servers. These are shared-use, shared-memory machines. The basic specifications are shown in Table A.1.

Table A.1: Bioinformatics Division server specifications.

Machine name	Processors	Number of cores	RAM
unix88	4× Intel Xeon X7350 @ 2.93GHz	16	128 GB
unix301	4× AMD Opteron 8435 @ 2.6GHz	24	256 GB
unix302	4× AMD Opteron 6174 @ 2.2GHz	48	512 GB
unix303	4× AMD Opteron 6176 @ 2.3GHz	48	512 GB
unix305	4× AMD Opteron 6276 @ 2.3GHz	64	512 GB

A.3 Software details

The analyses described in my thesis made extensive use of R (`R Under development (unstable) (2014-10-29 r66891)` and `R version 3.2.0 (2015-04-16)`) and Python (`v2.7`). All analyses were run on one of the servers described above. The scripts used to prepare the results for each chapter are available from `https://github.com/PeteHaitch/phd_thesis_analyses`.

I developed several pieces of software during my PhD. These are listed below, along with the version used in for analyses in my thesis:

- `methtuple` (`v1.4.0`)
- `GenomicTuples` (`v1.2.1`)
- `MethylationTuples` (`v0.3.0.9007`, commit `4e127d2`)
- `methsim` (`v0.5.0.9013`, commit `9162c8b`)

In addition, I made use of the R packages listed in Table A.2 and gratefully acknowledge the developers.

Table A.2: R packages used in thesis (as reported by `devtools::session_info()`)

package	version	source
BiocGenerics	0.14.0	Bioconductor
BiocParallel	1.2.1	Bioconductor
Biostrings	2.36.0	Bioconductor
BSgenome	1.36.0	Bioconductor
BSgenome.Hsapiens.UCSC.hg18	1.3.1000	Bioconductor
BSgenome.Hsapiens.UCSC.hg19	1.4.0	Bioconductor
BSgenome.Mmusculus.UCSC.mm10	1.4.0	Bioconductor
cmm	0.8	CRAN (R 3.2.0)
data.table	1.9.4	CRAN (R 3.2.0)
devtools	1.7.0	CRAN (R 3.2.0)
dplyr	0.4.1	CRAN (R 3.2.0)

GenomeInfoDb	1.4.0	Bioconductor
GenomicRanges	1.20.3	Bioconductor
GenomicTuples	1.2.1	Bioconductor
ggplot2	1.0.1	CRAN (R 3.2.0)
gridExtra	0.9.1	CRAN (R 3.2.0)
IRanges	2.2.1	Bioconductor
knitr	1.10	CRAN (R 3.2.0)
makeCGI	1.2	local
MethylSeekR	1.8.0	Bioconductor
mhsmm	0.4.14	CRAN (R 3.2.0)
mipfp	2.0	CRAN (R 3.2.0)
mvtnorm	1.0-2	CRAN (R 3.2.0)
numDeriv	2012.9-1	CRAN (R 3.2.0)
pryr	0.1	CRAN (R 3.2.0)
RColorBrewer	1.1-2	CRAN (R 3.2.0)
Rcpp	0.11.6	CRAN (R 3.2.0)
R.methodsS3	1.7.0	CRAN (R 3.2.0)
R.oo	1.19.0	CRAN (R 3.2.0)
RPushbullet	0.2.0	CRAN (R 3.2.0)
Rsamtools	1.20.1	Bioconductor
Rsolnp	1.15	CRAN (R 3.2.0)
rtracklayer	1.28.2	Bioconductor
R.utils	2.0.2	CRAN (R 3.2.0)
S4Vectors	0.6.0	Bioconductor
scales	0.2.4	CRAN (R 3.2.0)
sn	1.2-1	CRAN (R 3.2.0)
SNPlocs.Hsapiens.dbSNP.20120608	0.99.9	Bioconductor
stringr	1.0.0	CRAN (R 3.2.0)
tidyr	0.2.0	CRAN (R 3.2.0)
truncnorm	1.0-7	CRAN (R 3.2.0)
VariantAnnotation	1.14.0	Bioconductor

XVector	0.8.0	Bioconductor

Bibliography

Adams, J. M. and Cory, S. (1975). Modified nucleosides and bizarre 5'-termini in mouse myeloma mRNA. *Nature*, **255**(5503), 28–33.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* John Wiley & Sons.

Akalin, A., Garrett-Bakelman, F. E., Kormaksson, M., Busuttil, J., Zhang, L., Khrebtukova, I., Milne, T. A., Huang, Y., Biswas, D., Hess, J. L., Allis, C. D., Roeder, R. G., Valk, P. J. M., Löwenberg, B., Delwel, R., Fernandez, H. F., Paietta, E., Tallman, M. S., Schroth, G. P., Mason, C. E., Melnick, A., and Figueroa, M. E. (2012a). Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genetics*, **8**(6), e1002781.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012b). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, **13**(10), R87.

Akulenko, R. and Helms, V. (2013). DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human Molecular Genetics*, **22**(15), 3016–3022.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2007). *Molecular Biology of the Cell.* Garland Science.

Annunziato, A. (2008). DNA packaging: Nucleosomes and chromatin. *Nature Education.*

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika.*

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika*, **43**(3/4), 461–464.

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Xie, B., Daley, G. Q., and Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, **27**(4), 361–368.

Barrera, V. and Peinado, M. A. (2012). Evaluation of single CpG sites as proxies of CpG island methylation states at the genome scale. *Nucleic Acids Research*, **40**(22), 11490–11498.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, **12**(1), R10.

Benjamini, Y. and Heller, R. (2007). False Discovery Rates for Spatial Signals. *Journal of the American Statistical Association*, **102**(480), 1272–1281.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **57**(1), 289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple Hypotheses Testing with Weights. *Scandinavian Journal of Statistics*, **24**(3), 407–418.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, **40**(10), e72–e72.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**(4), 1165–1188.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, **93**(3), 491–507.

Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & Development*, **23**(7), 781–783.

Berman, B. P., Weisenberger, D. J., Aman, J. F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C. P. E., van Dijk, C. M., Tollenaar, R. A. E. M., Van Den Berg, D., and Laird, P. W. (2012). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nature Genetics*, **44**(1), 40–46.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., Fan, J.-B., and Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, **98**(4), 288–295.

Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, **40**(1), 91–99.

Bird, A. P. (2007). Perceptions of epigenetics. *Nature*, **447**(7143), 396–398.

Booth, M. J., Branco, M. R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, **336**(6083), 934–937.

Bostick, M., Kim, J. K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S. E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, **317**(5845), 1760–1764.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika*, **68**(1), 73–84.

Breslow, N. E. (1996). Statistics in epidemiology: the case-control study. *Journal of the American Statistical Association*, **91**(433), 14–28.

Bruder, C. E. G., Piotrowski, A., Gijsbers, A. A. C. J., Andersson, R., Erickson, S., Diaz de Ståhl, T., Menzel, U., Sandgren, J., von Tell, D., Poplawski, A., Crowley, M., Crasto, C., Partridge, E. C., Tiwari, H., Allison, D. B., Komorowski, J., van Ommen, G.-J. B., Boomsma, D. I., Pedersen, N. L., den Dunnen, J. T., Wirdefeldt, K., and Dumanski, J. P. (2008). Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *American Journal of Human Genetics*, **82**(3), 763–771.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M. B. (2013). Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research*, **41**(16), e155–e155.

Capra, J. A. and Kostka, D. (2014). Modeling DNA methylation dynamics with approaches from phylogenetics. *Bioinformatics*, **30**(17), i408–14.

Capuano, F., Mülleder, M., Kok, R., Blom, H. J., and Ralser, M. (2014). Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical chemistry*, **86**(8), 3697–3702.

Carroll, L. and Tenniel, J. (1897). *Through the Looking Glass: And what Alice Found There*. Altemus' illustrated young people's library. Henry Altemus.

Chatterjee, A., Stockwell, P. A., Rodger, E. J., and Morison, I. M. (2012). Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, **40**(10), e79–e79.

Chau, Y.-Y., Bandiera, R., Serrels, A., Martínez-Estrada, O. M., Qing, W., Lee, M., Slight, J., Thornburn, A., Berry, R., McHaffie, S., Stimson, R. H., Walker, B. R., Chapuli, R. M., Schedl, A., and Hastie, N. (2014). Visceral and subcutaneous fat have different origins and evidence supports a mesothelial source. *Nature cell biology*, **16**(4), 367–375.

Chen, Y., Ning, Y., Hong, C., and Wang, S. (2014a). Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. *Genetic Epidemiology*, **38**(1), 42–50.

Chen, Z., Huang, H., and Liu, Q. (2014b). Detecting differentially methylated loci for multiple treatments based on high-throughput methylation data. *BMC Bioinformatics*, **15**(1), 142.

Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J. A., Kuo, F., Kim, J., Cokus, S. J., Casero, D., Bernal, M., Huijser, P., Clark, A. T., Krämer, U., Merchant, S. S., Zhang, X., Jacobsen, S. E., and Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. *Nature*, **466**(7304), 388–392.

Clark, S. J., Harrison, J., Paul, C. L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Research*, **22**(15), 2990–2997.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, **452**(7184), 215–219.

Cooper, D. N., Taggart, M. H., and Bird, A. P. (1983). Unmethylated domains in vertebrate DNA. *Nucleic Acids Research*, **11**(3), 647–658.

Daxinger, L. and Whitelaw, E. (2010). Transgenerational epigenetic inheritance: more questions than answers. *Genome Research*, **20**(12), 1623–1628.

Deans, C. and Maggert, K. A. (2015). What do you mean, "epigenetic"? *Genetics*, **199**(4), 887–896.

Dedeurwaerder, S., Defrance, M., Bizet, M., Calonne, E., Bontempi, G., and Fuks, F. (2014). A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in Bioinformatics*, **15**(6), 929–941.

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, **14**(6), 390–403.

Doi, A., Park, I.-H., Wen, B., Murakami, P., Aryee, M. J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G. Q., and Feinberg, A. P. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nature Genetics*, **41**(12), 1350–1353.

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, **15**(1), 215.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010).

Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**, 587.

Ecker, J. R. (2010). Zeroing in on DNA methylomes with no BS. *Nature Methods*, **7**(6), 435–437.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haefliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, **38**(12), 1378–1385.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer New York, New York, NY.

Eddelbuettel, D., François, R., and Allaire, J. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*.

Ehrich, M., Correll, D., and van den Boom, D. (2006). Introduction to EpiTYPER for quantitative DNA methylation analysis using the massARRAY system. *Sequenome*.

Ehrlich, M. (2002). DNA methylation in cancer: too much, but also too little. *Oncogene*, **21**(35), 5400–5413.

Ehrlich, M., Gama-Sosa, M. A., Carreira, L. H., Ljungdahl, L. G., Kuo, K. C., and Gehrke, C. W. (1985). DNA methylation in thermophilic bacteria: N4-methylcytosine, 5-methylcytosine, and N6-methyladenine. *Nucleic Acids Research*, **13**(4), 1399–1412.

Elgar, G. and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics : TIG*, **24**(7), 344–352.

Fang, F., Hodges, E., Molaro, A., Dean, M., Hannon, G. J., and Smith, A. D. (2012). Genomic landscape of human allele-specific DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(19), 7332–7337.

Faust, G. G. and Hall, I. M. (2014). SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, **30**(17), 2503–2505.

Feinberg, A. P. and Irizarry, R. A. (2010). Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences of the United States of America*, **107 Suppl 1**, 1757–1764.

Feng, H., Conneely, K. N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, **42**(8), e69–e69.

Fisher, R. A. (1922). On the interpretation of $\chi 2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, **85**(1), 87–94.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, **98**(1), 39–82.

Fisher, S. R. A. (1936). *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., and Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, **7**(6), 461–465.

Freeman, M. F. and Tukey, J. W. (1950). Transformations Related to the Angular and the Square Root. *The Annals of Mathematical Statistics*, **21**(4), 607–611.

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(5), 1827–1831.

Fürst, R. W., Kliem, H., Meyer, H. H. D., and Ulbrich, S. E. (2012). A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. *The Journal of steroid biochemistry and molecular biology*, **130**(1-2), 96–104.

Gaile, D. P., Schifano, E. D., Miecznikowski, J. C., Java, J. J., Conroy, J. M., and Nowak, N. J. (2007). Estimating the arm-wise false discovery rate in array comparative genomic hybridization experiments. *Statistical applications in genetics and molecular biology*, **6**(1), Article32.

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology*, **196**(2), 261–282.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5**(10), R80.

Gokhman, D., Lavi, E., Prüfer, K., Fraga, M. F., Riancho, J. A., Kelso, J., Pääbo, S., Meshorer, E., and Carmel, L. (2014). Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan. *Science*, **344**(6183), 523–527.

Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C.-L., Zhang, X., Golic, K. G., Jacobsen, S. E., and Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, **311**(5759), 395–398.

Good, I. J. and Mittal, Y. (1987). The Amalgamation and Geometry of Two-by-Two Contingency Tables. *The Annals of Statistics*, **15**(2), 694–711.

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, **23**(12), 2126–2135.

Haldane, J. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, **20**(4), 309–311.

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, **43**(8), 768–775.

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, **13**(10), R83.

Hansen, K. D., Sabunciyan, S., Langmead, B., Nagy, N., Curley, R., Klein, G., Klein, E., Salamon, D., and Feinberg, A. P. (2014). Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization. *Genome Research*, **24**(2), 177–184.

Hauck, W. W. (1989). Odds ratio inference from stratified samples. *Communications in Statistics-Theory and Methods*, **18**(2), 767–800.

Hayatsu, H., Wataya, Y., and Kazushige, K. (1970). The addition of sodium bisulfite to uracil and to cytosine. *Journal of the American Chemical Society*, **92**(3), 724–726.

He, J., Sun, X., Shao, X., Liang, L., and Xie, H. (2013). DMEAS: DNA methylation entropy analysis software. *Bioinformatics*, **29**(16), 2044–2045.

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

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. (2014). A Genetic Atlas of Human Admixture History. *Science*, **343**(6172), 747–751.

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *The Journal of biological chemistry*, **175**(1), 315–332.

Houseman, E. A., Molitor, J., and Marsit, C. J. (2014). Reference-Free Cell Mixture Adjustments in Analysis of DNA Methylation Data. *Bioinformatics*, page btu029.

Huang, T. H., Perry, M. R., and Laux, D. E. (1999). Methylation profiling of CpG islands in human breast cancer cells. *Human Molecular Genetics*, **8**(3), 459–470.

Irizarry, R. A., Ladd-Acosta, C., Carvalho, B. S., Wu, H., Brandenburg, S. A., Jeddeloh, J. A., Wen, B., and Feinberg, A. P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research*, **18**(5), 780–790.

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S., and Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, **41**(2), 178–186.

Jaffe, A. E. and Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, **15**(2), R31.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., and Irizarry, R. A. (2012a). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology*, **41**(1), 200–209.

Jaffe, A. E., Feinberg, A. P., Irizarry, R. A., and Leek, J. T. (2012b). Significance analysis and statistical dissection of variably methylated regions. *Biostatistics (Oxford, England)*, **13**(1), 166–178.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim, S.-B., Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J.-S., Gunaratne, P., Godley, L. A., Darlington, G. J., Rao, A., Li, W., and Goodell, M. A. (2014). Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature Genetics*, **46**(1), 17–23.

Ji, L., Sasaki, T., Sun, X., Ma, P., Lewis, Z. A., and Schmitz, R. J. (2014). Methylated DNA is over-represented in whole-genome bisulfite sequencing data. *Frontiers in genetics*, **5**, 341.

Johnson, T. and Coghill, R. (1925). Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus1. *Journal of the American Chemical . . .*, **47**.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, **13**(7), 484–492.

Jones, P. A. and Liang, G. (2009). Rethinking how DNA methylation patterns are maintained. *Nature Reviews Genetics*, **10**(11), 805–811.

Jones, P. A. and Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science*, **293**(5532), 1068–1070.

Jurkowska, R. Z., Jurkowski, T. P., and Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. *Chembiochem : a European journal of chemical biology*, **12**(2), 206–222.

Kechris, K. J., Biehs, B., and Kornberg, T. B. (2010). Generalizing moving averages for tiling arrays using combined p-value statistics. *Statistical applications in genetics and molecular biology*, **9**(1), Article29.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*.

Khulan, B., Thompson, R. F., Ye, K., Fazzari, M. J., Suzuki, M., Stasiek, E., Figueroa, M. E., Glass, J. L., Chen, Q., Montagna, C., Hatchwell, E., Selzer, R. R., Richmond, T. A., Green, R. D., Melnick, A., and Greally, J. M. (2006). Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Research*, **16**(8), 1046–1055.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, **21**(3), 487–493.

Klotz, J. (1973). Statistical Inference in Bernoulli Trials with Dependence. *The Annals of Statistics*, **1**(2), 373–379.

Kraemer, H. C. (2006). Correlation coefficients in medical research: from product moment correlation to the odds ratio. *Statistical methods in medical research*, **15**(6), 525–545.

Kriaucionis, S. and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**(5929), 929–930.

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

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, **9**(2), 145–151.

Kunde-Ramamoorthy, G., Coarfa, C., Laritsky, E., Kessler, N. J., Harris, R. A., Xu, M., Chen, R., Shen, L., Milosavljevic, A., and Waterland, R. A. (2014). Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. *Nucleic Acids Research*, **42**(6), e43.

Lacey, M. R. and Ehrlich, M. (2009). Modeling dependence in methylation patterns with application to ovarian carcinomas. *Statistical applications in genetics and molecular biology*, **8**(1), Article 40.

Lacey, M. R., Baribault, C., and Ehrlich, M. (2013). Modeling, simulation and analysis of methylation profiles from reduced representation bisulfite sequencing experiments. *Statistical applications in genetics and molecular biology*, **12**(6), 723–742.

Laird, P. W. (2003). The power and the promise of DNA methylation markers. *Nature reviews. Cancer*, **3**(4), 253–266.

Laird, P. W. (2010). Principles and challenges of genomewide DNA methylation analysis. *Nature Reviews Genetics*, **11**(3), 191–203.

Landan, G., Cohen, N. M., Mukamel, Z., Bar, A., Molchadsky, A., Brosh, R., Horn-Saban, S., Zalcenstein, D. A., Goldfinger, N., Zundelevich, A., Gal-Yam, E. N., Rotter, V., and Tanay, A. (2012). Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genetics*, **44**(11), 1207–1214.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J. E., Ainscough, R., Beck, S., Bentley, D. R., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R. M., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M. C., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton,

R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F. S., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P. J., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4), 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10**(3), R25.

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., Bartlett, J. J., Pavlenok, M., and Gundlach, J. H. (2013). Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(47), 18904–18909.

Lawrence, M. and Morgan, M. (2014). Scalable Genomics with R and Bioconductor. *Statistical Science*, **29**(2), 214–226.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, **9**(8), e1003118.

Ledford, H. (2008). Language: Disputed definitions. *Nature*, **455**(7216), 1023–1028.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K. A., and Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**(10), 733–739.

Li, E., Bestor, T. H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, **69**(6), 915–926.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, **366**(6453), 362–365.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5), 589–595.

Li, H. and Durbin, R. M. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. R., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, J., Harris, R. A., Cheung, S. W., Coarfa, C., Jeong, M., Goodell, M. A., White, L. D., Patel, A., Kang, S.-H., Shaw, C., Chinault, A. C., Gambin, T., Gambin, A., Lupski, J. R., and Milosavljevic, A. (2012). Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. *PLoS Genetics*, **8**(5), e1002692.

Li, S., Garrett-Bakelman, F., Perl, A. E., Luger, S. M., Zhang, C., To, B. L., Lewis, I. D., Brown, A. L., D Andrea, R. J., Ross, M., Levine, R., Carroll, M., Melnick, A., and Mason, C. E. (2014). Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biology*, **15**(9), 472.

Li, W. (1997). The study of correlation structures of DNA sequences: a critical review. *Computers & chemistry*, **21**(4), 257–271.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., Zhang, H., Chen, Q., Luo, R., Chen, M., He, Y., Jin, X., Zhang, Q., Yu, C., Zhou, G., Sun, J., Huang, Y., Zheng, H., Cao, H., Zhou, X., Guo, S., Hu, X., Li, X., Kristiansen, K., Bolund, L., Xu, J., Wang, W., Yang, H., Wang, J., Li, R., Beck, S., Wang, J., and Zhang, X. (2010). The DNA methylome of human peripheral blood mononuclear cells. *PLoS biology*, **8**(11), e1000533.

Liang, K.-Y. and Self, S. G. (1985). Tests for homogeneity of odds ratio when the data are sparse. *Biometrika*, **72**(2), 353–358.

Lindqvist, B. (1978). A note on Bernoulli trials with dependence. *Scandinavian Journal of Statistics*, **5**, 205–208.

Lister, R. and Ecker, J. R. (2009). Finding the fifth base: genome-wide sequencing of cytosine methylation. *Genome Research*, **19**(6), 959–966.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**(3), 523–536.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**(7271), 315–322.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., Downes, M., Yu, R., Stewart, R., Ren, B., Thomson, J. A., Evans, R. M., and Ecker, J. R. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**(7336), 68–73.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. *Science*, **341**(6146), 1237905–1237905.

Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. (2012). Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biology*, **13**(7), R61.

Liu, Y., Li, X., Aryee, M. J., Ekström, T. J., Padyukov, L., Klareskog, L., Vandiver, A., Moore, A. Z., Tanaka, T., Ferrucci, L., Fallin, M. D., and Feinberg, A. P. (2014). GeMes, Clusters of DNA Methylation under Genetic Control, Can Inform Genetic and Epigenetic Analysis of Disease. *American Journal of Human Genetics*, **94**(4), 485–495.

Lun, A. T. L. and Smyth, G. K. (2014). De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly. *Nucleic Acids Research*, **42**(11), e95.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS biology*, **8**(11), e1000506.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**(4), 719–748.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, **17**(1).

Martin, T. C., Yet, I., Tsai, P.-C., and Bell, J. T. (2015). coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns. *BMC Bioinformatics*, **16**(1), 131.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, **342**(6159), 747–749.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, **33**(18), 5868–5877.

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Research*, **40**(17), e136.

Mohandas, T., Sparkes, R. S., and Shapiro, L. J. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*, **211**(4480), 393–396.

Moore, L. D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, **38**(1), 23–38.

Morgan, H. D., Sutherland, H. G., Martin, D. I., and Whitelaw, E. (1999). Epigenetic inheritance at the agouti locus in the mouse. *Nature Publishing Group*, **23**(3), 314–318.

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigó, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., von Niederhausern, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zdobnov, E. M., Zody, M. C., and Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915), 520–562.

Nan, X., Meehan, R. R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Research*, **21**(21), 4886–4892.

Nestor, C. E., Ottaviano, R., Reinhardt, D., Cruickshanks, H. A., Mjoseng, H. K., McPherson, R. C., Lentini, A., Thomson, J. P., Dunican, D. S., Pennings, S., Anderton, S. M., Benson, M., and Meehan, R. R. (2015). Rapid reprogramming of epigenetic and transcriptional profiles in mammalian culture systems. *Genome Biology*, **16**(1), 11.

Nunes, M. A. and Nason, G. P. (2009). A multiscale variance stabilization for binomial sequence proportion estimation. *Statistica Sinica*.

Olkin, I. (1960). *Contributions to Probability and Statistics*. Essays in Honor of Harold Hotelling. Stanford University Press.

Park, Y., Figueroa, M. E., Rozek, L. S., and Sartor, M. A. (2014). methylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. In *Proceedings of the Royal Society of . . .*, pages 240–242.

Pedersen, B. S., Schwartz, D. A., Yang, I. V., and Kechris, K. J. (2012). Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, **28**(22), 2986–2988.

Pedersen, B. S., Eyring, K., De, S., Yang, I. V., and Schwartz, D. A. (2014). Fast and accurate alignment of long bisulfite-seq reads. *arXiv.org*.

Peng, Q. and Ecker, J. R. (2012). Detection of allele-specific methylation through a generalized heterogeneous epigenome model. *Bioinformatics*, **28**(12), i163–i171.

Phipson, B. and Oshlack, A. (2014). DiffVar: a new method for detecting differential variability with application to methylation in cancer and aging. *Genome Biology*, **15**(9), 465.

Plongthongkum, N., Diep, D. H., and Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics*, **15**(10), 647–661.

Rackham, O. J. L., Dellaportas, P., Petretto, E., and Bottolo, L. (2015). WGBSSuite: simulating whole-genome bisulphite sequencing data and benchmarking differential DNA methylation analysis tools. *Bioinformatics*.

Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I., and Whitelaw, E. (2002). Metastable epialleles in mammals. *Trends in genetics : TIG*, **18**(7), 348–351.

Ratel, D., Ravanat, J. L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. *BioEssays : news and reviews in molecular, cellular and developmental biology*, **28**(3), 309–315.

Razin, A. and Cedar, H. (1991). DNA methylation and gene expression. *Microbiological reviews*, **55**(3), 451–458.

Rijlaarsdam, M. A., van der Zwan, Y. G., Dorssers, L. C. J., and Looijenga, L. H. J. (2014). DMRforPairs: identifying differentially methylated regions between unique samples using array based methylation profiles. *BMC Bioinformatics*, **15**(1), 141.

Robins, J., Breslow, N., and Greenland, S. (1986). Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics*, **42**(2), 311–323.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, **5**, 324.

Ruppel, W. G. (1899). Zur Chemie der Tuberkelbacillen. *Hoppe-Seyler´ s Zeitschrift für physiologische Chemie*.

Russo, V. E. A., Martienssen, R. A., and Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor monograph series. Cold Spring Harbor Laboratory Press.

Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M. A., Bibikova, M., and Esteller, M. (2011). Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : official journal of the DNA Methylation Society*, **6**(6), 692–702.

Scarano, E., Iaccarino, M., Grippo, P., and Parisi, E. (1967). The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proceedings of the National Academy of Sciences of the United States of America*, **57**(5), 1394–1400.

Schatz, P., Dietrich, D., and Schuster, M. (2004). Rapid analysis of CpG methylation patterns using RNase T1 cleavage and MALDI-TOF. *Nucleic Acids Research*, **32**(21), e167.

Schreiber, J., Wescoe, Z. L., Abu-Shumays, R., Vivian, J. T., Baatar, B., Karplus, K., and Akeson, M. (2013). Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(47), 18910–18915.

Schwartzman, A., Jaffe, A. E., Gavrilov, Y., and Meyer, C. A. (2011a). Multiple Testing of Local Maxima for Detection of Peaks in ChIP-Seq Data.

Schwartzman, A., Gavrilov, Y., and Adler, R. (2011b). Multiple Testing of Local Maxima for Detection of Unimodal Peaks in 1D. *Harvard University Biostatistics Working Paper Series*, page 131.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, **442**(7104), 772–778.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. *Molecular cell*, **48**(6), 849–862.

Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., and Reik, W. (2013). Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **368**(1609), 20110330.

Shapiro, R., Servis, R. E., and Welcher, M. (1970). Reactions of uracil and cytosine derivatives with sodium bisulfite. *Journal of the American Chemical Society*, **92**(2), 422–424.

Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T. A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M., and Koseki, H. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, **450**(7171), 908–912.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, **29**(1), 308–311.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010). Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Research*, **20**(7), 883–889.

Siegmund, K. D., Marjoram, P., Woo, Y.-J., Tavaré, S., and Shibata, D. (2009). Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(12), 4828–4833.

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, **11**(8), 817–820.

Smyth, G. (2005). *Limma: linear models for microarray data in: Bioinformatics and computational Biology Solutions.* New York: Springer.

Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L., and Baccarelli, A. A. (2013). A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics*, **29**(22), 2884–2891.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, **15**(1), 72–101.

Speed, T. P. (2005). Iterative proportional fitting. *Encyclopedia of Biostatistics*.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K., and Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**(7378), 490–495.

Stevens, M., Cheng, J. B., Li, D., Xie, M., Hong, C., Maire, C. L., Ligon, K. L., Hirst, M., Marra, M. A., Costello, J. F., and Wang, T. (2013). Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Research*, **23**(9), 1541–1553.

Stirzaker, C., Taberlay, P. C., Statham, A. L., and Clark, S. J. (2014). Mining cancer methylomes: prospects and challenges. *Trends in genetics : TIG*, **30**(2), 75–84.

Stockwell, P. A., Chatterjee, A., Rodger, E. J., and Morison, I. M. (2014). DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics*, **30**(13), 1814–1822.

Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **69**(3), 347–368.

Stouffer, S. A. (1949). The American Soldier: Adjustment during Army life.

Stricker, S. H., Feber, A., Engström, P. G., Carén, H., Kurian, K. M., Takashima, Y., Watts, C., Way, M., Dirks, P., Bertone, P., Smith, A., Beck, S., and Pollard, S. M. (2013). Widespread resetting of DNA methylation in glioblastoma-initiating cells suppresses malignant cellular behavior in a lineage-dependent manner. *Genes & Development*, **27**(6), 654–669.

Su, J., Yan, H., Wei, Y., Liu, H., Liu, H., Wang, F., Lv, J., Wu, Q., and Zhang, Y. (2013). CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data. *Nucleic Acids Research*, **41**(1), e4–e4.

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A., and Li, W. (2014). MOABS: model based analysis of bisulfite sequencing data. *Genome Biology*, **15**(2), R38.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**(5929), 930–935.

Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(6), 3740–3745.

Tange, O. (2011). *GNU Parallel—the command-line power tool*. The USENIX Magazine.

Teschendorff, A. E. and Widschwendter, M. (2012). Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions. *Bioinformatics*, **28**(11), 1487–1494.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**(1), 36–46.

Tycko, B. (2010). Allele-specific DNA methylation: beyond imprinting. *Human Molecular Genetics*, **19**(R2), R210–20.

van Eijk, K. R., de Jong, S., Boks, M. P. M., Langeveld, T., Colas, F., Veldink, J. H., de Kovel, C. G. F., Janson, E., Strengman, E., Langfelder, P., Kahn, R. S., van den Berg, L. H., Horvath, S., and Ophoff, R. A. (2012). Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, **13**, 636.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine,

A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, **291**(5507), 1304–1351.

Vischer, E., Zamenhof, S., and Chargaff, E. (1949). Microbial nucleic acids; the desoxypentose nucleic acids of avian tubercle bacilli and yeast. *The Journal of biological chemistry*, **177**(1), 429–438.

Waddington, C. H. (2012). The epigenotype. *International journal of epidemiology*, **41**(1), 10–13.

Wang, H.-Q., Tuominen, L. K., and Tsai, C.-J. (2011). SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, **27**(2), 225–231.

Warnecke, P. M., Stirzaker, C., Melki, J. R., Millar, D. S., Paul, C. L., and Clark, S. J. (1997). Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Research*, **25**(21), 4422–4426.

Warnecke, P. M., Stirzaker, C., Song, J., Grunau, C., Melki, J. R., and Clark, S. J. (2002). Identification and resolution of artifacts in bisulfite sequencing. *Methods*, **27**(2), 101–107.

Warton, D. I. and Hui, F. K. C. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, **92**(1), 3–10.

Weigel, D. and Colot, V. (2012). Epialleles in plant evolution. *Genome Biology*, **13**(10), 249.

Wickham, H. (2014). *Advanced R*. Chapman & Hall / CRC The R Series. Taylor & Francis.

Wu, H. and Zhang, Y. (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell*, **156**(1-2), 45–68.

Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A., and Feinberg, A. P. (2010). Redefining CpG islands using hidden Markov models. *Biostatistics (Oxford, England)*, **11**(3), 499–514.

Wu, H., Wu, X., Shen, L., and Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nature Biotechnology*, **32**(12), 1231–1240.

Xi, Y. and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, **10**, 232.

Xie, H., Wang, M., de Andrade, A., Bonaldo, M. d. F., Galat, V., Arndt, K., Rajaram, V., Goldman, S., Tomita, T., and Soares, M. B. (2011). Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Research*, **39**(10), 4099–4108.

Xie, Q., Liu, Q., Mao, F., Cai, W., Wu, H., You, M., Wang, Z., Chen, B., Sun, Z. S., and Wu, J. (2014). A Bayesian Framework to Identify Methylcytosines from High-Throughput Bisulfite Sequencing Data. *PLoS Computational Biology*, **10**(9), e1003853.

Xu, H., Podolsky, R. H., Ryu, D., Wang, X., Su, S., Shi, H., and George, V. (2013). A method to detect differentially methylated loci with next-generation sequencing. *Genetic Epidemiology*, **37**(4), 377–382.

Yu, M., Hon, G. C., Szulwach, K. E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., Min, J.-H., Jin, P., Ren, B., and He, C. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell*, **149**(6), 1368–1380.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology*, **24**(8), 1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology*, **22**(2), 170–185.

Zhang, D., Cheng, L., Badner, J. A., Chen, C., Chen, Q., Luo, W., Craig, D. W., Redman, M., Gershon, E. S., and Liu, C. (2010). Genetic control of individual differences in gene-specific methylation in human brain. *American Journal of Human Genetics*, **86**(3), 411–419.

Zhang, Y., Liu, H., Lv, J., Xiao, X., Zhu, J., Liu, X., Su, J., Li, X., Wu, Q., Wang, F., and Cui, Y. (2011). QDMR: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Research*, **39**(9), e58.

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**(7463), 477–481.

Zou, J., Lippert, C., Heckerman, D., Aryee, M., and Listgarten, J. (2014). Epigenome-wide association studies without the need for cell-type composition. *Nature Methods*, **11**(3), 309–311.

Author/s:
HICKEY, PETER

Title:
The statistical analysis of high-throughput assays for studying DNA methylation

Date:
2015

Persistent Link:
http://hdl.handle.net/11343/55699

File Description:
The statistical analysis of high-throughput assays for studying DNA methylation