Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Bioinformatics of DNA Methylation analysis

Kemal Akman
aus
Istanbul, Türkei

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Achim Tresch betreut und von Herrn Prof. Dr. Klaus Förstemann von der Fakultät für Chemie und Pharmazie vertreten.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 1.04.2014

....

Unterschrift des Autors

Dissertation eingereicht am: 01.04.2014

1. Gutachter: Prof. Dr. Klaus Förstemann

2. Gutachter: Prof. Dr. Achim Tresch

Mündliche Prüfung am: 19.05.2014

Summary

Methylation of cytosines in genomic DNA is a major epigenetic factor affecting gene expression and genomic stability [1]. Novel methods to analyse changes in DNA methylation in individual cells were developed. Both stochastic and programmed epigenetic changes may affect functionality throughout development, disease and aging. To further the understanding of changes in DNA methylation in human aging and disease, which are termed 'epimutations' throughout this thesis, a literature review on DNA methylation in aging and disease was conducted [2]. The review focused on mechanisms linking changes in DNA methylation to aging, and on potential therapeutic interventions.

DNA methylation status is usually obtained from pooled cells, which limits the understanding of cell-to-cell epigenomic variability and the significance of DNA methylation changes which may arise from individual cells. Single-cell BS-Seq allows for the analysis of functional states of individual cells by looking at cell-specific epigenomic changes [3]. Being able to identify cell-specific epigenomic variation helps elucidate epigenomic changes in aging and degenerative diseases which may originate from a single cell such as cancer [4]. As of today, only one study is known to the author that uses single-cell Reduced Representation BS-Seq (scRRBS) experiments [4] besides the yet unpublished single-cell Reduced Representation Bisulfite Sequencing (scRRBS) data analysed in this [5]. A novel methodology in Bioinformatics and statistic modeling for the analysis of DNA methylation changes in novel single cell BS-Seq experiments was developed in this thesis. By comparing genome-wide DNA methylation patterns of single mouse hepatocytes with total genomic liver DNA, genome-wide epimutation rates are established, as well as epimutation rates specific to features such as genes, intergenic regions, CpG islands and repeats.

A Bayesian Binomial Dirichlet (BD) mixture model model was derived for the robust estimation of CpG methylation levels, which gives detailed information about the methylation rate in a region of multiple CpG positions. It provides measures of confidence for this estimate and is able to test regions for high or low methylation status, which facilitates the easy definition of regional epimutation events. As part of this thesis, the R/Bioconductor package BEAT (BS-Seq Epimutation Analysis Toolkit), a novel tool for analyzing bisulfite-converted DNA sequences, was written and published [6]. It implements the BD mixture model and can aggregate consecutive cytosines into genomic regions in order to overcome limitations of experimental data of low genomic coverage. For each region, it calculates a posterior methylation probability distribution which can be used for the comparison of DNA methylation between samples. To the author's knowledge, it is the first tool providing a rigid statistical model for handling BS-Seq samples. It has an in-built correction for conversion errors and is therefore perfectly suited for handling heterogeneous BS-Seq experiments.

Contents

Acknowledgments

My gratitude goes to my advisor, Prof. Dr. Achim Tresch, who patiently provided the advice, practical support and encouragement for me to proceed through the process of completing my dissertation. I owe him my heartfelt appreciation for his guidance. Special thanks go to our collaboration partners, Silvia Gravina, PhD and Jan Vijg, PhD in New York, who provided the novel type of DNA methylation data that enabled me to work on my research questions of providing an analysis for a novel experimental protocol as well as focusing on research related to human aging, as well as Aubrey David Nicholas Jasper de Grey, PhD, who provided the vision and initial support for starting the important project of establishing a single-cell BS-Seq protocol to establish a technology necessary for understanding age- and disease-related changes in the epigenome. I also wish to thank my my family, especially my mother Angelika Akman for her moral support and my partner, Adriana Halina Weilandt, for her encouragement.

1.1 Introduction

1.1.1 Biological Background

1.1.1.1 DNA methylation and chromatin structure

DNA methylation plays a major role in the modulation of gene regulation for many organisms. In genomic DNA, cytosine bases followed by guanine (CpG) may carry a methyl group, which influences transcriptional activity and gene expression. DNA methylation encompasses the state of methylated CpG sites (5mC) as well as the modification of this methylation status is described as DNA methylation in the literature. In addition, non-CpG DNA methylation, where cytosines not following a guanine are methylated, also exists, but is rare in mammals, except for embryonic stem cells[7].

At the DNA level, gene expression is regulated by two basic mechanisms: DNA methylation and modification of chromatin. Patterns of DNA methylation and chromatin status are known to influence each other. For instance, certain proteins, such as the protein MeCP2, bind to CpG islands (CGIs) in a methylation-sensitive fashion. The interaction of MeCP2 with the chromatin-modifying Sin3/histone deacetylase co-repressor complex is an exemplary evidence of the direct link and influence between DNA methylation and chromatin modifications[8].

Consecutive regions that are rich in CpG sites, while being mostly unmethylated, are called CpG islands. These islands tend to recruit many methylation-sensitive transcription factors and appear to be protected from accumulating novel methylation by DNA methyltransferases. CpG islands make up about 1% of the genome and are distributed in a non-random fashion. Most often, they are found near the promoters of housekeeping genes[9].

CpG methylation is pervasive across the genome, with approximately 80% of mammalian genomes being methylated on average[10]. For the human genome, the overall methylation percentage of CpG sites in the human genome is estimated to be 68.4%, with the majority of the CpG islands being unmethylated [11].

The basic functions of DNA methylation and the associated chromatin modifications are the epigenetic regulation of gene expression and the transcriptional silencing of heterochromatin. Epigenetic changes are stable DNA modifications that are (partly) inherited in a Lamarckian (non-Mendelian) manner. These changes have important functions in development and reaction to environmental stimuli, as they enable cell differentiation by allowing cells to have an individual transcriptional profile. However, epigenetic changes may also arise due to external influences that cause pathologies and they accumulate stochastically with age, which is associated with several degenerative diseases, such as cancer [10].

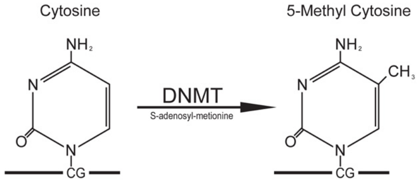

Figure 1.1: Taken from [12], this figure shows the difference between genomic cytosines and methylated cytosines (5mc). Methylation of genomic cytosines happens by enzymes known as DNA methyltransferases (DNMTs).

1.1.1.2 DNA methylation and transcription

CpG sites are sites of transcription initiation, including but not limited to those located at or near known promoters. Their methylation status regulates gene activity by the recruitment of methylation-sensitive transcription factors. In general, low methylation in CGIs destabilizes nucleosomes of the chromatin and attracts transcription factors and other proteins that help induce transcription of genes associated with the respective CGIs. Correspondingly, methylation of CGI sites, especially at gene promoters, leads to transcriptional silencing and is mediated by polycomb genes[13]. Another important function of the pervasive silencing of genomic regions by DNA methylation is the transcriptional silencing of human endogenous retroviruses (HERVs) and other self-replicating repetitive DNA elements, which makes transcriptional silencing by DNA methylation also necessary for genome stability [14] [1]. When CpG islands do accumulate methylation due to programmed or random processes, this often leads to gene silencing, a process which often can also be observed in tumor cells [9]. In addition, DNA methylation of intragenic regions, i.e. within exons and introns can help reduce transcriptional noise, while promoter methylation may be correlated with increased transcriptional noise [15].

1.1.1.3 Role of DNA methylation in development

DNA methylation is essential for the proper regulation of gene expression during mammalian embryonic development in addition to its role in differentiation of adult tissues [16]. In mammals, active changes in DNA methylation on a genome-wide scale occur only during early development of the embryo. In mice, the paternal genome is actively demethylated in the zygote early after fertilization, and then passively further loses methylation during cleavage. Later, it incurs de novo methylation in the inner cell mass (ICM) of the embryo, which is thought to occur after implantation of the zygote [17]. Long-term silencing of gene expression in a cell-specific manner which enables cell- and tissue differentiation during and after development [10]. It is known to play additional roles in development by inactivating one of the X-chromosomes in female offspring, as well as being an underlying mechanism of genomic imprinting. Imprinting is the monoallelic expression of a few genes in the genome, resulting in either exclusive maternal or exclusive paternal gene expression [18], which allows for inheritance of gene-specific traits in a parent-of-origin specific manner and occurs, for example, in the genes H19, Igf-2 and Igf-2r, which are related to insulin signaling. DNA methylation is the underlying mechanism that causes the permanent silencing of one of the alleles of imprinted genes during early embryonic development. It has been demonstrated that normal levels of DNA methylation at imprinted genes are necessary for proper differential expression of maternal- and paternal alleles [19].

1.1.1.4 Role of DNA methylation in aging and disease

Mammalian aging is associated with both increases and decreases in DNA methylation. Age-associated changes in DNA methylation in or near gene regulatory elements can suppress gene expression by affecting chromatin structure and DNA binding proteins and may carry pathological consequences, including, but not limited to neurodegeneration, cancer, osteoarthritis, sensory neuropathies or autoimmune diseases [2][20]. In this thesis, both changes in DNA methylation associated with aging and degenerative processes are termed ’epimutations’, while increases in methylation due to epimutations are denoted as ’methylating epimutations’ and corresponding decreases are denoted as ’demethylating epimutations’. In other recent literature, epimutations have been referred to as differentially methylated CpG dinucleotides (dmCpG)[21]. The frequency of epimutations may vary by tissue, genes or other genomic regions affected [20].

The most common known patterns of age-associated changes in DNA methylation are genome-wide hypomethylation and promoter-specific hypermethylation. For example, methylating epimutations in the promoter regions of tumor suppressor genes may silence such genes, leading to an increased tendency to develop malignancies. Epimutations represent potential biomarkers for diseases and degenerative age-related processes leading to disease [2]. Novel methods and models for the detection of epimutations and establishment of region-specific epimutation rates, as those established in this thesis, may therefore be helpful for establishing diagnostics related to identifying biomarkers associated with aging and disease. Some experimental drugs that can modify DNA methylation patterns exist, the most well-established of which is 5-azacytidine (5-azaC), which reduces DNA methylation, and has shown promising therapeutic efficiency in clinical trials for various cancers [2]. In addition, age- and diease-related changes of the maintenance DNA methyltransferase (DNMT) DNMT1, as well as of the de-novo DNA methyltransferases DNMT3A and DNMT3B, could potentially play a role in aging and disease [2].

Age-related changes at CpG sites have been shown to occur at clusters of correlated CpG markers whose sequence is related to targets of polycomb group genes. Functionally, several of the genes associated with aging-related changes in DNA methylation are related to neuronal development, -differentiation and neurogenesis [22]. The epigenetic regulation of gene expression has been found to be crucial for brain function, especially in the mammalian central nervous system, and pathological changes of DNA methylation are associated with a variety of neurodegenerative diseases and age-associated neurodegeneration[23]. For example, some genes identified in [22] are also down-regulated in Alzheimer's disease, implying a correlation of age-related DNA methylation changes in the disease. An age-related decline in skeletal muscle function has also been associated with age-related epimutations[21].

An association between genome-wide hypermethylation and age has been found in a study on human male aging [21]. In this study, epimutations were mostly observed within gene bodies related to neuronal function, and underrepresented in promoters, while being overrepresented in the middle and at the 3' end of genes. Additionally, low negative correlations between the gene expression of aged muscle cells and DNA methylation were found in general, but genes that showed intragenic hypermethylation with age were associated with increased gene expression. The study also suggested that a minimum threshold of epimutations were required in order for changes in gene expression to take place [21].

1.1.2 Technical Background

1.1.2.1 Biochemistry of Bisulfite conversion

Bisulfite conversion of DNA is a reproducible method for detecting and quantifying the methylation status of single cytosine nucleotides when used in combination with high-throughput re-sequencing, the combination of which is called Bisulfite sequencing (BS-Seq). An accurate treatment of genomic DNA with bisulfite, as described in [24] for mammalian cells, leads to deamination of unmethylated cytosines in the genome into uracil nucleotides, which are later converted into thymine residues in further BS-seq steps. In more detail, this chemical deamination reaction can be described as pH-dependent partial depurination [25]. When this reaction is allowed to continue for too long, it may degrade intact DNA [25].

Some of the most crucial parameters for achieving a correct bisulfite conversion reaction include keeping the correct acidic pH, the amount and quality of the starting DNA before amplification, as well as the bisulfite treatment time. Excessive bisulfite treatment can degrade DNA, which leaves less material to be amplified and sequenced and can therefore reduce sequence quality and genomic coverage. On the other hand, too short bisulfite treatment may result in incomplete conversion of unmethylated cytosines to thymine, which then results in unmethylated cytosines falsely appearing to be methylated on a single-nucleotide level, as well as detecting erroneously higher methylation rates in such samples [24] [26].

1.1.2.2 Sequencing of bisulfite converted reads

Following bisulfite conversion, the converted DNA is then amplified by PCR using strand-specific primers. Stand-specific PCR products are then either re-sequenced by pyrosequencing or other high-throughput-sequencing methods, either directly or after cloning. The complete process is called bisulfite sequencing (BS-Seq). Finally, the converted amplified sequence can then be compared to an existing reference genome. The methylation status of re-sequenced DNA can then be determined by counting thymine residues in the converted DNA that match cytosines in the reference genome as unmethylated cytosines, while counting matching cytosines as methylated cytosines [24] [27]. The detailed steps of bisulfite treatment prior to sequencing consist of alkaline denaturation of initial DNA material, followed by bisulfite conversion, which also referred to as deamination in the literature, and chemical desulfonation, neutralization and desalting of bisulfite-converted DNA, which is then buffered and stored at -20°C. Later, the stored DNA material is then amplified using PCR and cloned into appropriate vectors in preparation for high-throughput DNA sequencing [28].

Several optimizations of BS-Seq method have been performed. One of them consists of performing bisulfite conversion and PCR on agarose-embedded material to minimize the loss of DNA and to keep it in single-stranded form, which empirically leads to better bisulfite conversion results[29]. Bisulfite converted DNA can also be combined with general variants of sequencing, such as high-resolution melting (HRM), where the melting of PCR products with fluorescent nucleotides is monitored in real-time to detect single-base changes. Methylation-sensitive HRM (MS-HRM) is an application of this method to bisulfite converted DNA with a high accuracy of predicting methylation levels [30]. However, specific DNA sequencing methods come with their own issues, such as issues in differentiating the melting curves of similar primers in the case of HRM, which limits the applicability of specific variants

[31]. Several further optimizations of BS-Seq protocols, such as in primer design and cloning have also been used in order to improve the accuracy of the method in general [32]. Another variant for BS-Seq with comparable accuracy as the sequencing of agarose- or paraffin-embedded DNA, is the use of degenerate primers which fit to both cytosine and thymine residues in bisulfite treated DNA, where a cytosine is expected in the genome, in combination with pyrosequencing [33].

In addition to BS-Seq variants, alternative sequencing methods for assessing genome-wide methylation levels exist, which include MBD-Seq and MeDIP-Seq[34][35][30]. MeDIP-Seq avoids bisulfite treatment by immunoprecipitation with 5mC-sensitive antibodies and resulting fragments are sequenced, after which methylation levels can be estimated by the amount of fragments sequenced per region. Its advantage lies in a higher sensitivity to highly methylated CpG-rich sites, which may however also induce biases when sequencing larger regions or genomes[34]. Similarly, MBD-seq achieves enrichment of methylated DNA fragments by precipitating DNA with bead-immobilized recombinant versions of 5mC-binding proteins, typically MeCP2 or MBD2. MBD-Seq has drawbacks comparable to MeDIP-Seq and is most sensitive for highly methylated sites with moderate CpG densities and shares [35].

Finally, reduced-representation BS-Seq (RRBS) currently represents the most important variant of BS-Seq for experiments with small initial amounts of DNA, such as single-cell BS-Seq. In general, RRBS is recommended when the amount of sample DNA and/or replicates is limited. RRBS enriches the sequencing material for those parts of the genome that are rich in CpG sites from sample genomic DNA, which greatly improves coverage those parts even with small amounts of initial gDNA available (10 - 300 ng). By focusing on CpG-rich regions, most of the features of interest such as promoters, genes and short repeats, will still be covered by sequencing. The Reduced Representation is achieved by digesting the initial DNA material with an restriction endonuclease that cuts for a sequence specific to CpG-rich parts of the genome. Restriction fragments resulting from the most widely used MspI digest, which are typically 40-220 nt long, are then end-repaired, polyadenylated and ligated to adapters for Illumina sequencing. Other restriction enzymes that enrich for CpG-rich parts of the genome have been used, such as MseI or BglII, the latter resulting in 500-600 nt long fragments. Following the procedure, fragments are then subjected to size selection on a gel, treated with bisulfite and subjected to PCR amplification, after which they are ready for cloning and sequencing [26][36]. RRBS can also be combined with the embedding of sample material in paraffin or other material in order to further improve conversion results as mentioned above[29]. Given sufficient DNA material, RRBS can achieve a nearly complete bisulfite conversion rate (>99.9% conversion) [36].

1.1.2.3 Mapping of bisulfite converted reads

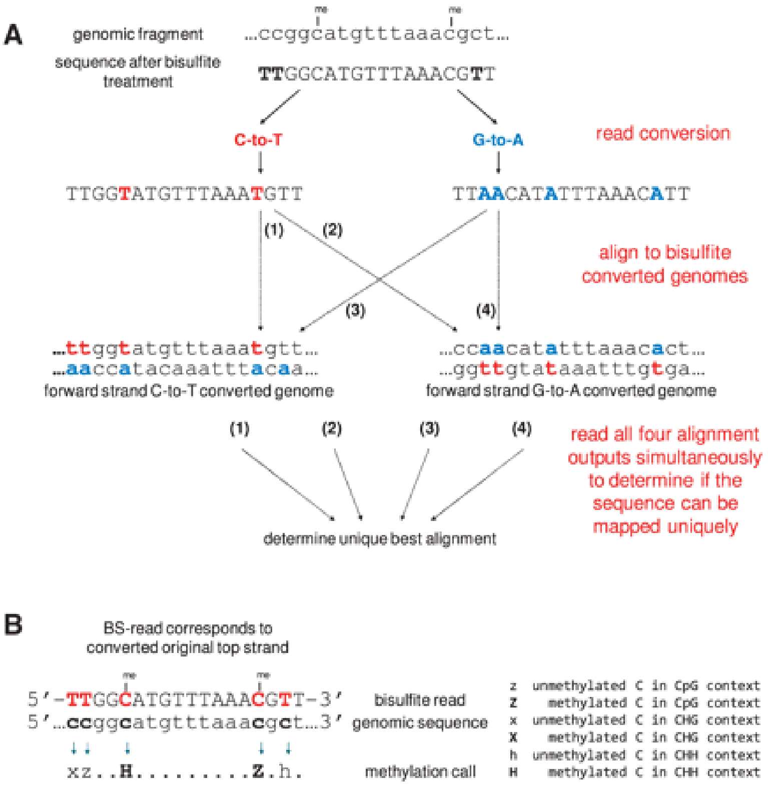
A
genomic fragment
...ccggcatgtttaaacgct...
me
me
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
(3)
(4)
align to bisulfite converted genomes
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment
B
BS-read corresponds to converted original top strand
me
me
5'-TTGGCATGTTTAAACGTT-3'
bisulfite read
5'...ccggcatgtttaaacgct...3'
genomic sequence
xz..H.........Z.h.
methylation call
z unmethylated C in CpG context
Z methylated C in CpG context
x unmethylated C in CHG context
X methylated C in CHG context
h unmethylated C in CHH context
H methylated C in CHH context

Figure 1.2: Taken from [37], this figure shows the mapping strategy of the alignment tool Bismark which was used to align all BS-Seq data in this thesis. In (A), bisulfite treated DNA sequence reads match their corresponding original genomic sequence with C->T mismatches at genomic cytosine positions. Bismark computationally precomputes C->T conversions on the reference genome's forward strand and G->A conversion for the complementary bases on the reverse strand. Bisulfite-converted reads are converted analogously and aligned to both converted genomic strands, after which the unique best alignment of a bisulfite-converted read against the converted genome is accepted if one exists, in this case on the forward strand. Subfigure (B) further illustrates the methylation calling process from reads successfully aligned to the genome, where C/C matches count as methylated cytosines, while C->T mismatches of genomic sequence to mapped reads count as unmethylated cytosines, both in CpG and non-CpG (CHG, CHH) contexts.

The process of aligning short reads to a reference genome is called mapping. In modern high-throughput sequencing such as Illumina-based pyrosequencing, some sequencing errors occur and must be permitted by the mapping tool. Empirically, sequence quality of Illumina reads deteriorates with length. For short reads in general, the mapping tool Bowtie is one of the fastest and most memory-efficient aligners for mapping high-throughput reads available. It uses a hash indexing algorithm based on the Burrows-Wheeler transformation, which consists of hashing the first few bases of each short read into a hash value for fast matching, referred to as seed. For each reference genome available, a Burrows-Wheeler index of all possible seeds is precomputed and the initial assignment of a short read to a genomic position is done via matching seeds of the short read with the genome index. This allows for aligning over 25 million reads per CPU hour. In addition, Bowtie's modification of the Burrows-Wheeler algorithm allows for mismatches in the seed region which allows for mapping reads with high sensitivity despite fast processing speeds [38].

Bisulfite converted reads can map to either forward- or reverse strands of a reference genome, as well as to the reverse complement of either strand, respectively. When a read maps to the reverse complement of a genomic strand, unmethylated cytosines are indicated by G-to-A-conversion instead of C-to-T conversion. All four mapping options need to be considered. Also, the matching of reads to the reference genome must allow for mismatches due to the bisulfite conversion[37]. BS-Seq mapping tools are aware of these issues and will produce counts of methylated- and unmethylated cytosines per genomic position covered by reads, in addition to the alignment. These include Bismark, BSMAP, RMAPBS and BS Seeker [11].

Bismark is a perl application that performs short read mapping and methylation calling as a front-end on top of Bowtie, therefore sharing its speed and efficiency advantages. Bismark processes both strands of the reference genome to generate CT- and GA-converted files for the forward and reverse strands and their respective reverse complements, as explained above. Mapping of a read to each of these 4 strands is done with parallel invocations of Bowtie. Methylation calling is performed with Bismark's methylation_extractor script, which can discriminate between CpG methylation and non-CpG methylation [37].

BSMAP is based on SOAP, a different NGS mapping tool, and uses bitwise masking and hashing based to generate seeds of 8 bits in length that allow for flexible C/T and G/A matching in converted reads. BSMAP uses bitwise AND matching between C (01) and T (11) and allows a user-specified degree of mismatches in order to increase sensitivity and flexibility[39].

RMAPBS is a mapper designed exclusively for single-end BS-Seq reads which uses an algorithm similar to that of BSMAP, but is more tolerant in case of sequence mismatches, as it allows for partial read matches of whole reads[11].

BS Seeker, which also uses Bowtie for mapping, converts the genome to a three-letter alphabet by changing all C to T. It also uses sequence tags specific to certain restriction enzymes, combined with post-processing in order to reduce mapping ambiguity. The authors of BS Seeker have demonstrated a higher speed and versatility in comparison to other tools such as BSMAP[40].

A recent comparison of the BS-Seq mapping tools Bismark, BSMAP and RMAPBS has shown that they all show comparable and expected genome-wide methylation levels, as well as comparable mapping efficiencies of approximately 50-65% mappable reads in case of standard BS-Seq experiments of high quality. In this review, Bismark was preferred as the mapping tool of choice since it was the fastest tools and due to additional criteria including overall efficiency and ease of interfacing with other software and of the extraction of methylation data [11].

1.1.3 Goals of this thesis

In this thesis, I developed a Biostatistics and Bioinformatics model and -pipeline for estimating methylation level and status of genomic regions from BS-Seq DNA methylation data. The novel and specific aim of the pipeline is to generate robust, position-specific methylation estimates and to provide a statistical method for detecting epimutations from BS-Seq data even in cases where high levels of false positive and false negative errors can be expected. This may be the case wherever low amounts of initial DNA are available, such as in the newly-emerging single-cell BS-Seq experiments, which require gentle bisulfite treatment to yield results, but in turn suffer from low conversion rates with an associated increase in error rates, which is an issue that has not been properly addressed by existing research so far. The practical use of the model and pipeline is demonstrated by the analysis of RRBS single-cell methylomes as a novel type of experiment and the determination of the rate of epimutations, i.e. the frequency of differences in methylation status between two BS-Seq samples.

1.2 DNA Methylation in Aging and Disease

In preparation of my research related to acquiring a better understanding about DNA methylation changes and how they relate to the human aging process, I co-authored a review on DNA methylation in aging and disease [2], which focused on the mechanisms linking changes in DNA methylation to aging, as well as potential interventions. Changes in DNA methylation play a major role in aging and age-related diseases, as initially mentioned in sec. 1.1.1.4. DNA hypomethylation and hypermethylation have been observed in aging, and their pivotal roles in age-related changes to the epigenome has been established [41]. Aging and age-related diseases include changes in 5-methylcytosine content and are generally characterized by genome-wide hypomethylation and promoter-specific hypermethylation. These changes in the epigenetic landscape represent potential disease biomarkers and are thought to contribute to age-related pathologies. However, epigenetic modifications are reversible and are therefore a prime target for therapeutic intervention. My paper concludes that while many interventions are possible, the most important question remains whether stochastic age-associated changes in mammalian DNA methylation do actively contribute to age-related diseases. This is a key question that the novel single-cell RRBS experiments and appropriate statistical analysis of their results, as presented in this thesis, help to address.

1.2.1 Mechanisms of age-related DNA methylation changes

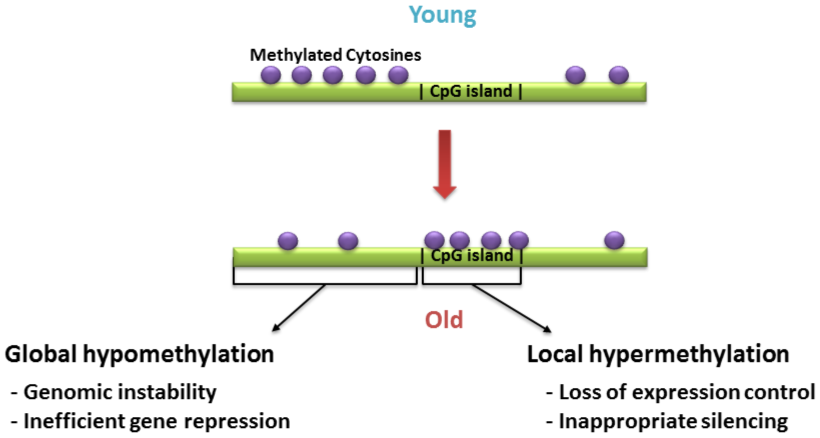

Figure 1.3: This figure, taken from [2], shows age-related changes in DNA methylation as they are assumed to occur in current scientific literature: while the genome tends to globally lose methylated cytosines with increasing age, which may lead to genomic instability, some regions rich in CpG islands tend to accumulate newly methylated cytosines, which may lead to promoter hypermethylation and consequently, inappropriate silencing of genes.

DNMTs, demethylases, and associated partners are dynamically shaping the methylome, and their activity is changing with age, which may have an impact on DNA methylation changes with age and therefore, also on age-related disease. Multiple DNA-methylation-mediated silencing pathways for DNMTs exist, which includes the Methyl-CpG binding proteins (MeCP) [42], whose expression also changes with age. Multiple prospective mechanisms for active demethylation have been discussed, which may have an additional impact [43]. DNMT activity has also been shown to be regulated by the recently discovered hydroxymethylation by the ten eleven translocation (TET) family of enzymes [44]. Regulation of DNMTs by protein methylation-, -phosphorylation, -acetylation, -SUMOylation and ubiquitination [45] [46].

1.2.2 Associations of DNA methylation with aging and disease

Epigenetic changes have been demonstrated in a large number of CpG sites and a diverse range of genes. Some of these changes, especially those in highly-affected genes, have been able to predict the age of an individual with an average error of approximately 5 years [47]. It shows that epimutations are liked to age-related disease by examples such as aberrant methylation in cancer[48], where hypermethylation of promoter regions of oncogenes can most often be observed, as well as in Alzheimer's disease, Type-2-Diabetes and renal disease [49].

1.2.3 Prospective preventive and therapeutic interventions

Regarding the mitigation of age-related pathology, my paper makes a case for caloric restriction (CR) for reducing age-related epimutations, and proposes a modulation of DNMT activity as a possible beneficial mechanism of CR in aging [50]. Today, therapies which focus on DNA methylation changes as mechanism are still very few, but some prospective drugs exist. Several drugs that specifically target DNMTs are being tested in ongoing clinical trials for a variety of cancers. My paper discusses some potential therapeutic interventions to age-related disease that may work by modifying DNA methylation in the genome nucleoside analogs, which include 5-azaC, zebularine and decitabine. These drugs have been shown to be able to reactivate tumor suppressor genes in vitro in some types of cancer [51]. Other drugs such as procain, procainamide and hydralazine, which are known to impact DNA methylation in a way that has been linked to treating some cancers, appear to inhibit DNMTs. However, their mechanism is not yet fully understood [46]. Methyl donors such as folic acid or S-adenosylmethionine, which are natural co-factors for DNA methylation in human biochemistry, might offer gentle treatment opportunities in diseases related to age-related changes in DNA methylation [52].

1.3 Single-Cell BS-Seq Analysis for Epimutation Rate Estimation

Single-cell sequencing technologies constitute a new trend that allows for new genome-related questions to be answered. It has potential to revolutionize genomics. Single-cell DNA sequencing may uncover cell lineage relationships, while single-cell transcriptomics may identify individual variation in gene expression. Single-cell BS-Seq, as used for the experiments analysed throughout this thesis, for the first time allows for the analysis of functional states of individual cells by looking at cell-specific epigenomic changes [3]. Being able to identify cell-specific epigenomic variation is especially important in the context of aging and degenerative diseases which may originate from a single cell such as cancer [2][4], where it is important to make a difference between programmed- vs. stochastic changes in DNA methylation and thus, in gene expression.

As of today, only one study is known to the author that uses single-cell Reduced Representation BS-Seq (scRRBS) experiments [4], except for the yet unpublished scRRBS data analysed in this thesis [5]. While the mentioned study is limited to sperm and analyses the rate of demethylation, the single cells analysed in this thesis are mouse liver cells, for which region-based methylation levels are modeled using a Bayesian Binomial-Dirichlet model, and both demethylating- and methylating epimutation rates are determined from these model estimates. As discussed in [2], epimutations are associated with a diverse bandwidth of physiological effects relevant to many diseases, of which many may originate from individual cells.

A high load of epimutations in an individual cell could seriously impedact cell function, but such epimutations would not show up in a whole-tissue analysis because pooling individually different cells would hide many differences, as elucidated in Fig. 1.4. Hence, only single-cell BS-Seq can uncover the true epimutation load. This was the rationale and motivation for the following single cell BS-Seq experiments and their analysis using Bioinformatics and statistics as described below.

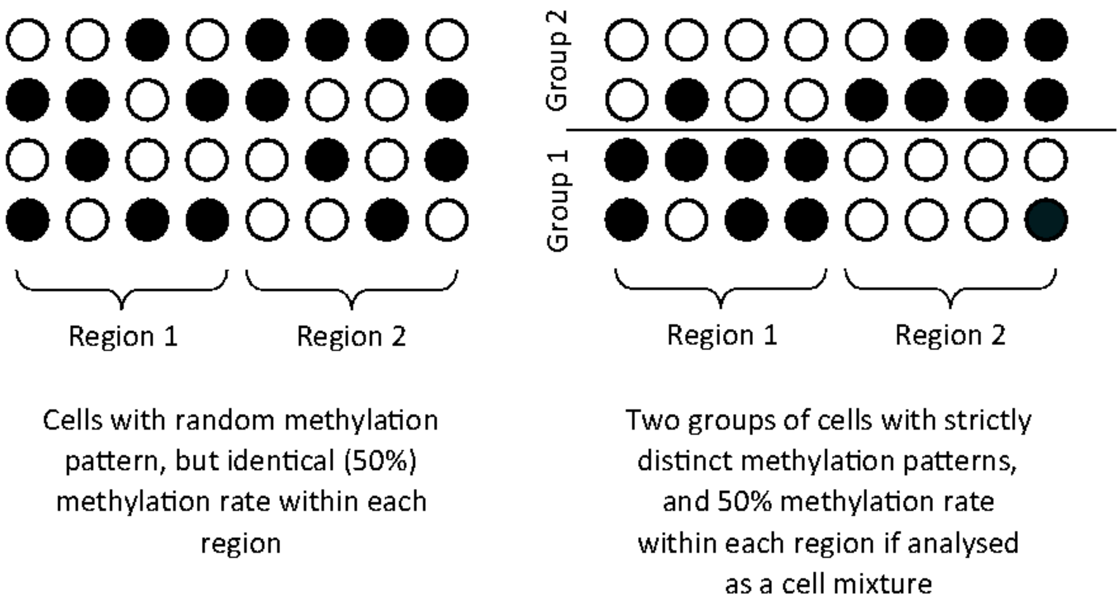
Region 1
Region 2
Cells with random methylation pattern, but identical (50%) methylation rate within each region
Group 2
Group 1
Region 1
Region 2
Two groups of cells with strictly distinct methylation patterns, and 50% methylation rate within each region if analysed as a cell mixture

Figure 1.4: Schematic view of methylation patterns in two genomic regions. Black dots represent methylated regions while white dots represent unmethylated regions. One row represents a single cell. The left subfigure shows two regions with an equal methylation rate of 50% in both regions if cells are pooled. However, every individual cell shows vastly different methylation patterns for each region. The right subfigure shows the same for two groups of two cells each with similar methylation patterns. In both cases, the individual variation in DNA methylation patterns cannot be detected when all cells are pooled and sequenced together, but only from single-cell DNA methylation experiments.

1.3.1 Specific issues of single cell BS-Seq

The most critical issue in BS-Seq experiments in general is to be able obtain a complete bisulfite conversion of the sample DNA material, while avoiding erroneous conversion of methylated cytosine to uracil is also important. To achieve complete conversion, the two most critical parameters are incubation time and incubation temperature. While maximum bisulfite conversion occurs at either 95°C incubation temperature for a short incubation time of 1 hour, or 55°C for longer incubation of 4-18 hours, these conditions are harsh for DNA stability, as they lead to a degradation of 84-96% of available DNA [28]. The main issue with single-cell BS-Seq experiments, such as those presented in this thesis, is that the current standard BS-Seq protocols, which primarily aim for complete bisulfite conversion, have an unacceptably high rate of DNA degradation for single-cell- and other experiments where little initial DNA material is available per sample [29]. Other quality issues of BS-Seq are sensitivity and reproducibility of the method, which are an important factor due to the often small amounts of initial original DNA material available for analysis [28].

Single-cell analysis is necessary since DNA methylation information from pooled cells only represents average values which do not represent individual differences between cells. such cell-specific differences are relevant in the case of random epimutations in aging and diseases such as cancer, as well as in cell differentiation in embryonic and adult tissues [51]. The minimum amount of DNA required for previously existing BS-Seq protocols is 30 ng, which typically requires the use of populations of a few thousand cells, as single cells do not supply sufficient DNA material when using standard BS-Seq protocols [36]. Another issue affecting single-cell experiments more seriously is the excessive DNA fragmentation induced by the bisulfite treatment, which prevents analysis of a large portion of the treated DNA material[25]. In theory, false conversion of 5mC to uracil is possible with prolonged or too harsh treatment, however, the practical problem of prolonged bisulfite treatment is mostly destruction of DNA, resulting in less sample material available for sequencing [53].

1.3.2 A statistical model for BS-Seq data

1.3.2.1 Sources of bias and variation in BS-Seq

The amount of unmethylated cytosines which are not converted by bisulfite treatment is referred to as ’non-conversion-’ or ’false methylation’ rate in this thesis. This rate depends on the completeness of bisulfite conversion, which has been expected to be nearly complete in existing studies, but may be significantly lower for more gentle bisulfite conversion treatment, as is needed by the single-cell BS-Seq experiments presented here [26]. On the other hand, the erroneous conversion of 5mC residues to uracil is referred to in this thesis as ’inappropriate conversion’. Empirically, even under harsh conditions of a typical bisulfite treatment that leads to near-complete conversion, not more than 6% 5mC residues were detected as inappropriately converted, i.e. unmethylated, by a study that investigated this type of error[28]. As in genome sequencing in general, low genomic coverage, due to low mapping efficiency and/or due to insufficient DNA material, is also an issue in BS-Seq that can increase variance and decrease reliability of experiments, an issue which can be mitigated somewhat by replicates and modeling. Additionally, sequencing errors, which in properly performed experiments should occur at rates substantially below 10% in Illumina sequencing, can bias the data, when occurring at cytosine residues, increasing both fundamental error types, false methylation and inappropriate conversion [54]. When errors in BS-Seq occur that are not corrected by appropriate statistical modeling, empirical methylation levels even between sequencing runs of technical replicates have in some cases been found to vary by over 100%, which may make analysis of empirical BS-Seq data without modeling impossible in some cases [55].

1.3.2.2 A Bayesian Binomial-Dirichlet (BD) model for BS-Seq data

For analyzing CpG methylation, a Bayesian statistical model was derived which gives detailed information about the methylation rate in a region of multiple CpG positions which is described below. Apart from estimating the methylation rate, it provides measures of confidence for this estimate, it can test regions for high or low methylation. On the basis of of these tests it is later possible to give a precise definition of a regional epimutation event. A region-based approach was favored over single-position analysis because single-position methylation values inevitably display a large cell-to-cell variance of methylation at the single-position level, as methylation rates in cell mixtures vary widely, ranging between 10% and 90% [56]. This makes the identification of regional changes in methylation levels is functionally more relevant and statistically more robust. By comparing single position versus region-based epimutation calling, I was able to show a reduction in false positive and false negative methylation calling rates, as detailed under paragraph 1.3.2.5.4.

1.3.2.3 Methylation estimation and Epimutation calling

For multi-cell samples, it is assumed that all counts at a single CpG position were obtained from pairwise different bisulfite converted DNA template strands and represent independent observations. This certainly holds in good approximation, because the number of available DNA template strands typically supersedes the read coverage at this position by far. For single cell samples, the opposite situation is encountered: There are at most two template DNA strands available, and for many CpG positions this number is reduced further through DNA degradation. Multiple reads covering one CpG position are therefore highly dependent. Multiple counts at one position were combined to one single (non-)methylation call. For different CpG position, these calls are then independent observations. First, fix one region, i.e.

some set of CpG positions. The number of counts at a given position is the number of reads mapping to that position. Let n denote the total number of counts at all CpG positions in the given region, and let k (respectively $n-k$) of them indicate methylation (respectively non-methylation). Let r be the (unknown) methylation rate at the given position. Then, assuming independence of the single counts as mentioned above, the actual number j of counts originating from methylated CpGs in this region follows a binomial distribution,

$$P(j \mid n, r) = Bin(j; n, r) \tag{1.1}$$

Let the false positive rate p_+ be the global rate of false methylation counts, which is identical to the non-conversion rate of non-methylated cytosines. Conversely, let the false negative rate p_- be the global rate of false non-methylation counts, which is identical to the inappropriate conversion rate of methylated cytosines. One can find an upper bound for p_+ by considering all methylation counts at non-CpG positions as false positives (resulting from non-conversion of presumably unmethylated cytosines). Bear in mind that in single cell bisulfite experiments, the limited DNA amount requires a particularly mild bisulfite treatment, which increases the false positive rate relative to standard bisulfite sequencing procedures. In the literature, false negative rates were not described, an estimate of $p_- = 0.01$ is reported [26]. A conservative value of $p_- = 0.2$ was chosen, which takes into account potential errors originating from mapping artifacts or sequencing errors. Due to failed or inappropriate conversion, the number k of counts indicating methylation differs from the actual number j of counts originating from methylated CpGs. Given the true number of methylation counts j, the the observed methylation counts k are the sum of the number m of correctly identified methylations and the number $k-m$ if incorrectly identified methylations (false positives). Hence, the probability distribution of k is a convolution of two binomial distributions,

$$\begin{aligned} P(k \mid j, n; p_+, p_-) &= \sum_{m=0}^{k} P(m \mid j, 1-p_-) \cdot P(k-m \mid n-j, p_+) \\ &= \sum_{m=0}^{k} \underbrace{Bin(m; j, 1-p_-)}_{=:C^1_{m,j}} \cdot \underbrace{Bin(k-m; n-j, p_+)}_{=:C^2_{n-j,k-m}} \end{aligned} \tag{1.2}$$

In (1.2), by convention, $Bin(m; j, p) = 0$ whenever $m > j$. Thus, given n reads, k methylation counts, the likelihood function for r is a mixture of Bionomial distributions,

$$\begin{aligned}
P(k \mid n, r;\, p_+, p_-) &= \sum_{j=0}^{n} P(k, j \mid n, r, p_+, p_-) \\
&= \sum_{j=0}^{n} P(k \mid j, n, r, p_+, p_-) \cdot P(j \mid n, r, p_+, p_-) \\
&= \sum_{j=0}^{n} P(k \mid j, n, p_+, p_-) \cdot P(j \mid n, r) \\
&\overset{(1.1,1.2)}{=} \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot Bin(j; n, r) \qquad (1.3)
\end{aligned}$$

1.3.2.4 Parameter estimation in the BD-model

In the Bayesian approach, a prior needed to be specified for r to calculate the posterior distribution of r. Recall the beta distribution(s), a 2-parameter family of continuous probability distributions defined on the unit interval $[0, 1]$,

$$Beta(r; \alpha, \beta) \propto r^{\alpha-1}(1-r)^{\beta-1} \text{ , for } \alpha, \beta > 0,\ r \in (0,1) \text{ ,}$$

It was assumed that a fraction of λ_m positions are essentially methylated, and that their rate r follows a $Beta(r; \alpha = r_m \cdot w_m, \beta_m = (1-r_m) \cdot w_m)$ distribution, having an expectation value for r of $\frac{\alpha_m}{\alpha_m+\beta_m} = r_m$. λ_m was set to 0.7, resembling the expected level of mammalian genomic methylation [10]. The additional parameter w_m weights the strength of the prior relative to the strength of the likelihood. Since the confidence into/ the knowledge about the prior distribution of methylation rates is rather weak, the procedure should be strongly data-driven, therefore a a low w_m was chosen, $w_m = 0.5$. A fraction of $\lambda_u = 1 - \lambda_m$ is essentially unmethylated, and their rate is assumed to follow a $Beta(r; \alpha_u = r_u \cdot w_u, \beta_u = (1 - r_u) \cdot w_u)$ distribution, having an expectation value for r of $\frac{\alpha_u}{\alpha_u+\beta_u} = r_u$, which was set to $r_u = 0.2$ and $w_u = 0.5$. Thus, the prior distribution $\pi(r)$ is a 2-Beta mixture distribution,

$$\pi(r; \alpha_m, \beta_m, \alpha_u, \beta_u, \lambda_m) = \sum_{s \in \{m,u\}} \lambda_s Beta(r; \alpha_s, \beta_s) \qquad (1.4)$$

The pragmatic reason for choosing a Beta mixture as a prior distribution is the fact that the Beta distribution is the conjugate prior of the Binomial distribution[57], such that for some normalizing constant $D^{\alpha,\beta}_{j,n}$,

$$Bin(j; n, r) \cdot Beta(r; \alpha, \beta) \;=\; D_{j,n}^{\alpha,\beta} \cdot Beta(r; ; j + \alpha, n - j + \beta) \tag{1.5}$$

By virtue of Equation (1.5), the posterior distribution of r can be written down analytically (Equation 1.7). This has the advantage that all questions on the posterior distribution of r can be answered efficiently and up to an arbitrary precision. Efficiency is an issue, because the posterior distributions for all regions needed to be calculated, which can easily amount to millions.

$$\begin{aligned}
&P(r \mid k, n;\, p_+, p_-; \alpha_m, \beta_m, \alpha_u, \beta_u, \lambda_m) \\
&\quad = N^{-1} \cdot P(k \mid n, r;\, p_+, p_-) \cdot \pi(r; \alpha_m, \beta_m, \alpha_u, \beta_u) \qquad (1.6) \\
&\quad \overset{(1.3,1.4)}{=} N^{-1} \cdot \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot Bin(j; n, r) \cdot \sum_{s \in \{m,u\}} \lambda_s Beta(r; \alpha_s, \beta_s) \\
&\quad \overset{(1.5)}{=} N^{-1} \cdot \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot \qquad (1.7) \\
&\left(\sum_{s \in \{m,u\}} \lambda_s D_{j,n}^{\alpha_s,\beta_s} Beta(r; j + \alpha_s, n - j + \beta_s) \right)
\end{aligned}$$

In the above equation, N is a normalization constant,

$$N = \sum_{j=0}^{n} \sum_{m=0}^{k} C^1_{m,j} C^2_{n-j,k-m} \cdot \sum_{s} \lambda_s D_{j,n}^{\alpha_s,\beta_s} \tag{1.8}$$

The ingredients for the construction of the posterior distribution are visualized in Figure (Fig. 1.5). The upper bound for false methylation was estimated as $p_+ = 0.2$, 0.51, 0.41, 0.44, 0.39 in the experiments Liver, H1, H2, H3 and H4, respectively, as described under sec. 1.3.2.3.

1.3.2.5 Evaluation of precision, sensitivity and specificity of the BD-model in simulation studies

1.3.2.5.1 Testing for epimutation events In order to validate the model, the calling of epimutation events was tested on simulated data. If a highly methylated region in the wild type sample is affected by several demethylation events, its methylation rate will drop. Biologically relevant changes were expected that are called a (demethylating) epimutation event when this region shows decreased methylation in a test sample (i.e., in the case the single cell samples). Vice versa, a region which

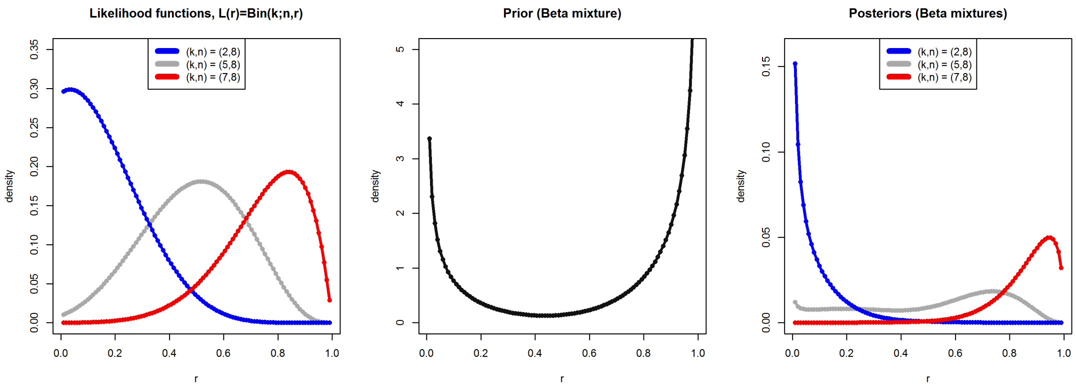

Figure 1.5: Plot of the likelihood functions for three different observations (k, n) (left), the Beta mixture prior distribution (middle) and the corresponding three posteriors (right). The number n of counts is set to 8, of which $k = 2$ (blue), $k = 5$ (grey) and $k = 7$ (red) are methylation counts. $\lambda_m$was set to 0.7 to resemble the expected level of mammalian genomic methylation [10]. The unknown parameter $p_+$was determined empirically from the false non-CpG methylation, which reflects the incomplete conversion rate, as follows: L1: 0.2, H1: 0.51, H2: 0.41, H3: 0.44, H4: 0.39. $p_-$was set to 0.2 as a conservative choice. The beta mixture prior was set as described in the text.

is called sparsely methylated in the wild type sample and which shows increased methylation in the test sample defines a (methylating) epimutation event. For each pair (k, n), Fig. 1.14 shows the estimated methylation rate, and the corresponding methylation calls. Note that according to the model, (strict) high methylation calls can only be made if $n \geq 5$. The number of positions with at least this coverage however is very small, so it is not advisable to apply epimutation calling to single positions. Instead, it was preferred to pool single position counts in regions of appropriate size, i.e., regions containing sufficiently many CpG positions that have a positive read count number in both the reference and the single cell sample ("shared" CpG positions). The method has then sufficient power to reliably detect epimutation events affecting these regions. The genome was tiled into disjoint consecutive regions of fixed length d, $d = 50, 250, 1000, 2500$ and d was finally chosen as $d = 1000$, because the distribution of shared counts per region (Fig. 1.7) was most suitable for the epimutation calling procedure. As epimutation calls were used only for exploratory purposes, multiple testing problems were not encountered and no need for multiple testing correction for the number of regions that were investigated was given.

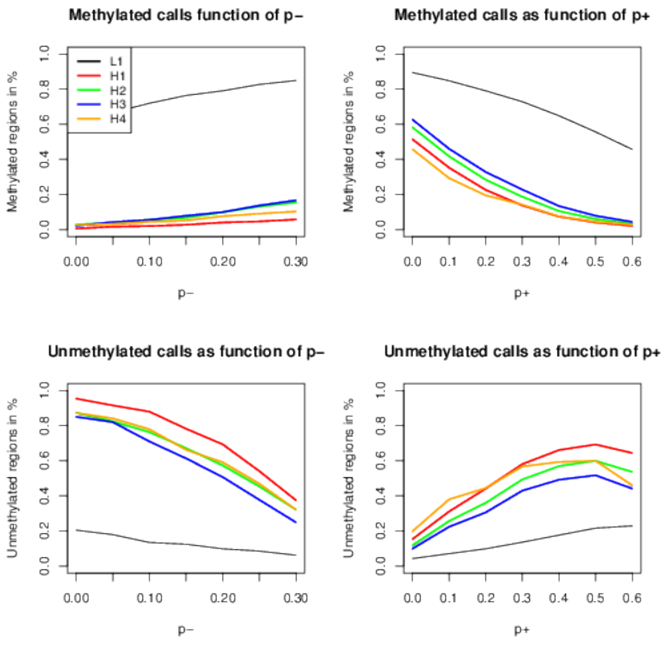

Figure 1.6: The (relative) number of 10k regions which were called showing increased methylation (top row) or showing decreased methylation (bottom row) as a function of the false negative rate p_- (left column) and the false positive rate (right column). Colors indicate results for different samples. black: L1, red: H1, green: H2, blue: H3, yellow: H4. When p_- was varied, p_+ was set to 0.4. When p_+ was varied, $p_-$was set to 0.2.

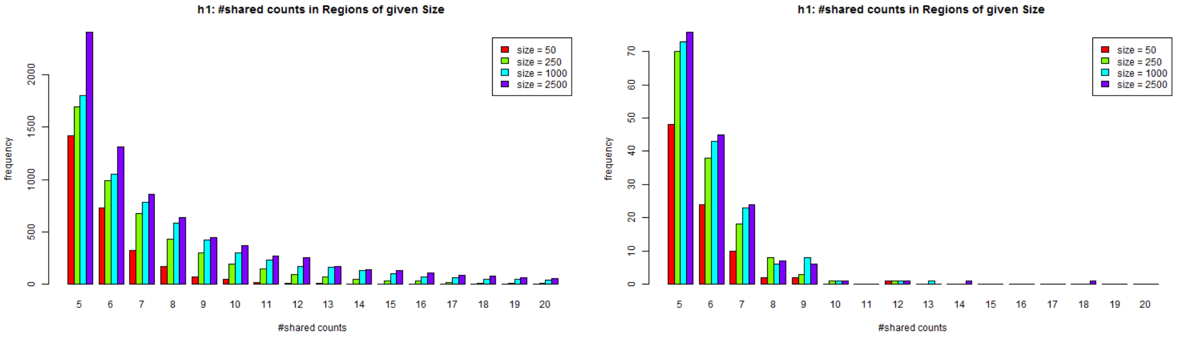

Figure 1.7: Left: Distribution of shared CpG positions between the livy reference and the H1 single cell sample, using all mappable reads, for different region sizes. Right: Distribution of shared CpG positions between the reference and the Liver single cell sample using only reads filtered for 99% bisulfite conversion.

Previously, genome-wide strategies for DNA methylation analysis filtered for reads that did not show non-CpG methylation[26]. However, while this removes presumably incorrect non-CpG methylation calls, it does by no means remove excessive CpG methylation calls. Moreover, it also removes a vast number of reads. While this is generally not a problem with fairly large amounts of genomic DNA, it is a problem when dealing with single cells in which the impact of bisulfite-mediated DNA degradation is much greater. Indeed, in our data, this strategy would have removed over 90% of all mappable reads. To access the vast amount of incompletely converted DNA, the probabilistic model was developed, allowing not only to obtain more reliable methylation estimates, but also assign realistic probabilities to all differential methylation calls.

Methylation/non-methylation counts for a single CpG position originating from a single template DNA strand will always be prone to unavoidable errors from bisulfite conversion, DNA amplification and sequencing. Since false positive and false negative methylation call rates are unequal, these errors have both a stochastic component (noise) and a systematic component (bias). Thus, statements about methylation states at individual CpG positions in single cell experiments are inherently uncertain, and may even be misleading.

It is therefore necessary to pool evidence about methylation across multiple CpG positions. Using an appropriate statistical model, this is also sufficient for single cell bisulfite sequencing to answer the important question whether epimutation events tend to occur stochastically dispersed along the genome, or if they co-occur in regions that are simultaneously de-methylated or hyper-methylated relative to a wild type sample. More generally, arbitrary sets of CpG positions were considered and tested them for epimutation linkage, i.e., the fact that epimutations in these regions occur in a coupled manner. Epimutation linkage might be due to genomic distance. Since regular methylation marks are typically grouped in regions[58][59], it is likely that the erasure respectively additional placement of methylation marks is also spatially dependent.

Another important cause for epimutation linkage might be functional similarity. E.g., CpG positions in the vicinities of a transcription factor's binding sites might show coupled epimutation events upon permanent activation of this transcription factor. The larger the sets of CpG positions are, the higher the ability to detect epimutation linkage. In other words, statistical power had to be traded against spatial resolution for the detection of epimutation linkage.

1.3.2.5.2 Computational validation of epimutation calling robustness and sensitivity To ensure that the biological differences in epimutation rates of different genomic regions (promoters, genes, CG islands, repeats) are no artifacts of the model parametrization (namely the false positive and false negative rates $p_+$and

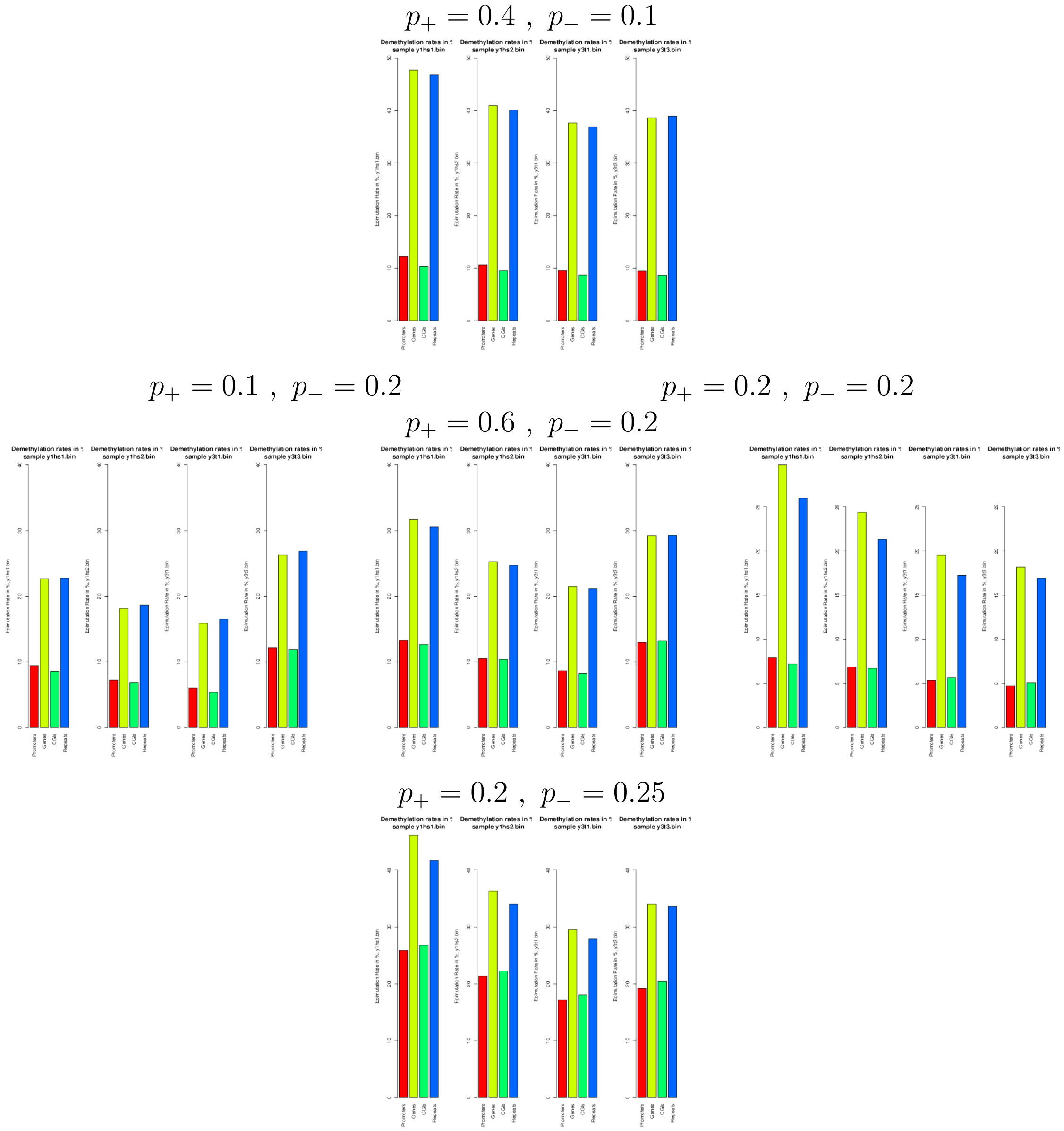

Figure 1.8: Each plot shows the epidemethylation rates in different genomic regions (red =promoters, yellow = genes, green = CG islands, blue = repeats) and for different single cell samples (H1-H4, from left to right). Each plot was generated using different parameters for the false positive and false negative methylation calling rates p_+ and p_-(see header of the plots).

p), the whole epimutation calling procedure was ran for 5 different parameter sets which explore extreme values of p_+ and p_-. The results for the epidemethylation and epimethylation calls are shown in Fig. 1.8 and Fig. 1.9, respectively. The main findings are: The epidemethylation rates in gene and repeat regions are consistently higher than in promoters and CG islands, and the epimethylation rates behave inversely. Although the false positive rate of the calls was controlled, the non-conversion rate p_{+} results in a reduced sensitivity in epidemethylation calling and excessive epimethylation events in the less converted sample. This is due to the fact that in the less converted sample, it is more difficult to distinguish increased methylation from random fluctuations in the number of non-converted positions.

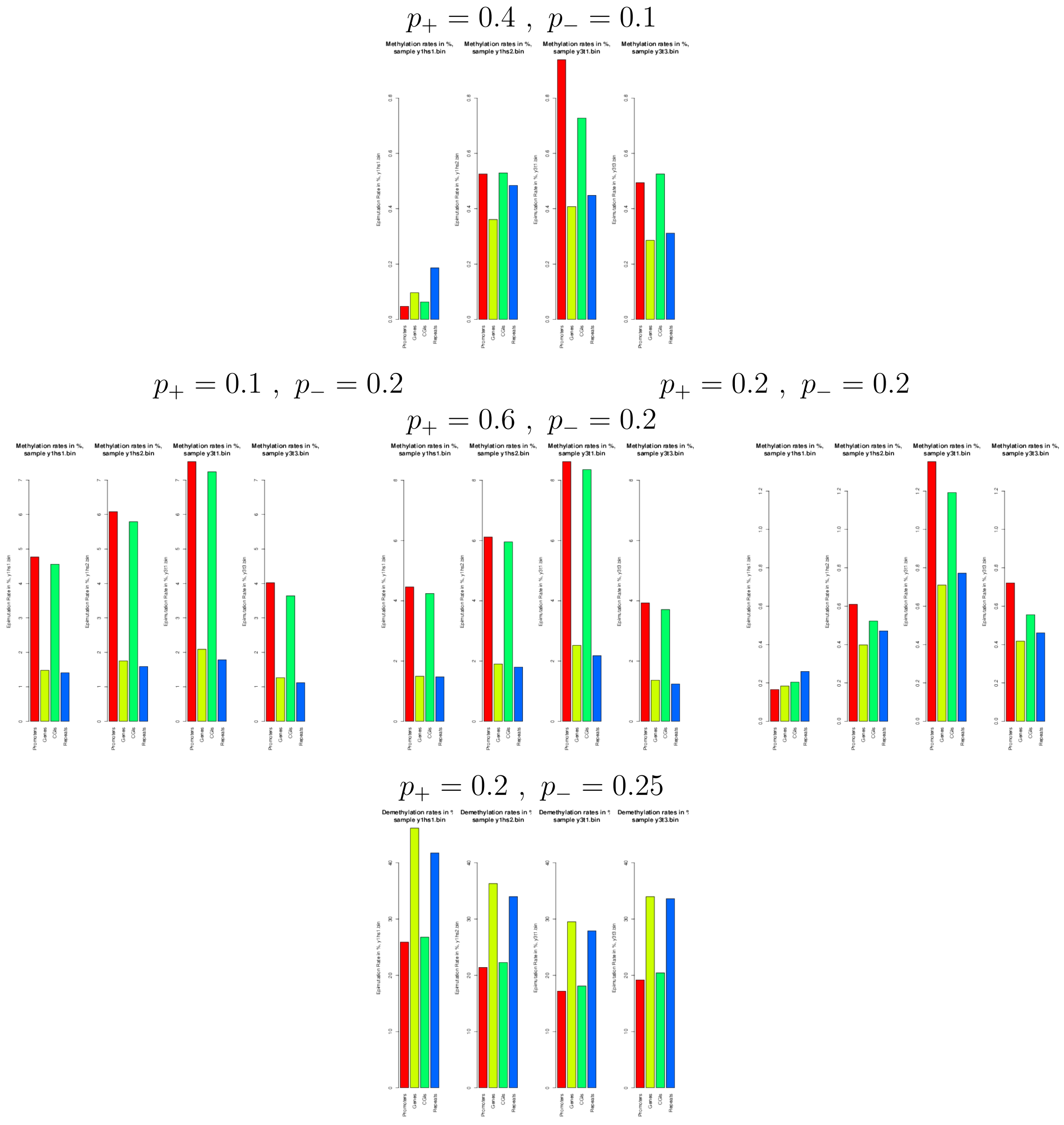

Figure 1.9: Each plot shows the epimethylation rates in different genomic regions (red =promoters, yellow = genes, green = CG islands, blue = repeats) and for different single cell samples (H1-H4, from left to right). Each plot was generated using different parameters for the false positive and false negative methylation calling rates p_+ and p_-(see header of the plots).

1.3.2.5.3 Epimutation calling by a hidden Markov model A second strategy for epimutation calling was considered, which resembles methods used for calling copy number variants from SNP chip data [60]. A Hidden Markov Model[61] is trained on the sequence of methylation / non-methylation counts along the genome. The hidden states of the HMM can assume three values, indicating no change in methylation, epimethylation, or epidemethylation. After learning of the HMM, the most probable hidden state sequence (the Viterbi path) can be used to annotate the genome, i.e., to call epimutations. In contrast to the test-based calling of epimutation events, the HMM model does not require a pre-defined partitioning into windows of fixed size to achieve a sufficient statistical power. It might therefore have a better positional resolution in detecting epimutated regions. On the other hand, the HMM-based epimutation calls do not offer any control over the rate of false positive epimutation calls. Control over the false positives is a desirable property if one seeks to validate individual epimutated regions.

First, the methylation-/non-methylation counts in the reference and the sample had to be pooled in identical windows of size $w = 1000$. This window size was chosen in order to yield results that can be directly compared to the test-based epimutation calling procedure. Within each window j, the total counts n^j_{ref} for the reference (resp. n^j_{sam} for the single cell sample) and the methylation counts k^j_{ref} for the reference (resp. k^j_{sam} for the single cell sample) was calculated, and afterward the quantities

$$\begin{aligned}
q^j_{epimeth} &= P(r_{ref} < 0.3 | k^j_{ref}, n^j_{ref}; \, p_{+ref}, p_{-ref}) \\
&\quad \cdot P(r_{sam} > 0.5 | k^j_{sam}, n^j_{sam}; \, p_{+sam}, p_{-sam}) \\
q^j_{epidemeth} &= P(r_{ref} > 0.7 | k^j_{ref}, n^j_{ref}; \, p_{+ref}, p_{-ref}) \\
&\quad \cdot P(r_{sam} < 0.5 | k^j_{sam}, n^j_{sam}; \, p_{+sam}, p_{-sam}) \\
q^j_{no\,call} &= 1 - q^j_{epimeth} - q^j_{epidemeth} \qquad (1.9)
\end{aligned}$$

These three probabilities were converted into the emission probabilities

$$\begin{aligned}
p^j_{epimeth} &= P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam} \mid \text{epimethylation event}) = q^j_{epimeth} \\
&\quad \cdot P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam}) / P(\text{epimethylation event}) \\
p^j_{epidemeth} &= P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam} \mid \text{epidemethylation event}) = q^j_{epimeth} \\
&\quad \cdot P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam}) / P(\text{epidemethylation event}) \\
p^j_{no\,call} &= P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam} \mid \text{no call}) \\
&= q^j_{epimeth} \cdot P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam}) / P(\text{no call}) \qquad (1.10)
\end{aligned}$$

Here, an uninformative prior was assumed

$$P(\text{epimethylation event}) = P(\text{epidemethylation event}) = P(\text{no call}) = \frac{1}{3}.$$

The probability $P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam})$ affects all three quantities in the same way, so it does neither influence the learning of the HMM by the standard Baum-Welch algorithm nor the calculation of the Viterbi path. Therefore, the factor $P(k^j_{ref}, n^j_{ref}, k^j_{sam}, n^j_{sam})$ in the Equations (1.10) was ignored. The position-specific emission probabilities are used for the training of an HMM whose hidden states can assume one of the three epimutation states *epimethylated*, *epidemethylated*, *no call*. The Baum-Welch algorithm learns the hidden state transition matrix A and the probability distribution of the initial states π. The Viterbi path (the maximum likelihood sequence of epimutation calls) is afterward calculated based on these parameters.

Table Tab. 1.1 shows a comparison of the test-based and the HMM-based annotation for the epimutation calls for all four single cell samples. It is evident that HMM-based calling is anticonservative; it systematically calls more epimutation events than the test-based procedure. This was expected, because the HMM procedure is not designed to keep any type 1 error levels rather than providing a good fit to the epimutation rates observed in consecutive regions. I decided against the use of HMMs in the specific context of single cell epimutation analysis, for a simple reason: the basic assumption of an HMM, namely that the observations were generated from a hidden Markov sequence, is strongly violated. The regions that have enough counts to provide any evidence for or against an epimutation event are relatively sparse. The hidden state sequence of the HMM is therefore not really a Markov chain but a sequence of independent observations. This means that the HMM cannot exploit its advantage, namely the modeling of the (putative) spatial correlation of the epimutation rates along the genome.

1.3.2.5.4 Comparison of single position vs. region-based epimutation calling

For the sake of comparability with previous results in the literature, the epimutation calling procedure was compared with a naive approach. Highly methylated positions were defined as those with methylation levels $> 90\%$, while positions with methylation levels$< 10\%$ were classified as sparsely methylated. Since the coverage in the samples is less than one, this essentially amounts to checking whether a methylation mark is set or not. Demethylating epimutation events were then called if sparsely methylated positions in a single cell sample were highly methylated in the wild type (i.e. reference cell pool). Methylating epimutations are called in an analogous manner. These results are given in Fig. 1.8 and Fig. 1.9. I quantitatively investigated the false positive and false negative epimutation calling rates by a graphical model (1.11), which for each nucleotide position consists of four binary variables: early, reference - the true resp.the observed methylation state of this position in the refera)

Test \ HMM	epidemeth.	epimeth.	no call	row sum
epidemeth.	85198	24	4010	89232 (48.5%)
epimeth.	0	361	127	488 (0.2%)
no call	31532	2723	59766	94021 (51.17%)
column sum	116730 (63.52%)	3108 (1.69%)	63903 (34.77%)	183741 (100%)

b)

Test \ HMM	epidemeth.	epimeth.	no call	row sum
epidemeth.	64371	3	1087	65461 (39.38 %)
epimeth.	0	784	214	998 (0.6 %)
no call	29867	2663	67214	99744 (60.01 %)
column sum	94238 (56.70 %)	68515 (41.22 %)	3450 (2.07 %)	166203 (100 %)

c)

Test \ HMM	epidemeth.	epimeth.	no call	row sum
epidemeth.	49275	2	1043	50320 (34.69 %)
epimeth.	0	692	166	858 (0.59 %)
no call	28690	2561	62596	93847 (64.71 %)
column sum	77965 (53.75 %)	3255 (2.24 %)	63805 (43.99 %)	145025 (100 %)

d)

Test \ HMM	epidemeth.	epimeth.	no call	row sum
epidemeth.	41510	2	537	42049 (39.24 %)
epimeth.	0	373	98	471 (0.43 %)
no call	20080	1407	43151	64638 (60.32 %)
column sum	61590 (57.47 %)	1782 (1.66 %)	43786 (40.86 %)	107158 (100 %)

Table 1.1: Comparison of the test-based epimutation calls and the HMM-based epimutation calls for the samples H1 (a), H2 (b), H3 (c) and H4 (d) relative to the multi-cell reference L1.

ence sample, late, sample - the true resp. observed methylation state of this position in the single cell sample.

$$\begin{tikzcd}
\text{early} & \bullet \arrow[r] \arrow[d] & \bullet & \text{reference} \\
\text{late} & \bullet \arrow[r] & \bullet & \text{sample}
\end{tikzcd} \tag{1.11}$$

The graphical model gives rise to a factorization of the joint probability distribution of (early,late,reference,sample):

$$P(\text{early,late,reference,sample}) = P(\text{sample} \mid \text{late}) \cdot P(\text{late} \mid \text{early}) \cdot P(\text{reference} \mid \text{early}) \cdot P(\text{early}) \tag{1.12}$$

The factors on the right hand side of Equation (1.12) are defined in Table (Tab. 1.2).

P(sample | late)	sample = -	samples = +
late = -	$1 - p_+^{sam}$	p_+^{sam}
late = +	p_-^{sam}	$1 - p_-^{sam}$

P(reference | early)	reference = -	reference = +
early = -	$1 - p_+^{ref}$	p_+^{ref}
early = +	p_-^{ref}	$1 - p_-^{ref}$

P(late | early)	late = -	late = +
early = -	$1 - r_+$	r_+
early = +	r_-	$1 - r_-$

P(early)	
early = -	m
early = +	$1 - m$

Table 1.2: The probability distributions on the right hand side of Equation (1.12) are parametrized by 7 parameters: m, the average methylation rate of the sequences under consideration, r_+ and r_-, the methylating resp. demethylating epimutation rates in these sequences, p_+^{ref}, p_-^{ref}, the false methylation resp. non-methylation call rates in the reference, and p_+^{sam}, p_-^{sam}, the false methylation resp. non-methylation call rates in the single cell sample.

A methylating epimutation occurs when early = - and late = +. A methylating epimutation was called if reference = - and sample = +. Vice versa, a demethylating mutation occurs when early = + and late = -. A demethylating epimutation was called if reference = + and sample = -. The false positive rates in single position epimethylation/ epidemethylation calling are then obtained as

$$\begin{aligned} FP_{meth} &= P(\text{reference} = -, \text{sample} = +, \sim (\text{early} = -, \text{late} = +)) \\ FP_{demeth} &= P(\text{reference} = +, \text{sample} = -, \sim (\text{early} = +, \text{late} = -)) \quad (1.13) \end{aligned}$$

Figure Fig. 1.10 shows the false positive rates that are achieved with single position epimutation calling in a large range of scenarios. The false positive rates are exceedingly high, which makes single position calls useless for drawing reliable conclusions. In particular, epimutation rates that were obtained from single position calls must be considered way too high, e.g., a false positive epimethylation rate of 90% implies that the epimethylation rates are over-estimated by a factor of 10. For these reasons, single position epimutation calling is advised against.

On the other hand, a proof for the utility of the region-based epimutation calling approach was required. It is not sensible to plot the single nucleotide false positive / false negative epimutation calls for regions that were identified as epimutated. This would lead to almost the same results as in Figure Fig. 1.10, because single nucleotide calling cannot be improved. However, it is meaningful to compare the true methylation rate differences (late - early) of the regions that were called epimutated with the differences of regions that were not called epimutated. This was done in

Fig. 1.11. It can be shown that all the regions called show a true(!) difference in methylation, and this difference is substantial. In particular, no false epimethylation calls were obtained at all when a demethylating scenario was considered, and vice versa. Therefore, I believe that the epimutation calling is not only biologically more meaningful, it is also robust and reliable.

1.3.3 Methylation Analysis of mouse liver cells

1.3.3.1 Sample preparation

Underlying the Bioinformatics analysis is a novel procedure for DNA methylation analysis of single cells using a combination of optimized bisulfite treatment and whole genome amplification, which is described as follows [5]. Single cells were collected from populations of mouse embryonic fibroblasts (MEFs) or hepatocytes, under an inverted microscope by hand-held capillaries, and either frozen or immediately subjected to heat DNA denaturation, followed by bisulfite treatment. The converted DNA was subsequently subjected to whole genome amplification using multiple displacement amplification (MDA), based on phi29 DNA polymerase and random hexamer primers in an isothermal reaction. The bisulfite-converted, amplified material was then used as template for conversion-specific PCR, targeting regions of interests, or to make reduced-representation libraries for Next Generation Sequencing (NGS). PCR primers were designed to amplify only converted sequences, that is, sequences in which non-methylated cytosines are replaced by thymines. To increase specificity, a nested PCR approach was used. As positive controls, collections of 100 MEFs were used, bisulfite-treated and MDA-amplified in the same way as the single cells, as well as 800 ng of bisulfite-treated unamplified DNA from the same MEF population. Non-bisulfite-treated, non-amplified, genomic DNA served as negative control to verify PCR specificity for only fully converted DNA.

Four single mouse hepatocytes (H1, H2, H3 and H4) were denatured, bisulfite-converted and MDA-amplified according to the protocol. In parallel, 200 ng of total genomic liver DNA was subjected to the same procedure. Amplification products were then digested with the restriction enzyme MseI, which cuts at TTAA sites. After digestion, a size selection (250-300bp) was performed, which limited the analysis to ~10% of the genome. An in silico digestion of the converted mouse genome indicated that this size range would allow me to interrogate ~1.2 million CpG sites. The size-selected DNAs from the 4 single cell samples and the liver sample were end-repaired, A-tailed and ligated to Illumina adapters, and the completed libraries were sequenced to 120bp on the Illumina HiSeq 2000. The first computation steps in the pipeline following sequencing are described in sec. 1.4.0.6.

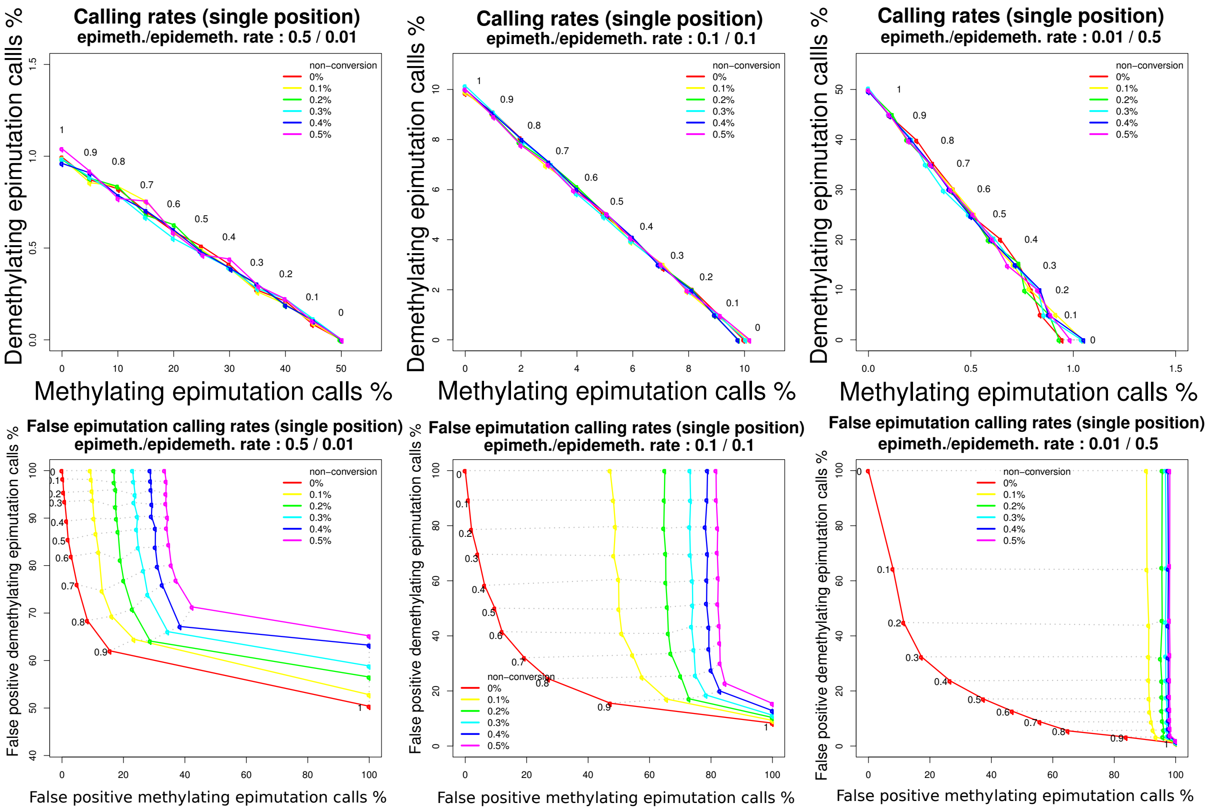

Figure 1.10: Top panel: Epimutation calling rates as a function of the average methylation rate $m \in \{0, 0.1, 0.2, ..., 1\}$ of the reference region, and the false methylation calling rate $p_+^{sam} \in \{0, 0.1, ..., 0.5\}$ of the single cell sample. Three epimutation scenarios were investigated: pervasive epimethylation ($r_+ = 0.5$, $r_- = 0.01$, left plot), no preferential / spurious epimutations ($r_+ = 0.1$, $r_- = 0.1$, middle plot), and pervasive demethylation ($r_+ = 0.5$, $r_- = 0.01$, right plot). The remaining parameter were fixed at $p_+^{ref} = 0.1$, $p_-^{ref} = 0.01$, $p_-^{sam} = 0.1$. Shown are the rates of epimethylating calls (x-axis) and the rates of epidemethylating calls (y-axis). The different rates of false methylation calls (= non-conversion rates) are color-coded, ranging from $p_+^{sam} = 0$ (red line) to $p_+^{sam} = 0.5$ (purple line). The dots of the red line are labeled with their respective average methylation rates m. Grey dotted lines connect scenarios with the same methylation rate. Note that the scales of the three plots are not identical. Bottom panel: Assessment of false positive epimutation calls for the same scenarios as in the top panel. Shown are the rates of false positive epimethylation calls (x-axis) and the false positive rates of epidemethylation calls (y-axis).

1.3.3.2 Statistics on methylation estimation in mouse BS-Seq data

This section provides basic statistics about the length distribution of the parts of the reads that could be mapped after trimming, and investigates the number of CH and

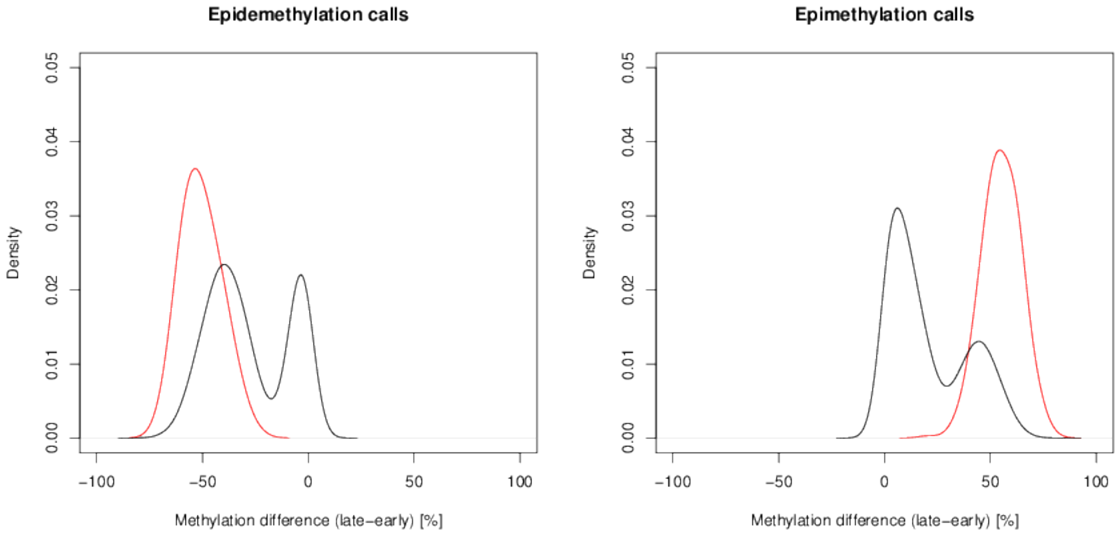

Figure 1.11: Validation of the region based epimutation calling procedure by simulation. The parameter settings were as follows: All false conversion rates were set to a relatively high level of 0.2 (to generate a statistically challenging situation). The methylation rates m for each region were chosen from $\{0.1, 0.7, 0.9\}$ with equal probability, to roughly mimic the real situation in the genome. Two scenarios were considered: A demethylating scenario (left, $r_+ = 0.01$, $r_- = 0.5$) and a methylating scenario (right, $r_+ = 0.5$, $r_- = 0.01$). The red density curve shows the methylation differences of those regions that were called epidemethylated (left) respectively epimethylated (right). The black density curve shows the methylation differences of the regions that were not called epimutated. Importantly, there never were epimutation calls in the false direction, i.e., regions were never called epimethylated (epidemethylated) in the left (right) scenario.

CG positions present in these reads. Since the read length distributions are rather tight (Fig. 1.12), the absolute numbers of CH and CG positions are reported, without scaling them to relative frequencies. Relative frequencies would assume only a few discrete numbers and would be difficult to present as histograms, because any binning of the histogram bars can easily produce discretization artifacts. The relative abundance of CG and CH positions does not resemble that in the mouse genome, which is due to the library preparation by Reduced Representation Bisulfite Sequencing (RRBS). The distribution of the methylated CG positions per read resembles that of the total CG positions per read, whereas the distribution of methylated CH position per read has a substantially lower mean than the distribution of total CH positions per read. This is in agreement with the expectation of a high CG methylation rate and a low CH methylation rate.

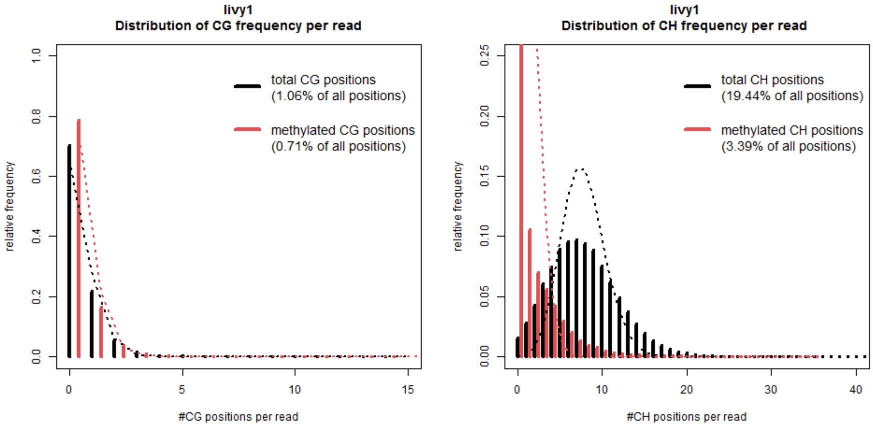

Figure 1.12: Descriptive statistics for the reads from the Liver reference. The median lengths of the mapped parts of the reads were 30, 47, 48 and 55 in the different samples. Frequency of CG (left) and CH (right) positions per read, for the total number of positions (black) and for the the number of positions that were called methylated (red). Dotted lines: corresponding maximum likelihood fits for a mixture of binomial distributions with their means equal to the empirical mean of the distribution. The mixture is taken over the distribution of the total read lengths. Compared to the idealized distribution, the empirical distributions are over-dispersed, i.e., their variance is higher than in the binomial fit. This can be explained by the fact that CH resp. CG positions do not occur independently.

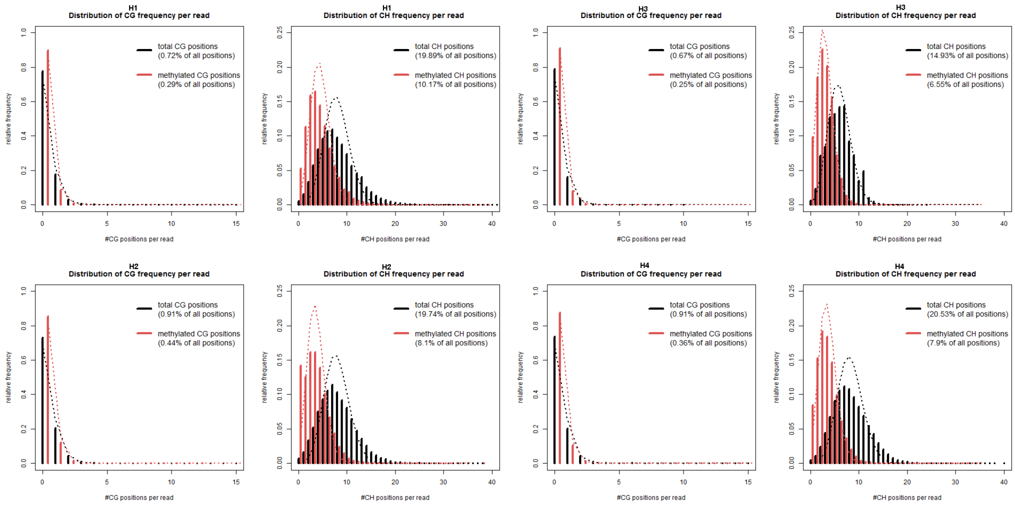

Figure 1.13: CG and CH read statistics of the single sample experiments. Colors and lines are the same as in Fig. 1.12.

1.3.3.3 Epimutation calling in mouse BS-Seq data

1.3.3.3.1 Methylation statistics derived from read counts For each region under consideration, an individual posterior distribution $P(r \mid k, n, p_+, p_-)$ is obtained. With this posterior at hand, it is an easy task to calculate the expected methylation rate $\hat{r}$ in the corresponding region,

$$\hat{r} = \int_0^1 r \cdot P(r \mid k, n, p_+, p_-) \, dr \tag{1.14}$$

It is customary to provide a Bayesian measure of uncertainty of this estimate, a so-called credible interval. A credible interval is an interval which contains the estimate ($\hat{r}$) and in which a prescribed probability mass of the posterior is located. One can construct a 90% credible interval $[m, M]$ as the shortest interval containing $\hat{r}$ such that $P(r \in [n, M] \mid k, n, p_+, p_-) = 0.9$. Moreover, a region is called *highly methylated* if

$$P(r > 0.7 \mid k, n, p_+, p_-) > c \tag{1.15}$$

for some stringency level c which was set to 0.75 here. The false negative methylation calling rates were set to $p_- = 0.1$ for all samples, and the false positive calling rates were determined by $p_+ = 1 -$ CH methylation rate for each sample separately. A region is said to show *increased methylation* if

$$P(r > 0.5 \mid k, n, p_+, p_-) > c$$

Analogously, a region is called *sparsely methylated* if

$$P(r < 0.3 \mid k, n, p_+, p_-) > c$$

and a region with *decreased methylation* satisfies

$$P(r < 0.5 \mid k, n, p_+, p_-) > c \tag{1.16}$$

By definition, any highly methylated region has increased methylation, and every sparsely methylated region shows decreased methylation. For $c > 0.5$, high and sparse methylation calls are mutually exclusive. Regions that are neither highly nor sparsely methylated are called *ambiguous.*

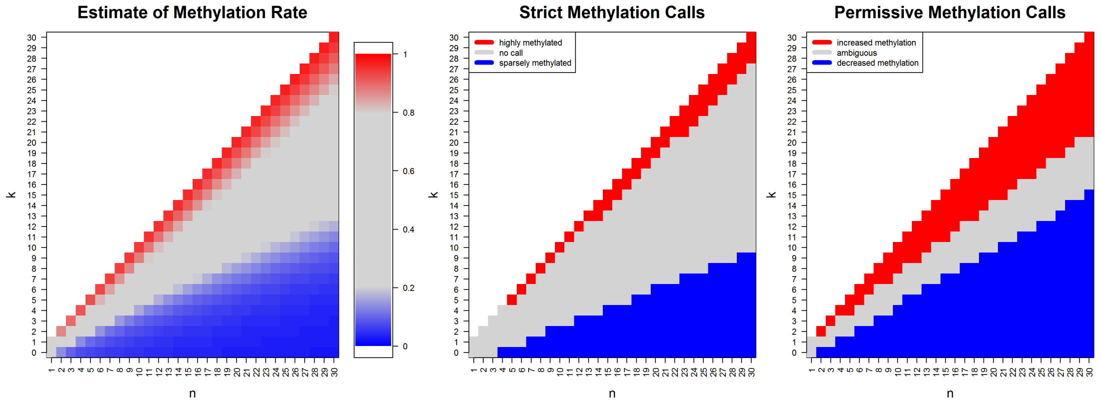

Figure 1.14: Illustration of the the results of the statistical modeling applied to regions of size $d = 1000$ in the Liver sample. In each plot, n (on the x-axis) denotes the total number of counts mapping to that region, of which k (on the y-axis) are counts indicating methylation. Left: Using the Liver-specific estimates of the false positive rate $p_+ = 0.2$ and the false negative rate $p_- = 0.1$ and the methylation prior in Equation (1.4), for each admissible pair (k, n) a methylation rate estimate $\hat{r}$ from equation (1.14) was obtained. Colors correspond to methylation rate, ranging from deep blue (zero methylation) to deep red (full methylation). Middle: The red respectively blue area defines the pairs (k, n) which satisfy the criteria for high respectively sparse methylation. Right: The red respectively blue area defines the pairs (k, n) which satisfy the criteria for increased respectively decreased methylation. Note that strict methylation calls are only made when at least $n = 5$ counts were observed.

The time-critical step is the calculation of the region-specific posterior distribution $P(r \mid k, n, p_+, p_-)$, and the quantities related to it (Equations 1.14-1.16). Since k and n vary for each region, and the number of regions is large, a lot of time was saved by pre-calculating all required quantities for a set of values $n = 1, ..., 45$, $k = 0, ..., n$. The statistics for, on average, 80% of all regions can then be looked up and do not need to be re-computed. The running times for $d = 250, 500, ...$ on the mouse genome took less than a minute plus $t = 25$ min for the pre-computing of each of the five samples, which did not vary substantially with region size.

1.3.3.3.2 Robustness of methylation estimation (computational validation)

The accuracy of the methylation rate estimates were compared when using 1k or 10k regions. This revealed that the methylation rate distributions estimated from 1k and 10k regions that contain at least 30 counts are virtually identical (Figure Fig. 1.15 a) and therefore can be assumed to correctly reflect the true methylation rate distribution. Figure Fig. 1.15 c) confirms this by plotting the methylation rate estimate of each 1k region with at least 30 counts against the estimate of the corresponding 10k region. The apparent bias towards lower methylation estimates can be explained by the fact that (rather rare) highly methylated 1k regions are compared to larger 10k regions in which this high methylation rate is averaged with 9 other 1k regions likely to have lower methylation rates. However, 1k regions containing a minimum number of 5 counts (the minimum number to cast a strict methylation call according to the standard parameter choice, see Figure Fig. 1.14), the methylation rate estimates are still dominated by sampling variance . The erratic bars at methylation values larger than 0.2 indicate that using 1k regions for methylation rate estimation might be unstable (Figure Fig. 1.15 b). For 10k regions with a least 5 counts, the corresponding rate distribution resembles that for minimum 5 counts. Therefore, I decided to use windows of size 10k for the analyses.

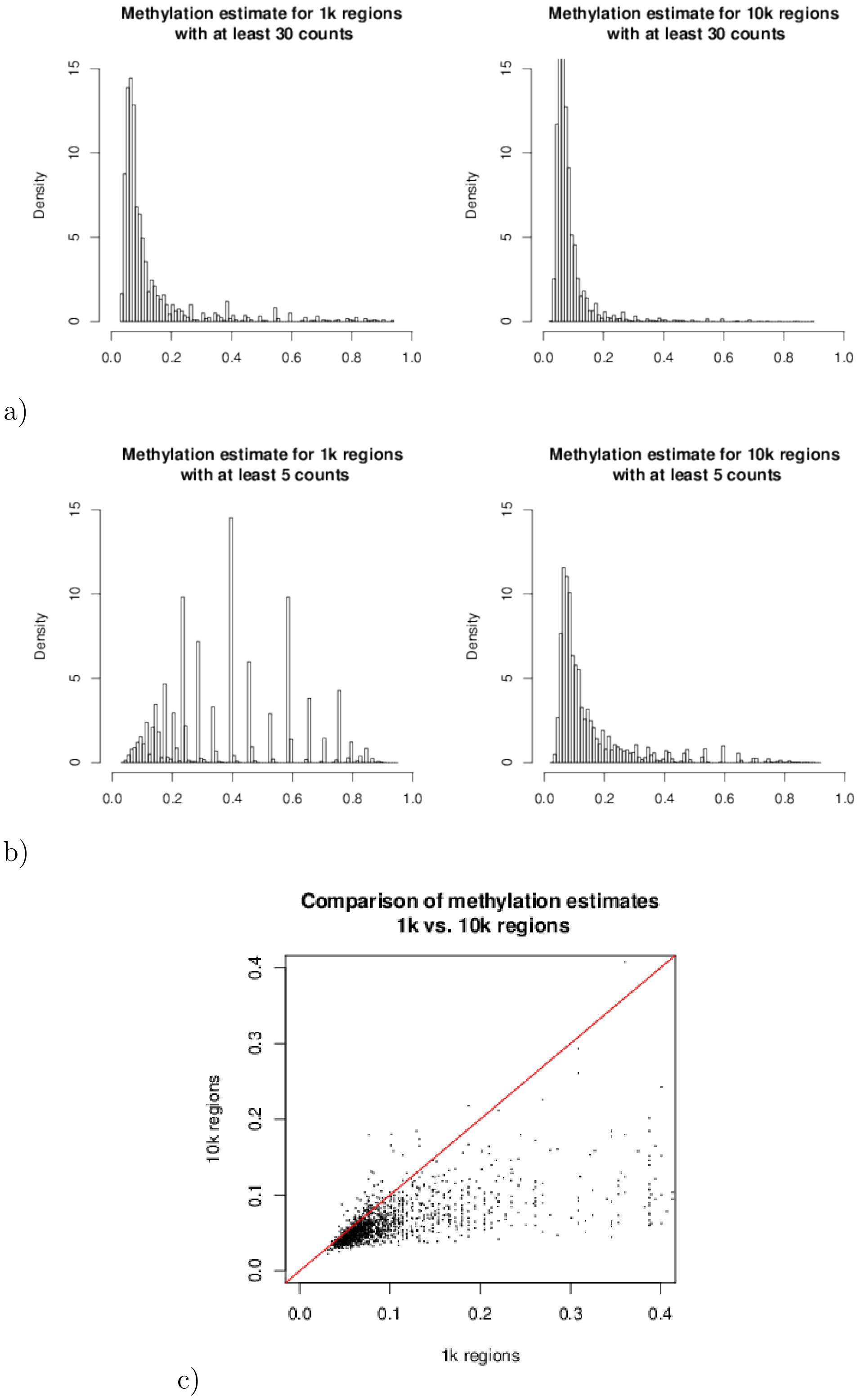

Figure 1.15: Comparison of methylation rate estimates for regions of size 1k and 10k. a) Methylation rate distribution for regions having at least 5 counts, for regions of size 1k (left) and of size 10k (right). b) Methylation rate distribution for regions having at least 30 counts, for regions of size 1k (left) and of size 10k (right). c) Scatter plot comparing the methylation rate estimate for a region of size 1k and minimum 30 counts (x-axis) with the methylation rate estimate for the 10k region in which the 1k region is contained (y-axis). The few (<50) points having larger estimates than 40% methylation rate were omitted from the plot.

Furthermore, I systematically explored the dependence of the methylation calling procedure on the estimates on the false positive and false negative rates p_+ and p_-. An increased false negative rate makes the identification of (truly) demethylated regions more difficult. I therefore expected and observed a decrease in the number of regions that are called methylated (Figure Fig. 1.6, bottom left). On the other hand, the probability of a region with high methylation counts to be truly highly methylated is even increased, given a larger false negative rate . Therefore, the number of regions called methylated increases with p_-(Figure Fig. 1.6, top left). The opposite reasoning applies to the false positive rate p_+. It becomes more difficult to detect methylated regions, as they might be due to false positive methylation counts. The number of regions called methylated therefore decreases with p_+ (Figure Fig. 1.6, top right). Conversely, the number of regions called unmethylated increases with with the false positive rate, because low methylation count numbers become less likely (Figure Fig. 1.6, bottom right).

It is evident from Figure Fig. 1.6 that the methylation calls vary considerably with p_+ and p_-. A wrong estimate of these critical parameters could potentially bias the analysis. However, there is no reason to believe that the estimates for $p_+$and p_- are incorrect. The false negative rate was set to 0.2, although it is probably much lower. As one can see from Figure Fig. 1.6, this hardly affects both increased and decreased methylation calls. The false positive rates were estimated by the CH methylation rate, because CH methylation is practically absent in mammals [62]. Since there are hundreds of thousands of CH (non-)methylation calls, the p_+ estimate is very precise.

1.3.3.3.3 Robustness against false positives (experimental validation) The non-conversion of unmethylated cytosines was the largest source of systematic errors (false positives). In order to verify that the model corrects for this error, a mixture of mouse liver DNA was split into two aliquots and carried out two bisulfite conversions with different (high and low) conversion efficiency. I.e., apart from different conversion times, the two aliquots were treated and processed identically. In total, 94,376,609 (326,183,300) reads were obtained from the high (low) conversion sample, 20,033,321 (36,149,298) of which could be successfully mapped to the genome, corresponding to a mapping efficiency of 21,22% (11,08%). The higher coverage for the low conversion sample was expected as a consequence of the lower DNA degradation. The false positive rates were estimated from CH-conversion as described before, which led to $p_+ = 0.1242$ (87.58% BS-conversion) for the high conversion sample and $p_+ = 0.5710$ (42.90% BS-conversion) for the low conversion sample.

Error-corrected methylation rate estimates give a better agreement between the two samples and have a more realistic distribution than empirical methylation rates (the number of methylation counts divided by the total number of counts in a region). Figure SFig. 1.16 shows the distribution of the empirical methylation rates and a scatter plot comparing the estimates for 10k regions with sufficiently many ($\geq$10) methylation counts in both samples. It is evident that the low conversion sample is strongly biased towards high empirical methylation rates; there are almost no regions with less than 50% methylation counts.

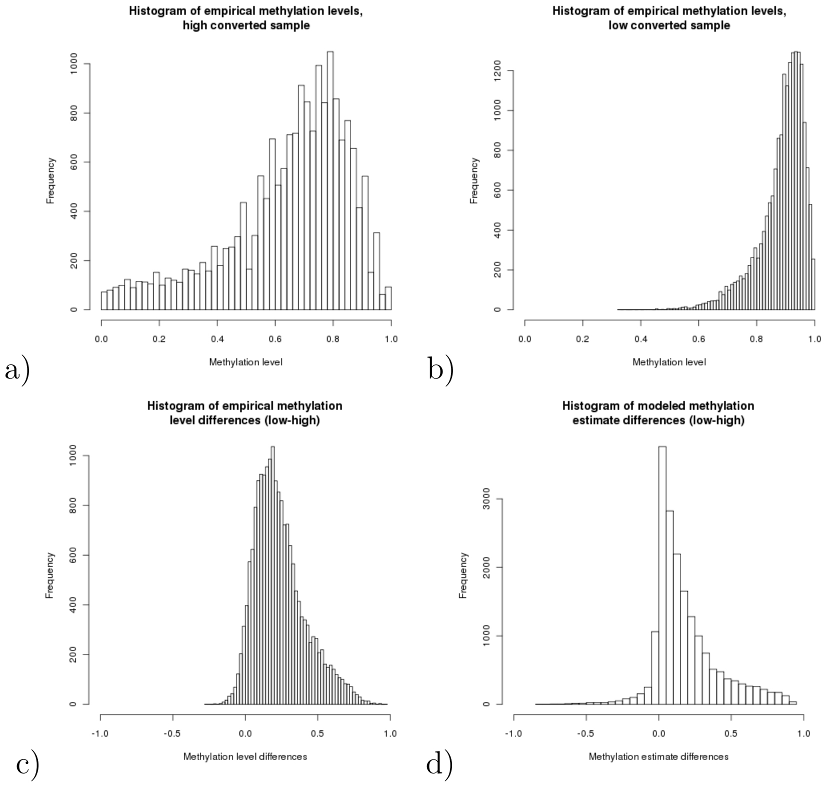

Figure 1.16: The distribution of the empirical methylation rates for the high conversion (a) and the low conversion (b) sample c) Distribution of the differences of the empirical methylation rates d) Distribution of the differences of the model estimated methylation rates

Fig. 1.17 shows the two "transfer functions" which convert observed methylation rates into the methylation rate estimates for the low and the high conversion sample. Most importantly, regions in the low conversion sample with an empirical methylation rate as high as 50% are estimated to have a low methylation rate (<10%). This is intuitive because one expects to see at least around 57.10% of the counts methylated due to false positives alone.

Fig. 1.18 shows the distribution of methylation rate estimates obtained by the model and a scatter plot comparing the estimates for 10k regions with sufficiently many (≥10) counts in both samples. As can be seen particularly from the rate distribution in the low conversion sample, the bias has been reduced substantially. The scatter plot demonstrates that the agreement between the estimates has increased, as many points now cluster not only at $(1, 1)$, but also at (0,0). This means in particular that regions of low methylation in the low conversion sample were correctly identified.

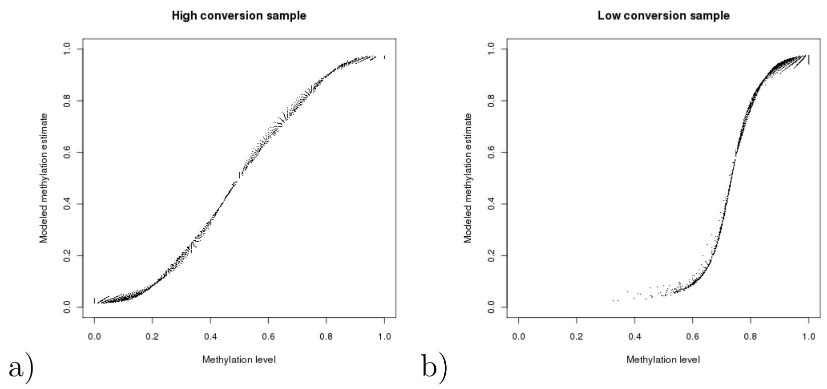

Figure 1.17: Conversion of empirical methylation rates into methylation rate estimates for the high conversion sample (a) and the low conversion sample (b). Note that tuples (k, n) of k observed methylated and a total of n observed counts are converted into one rate estimate $\hat{r}$. k/n is plotted on the x-axis vs. $\hat{r}$ on the y-axis. Since two tuples (k_1, n_1), (k_2, n_2) with identical quotient k_j/n_j may lead to slightly different estimates $\hat{r}_j$, the relation between empirical methylation rate and estimated rate is, strictly speaking, not a function.

The implicit bias correction leads to methylation rate estimates as low as they are almost never observed directly in the low conversion sample.

Fig. 1.18 c) still shows a bias towards high methylation values in the low conversion sample. This is a necessary consequence of the fact that the posterior distribution for the methylation rate r inevitably has a larger variance in the low conversion sample, simply because of the larger errors that blur the experimental evidence. However, one should bear in mind that the primary goal of inference is the detection of epimutation events, which relies on the accurate calling of regions with low/high methylation. Here, I exploited the advantages of the Bayesian approach in the calling procedure for sparsely/highly methylated regions. Namely, the methylation calls are not based on a point estimate (like the methylation rate estimate $\hat{r}$ in Equation 1.14), they are formulated as conditions on the full posterior distribution (Equations 1.15 and 1.16). They implicitly account for a higher variance in the posterior distributions of r for the low conversion sample. Thus, methylation calls in the low conversion sample are automatically more conservative. Fig. 1.16 demonstrates that the agreement between high/low respectively increased/decreased methylation calls is satisfactorily high.

1.3.3.4 Additional single-cell experiments

In addition to the single-cell samples H1-H4, six additional single hepatocyte samples were sequenced and analyzed. Basic statistics for them are shown in Tab. 1.3, corresponding to the description for the main samples H1-H4 in . The reference cell

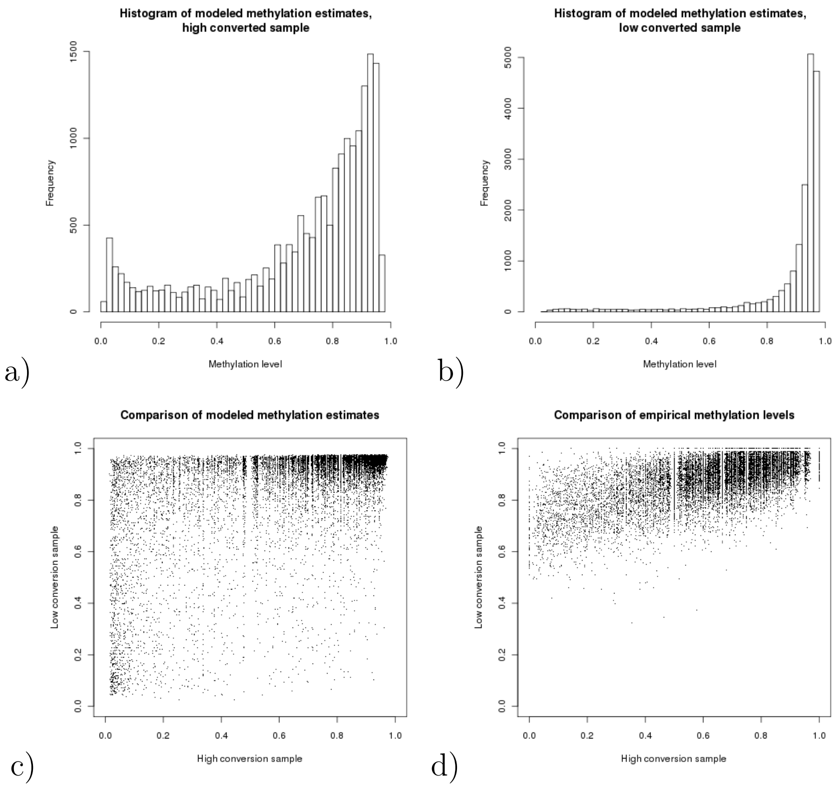

Figure 1.18: The distribution of the (model-based) methylation rate estimates for the high conversion (a) and the low conversion (b) sample c) Distribution of the differences of the estimated methylation rates d) Scatter plot comparing the empirical methylation rates for regions having at least 10 counts in both samples.

pool used for these samples was also L1. Plots of all epimutation rates including these samples are shown in Fig. 1.19.

Sample	Bisulfite conversion rate	Total reads	Mapped reads	Unique CpG positions
H5	58,13%	129,471,967	2,328,930	126,570
H6	57,09%	128,263,779	2,259,116	131,242
H7	54,00%	131,379,837	2,712,605	213,542
H8	57,82%	129,313,949	7,189,642	213,638
H9	56,54%	117,318,027	3,737,191	132,270
H10	56,31%	132,060,126	5,7109,21	209,971

Table 1.3: General statistics of RRBS of 6 additional hepatocytes. Table format corresponds to Table 2 in the main text.

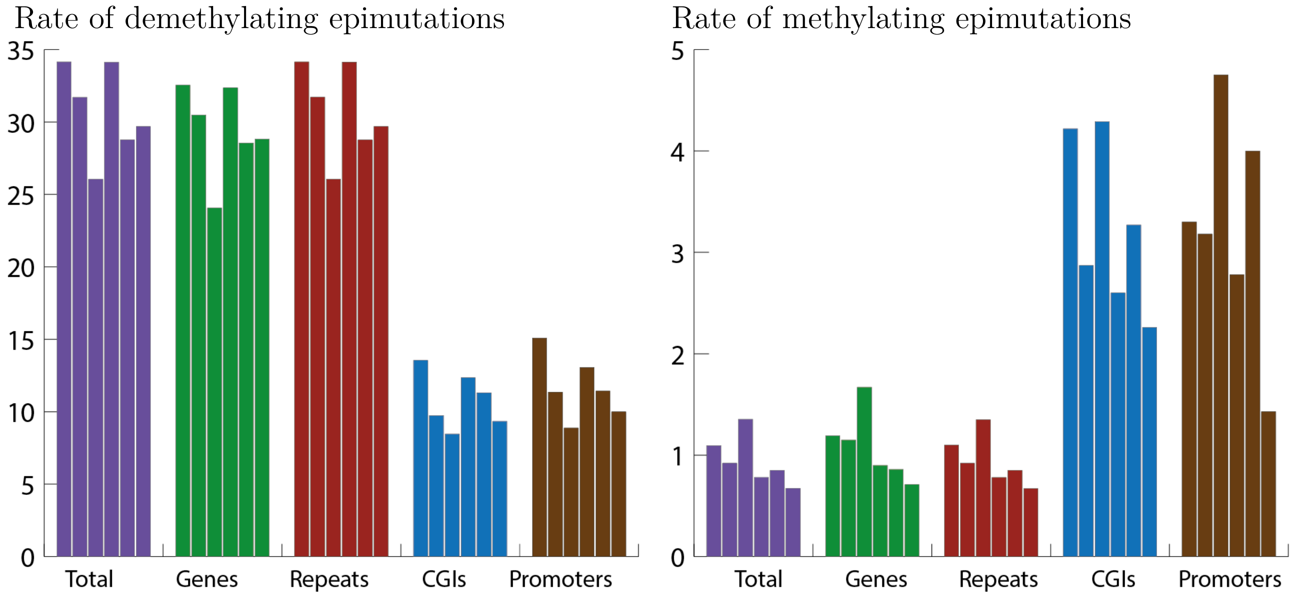

Figure 1.19: Genome-wide demethylating (left) and methylating (right) epimutation rates for the 6 additional cells as shown in Table 3. For each feature, samples shown from left to right are: H5, H6, H7, H8, H9, H10. Figure corresponds to Figure 2 a-d) for samples H1-H4 in the main text.

1.3.3.5 Results and discussion of biological results

Using the model, methylation data from the four single cells H1, H2, H3 and H4, DNA methylation patterns were compared with that of the liver tissue sample L1 as reference. Epimutation rates were 1.9 % - 4.7% for demethylating events and 0.2% - 1 % for methylating events (Tab. 1.4). Demethylating events were more common than methylating events, perhaps due to errors generated by DNMT1 in the propagation of methylation patterns during cell division. The analysis reveals demethylating epimutations being enriched in gene bodies and repeat regions while promoter regions are much more stable than the rest of the genome. Conversely, methylating epimutations are homogeneously distributed across the genome. This high propensity for repeats and gene bodies to accumulate demethylating epimutations could arise because such genomic features are highly methylated and therefore more prone to accumulate demethylating epimutations. Such events could occur as errors in the activity of DNMT1 in propagating DNA methylation patterns. This finding is in accordance with prior observations of mammalian methylomes in aging, which found that most aberrant changes in coding regions consist of demethylation, with aberrant methylation occurring mostly in CGIs and promoters [63]. As expected, epimutation events are distributed heterogeneously among chromosomes, pointing toward a stochastic mechanism, similar to DNA sequence mutations (data not shown). While only four cells were analyzed, these results also indicate considerable cell-to-cell variation. To my knowledge, this is the first procedure that is able to accurately analyze DNA methylation patterns at a single-cell resolution. The main new insight that can be derived from the analyses is that at 0.2-4.7% DNA epimutation rate is 2-3 orders of magnitude higher than the DNA sequence mutation rate [64]. This finding is in accordance with recent results suggesting that spontaneous transgenerational epigenetic changes in the Arabidopsis thaliana methylome are three orders of magnitude more frequent than DNA mutations [65][66]. The high propensity of accumulating demethylating epimutations, particularly in repetitive elements, is in keeping with recent findings suggesting an age-dependent global hypomethylation of transposed elements such as Alu elements [67] [41]. The presented single-cell bisulfite Reduced Representation approach can be applied not only to basic research on phenotypic diversity within organs and tissues in relation to disease states, but also to improve diagnostic and prognostic assays that sample very small numbers of cells from affected areas of diseased tissues. One major clinical application is to assess DNA methylation patterns in promoter regions of tumor suppressor genes in circulating tumor cells [68]. I anticipated that having access to such sensitive technology will radically change the way biomedical science is practiced, shifting emphasis from average endpoints towards the description of cell populations, tissues and organs through their individual parts at single-cell resolution.

Sample	Positions	Fully methylated	Fully unmethylated	Epimethylation (%)	Epidemethylation (%)
H1	12,635	57	12,268	27 (0.21)	596 (4.71)
H2	11,333	162	10,682	84 (0.74)	431 (3.83)
H3	7,417	150	6,808	81 (1.09)	186 (2.50)
H4	4,150	28	4,034	14 (0.33)	79 (1.90)

Table 1.4: Results from methylation analysis per sample, all reads, model-adjusted. The first column shows the number of common CpG positions overlapping between the reference L1 and each of the samples, followed by two columns showing the amount of these positions detected as fully methylated and -unmethylated, respectively, by our model. The last two columns show the number of positions with epimethylation and epidemethylation and the epimutation rates, respectively.

1.4 The BEAT R/Bioconductor Package

As part of this thesis, I wrote and published the R/Bioconductor package BEAT (BS-Seq Epimutation Analysis Toolkit) and a corresponding Bioinformatics Applications Note [6]. BEAT implements the Bayesian-Dirichlet mixture model and can aggregate consecutive cytosines into genomic regions with a minimum coverage threshold in order to overcome the aforementioned limitations in the estimation of methylation rates. It can precompute and re-use model-adjusted values of empirical data sets, call epimutations and generate related statistics. BEAT represents a novel tool for analyzing bisulfite-converted DNA sequences. For each region, it calculates a posterior methylation probability distribution which can be used for the comparison of DNA methylation between samples. The BEAT package delivers methods for the estimation of methylation levels, methylation status and for calling epimutation events in a two-sample comparison. To the author's knowledge, it is the first tool providing a rigid statistical model for handling BS-Seq samples. It has an in-built correction for conversion errors and is therefore perfectly suited for handling BS-Seq samples with possibly different BS-conversion rates.

1.4.0.6 Data preprocessing, quality control

FastQ data was first subjected to quality control by FastQC 0.10[69] to identify overrepresented sequences and low-quality ends. Several overrepresented sequences per sample were identified. Foreign sequence at the start or end of a read was treated as adapter contamination and clipped from the ends of all reads using cutadapt 1.0[70]. Using a Phred score of 20 as minimum quality threshold, ends of all reads were trimmed accordingly using the FASTX-Toolkit 0.13 (available at http://hannonlab.cshl.edu/fastx_toolkit/, accessed 10/04/2012). Mapping and methylation calling was then performed on all quality-corrected reads using the BS-methylation caller bismark 0.7.3[37] with the alignment tool Bowtie2 2.0[71]. Mapping options were standard, apart from an option for non-directional libraries, allowing a single mismatch in each seed region identified by Bowtie2 and permissive mismatch handling by allowing a decrease in 1.5 sequencing score points per base. Mapped reads were postprocessed by removing PCR duplicates (multiple, identical reads) using samtools 0.1.18 [72]. Annotation-specific results were generated by a genome-wide survey of promoter-, gene-, repeat regions and CGIs based on the mapped position of CG sites. Genome-wide promoter regions were obtained from the 2012 revision of the Eukaryotic Promoter Database (EPD), annotating a total of 9773 promoters[73], the current RepeatMasker annotation (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010, unpublished, available at http://www.repeatmasker.org, accessed 10/04/2012) was used to determine repetitive elements, and the 2012 UCSC annotations for mm10 were used to discern CGIs and genes for mapped reads[74].

1.4.0.7 Verification of preprocessing and mapping accuracy using BLAST

To test for biases related to preprocessing and read mapping, I used BLAST+[75] to perform local alignments of randomly sampled reads. In order to simulate methylation-aware mapping, C-to-T and G-to-A conversions were done on both query sequence and the mm10 reference genome and cross-aligned accordingly for both strands. Average mapping efficiency from Bismark, which was 10.7% for single-cells, was compared with BLAST results and found to be as good or better. BLAST could map 6.8% of reads with the standard parameter settings (6.2% if only hits with a maximum of 4 mismatches was allowed.) The higher mapping yield of Bismark/Bowtie2 in comparison to BLAST could be due to Bismark's seed-match mapping strategy, the use of FASTQ-scores, and the ability to tolerate missing bases, shown in the sequence as N's[71]. For further investigation, I BLASTed some randomly chosen reads that were previously unmapped by both BLAST and Bismark against the nucleotide collection to find the closest match from mouse sequence or any contaminating DNA. I concluded that low mapping efficiency results from lower sample quality and not from biases in the preprocessing or mapping. The comparably low yield using a local alignment also means that trimming and clipping of read borders cannot improve the mapping yield from the data further. To test for biases related to preprocessing and read mapping, I used BLAST+[71] to perform local alignments of randomly sampled reads. In order to simulate methylation-aware mapping, C-to-T and G-to-A conversions were done on both query sequence and the reference genomes and cross-aligned accordingly for both strands. Average mapping efficiency from Bismark was compared with BLAST results and found to be as good or better, as follows: BLAST could map 6.8% of reads in this fashion and 6.2% if hits were limited by allowing a maximum of 4 mismatches.

The fact that an even lower mapping yield from BLAST, as compared to Bismark/Bowtie2, was observed, could be attributed to its lack of a seed-match mapping strategy, lack of awareness of FASTQ-scores, inability to tolerate N's and BLAST's inexact matching algorithm. To investigate the potential origin of unmappable sequences, BLAST queries of some random previously unmapped sequences against the nucleotide collection were made to find the closest match. Some of the matches to the reads were bacteria such as Leptosphaeria, matches to human and ape DNA (suggesting reads with degenerated mouse sequence) and phage X174 (whose DNA was used for spike-in material). Therefore, this demonstrated that low mapping efficiency stems from base-calling quality and is not due to biases in preprocessing or mapping options or software. The comparably low yield using a local alignment also means that trimming and clipping of read borders could not have been improved further to yield better results from the data.

1.4.0.8 A sample run of BEAT

The BEAT package can be used for estimating the true methylation levels of BS-Seq samples and for the calling of epimutations, which are differences in methylation states of a region in the genome. The input for the package BEAT consists of count data in the form of counts for unmethylated and methylated cytosines per genomic position, which are then grouped into genomic regions of sufficient coverage in order to allow for low-coverage samples to be analyzed. Methylation rates of each region are then modeled using the BD-model in order to adjust for experimental bias. Pooling single CG counts into regions can be done with the function positions to regions, which reads a comma separated file (CSV) and outputs a data.frame. The latter is the input to the BEAT model, which is accessed via the core function generate_results. It was assumed that all counts at a single CG position were obtained from pairwise different bisulfite converted DNA templates, representing independent observations. Some of the most important parameters of the model are the false positive and false negative conversion rates. With the resulting estimates of methylation levels and methylation status from the model, which are returned as a data.frame, the function epimutation calls can then determine epimutation differences between two samples and compute rates for demethylating and methylating epimutations.

The package BEAT expects as input one csv per sample with counts for unmethylated and methylated cytosines per genomic position. Such data can be obtained from the output of BS-Seq mapping tools such as bismark[37], followed by simple script-based data processing. BS-Seq data for methylated- and unmethylated counts per genomic position needs to be formatted into a data.frame object with the columns: 'chr', 'pos', 'meth' and 'unmeth' (signifying chromosome name, chromosomal position, as well as methylated and unmethylated cytosine counts at that chromosomal position).

1.4.0.8.1 Reading input files

```
# Set working path

localpath <- system.file('extdata', package = 'BEAT')

# Load sample data. Alternatively, sample data is available via command: data(BEAT)

positions <- read.csv(file.path(localpath, "sample.positions.csv"))

head(positions)

chr pos meth unmeth

1 chr1 100001310 1 0

2 chr1 100002648 0 1

3 chr1 100002688 0 1

4 chr1 100002802 1 0

5 chr1 100004564 0 1

6 chr1 100004606 0 1
```

1.4.0.8.2 Configuration and parameters BEAT can process multiple samples at a time. For each sample specified under 'sampNames', the package BEAT reads this data frame from a csv from the working directory, which is specified by the parameter 'localpath'. The sample should be called <samplename>.positions.csv and should contain the aforementioned csv under the name 'positions'. Result samples written by BEAT will have the names <samplename>.results.RData. A distinction is made between two samples whose methylation status is compared against that of the reference. For each sample, the assignment of reference or non-reference status to samples is done via the vector ’is.reference’, which is indexed by the ’sampNames’ vector and contains one TRUE entry for the reference and one to many FALSE entries for samples to be compared against the reference. The example data set contains two samples, one named ’reference’ for a reference consisting of a mixture of cells, and one named ’sample’ for a single-cell sample to be compared against the reference.

```
# Set sample names and prefix of data files
sampNames <- c(’reference’,’sample’)
# Set reference vs. non-ref status per sample
is.reference <- c(TRUE, FALSE)
```

For the modeling part of BEAT, Bisulfite conversion rates have to be specified per sample using the vector ’convrates’, which is indexed by ’sampNames’. These conversion rates need to be specified manually by the user. Practically, for mammalian somatic cell samples, these rates can be estimated for each sample by looking at the non-CpG methylation rate per sample and using the inverse value as estimated CpG methylation rate, because non-CpG methylation in these types of cells is expected to be near zero. For example, if non-CpG methylation in a sequenced sample was measured to be 0.1 then BS-conversion rate for that sample would be set to 0.9 when near-zero non-CpG methylation can be expected, as is the case in somatic mammalian cells.

```
# Set BS-conversion rate per sample
pplus <- c(0.2, 0.5)
convrates <- 1 - pplus
```

The aforementioned values are then set in a parameter object, which is used throughout the further workflow in this package. It provides the following additional options:

- 1-convrates represents the fraction of unmethylated counts that are falsely called as methylated due to incomplete BS-conversion. This parameter is also referred to as ’pplus’ in the statistical model.
- ’pminus’ represents the fraction of methylated counts that are falsely called as unmethylated.
- ’regionSize’ is the size of regions into which genomic positions are grouped. In single cells, the number of positions with sufficient coverage for reliable statistical predictions may be very small. Therefore, the pipeline applies epimutation calling to regions instead of single positions. Single positions are pooled into regions of appropriate size, i.e., regions containing sufficiently many CpG positions that have a positive read count number in both the reference and the single cell sample. The method has then sufficient power to reliably detect epimutation events affecting these regions.

- After pooling CG positions to regions, there may still be regions with low count numbers that do not allow for reliable downstream analysis. Regions with less than 'minCounts' counts will be removed from further processing, potentially saving significant processing time in further analysis steps.

The following parameters are of minor importance, for a standard analysis, they can be left at their pre-set values.

- 'verbose' is an option that prints additional information during computation when set to TRUE.
- 'computeRegions' is an option that will recompute the regions from given positional input if set to TRUE; otherwise, it will depend on existing region samples already present in the 'localpath' directory (sample names ending in regions.RData).
- 'computeMatrices' is an option that will recompute the model data from given regions if set to TRUE; otherwise, it will depend on existing model output samples already present in the 'localpath' directory (sample names ending in convMat.RData and results.RData.
- 'writeEpicallMatrix' is an option that can be set to TRUE to generate epimutation calling output in the form of matrices (one row per genomic position, output format is CSV).

# Create parameter object

params = makeParams(localpath, sampNames, convrates, is.reference, pminus = 0.2, regionSize = 10000, minCounts = 5)

1.4.0.8.3 Pooling CG positions into regions The supplied methylation counts for individual CG positions are grouped into regions by BEAT for modeling according to the specified 'regionSize' and 'minCount' parameters. The function positions to regions takes as input the samples as csv objects located under <samplename>.positions.csv in the working directory, as discussed above. The output of the function is a list of data.frames of resulting counts per genomic regions. Each data.frame object contains a list of genomic regions covered by the given samples, consisting of the columns: 'chr', 'start', 'stop', 'meth', 'unmeth' (signifying chromosome name, start of region by chromosomal position, end of region by chromosomal position, as well as methylated- and unmethylated cytosine counts at that chromosomal position). Additionally, the output for each individual sample is saved in the working directory under <samplename>.regions.<regionSize>.<minCounts>.RData, with samplename, regionSize and minCounts replaced by the sample name and the respective parameters given.

# Pool CG positions into genomic regions

positions_to_regions(params)

```
Sample: reference.positions.csv regionSize: 10000 minCounts: 5
Processed reference.positions.csv, yielding 16728 regions of 10000 nt 5
Sample: sample.positions.csv regionSize: 10000 minCounts: 5
Processed sample.positions.csv, yielding 15948 regions of 10000 nt
# Model methylation levels and -status
results <- generate_results(params)
Sample: reference Generating conversion matrix...
Adding methylation states and 9 matrices...
Sample: sample Generating conversion matrix...
Adding methylation states and 9 matrices...
```

1.4.0.8.4 Calling epimutations Finally, epimutations are called by comparing two results objects from the previous step. The function 'epimutation calls' compares one sample to a reference sample. A methylating epimutation is called for each genomic region common to both samples when it was called as unmethylated in the reference and as methylated in the other sample by the model. Accordingly, the region is called as demethylating epimutation when called as methylated in the reference and as unmethylated in the other sample. Epimutation rates are then computed as the frequency of epimutation events in relation to the total regions shared between each sample and the reference. The resulting epimutations are saved as data.frames, for each sample one object for methylating epimutations and one for demethylating epimutations in the working directory as RData objects. Each data.frame object contains a list of genomic regions covered by the given samples, consisting of the columns: 'chr', 'pos', 'endpos', 'meth', 'unmeth', 'methstate', signifying chromosome name, start of region by chromosomal position, end of region by chromosomal position, methylated- and unmethylated cytosine counts at that chromosomal position and the methylation state, which is 1 for methylated positions and -1 for unmethylated positions.

```
# Call epimutations
epiCalls <- epimutation_calls(params)
Statistics for sample: sample (min. coverage: 5 reads/site)
Total shared CG sites(=regions!) between sample and reference/CTRL: 14492
Median CG sites of these total shared regions (REFERENCE): 23
Median CG sites of these total shared regions (SINGLE CELL): 13
Control has 10164 fully methylated and 967 fully unmethylated sites
Methylating Epimutation Rate: 0.00283
```

```
Demethylating Epimutation Rate: 0.47661
Total Epimutation Rate: 0.47944
Fully meth in single: 606, Fully unmeth in single: 9985
Written epi-methylation calls to matrix: sample.methEpicalls.10000.5p+=0.5p-=0.2.RData
Written epi-demethylation calls to matrix: sample.demethEpicalls.10000.5p+=0.5p-=0.2.RData
```

1.4.0.9 Summary statistics and visualization of the results

Fig. 1.20 graphically illustrates the output of BEAT by showing methylation estimates and epimutation calls for a sample of regions.

Table of Abbreviations

5mC 5'-methyl-cytosine, methylated cytosine

BS-Seq Bisulfite sequencing

CpG Cytosine followed by guanine, the most frequent genomic nucleotide combination in which methylated cytosines are observed

CG see CpG

CGI CpG island

DNMT DNA methyltransferase

RRBS Reduced Representation Bisulfite sequencing

scRRBS single-cell Reduced Representation Bisulfite sequencing

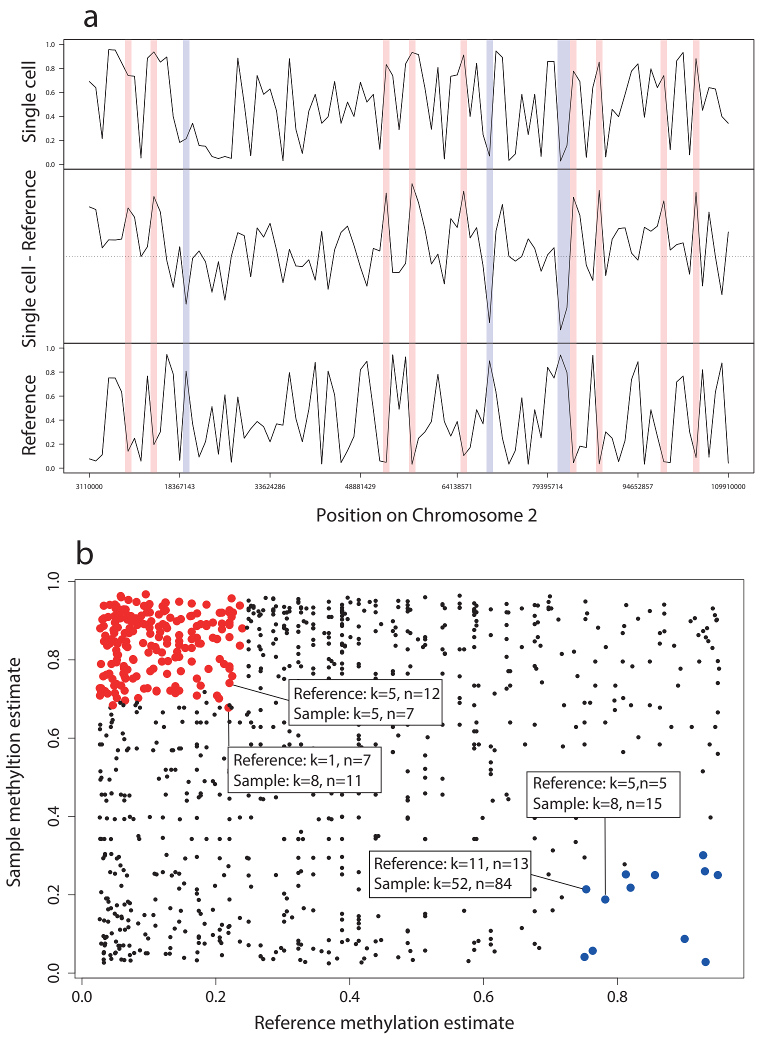

Figure 1.20: a) Methylation estimates and epimutation calls on a DNA segment. For all regions with sufficient read coverage, the black curves show the methylation estimates for a single cell sample (top), a reference sample (bottom), and their difference (middle). Regions with methylating epimutations are marked in red, while regions with demethylating epimutations are marked in blue. Samples used for our analysis in this paper were obtained from neuronal cells of young mice (data unpublished). b) Scatter plot of methylation estimates of a multi-cell reference sample (x-axis) versus those of a sample (y-axis) for all common regions with sufficient coverage. Each dot represents a single region that is covered by both samples. Red dots indicate methylating epimutations in the sample, while blue dots indicate demethylating epimutations in the sample. Four dots representing exemplary regions with epimutations at the corresponding boundary value ranges for demethylating- and methylating epimutations have been annotated with their values of methylated (k) and total (n) counts. Note that there exists no boundary line separating the red respectively the blue region, because our Bayesian model assigns different methylation estimates to tuples (k1,n1), (k2,n2) with equal empirical methylation level k1/n1=k2/n2.

Bibliography

[1] Lavie, Laurence and Kitova, Milena and Maldener, Esther and Meese, Eckart and Mayer, Jens (2005) CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2) Journal of Virology 79(2): p 876–883 PMID: 15613316.

[2] Johnson, Adiv A. and Akman, Kemal and Calimport, Stuart R.G. and Wuttke, Daniel and Stolzing, Alexandra and de Magalhaes, Joao Pedro (2012) The role of DNA methylation in aging, rejuvenation, and age-related disease Rejuvenation Research 15(5): p 483–494 PMID: 23098078 PMCID: PMC3482848.

[3] Shapiro, Ehud and Biezuner, Tamir and Linnarsson, Sten (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science Nature reviews Genetics 14(9): p 618–630 PMID: 23897237.

[4] Guo, Hongshan and Zhu, Ping and Wu, Xinglong and Li, Xianlong and Wen, Lu and Tang, Fuchou (2013) Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing Genome research 23(12): p 2126–2135 PMID: 24179143 PMCID: PMC3847781.

[5] Gravina, Silvia and Vijg, Jan Personal communication with silvia gravina and jan vijg Tech. rep. Department of Genetics, Albert Einstein College of Medicine Bronx, NY 10461, USA.

[6] Akman, Kemal and Haaf, Thomas and Gravina, Silvia and Vijg, Jan and Tresch, Achim (2014) Genomewide, quantitative analysis of DNA methylation from bisulfite sequencing data Bioinformatics Oxford England PMID: 24618468.

[7] Ramsahoye, Bernard H. and Biniszkiewicz, Detlev and Lyko, Frank and Clark, Victoria and Bird, Adrian P. and Jaenisch, Rudolf (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a Proceedings of the National Academy of Sciences of the United States of America 97(10): p 5237–5242 PMID: 10805783 PMCID: PMC25812.

[8] Ng, H H and Bird, A (1999) DNA methylation and chromatin modification Current opinion in genetics development 9(2): p 158–163 PMID: 10322130.

[9] Caiafa, Paola and Zampieri, Michele (2005) DNA methylation and chromatin structure: the puzzling CpG islands Journal of cellular biochemistry 94(2): p 257–265 PMID: 15546139.

[10] Jaenisch, Rudolf and Bird, Adrian (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals Nature genetics 33 Suppl: p 245–254 PMID: 12610534.

[11] Chatterjee, Aniruddha and Stockwell, Peter A. and Rodger, Euan J. and Morison, Ian M. (2012) Comparison of alignment software for genome-wide bisulphite sequence data pgks150 PMID: 22344695.

[12] Trzyna, Elzbieta and Duleba, Marcin and Faryna, Marta and Majka, Marcin (2012) Regulation of transcription in cancer Frontiers in bioscience Landmark edition 17: p 316–330 PMID: 22201746.

[13] Deaton, Aimee M. and Bird, Adrian (2011) CpG islands and the regulation of transcription Genes Development 25(10): p 1010–1022 PMID: 21576262.

[14] Kass, Stefan U. and Pruss, Dmitry and Wolffe, Alan P. (1997) How does DNA methylation repress transcription? Trends in Genetics 13(11): p 444–449.

[15] Huh, Iksoo and Zeng, Jia and Park, Taesung and Yi, Soojin V. (2013) DNA methylation and transcriptional noise Epigenetics Chromatin 6(1): p 9 PMID: 23618007.

[16] Smith, Zachary D. and Meissner, Alexander (2013) DNA methylation: roles in mammalian development Nature Reviews Genetics 14(3): p 204–220.

[17] Santos, Fatima and Hendrich, Brian and Reik, Wolf and Dean, Wendy (2002) Dynamic reprogramming of DNA methylation in the early mouse embryo Developmental Biology 241(1): p 172–182.

[18] Paulsen, Martina and Ferguson-Smith, Anne C. (2001) DNA methylation in genomic imprinting, development, and disease The Journal of Pathology 195(1): p 97–110.

[19] Li, En and Beard, Caroline and Jaenisch, Rudolf (1993) Role for DNA methylation in genomic imprinting Nature 366(6453): p 362–365.

[20] Richardson, Bruce (2003) Impact of aging on DNA methylation Ageing Research Reviews 2(3): p 245–261.

[21] Zykovich, Artem and Hubbard, Alan and Flynn, James M. and Tarnopolsky, Mark and Fraga, Mario F. and Kerksick, Chad and Ogborn, Dan and MacNeil, Lauren and Mooney, Sean D. and Melov, Simon (2013) Genome-wide DNA methylation changes with age in disease-free human skeletal muscle p1474–9726.

[22] Horvath, Steve and Zhang, Yafeng and Langfelder, Peter and Kahn, RenÃ© S. and Boks, Marco PM and Eijk, Kristel van and Berg, Leonard H. van den and Ophoff, Roel A. (2012) Aging effects on DNA methylation modules in human brain and blood tissue Genome Biology 13(10): p R97 PMID: 23034122.

[23] Irier, Hasan A and Jin, Peng (2012) Dynamics of DNA methylation in aging and alzheimer's disease DNA and cell biology 31 Suppl 1: p S42–48 PMID: 22313030.

[24] Clark, Susan J and Statham, Aaron and Stirzaker, Clare and Molloy, Peter L and Frommer, Marianne (2006) DNA methylation: bisulphite modification and analysis Nature protocols 1(5): p 2353–2364 PMID: 17406479.

[25] Raizis, A M and Schmitt, F and Jost, J P (1995) A bisulfite method of 5-methylcytosine mapping that minimizes template degradation Analytical biochemistry 226(1): p 161–166 PMID: 7785768.

[26] Meissner, Alexander and Gnirke, Andreas and Bell, George W and Ramsahoye, Bernard and Lander, Eric S and Jaenisch, Rudolf (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis Nucleic acids research 33(18): p 5868–5877 PMID: 16224102.

[27] Frommer, M. and McDonald, L. E. and Millar, D. S. and Collis, C. M. and Watt, F. and Grigg, G. W. and 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 89(5): p 1827–1831 PMID: 1542678.

[28] Grunau, C. and Clark, S. J. and Rosenthal, A. (2001) Bisulfite genomic sequencing: systematic investigation of critical experimental parameters Nucleic Acids Research 29(13): p e65–e65 PMID: 11433041.

[29] Olek, A and Oswald, J and Walter, J (1996) A modified and improved method for bisulphite based cytosine methylation analysis. Nucleic Acids Research 24(24): p 5064–5066 PMID: 9016686 PMCID: PMC146326.

[30] Wojdacz, Tomasz K. and Dobrovic, Alexander (2007) Methylation-sensitive high resolution melting (MS-HRM): a new approach for sensitive and high-throughput assessment of methylation Nucleic Acids Research 35(6): p e41 PMID: 17289753 PMCID: PMC1874596.

[31] Bianco, T and Hussey, D and Dobrovic, A (1999) Methylation-sensitive, single-strand conformation analysis (MS-SSCA): a rapid method to screen for and analyze methylation Human mutation 14(4): p 289–293 PMID: 10502775.

[32] Zhang, Yingying and Rohde, Christian and Tierling, Sascha and Stamerjohanns, Heinrich and Reinhardt, Richard and Walter, Joern and Jeltsch, Albert (2009) in: , DNA Methylation no p 177–187.

[33] Tost, Joerg and Dunker, Jenny and Gut, Ivo Glynne (2003) Analysis and quantification of multiple methylation variable positions in CpG islands by pyrosequencing BioTechniques 35(1): p 152–156 PMID: 12866415.

[34] Li, Ning and Ye, Mingzhi and Li, Yingrui and Yan, Zhixiang and Butcher, Lee M and Sun, Jihua and Han, Xu and Chen, Quan and Zhang, Xiuqing and Wang, Jun (2010) Whole genome DNA methylation analysis based on high throughput sequencing technology Methods San Diego Calif 52(3): p 203–212 PMID: 20430099.

[35] Aberg, Karolina A and McClay, Joseph L and Nerella, Srilaxmi and Xie, Lin Y and Clark, Shaunna L and Hudson, Alexandra D and Bukszar, Jozsef and Adkins, Daniel and Consortium, Swedish Schizophrenia and Hultman, Christina M and Sullivan, Patrick F and Magnusson, Patrik K E and van den Oord, Edwin J C G (2012) MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case–control samples Epigenomics 4(6): p 605–621 PMID: 23244307.

[36] Gu, Hongcang and Bock, Christoph and Mikkelsen, Tarjei S and Jaeger, Natalie and Smith, Zachary D and Tomazou, Eleni and Gnirke, Andreas and Lander, Eric S and Meissner, Alexander (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution Nature methods 7(2): p 133–136 PMID: 20062050.

[37] Krueger, Felix and Andrews, Simon R. (2011) Bismark: a flexible aligner and methylation caller for bisulfite-seq applications Bioinformatics 27(11): p 1571–1572 PMID: 21493656.

[38] Langmead, Ben and Trapnell, Cole and Pop, Mihai and Salzberg, Steven L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome Genome Biology 10(3): p 25.

[39] Xi, Yuanxin and Li, Wei (2009) BSMAP: whole genome bisulfite sequence MAPping program BMC Bioinformatics 10(1): p 232 PMID: 19635165.

[40] Chen, Pao-Yang and Cokus, Shawn J. and Pellegrini, Matteo (2010) BS seeker: precise mapping for bisulfite sequencing BMC Bioinformatics 11(1): p 203 PMID: 20416082.

[41] Fraga, Mario F. and Ballestar, Esteban and Paz, Maria F. and Ropero, Santiago and Setien, Fernando and Ballestar, Maria L. and Heine-Suner, Damia and Cigudosa, Juan C. and Urioste, Miguel and Benitez, Javier and Boix-Chornet, Manuel and Sanchez-Aguilera, Abel and Ling, Charlotte and Carlsson, Emma and Poulsen, Pernille and Vaag, Allan and Stephan, Zarko and Spector, Tim D. and Wu, Yue-Zhong and Plass, Christoph and Esteller, Manel (2005) Epigenetic differences arise during the lifetime of monozygotic twins Proceedings of the National Academy of Sciences of the United States of America 102(30): p 10604–10609 PMID: 16009939.

[42] Fraga, Mario F and Ballestar, Esteban and Montoya, Guillermo and Taysavang, Panya and Wade, Paul A and Esteller, Manel (2003) The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties Nucleic acids research 31(6): p 1765–1774 PMID: 12626718 PMCID: PMC152853.

[43] Wu, Susan C. and Zhang, Yi (2010) Active DNA demethylation: many roads lead to rome Nature Reviews Molecular Cell Biology 11(9): p 607–620.

[44] Ito, Shinsuke and Shen, Li and Dai, Qing and Wu, Susan C and Collins, Leonard B and Swenberg, James A and He, Chuan and Zhang, Yi (2011) Tet proteins can

convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine Science New York NY 333(6047): p 1300–1303 PMID: 21778364 PMCID: PMC3495246.

[45] Qin, Weihua and Leonhardt, Heinrich and Pichler, Garwin (2011) Regulation of DNA methyltransferase 1 by interactions and modifications Nucleus Austin Tex 2(5): p 392–402 PMID: 21989236.

[46] Lee, Byron H and Yegnasubramanian, Srinivasan and Lin, Xiaohui and Nelson, William G (2005) Procainamide is a specific inhibitor of DNA methyltransferase 1 The Journal of biological chemistry 280(49): p 40749–40756 PMID: 16230360 PMCID: PMC1989680.

[47] Bocklandt, Sven and Lin, Wen and Sehl, Mary E. and Sanchez, Francisco J. and Sinsheimer, Janet S. and Horvath, Steve and Vilain, Eric (2011) Epigenetic predictor of age PLoS ONE 6(6): p e14821.

[48] Ahuja, N and Issa, J P (2000) Aging, methylation and cancer Histology and histopathology 15(3): p 835–842 PMID: 10963127.

[49] Mastroeni, Diego and McKee, Ann and Grover, Andrew and Rogers, Joseph and Coleman, Paul D. (2009) Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for alzheimer's disease PLoS ONE 4(8): p e6617.

[50] Chouliaras, Leonidas and van den Hove, Daniel L A and Kenis, Gunter and Keitel, Stella and Hof, Patrick R and van Os, Jim and Steinbusch, Harry W M and Schmitz, Christoph and Rutten, Bart P F (2012) Age-related increase in levels of 5-hydroxymethylcytosine in mouse hippocampus is prevented by caloric restriction Current Alzheimer research 9(5): p 536–544 PMID: 22272625 PMCID: PMC3561726.

[51] Egger, Gerda and Liang, Gangning and Aparicio, Ana and Jones, Peter A (2004) Epigenetics in human disease and prospects for epigenetic therapy Nature 429(6990): p 457–463 PMID: 15164071.

[52] Batra, Vipen and Sridhar, Swathi and Devasagayam, Thomas Paul Asir (2010) Enhanced one-carbon flux towards DNA methylation: Effect of dietary methyl supplements against gamma-radiation-induced epigenetic modifications Chemico biological interactions 183(3): p 425–433 PMID: 19931232.

[53] Genereux, Diane P and Johnson, Winslow C and Burden, Alice F and Stoeger, Reinhard and Laird, Charles D (2008) Errors in the bisulfite conversion of DNA: modulating inappropriate- and failed-conversion frequencies Nucleic acids research 36(22): p e150 PMID: 18984622 PMCID: PMC2602783.

[54] Krueger, Felix and Kreck, Benjamin and Franke, Andre and Andrews, Simon R (2012) DNA methylome analysis using short bisulfite sequencing data Nature methods 9(2): p 145–151 PMID: 22290186.

[55] Dinh, Huy Q. and Dubin, Manu and Sedlazeck, Fritz J. and Lettner, Nicole and Mittelsten Scheid, Ortrun and von Haeseler, Arndt (2012) Advanced methylome

analysis after bisulfite deep sequencing: An example in arabidopsis PLoS ONE 7(7): p e41528.

[56] Schneider, Eberhard and Pliushch, Galyna and El Hajj, Nady and Galetzka, Danuta and Puhl, Alexander and Schorsch, Martin and Frauenknecht, Katrin and Riepert, Thomas and Tresch, Achim and Mueller, Annette M and Coerdt, Wiltrud and Zechner, Ulrich and Haaf, Thomas (2010) Spatial, temporal and interindividual epigenetic variation of functionally important DNA methylation patterns Nucleic acids research 38(12): p 3880–3890 PMID: 20194112 PMCID: PMC2896520.

[57] Gelman, Andrew and Carlin, John B. and Stern, Hal S. and Rubin, Donald B. (2003) Bayesian Data Analysis, Second Edition CRC Press.

[58] Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes Journal of Molecular Biology 196(2): p 261–282.

[59] Song, Fei and Smith, Joseph F and Kimura, Makoto T and Morrow, Arlene D and Matsuyama, Tomoki and Nagase, Hiroki and Held, William A (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression Proceedings of the National Academy of Sciences of the United States of America 102(9): p 3336–3341 PMID: 15728362.

[60] Wang, Kai and Li, Mingyao and Hadley, Dexter and Liu, Rui and Glessner, Joseph and Grant, Struan F A and Hakonarson, Hakon and Bucan, Maja (2007) PennCNV: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data Genome research 17(11): p 1665–1674 PMID: 17921354.

[61] Rabiner, L. and Juang, B. (1986) An introduction to hidden markov models ASSP Magazine IEEE 3(1): p 4 –16.

[62] Ziller, Michael J and Mueller, Fabian and Liao, Jing and Zhang, Yingying and Gu, Hongcang and Bock, Christoph and Boyle, Patrick and Epstein, Charles B and Bernstein, Bradley E and Lengauer, Thomas and Gnirke, Andreas and Meissner, Alexander (2011) Genomic distribution and inter-sample variation of non-CpG methylation across human cell types PLoS genetics 7(12): p e1002389 PMID: 22174693.

[63] Heyn, Holger and Li, Ning and Ferreira, Humberto J. and Moran, Sebastian and Pisano, David G. and Gomez, Antonio and Diez, Javier and Sanchez-Mut, Jose V. and Setien, Fernando and Carmona, F. Javier and Puca, Annibale A. and Sayols, Sergi and Pujana, Miguel A. and Serra-Musach, Jordi and Iglesias-Platas, Isabel and Formiga, Francesc and Fernandez, Agustin F. and Fraga, Mario F. and Heath, Simon C. and Valencia, Alfonso and Gut, Ivo G. and Wang, Jun and Esteller, Manel (2012) Distinct DNA methylomes of newborns and centenarians p201120658 PMID: 22689993.

[64] Busuttil, Rita A and Garcia, Ana Maria and Reddick, Robert L and Dolle, Martijn E T and Calder, Robert B and Nelson, James F and Vijg, Jan (2007)

Intra-organ variation in age-related mutation accumulation in the mouse PloS one 2(9): p e876 PMID: 17849005 PMCID: PMC1964533.

[65] Schmitz, Robert J. and Schultz, Matthew D. and Lewsey, Mathew G. and O Malley, Ronan C. and Urich, Mark A. and Libiger, Ondrej and Schork, Nicholas J. and Ecker, Joseph R. (2011) Transgenerational epigenetic instability is a source of novel methylation variants Science 334(6054): p 369–373 PMID: 21921155.

[66] Becker, Claude and Hagmann, Joerg and Mueller, Jonas and Koenig, Daniel and Stegle, Oliver and Borgwardt, Karsten and Weigel, Detlef (2011) Spontaneous epigenetic variation in the arabidopsis thaliana methylome Nature 480(7376): p 245–249.

[67] Rodriguez, Jairo and Vives, Laura and Jorda, Mireia and Morales, Cristina and Munoz, Mar and Vendrell, Elisenda and Peinado, Miguel A. (2008) Genome-wide tracking of unmethylated DNA alu repeats in normal and cancer cells Nucleic Acids Research 36(3): p 770–784 PMID: 18084025 PMCID: PMC2241897.

[68] Pantel, Klaus and Alix-Panabieres, Catherine (2010) Circulating tumour cells in cancer patients: challenges and perspectives Trends in molecular medicine 16(9): p 398–406 PMID: 20667783.

[69] Schmieder, Robert and Edwards, Robert (2011) Quality control and preprocessing of metagenomic datasets Bioinformatics Oxford England 27(6): p 863–864 PMID: 21278185.

[70] Martin, Marcel (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads EMBnetjournal 17(1): p pp. 10–12.

[71] Langmead, Ben and Salzberg, Steven L (2012) Fast gapped-read alignment with bowtie 2 Nature methods 9(4): p 357–359 PMID: 22388286.

[72] Li, Heng and Handsaker, Bob and Wysoker, Alec and Fennell, Tim and Ruan, Jue and Homer, Nils and Marth, Gabor and Abecasis, Goncalo and Durbin, Richard and 1000 Genome Project Data Processing Subgroup (2009) The sequence Alignment/Map format and SAMtools Bioinformatics Oxford England 25(16): p 2078–2079 PMID: 19505943.

[73] Perier, Rouaida Cavin and Praz, Viviane and Junier, Thomas and Bonnard, Claude and Bucher, Philipp (2000) The eukaryotic promoter database (EPD) Nucleic Acids Research 28(1): p 302–303 PMID: 10592254.

[74] Fujita, Pauline A and Rhead, Brooke and Zweig, Ann S and Hinrichs, Angie S and Karolchik, Donna and Cline, Melissa S and Goldman, Mary and Barber, Galt P and Clawson, Hiram and Coelho, Antonio and Diekhans, Mark and Dreszer, Timothy R and Giardine, Belinda M and Harte, Rachel A and Hillman-Jackson, Jennifer and Hsu, Fan and Kirkup, Vanessa and Kuhn, Robert M and Learned, Katrina and Li, Chin H and Meyer, Laurence R and Pohl, Andy and Raney, Brian J and Rosenbloom, Kate R and Smith, Kayla E and Haussler, David and

Kent, W James (2011) The UCSC genome browser database: update 2011 Nucleic acids research 39(Database issue): p D876–882 PMID: 20959295 PMCID: PMC3242726.

[75] Camacho, Christiam and Coulouris, George and Avagyan, Vahram and Ma, Ning and Papadopoulos, Jason and Bealer, Kevin and Madden, Thomas L (2009) BLAST+: architecture and applications BMC bioinformatics 10: p 421 PMID: 20003500.