Freie Universität Berlin

Computational Analysis of Genome-wide Methylation Enrichment Experiments

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Matthias Lienhard

Berlin
2017

Betreuer:	Prof. Dr. Martin Vingron
Gutachter:	Prof. Dr. Dr. Michal-Ruth Schweiger
Tag der Disputation:	09.10.2017

PREFACE

COLLABORATIONS AND PUBLICATIONS

The methods and analysis described in this thesis emerged from collaboration with other researchers, and have been published in several peer-reviewed articles. The transformation of relative enrichment to absolute levels of methylation, described in Chapter 4, was published in [Lienhard et al., 2016]. The application of likelihood ratio tests based on nested generalized linear models to detect differentially methylated regions, described in Chapter 5, has been introduced in [Lienhard et al., 2014]. Furthermore, this method has been adapted to detect differentially enriched histone ChIP seq signals, which is explained in a detailed protocol [Lienhard and Chavez, 2016]. The development of the methods was mainly driven by two studies, in which I contributed to the analysis. The first study used methylation enrichment to analyze the methylome of mouse intestinal adenoma, to find epigenetic mechanisms involved in human colon cancer genesis [Grimm et al., 2013]. In the second study, the same assay was used to profile the methylome of lung cancer patients to predict therapy resistance [Grasse et al., in preparation]. I gratefully acknowledge all project partners, in particular Christina Grimm and Sabrina Grasse, who performed the methylation experiments for the two studies respectively. These studies and further applications of the methods developed in this theses are presented in Chapter 7. A complete list of publications, including those not related to this thesis, are listed in the curriculum vitae in the Appendix.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my thesis adviser, Ralf Herwig, for support and scientific advice during my time as PhD candidate in his group at the Max Planck Institute for Molecular Genetics in Berlin. The profound methodological expertise of his group as well as the diverse collaborations emerging from his exceptional professional network provided the basis to the excellent and motivating learning and working environment. I am grateful to my supervisor, Martin Vingron, for the stimulating discussions but also for the trust and freedom he granted me to develop my ideas. Furthermore, I would like to thank him, as well as Michal Ruth Schweiger, for reviewing the thesis.

I thank my fellow students as well as all other researchers at the Department of Computational Biology for the scientific atmosphere and fruitful discussions during group meetings and institute seminars, but also during lunch and coffee breaks. I would also like to also acknowledge Kirsten Kelleher and Fabian Feutlinske for their friendly support in many organizational issues, and in particular for hosting the annual IMPRS retreats.

I feel deeply obliged to Anjana Rao for having given me the great opportunity to work in her group at the La Jolla Institute for Allergy and Immunology, San Diego, CA, USA. Special thanks goes to Lukas Chavez for supervising my research stay and for his continuous scientific support, advice, and fruitful collaboration beyond this time.

I am grateful for financial support by a IMPRS-CBSC scholarship from the Max Planck Society, by the German Federal Ministry of Education and Research within the e:BIO research project "epitreat", and by the European Commission within the 7th framework project "HeCaToS".

I would like to thank my friends Jacob Taylor and Bastian Kayser for proofreading and commenting on the thesis manuscript, and the anonymous reviewers of the publication manuscripts for their critiques, which helped improve the quality of my research.

It would never have been possible for me to prepare this work without the support of my dear family and friends, whom I owe my deepest gratitude – none more so than my beloved wife Nora Tandetzki and my beautiful daughter Frida, who are always by my side.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BS	bisulfite
CNV	copy-number variation
CDF	cumulative distribution function
DMR	differentially methylated region
ETI	equal-tailed interval
FDR	false discovery rate
GLM	generalized linear model
HDI	highest posterior density interval
LHB	large hypomethylated block
LR	likelihood ratio
MBD	methyl-CpG binding domain
MeDIP	methylated DNA immunoprecipitation
NSCLC	non-small cell lung cancer
PDX	patient derived xenograft
qPCR	quantitative polymerase chain reaction
QSEA	quantitative sequencing enrichment analysis
RFR	random forest regression

TCGA the cancer genome atlas

TSG tumor suppressor gene

1

INTRODUCTION

1.1 MOTIVATION

Each cell of an organism contains a complete copy of its genetic information. This information is stored in DNA molecules, encoded in the sequence of the 4 different bases, adenosine (A), cytosine (C), guanine (G) and tyrosine (T). These molecules contain instructions to build the biomolecules the cells need to function. The general flow of genetic information is described by the central dogma of molecular biology, which states that the sequence information of genes is transcribed from DNA to RNA molecules, and then translated into proteins. The set of proteins in a cell defines the function of the cell.

Regulation of gene expression is crucial for the development and maintenance of cell function, and enables the cell to react to intrinsic as well as extrinsic signals. A multicellular organism consists of many different cells of different types, fulfilling different functions and having different expression patterns. Given the fact that genetic information is static between all these cells, the dynamic properties cannot be explained by the genetic information alone.

Epigenetics refers to a set of regulatory mechanisms that are not encoded in the DNA sequence. One of these mechanisms is DNA methylation, which is the covalent addition of methyl groups to the bases of the DNA. This modification can regulate gene expression, for example by recruiting regulatory factors, which specifically bind methylated DNA, or by influencing the binding affinity of transcription factors at the modified genomic region.

In mammals, DNA methylation occurs at cytosines, which are followed by guanines (CpG). This reverse compatible sequence structure allows specific enzymes to copy methylation marks of the template to the newly synthesized daughter strand. This process maintains DNA methylation patterns during replication and conserves the cellular identity in the next generation of cells.

DNA methylation is closely controlled during development and plays a determining role in cell fate decisions. For example, targeted demethylation of a MyoD distal enhancer promotes reprogramming of fibroblasts into myoblasts and facilitates myotube formation [X. S. Liu et al., 2016]. Aberrant DNA methylation has been identified as a hallmark of many diseases, in particular cancer. For example, it has been shown in several human cancers that hypermethylation of the CDKN2A promoter silences the expression of the tumor suppressor gene [Herman et al., 1995]. In most cases however, the general principles and underlying mechanisms are unknown, and subject to current research.

DNA methylation assays based on enrichment of methylated DNA fragments provide genome-wide information on DNA methylation at reasonable costs. Therefore this approach is especially attractive for profiling DNA methylation in large sets of samples, such as clinical studies. However, dependence of the enrichment on local sequence composition, as well as experimental co-factors require specific normalization procedures. More importantly, in contrast to alternative assays, enrichment provides a relative measurement for local DNA methylation levels. This restricts the application on comparative studies, and limits interpretability of the results. Lack of credible and efficient computational methods overcoming these limitations has caused reduced trust in the quality and value of enrichment based assays compared to more expensive alternative approaches.

1.2 RESEARCH OBJECTIVE

This thesis aims at developing and establishing reliable computational methods for enrichment based methylation analysis, suitable for the application to large clinical studies, allowing both absolute and comparative interpretation of DNA methylation. To meet this aim, I address the following objectives: in order to analyze the relation of fragment enrichment and absolute levels of methylation and the dependence on further co-factors, I compare local fragment enrichment with absolute methylation levels measured by alternative approaches. Based on these observations, I induce a specific normalization procedure and a statistical model for fragment enrichment. Subsequently, I apply this model to derive estimators for the absolute methylation levels, based on a Bayesian approach. Next, I propose different methods to detect differentially methylated regions between groups of samples. I test the accuracy and practicality of all methods with clinically relevant datasets and showcase the application by presenting relevant studies.

1.3 THESIS OUTLINE

In Chapter 2, I provide a general introduction to the concept of epigenetics and present mechanisms controlling the structure and transcriptional activity of DNA by epigenetic marks, in particular histone modifications and DNA methylation. I focus on functional principles depending on these marks, as well as the enzymes and mechanisms involved in maintaining and changing patterns of epigenetic marks. More specifically, I introduce the functional role of epigenetic marks in development and disease.

Chapter 3 provides the foundation for the methods developed in this thesis. First, this chapter provides an overview of the two different experimental principles for the analysis of DNA methylation: i) conversion of cytosines by bisulfite treatment, and ii) enrichment of methlyated DNA. Next, I explain the methodological concepts that are important for the developed methods for the analysis of enrichment based DNA methylation assays. Furthermore, I introduce a comprehensive, practically relevant dataset, which is used throughout the thesis to analyze the properties of the data, to derive the models, and to assess the performance of the methods.

In Chapter 4, I propose an approach to estimating absolute levels of methylation from the enrichment which greatly enhances interpretability of the results and allows comparison with bisulfite based methylation assays. For this transformation, sample specific enrichment characteristics are modeled. Based on the enrichment characteristics, I construct a statistical model of the read counts within a genomic region, and derive Bayesian estimators for absolute methylation levels. Furthermore, I compare my approach to two alternative methods from the literature and show that it is highly accurate and outperforms these methods. Most importantly, in contrast to the alternatives, my approach is able to correctly quantify individual differences in methylation levels between patient samples, which is a prerequisite for clinical applicability of the technique. The method has been published in [Lienhard et al., 2016].

Chapter 5 presents two statistical approaches to detect differentially methylated regions (DMRs) between groups of samples: a general non-parametric approach, and an approach based on explicit statistical modeling of the read counts. I compare the capability of the two approaches by reference to their performance in detecting DMRs in a clinically relevant example dataset. The application of the methods on enrichment based sequencing data have been presented in several publications [Lienhard et al., 2014; Lienhard and Chavez, 2016; Lienhard et al., 2016].

Chapter 6 focuses on practical aspects of the analysis workflow, and the implementation of the relevant methods in two R/bioconductor packages MEDIPS and QSEA. I introduce metrics for quality control and show how the data can be depicted using exploratory analysis methods. Furthermore, I present implemented methods that help in assessing the functional effect of differentially methylated regions.

In Chapter 7, I demonstrate the practical applicability of the work-flows by presenting 4 studies, where the analysis was performed with the methods described here. For each study, I summarize the main findings and outline the relevance in the context of this thesis.

Taken together, this thesis provides a comprehensive collection of computational methods for the analysis of enrichment based methylation assays, that outperform alternative approaches. The methods have been intensively tested and validated with independent data and have become the default analysis work-flow for these assays within and beyond the Max Planck Institute for Molecular Genetics.

2

BIOLOGICAL BACKGROUND: EPIGENETICS

This chapter provides an introduction to the concept of epigenetics, and explains the molecular basis of two epigenetic mechanisms, histone modification and DNA methylation. Further, I describe experimental methods which allow the assessment of these epigenetic mechanisms.

Epigenetics has been defined as the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" [Russo et al., 1996]. Conceptually, epigenetics provides an explanation of how cells can interpret their genome in different contexts. For example, the cells of a multicellular organism, which are carrying identical DNA, can act as different cell types, or adapt to changes of the environmental condition. Epigenetics thus connects the genotype as well as environmental influences to the cellular phenotype.

Epigenetic regulation can be stable or dynamic. For example, in the early female embryo, hundreds of genes on one of the X chromosomes are inactivated. This event occurs randomly either on the paternal or maternal X chromosome, and the decision is stably maintained during development throughout the life-span of the organism. On the other hand, epigenetic mechanisms allow the cell to react to developmental or environmental stimuli, by temporarily changing gene expression. These changes can be inherited by the next cell generations and maintained over a certain time, in absence of the primary stimulus, before the previous state is restored. This dynamic behavior is described as epigenetic memory.

This functionality is implemented by multiple regulatory mechanisms, often involving chromatin based changes, such as DNA methylation, and histone modifications. These mechanisms include processes to set marks on the chromatin, and later recognize these marks and change the expression of genes accordingly. In order to accomplish stability, further processes ensuring maintenance and inheritability are required.

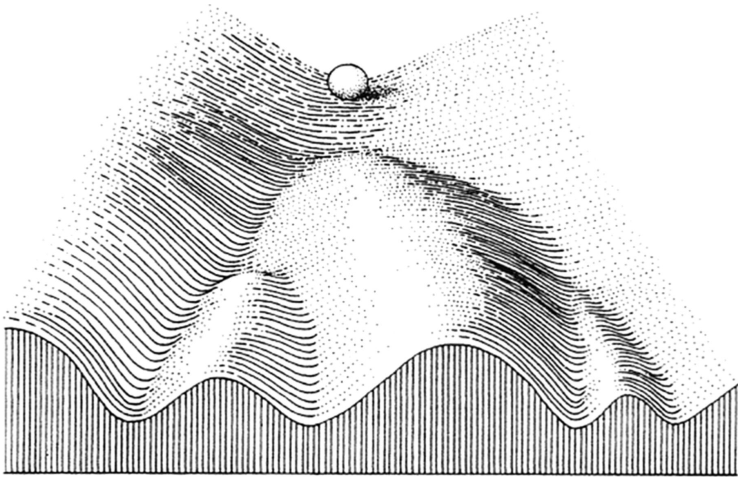


Figure 1: The epigenetic landscape, modified from [Waddington, 1957]

Recently, international efforts like the Roadmap Epigenomics Project [Kundaje et al., 2015], which collected and integrated different epigenetic marks in several tissues across the entire genome, have led to great progress in understanding the underlying processes of epigenetic regulation and its role in development and disease.

2.1 EPIGENETICS IN DEVELOPMENT

Starting as a single cell (zygote) formed by fertilization, the embryo develops to a mature multicellular organism consisting of a range of different cell types. During this development, epigenetic mechanisms guide the cells through various developmental pathways from totiplutent stage towards differentiation, in a highly ordered and reproducible manner. A precise description of epigenetic processes is central to a molecular description and understanding of embryo development.

In his classic illustration, Conrad H Waddington [Waddington, 1957] depicted the process of regulation during development as a landscape: the "epigenetic landscape" consists of a series of ridges and valleys, through which the cell traverses on its way from undifferentiated to fully differentiated tissue (Figure 1). A valley junction represents a decision point, at which the fate of a precursor cell is determined by choosing one or the other valley. The stability of this cell fate decision is reflected by the height of the ridge between the valleys.

Accordingly, for most cell types, the pattern of epigenetic marks form a stable characteristic epigenetic signature. However, at specific points in development, some cells undergo a major deletion of epigenetic marks, followed by establishment of a different set of marks. This ’epigenetic re-programming’ first happens during development of primordial germ cells (PGC), the precursor cells of gametes (Figure 2, [Patra et al., 2010; Cantone and Fisher, 2013]). At this stage, which marks the formation of the next generation, epigenetic marks involved in the process of imprinting are deleted and reset: while most genes are expressed from both paternal and maternal alleles a set of ’imprinted’ genes is expressed exclusively from one allele, in a parent-of-origin manner. For example, the gene of Insulin-like growth factor 2 (IGF2) is maternally imprinted, and thus expressed exclusively from the paternally inherited allele. Maintaining a parent-specific expression pattern of imprinted genes requires erasing the parental regulatory marks and re-establishing a specific maternal and paternal pattern within this first round of epigenetic re-programming, during formation of oocytes and sperm cells. This process also affects the paternally derived X chromosome in females, which is silenced in the zygote and during the first stage of development. However, in the zygote, a second round of epigenetic re-programming is initialized, resulting in genome-wide removal of DNA-methylation and other epigenetic marks, in order to re-acquire totipotency. After the first divisions, cells differentiate to form the first lineages: the trophoblast, which will form the placenta, and the inner cell mass (ICM), which will give rise to the definite structures of the fetus. While in the trophoblast the paternal X chromosome remains imprinted, in ICM removal of DNA-methylation leads to re-activation of the paternal X chromosome. Then, during gastulation, each of the cells randomly and independently deactivates one copy of the X chromosome. This deactivation is maintained during the lifetime of the cell and transferred to all descending cells.

2.2 EPIGENETICS IN DISEASE

Given the central role of epigenetic mechanisms in development, it is not surprising that alterations in the epigenome are associated with disease. Accordingly, aberrant epigenetic marks have been identified as a hallmark of many diseases, in particular cancer, neurological diseases and immune disorders [Falkenberg and Johnstone, 2014]. For example, down-regulation of tumor suppressor genes caused by focal hypermethylation of their promoters is a well described mechanism in the development of many cancer types [Sharma et al., 2010]. Another condition involving epigenetic mechanisms is fragile X syndrome, where a change in copy number and successive DNA hypermethylation of a repetitive region in the promoter of the FMR1 gene on the X chromosome lead to silencing of this gene, which in turn causes mental retardation [Verkerk et al., 1991]. Epigenetic mechanisms are also involved in Prader-Willi and Angelman syndromes, caused by a deletion on chromosome 15, which is controlled by imprinting: depend-

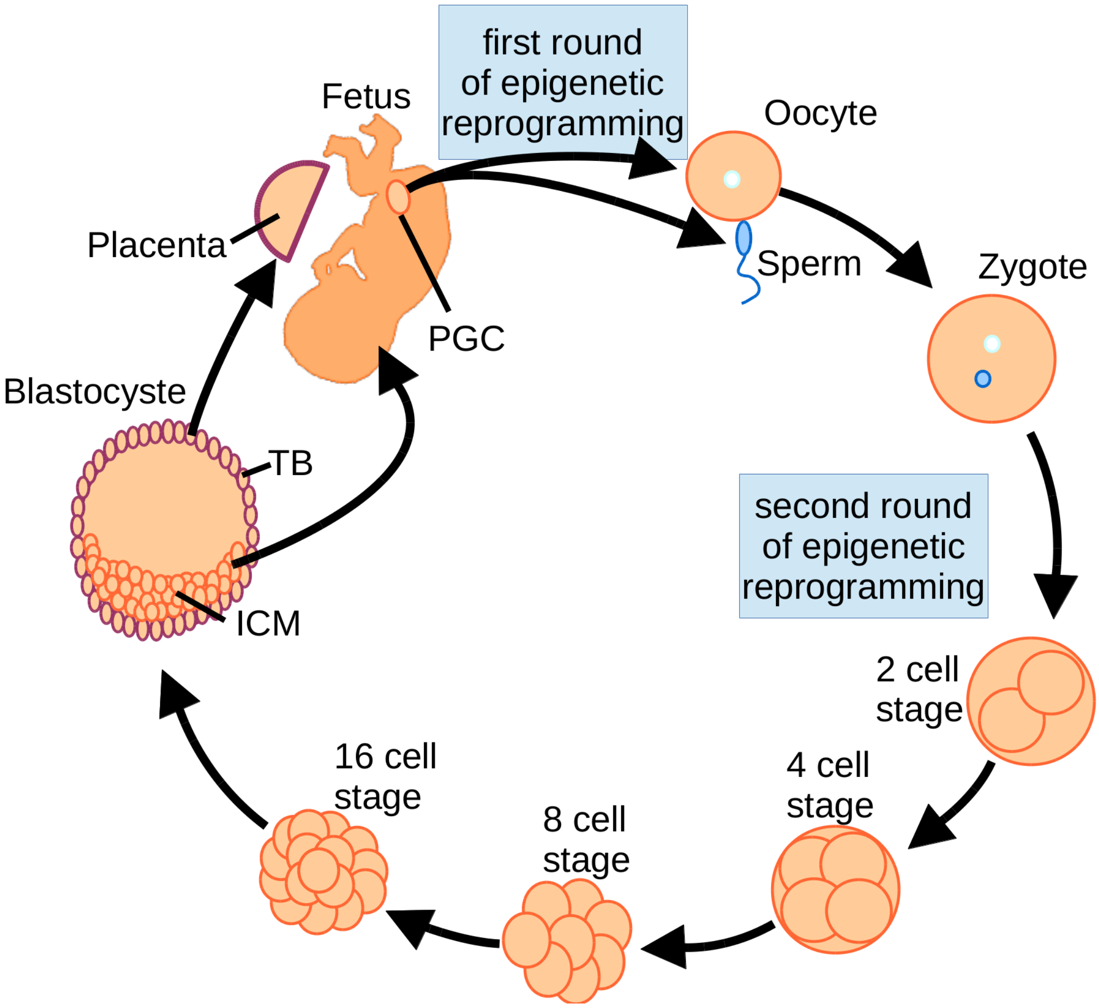


Figure 2: The two rounds of epigenetic reprogramming during embryo development. The first round is initiated during development of primordial germ cells (PGC), the second round in the zygote and during first cell divisions. At the blastocyst stage cells differentiate to the trophoblast lineage (TB), which will form the placenta, and the inner cell mass (ICM), which will give rise to the definite structures of the fetus.

ing on whether the deletion is inherited paternally or maternally, the phenotype of the deletion is either Prader-Willi or Angelman syndrome.

These diseases provide model systems for better understanding the role of epigenetic mechanisms in normal development. For example, specific hypermethylation of CpG rich promoters in cancer patients helps in assessing the influence of promoter hypermethylation on gene expression in general. Beyond revealing details about normal regulatory mechanisms, these studies also highlight pathologic mechanisms involved in specific diseases. For example, screening cancer patients for hypermethylated promoter regions allows the identification of candidate tumor suppressor genes. These approaches help further the understanding of the molecular basis of these diseases and the functional implication of epigenetic (dis)regulation.

Understanding the epigenetic basis of diseases has direct clinical impact: since the regulatory marks can potentially be observed before a phenotype is expressed, aberrant epigenetic marks in disease cohorts can be used as biomarkers for early diagnosis. For example, epi proColon provides a method for noninvasive early colon cancer diagnosis based on detection of DNA methylation at promoter of SEPT9 in floating tumor DNA purified from blood samples [Tetzner et al., 2009]. Furthermore, epigenetic biomarkers can be used to stratify patients in risk groups, to predict disease outcome, patient survival, or success chances of treatment alternatives [Heyn and Esteller, 2012].

However, the clinical potential of epigenetics goes beyond risk assessment and diagnosis: In contrast to genetic mutations, epigenetic changes are in principle reversible and thus bear a potential for therapy. So far, most agents modifying the epigenome have a broad genome-wide effect and are therefore of limited use in clinical practice. One of the rare examples of clinical application of genome-wide acting epigenetic modifiers is 5-aza-2'deoxycitidine (decitabine). This compound is a chemical analog of cytidine, which is built into the DNA and inhibits DNA methyltrasferase, causing genome-wide de-methylation, especially in rapidly proliferating cells. It is used in the chemotherapy of myelodysplastic syndrome [Kornblith et al., 2002], a type of cancer affecting precursors of blood cells.

Yet the development of sequence specific modifiers of epigenetic marks promises far more versatile use of epigenetics in clinics. Specificity can be achieved by fusing proteins capable of setting or removing epigenetic marks to zinc finger nucleases (ZFNs) [Beerli et al., 1998] or transcription activator-like effectors (TALEs) [F. Zhang et al., 2011], designed to bind a specific target sequence. However, engineering of sequence specific TALE or ZNF variants is a laborious process, and unspecific binding may lead to off target effects, opposing the safe use of these targeted modifiers as drugs. Recently, progress has been made by fusing nuclease deficient Cas9 to epigenetic modifiers [X. S. Liu et al., 2016]. Cas9 is guided by reverse complementary guide RNA, and thus can be easily targeted to a specific DNA sequence. With this promising technology at the doorstep, the systematic investigation of aberrant epigenetic marks in patient cohorts can also be considered as the search for potential modes of action for epigenome modifying drugs.

2.3 CHROMATIN STRUCTURE

In eucaryotic cells, genomic DNA forms a complex with proteins which is called chromatin. This complex fulfills several functions: it allows to condense the DNA in order to package DNA within the small volume of the nucleus and to preserve the integrity of the genome. The structure of chromatin also controls the function of the genome, as structurally loose chromatin ensures local accessibility of actively transcribed genes. The basic unit of the chromatin structure is the nucleosome; it consists of four dimeres of the core histone proteins, H2A, H2B, H3 and H4 (Figure 3A). These protein complexes are wrapped in approximately 146 nucleotide pairs of DNA, and are connected by linker DNA [Kornberg, 1977]. In the loose, open state, chromatin appears like 'beads on a string' (Figure 3C). This state is called euchromatin and DNA is accessible for transcription. For inactive regions of the genome, the chromatin is coiled in dense structure called heterochromatin (Figure 3D). This structure is not accessible to the transcriptional machinery, and thus the contained genes are silent.

In addition to the local structure, the spatial organization of the genome plays an important role in gene regulation. Enhancers are regulatory regions of the genome at which transcription factors can bind, and, in consequence, promote the expression of a target gene. The interaction of enhancer and promoter requires physical contact, which makes the regulation dependent on the 3D architecture of the genome. Recently, the development of experimental methods like Hi-C, which probes genomic contacts, has allowed the reconstruction of the 3D structure of the genome [Rao et al., 2014]. These studies suggest an organization of the genome in topological contact domains, formed by chromatin loops, which are thought to be mediated by CTCF binding at the boundaries of the domains [Splinter et al., 2006]. While the loops enable or reinforce interaction of regulatory elements and targets, the interaction to elements outside the loop are blocked. Therefore CTCF controls enhancer activity, either by bringing it to proximity of the target gene, or by insulating the interaction.

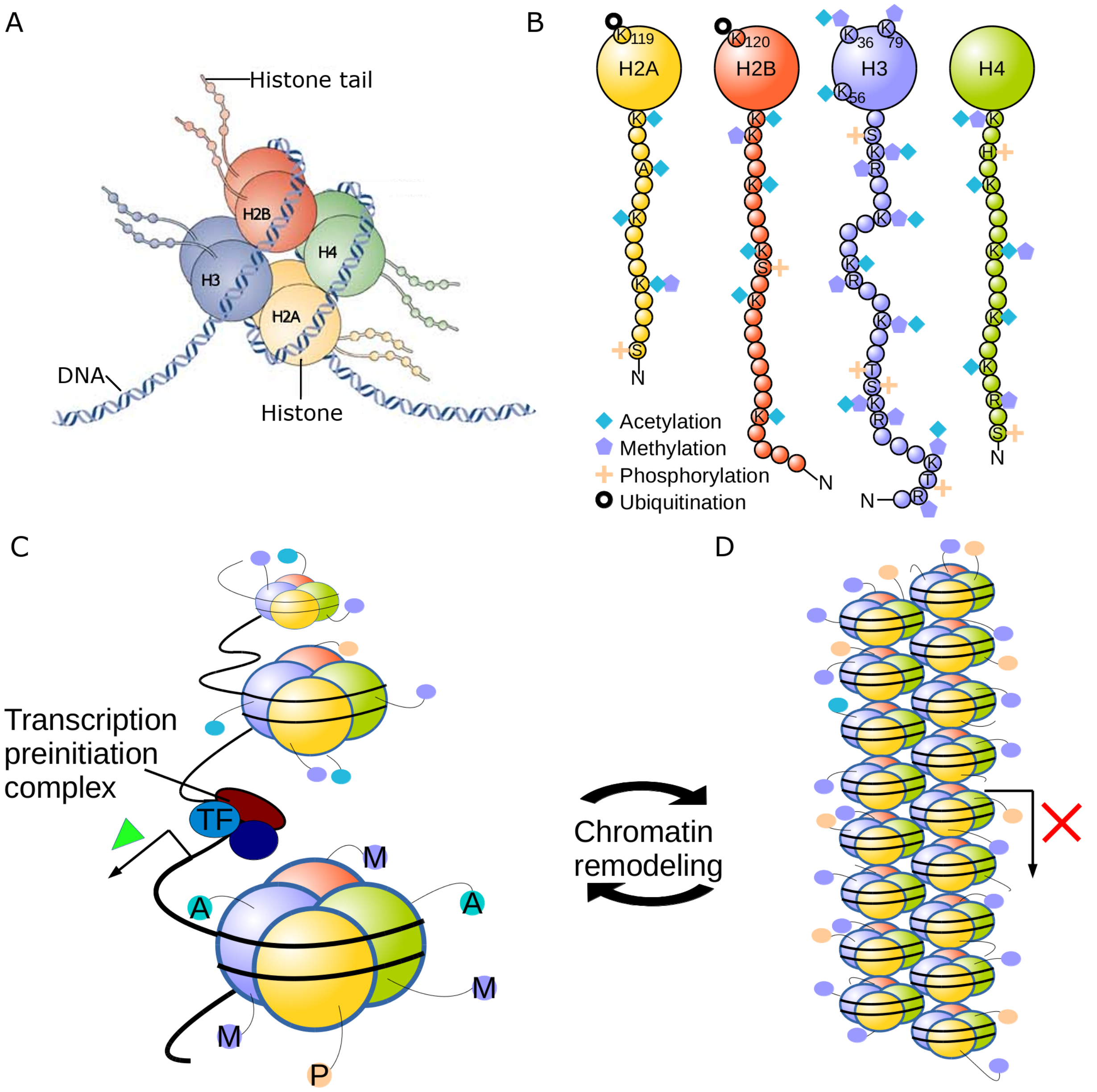


Figure 3: The chromatin structure of DNA. (A) The nucleosome, consisting of the histone octamer, wrapped in DNA. (B) Known and characterized modifications of the histone tails. (C) Euchromatin is accessible for transcription factor (TF) binding and gene expression can be initiated. (D) Heterochromatin is inaccessible for the transcriptional machinery, therefore gene expression is repressed. Subfigure A is modified from [Tsankova et al., 2007]

2.4 HISTONE MODIFICATION

The local chromatin structure is controlled by covalent epigenetic modifications on the chromatin, particularly modifications of histone tails and methylation of nucleotides of the DNA. Specific modifications are either associated with the accessible, open and active euchromatin structure, or with the dense, inactive heterochromatin structure. For each type of modification, specific regulatory proteins set or remove the epigenetic marks, or recognize them by specific binding. Furthermore, in order to maintain epigenetic profiles during proliferation, specific mechanisms ensure inheritability of epigenetic modifications of the chromatin.

The tails of amino acid chains of histones are at the surface of the proteins, and can be subject to different types of post translational modifications. This gives rise to a diverse range of epigenetic marks: the amino-terminal residuals of histone proteins can undergo acetylation, methylation, phosphorylation, ubiquitylation and SUMOylation [Berger, 2007]. The most common and best studied histone modifications are acetylation and methylation of lysine (K), phosphorylation of serine (S), and methylation of arginin (R) terminals (Figure 3B).

The common nomenclature of histone modifications is composed of the histone name (e.g. H2A, H2B, H3 or H4), the amino acid residual (e.g. K for lysine), the position in the amino acid chain, and the type and number of modifications. Depending on the position relative to the gene or regulatory region, a combination of histone modifications is associated with the activity of the region. The interplay between the regulatory functions of modifications is referred to as histone code hypothesis [B. M. Turner, 2000]. For example, H3K4me3, which is trimethylation of lysine at position 4 of core histone H3, is found at promoter regions of actively transcribed genes, but absent at intergenic regions and promoters of inactive genes. Monomethylation of the same residual (that is H3K4me1) is associated with active enhancers, cis regulatory elements which promote the expression of nearby or distant genes. Histone acetylation of lysines 9 and 27 on histone H3 (H3K9ac and H3K27ac) is also associated with open chromatin, active promoters as well as enhancers. The same modifications at different residuals show opposite effects: for example, H3K27me3, H3K9me3 and H3K79me3 are correlated with suppressed transcription. Co-occurrences of opposing histone modifications on the same nucleosome, such as H3K4me3 and H3K27me3, are thought to mark genes poised for expression [Bernstein et al., 2006]. Histone modifications are not limited to gene regulatory function: for example, ubiquitination of H2A and monomethylation of lysine 79 at H3 mark DNA damage and help recruit repair factors [Rossetto et al., 2010].

Several distinct classes of enzymes can recognize specific histone modifications, or modify specific histone amino acid residuals. For example, proteins containing a bromodomain recognize acetylated lysine residues. These 'readers' of the histone code ultimately determine the functional outcome of the modification. Other proteins, like those from the family of histone acetyltrasfereases (HATs) and histone methyltransferases (HMTs), add specific epigenetic marks; still other proteins from the histone deacetylase (HDAC) and demethylase (KDM) family remove them. Through mutual interaction of histone modifications with histone readers, writers and erasers, functionally associated marks are attracted, and opposing marks are repressed [T. Zhang et al., 2015]. This interaction yields bistable autoregulation of the histone code, which leads to propagation and spreading of the local chromatin structure.

Like other epigenetic marks, histone modifications are characteristic of developmental stages and cell types, and stably maintained during replication. However, the exact mechanism of heredity of histone modification to descending cells has still not been fully explored. For the DNA, complementary basepairing allows semiconservative replication, which ensures integrity of the newly synthesized strands. In contrast, parental nucleosomes are dispersively segregated to both sides of the replication fork. This process leaves each copy with half of the parental nucleosomes, still carrying the modification. The remaining nucleosomes are replaced by new, unmodified histone complexes. Computational modeling suggests that cooperative and competitive interactions between histones and histone modifying proteins might be able to carry over histone modifications from the remaining to the new nucleosomes and thus restore the histone code [Margueron and Reinberg, 2010].

2.5 DNA METHYLATION

DNA methylation is the first recognized and best characterized epigenetic modification of the chromatin. In contrast to histone modifications, DNA methylation is a modification of the DNA itself. Two of the four nucleotides of DNA, adenosine and cytosine, can be methylated. The type, context and fraction of methylated nucleotides in the genome varies widely between species: in plants, such as *Arabidopsis thaliana*, adenosine methylation of mRNA is thought to play a role in post-transcriptonal regulation [Bodi et al., 2012], and cytosine methylation occurs widely independent of the sequence context. For other species, like the baker's yeast *Saccharomyces cerevisiae*, or the invertebrate model organism *Caenorhabditis elegans* cytosine methylation is considered completely absent, and adenosine methylation is restricted to specific sites [Capuano et al., 2014; Greer et al., 2015]. In mammalian cells, methylation occurs at cytosines, at the 5' position of the pyrimidine ring (5mC). For these species, 5mC is typically found in the symmetric CpG context: since the C and G are complementary, the CpG dinucleotide is present on both forward and reverse strands of the DNA, and methylation occurs either on both of the strands or none of them. Apart from phases of epigenetic re-programming, where DNA methylation is removed to a great extent, for most regions of the mammalian genome, more than 80 % of the CpGs are methylated [Stadler et al., 2011]. Methylated cytosines can spontaneously deaminate to form thymidine. As a result, the transition rate of CpG dinucleotides to TpG (and CpA on the reverse strand) is 10 to 50 fold higher than the rate of other transitional changes [Sved and Bird, 1990]. Due to this mechanism, the CpG dinucleotide is present at approximately 20% of the expected frequency in the mammalian genome, in contrast to organisms with little or no cytosine methylation, such as the invertebrates *Drosophila melanogaster* and *Caenorhabditis elegans*, which show expected levels of CpG throughout the genome. Consequently, variation in methylation level within the genome causes variation in the local mutation rate, and, as a result, in the local CpG density. This effect can be observed most prominently in mammalian genomes, on a small fraction of the DNA that escapes methylation. In turn, the CpG density in these regions is in accordance with the expected number of CpG dinucleotides, which allows distinction from the CpG depleted remaining part of the genome. These so called 'CpG islands' (CGI) have been found in 72% of promoter regions of human genes [Saxonov et al., 2006] including a large proportion of tissue specific and developmental regulator genes. This observation, together with the evolutionary conservation of the regions, indicates the functional importance of DNA methylation.

Methylation of CGI promoters is correlated with down regulation of the corresponding gene. This coherence can be observed in imprinted genes as well as genes suppressed by X inactivation, suggesting that CGI methylation is involved in regulation of these developmental mechanisms. Further evidence for a role of methylation at CGI in gene regulation is provided by the observation in cancer cells, where hypermethylation of CGI promoters of tumor suppressor genes is frequently associated with down-regulation of the genes.

This down regulation can be explained by two alternative models describing the regulatory mechanism: DNA methylation can directly influence sequence specific binding affinity of transcription factors. Exemplary for this first, direct mechanism is the methylation sensitivity of CTCF: methylation of the CpG sites within CTCF binding motif disrupts binding of the factor [H. Wang et al., 2012]. Since CTCF binding impacts the 3D structure of the genome, and thereby controls the activity of enhancers, DNA methylation of CTCF sites might play a role in cell type specific enhancer function. The second regulatory activity is mediated by proteins that contain a methyl-CpG-binding-domain (MBD) and upon binding can recruit histone modifiers or chromatin remodeling proteins. For example, MECP2, a member of the MBD protein family, binds methylated DNA and recruits histone deacetylases, which in turn promote a repressive chromatin state [Jones et al., 1998].

Accordingly, genes effected by imprinting and X inactiviation are found methylated on the silenced allele. However, it is important to understand that, at least for those mechanisms, DNA methylation does not directly initiate transcriptional repression, but rather locks the gene in a silent state, which has been initiated by other factors [Deaton and Bird, 2011a]. For example, during initialization of X-chromosome inactivation, the noncoding RNA Xist recruits polycomb repressive complex 2 (PRC2) to CGI promoters. Binding of PRC2 leads to trimethylation of H3K27 and ubiquitination of histone H2A, and, in consequence, to repressive chromatin condensation and inhibition of transcriptional elongation. However, while the initiation of this process is independent of DNA methylation, it is essential for stable maintenance of X-chromosome inactivation: knockout of Dnmt1, a gene responsible for maintenance of CpG methylation, leads to re-activation of imprinted genes in mice [Sado et al., 2000].

While regulatory potential of DNA methylation at CGI promoters is well studied, approximately half of the mammalian CGIs are not associated with annotated promoters. To express the functional uncertainty about these CGIs, which show higher methylation levela in somatic cells compared to promoter CGIs, they have been termed 'orphan' CGIs. Evidence for transcriptional activity at the sites suggest that orphan CGIs mark previously unknown promoters of non-coding RNA [Illingworth et al., 2010]. These RNA might have cis-regulatory effects, such as the ncRNA Air, which is involved in paternal imprinting of Igf2r and other nearby genes.

At genomic regions with depleted CpG content, outside CGIs, the functional role of DNA methylation remains largely unknown. Recently, hypomethylation of large genomic blocks have been identified as a common epigenetic alteration in several different tumor types [Timp et al., 2014]. These regions tend to be co-localized with specific large heterochromatin structures termed large organized chromatin lysine modifications (LOCKs) and lamina associated domains (LADs) This co-localization suggests DNA methylation plays a role in chromosome organization within the nucleus. Furthermore, genome-wide demethylation with decitabine is associated with genomic aberrations and structural variants. Based on this observation, the prevalent DNA methylation outside CGIs has been proposed to be an important factor in maintaining genome stability [Vilain et al., 1999].

In mammals, two active types of DNA methyltransferases (DNMTs) establish and maintain CpG methylation: the ubiquitous maintenance methyltransferase DNMT1 recognizes hemi-methylated CpGs, and methylates the unmethylated strand. This function provides a mechanism for maintaining the DNA methylation pattern during cell divisions. Upon semiconservative replication, the template strand keeps methylation information, but the newly synthesized strand is completely unmethylated. Subsequently, DNMT1 binds to the hemimethylated CpGs and restores the methylation pattern from the template strand. Members of the second family of methyltransferases, DNMT3A and DNMT3B, are responsible for *de novo* methylation of CpGs. These enzymes are essential during the re-establishment of DNA-methylation after the phases of epigenetic reprogramming, but they also play important functional and pathological roles in development and disease [B.-F. Chen and Chan, 2014].

Methylated CpGs can be removed either by a passive or active process. Passive de-methylation is realized by transcriptional or functional inhibition of maintenance DNMT1. In the absence of functional DNMT1, the newly synthesized strand remains unmethylated, and thus, during proliferation, DNA methylation is gradually lost. Since this process is sequence independent, it results in genome-wide de-methylation. Accordingly, both phases of epigenetic re-programming depend on passive de-methylation. Treatment with decitabine, e.g. in cancer chemotherapy, functionally inhibits DNMT1 and induces passive genome-wide de-methylation [Kornblith et al., 2002].

In contrast, active de-methylation may occur locally upon targeted recruitment of enzymatic factors which specifically modify methylated cytosines. Ten-eleven-translocation methylcytosine dioxygenase 1 (TET1) catalyzes the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC) [Tahiliani et al., 2009]. This enzymatic modification has been postulated as the first step of an active de-methylation pathway [Ito et al., 2011]: 5hmC serves as an intermediate product within this pathway, which is further oxidized to 5-formylcytosine (5fC) and subsequently to 5-carboxylcytosine (5caC), in reactions that are potentially also catalyzed by TET1 and other TET family members. 5fCs and 5caCs are recognized by the DNA repair mechanisms and replaced with unmodified cytosines. Since this pathway is independent of replication, it provides a model for the de-methylation processes observed in zygote formation and the germ-cell lineage.

Even though the details and the exact functional role of individual TET family members remains unclear, the erasing function of TET protein on cytosine methylation is evident. This function has been used to induce targeted de-methylation, by fusing the epigenetic modifiers to nuclease deficient Cas9 constructs, which is directed to the target sequence by a guide-RNA [X. S. Liu et al., 2016].

Besides the role as an intermediate product in the active de-methylation pathway, increased levels of 5hmC in specific cell types indicate that this modification might act as a distinct stable epigenetic mark on its own. For example, in mouse cerebellum an age and expression level dependent enrichment of 5-hmC has been observed, suggesting a functional role of this modification [Song et al., 2011].

3

FUNDAMENTAL PRINCIPLES

In this chapter I present different experimental methods for high throughput profiling of epigenetic modifications. For DNA methylation measurement I describe two experimental principles: one principle is based on methylation dependent chemical conversion of cytosines, the other on enrichment of methylated DNA fragments. The methods presented in this thesis focus on enrichment based DNA methylation experiments. To this aim, I introduce the methodological concepts that provide the foundation for the specific methods I developed. Finally, I describe a comprehensive clinical dataset which contains DNA measurements from both types of experiments and which is used throughout the thesis to develop specific methods and compare their performance.

3.1 APPROACHING THE EPIGENOME: EXPERIMENTAL PRINCIPLES

In recent years several experimental protocols have been established for high throughput profiling of epigenetic modifications. These protocols are based either on direct detection and quantification of the epigenetic mark, or on enrichment of the DNA fragments marked by the epigenetic modification of interest. Both approaches can be quantified using DNA micro-arrays or high throughput sequencing (HTS).

3.1.1 *Bisulfite Conversion*

DNA methylation can be measured directly by specific conversion of unmethylated cytosines: DNA treated with sodium bisulfite converts unmethylated cytosines to uracil but does not affect methylated cytosines. This allows the classification of methylated and unmethylated sites, either by DNA microarrays with specifically designed probes covering CpG sites [Weisenberger et al., 2008], or by sequencing (BS-seq) [Lister et al., 2009]. Both quantification measures reveal the fraction of unconverted (and thus methylated) cytosines. This direct approach allows the investigation of absolute DNA methylation levels at base resolution. The principle is not limited to CpG context, and the approach can thus be applied to non-mammalian genomes as well.

Microarrays, designed to quantify DNA methylation levels based on bisulfite conversion, contain probes matching the fully converted fragment as well as the fragment with unconverted CpG sites. Current array designs cover roughly 450,000 genomic regions, selected for their known regulatory potential or their aberrant state in disease.

High throughput sequencing (HTS) provides a whole-genome quantification method for bisulfite treated DNA (WGBS). Short reads are aligned to the reference genome by a specific alignment algorithm that tolerates C/T mismatches due to the bisulfite conversions. At each cytosine in the reference genome, the C over T ratio in the reads is proportional to the fraction of methylated cytosines. In order to reliably estimate methylation level, BS-seq requires sufficient read coverage of individual CpGs. To this end, BS-seq depends on deep sequencing in order to cover most of the CpGs in the genome. It has been suggested that for the human genome an average of 30 fold coverage, or 800 million paired end reads of length 101 bases, is necessary [Ziller et al., 2015]. While this is feasible for individual samples, with current HTS technology the sequencing costs remain a limiting factor for the analysis of larger sample groups.

For this reason, in the context of larger studies, bisulfite sequencing has been performed mainly as a targeted approach. In this case, sequencing is confined to genomic regions of primary interest, for example with Methyl-seq [Deng et al., 2009] and reduced representation bisulfite sequencing (RRBS) [Meissner et al., 2005]. Like microarrays, these approaches are limited to their respective target regions and are not informative for the discovery of epigenetic mechanisms outside the covered genomic subset.

Bisulfite conversion is not only prevented by methyl cytosine, hydroxymethylcytosine is also protected from conversion. Therefore, bisulfite based approaches cannot distinguish 5mC and 5hmC. In order to specifically detect 5hmC, an additional step before bisulfite treatment can be performed, to oxidate 5hmC to 5fmC (OxBS). In contrast to 5hmC, 5fmC is sensitive to the bisulfite conversion. By comparing BS with OxBS, 5hmC levels can be inferred [Booth et al., 2013].

3.1.2 *Enrichment Based Methylation Assays*

Alternatively, DNA methylation can be detected by enriching fragments with methylated cytosines. To this end, genomic DNA is first purified and fragmented. From this 'input' sample, methylated DNA fragments are enriched in a second step. DNA fragment enrichment relative to the input provides a measure for the relative abundance of methylated cytosines at a genomic locus.

Two similar techniques enrich DNA fragments containing methylated cytosines: methylated DNA Immuno-Precipitation (MeDIP) uses an antibody specific for 5mC [Weber et al., 2005a], and MBD protein capture [Serre et al., 2010] uses the methyl-CpG binding domain of the protein to bind methylated cytosines. In both cases, the enrichment of genomic regions can be assessed by DNA microarray or high throughput sequencing (HTS). In the first case, the DNA microarray is spotted with oligonucleotide probes that cover a representative fraction of the genome. Input and methylation enriched fragments are labeled with different fluorescent dyes and exposed to the array. The fragments bind complementary probe sequences on the array. Relative fluorescence provides a measure of the level of DNA methylation at the genomic positions of the probes. In contrast, HTS provides a whole-genome method of quantification: input and methylation enriched samples are sequenced, typically resulting in 40-100 million short reads, either from one or both ends of the fragment. Alignment of the reads to a reference genome provides the local read density, which after normalization corresponds to the level of DNA methylation.

Compared to bisulfite sequencing based approaches, the enrichment based assays require substantially less sequencing depth, while still targeting the whole genome, and thus not restricting the analysis to predefined sites. Hence, those assays are an attractive alternative for studies with large numbers of samples across several conditions. However, the resolution of enrichment based experiments is limited by the fragment size, which is typically 250 bp on average, as opposed to single base resolution in bisulfite based approaches. Furthermore, MBD enrichment is insensitive for non-CpG methylation, and MeDIP cannot distinguish different sequence contexts. Therefore, enrichment based approaches are best suited for organisms where methylation is restricted to the CpG context.

One DNA fragment potentially covers multiple methylated cytosines, and therefore multiple potential binding sites for the antibody or MBD protein. The number of potentially methylated cytosines, which in mammals is equivalent to the number of CpGs within the fragment, has an effect on the enrichment efficiency (Figure 4). For differ-

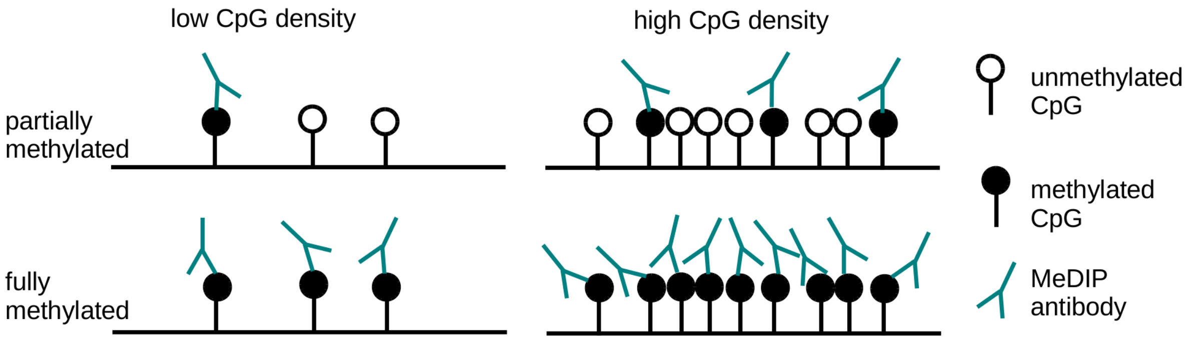


Figure 4: Influence of CpG density on enrichment efficiency. Fragments featuring few CpGs (left) provide fewer potential targets for the antibody and therefore require a higher fraction of methylated CpGs in order to become similarly enriched to fragments featuring higher numbers of CpGs (right).

ential analysis, where a given genomic window is compared across samples, the CpG density usually is the same and can be neglected. However, the inference of absolute levels of methylation - obtained, for example, from bisulfite based approaches - requires consideration of the CpG density. Furthermore, alterations of DNA copy number, which occur frequently in tumors and tumor derived cell lines, influence the fragment enrichment and require normalization. The focus of this thesis is on the development and assessment of methods for the analysis enrichment based methylation experiments. In the following chapters, I present the methodology for the individual steps of the analysis, assess different general concepts and develop specific solutions.

3.1.3 *Chromatin Immunoprecipitation*

The enrichment based approach is not limited to DNA methylation. Chromatin Immunoprecipitation (ChIP) is a general technique to detect interactions between DNA and proteins by enriching DNA fragments bound by the protein of interest using antibodies specific for that protein. In particular, ChIP can be used to detect histone modifications. To this end, specific antibodies have been developed for several histone modifications, qualifying this technique for a comprehensive analysis of the histone code.

Again, the enrichment of fragments can be detected either by DNA microarrays (ChIP-chip) or high throughput sequencing (ChIP-seq). Depending on the experimental design, different types of analysis have been established for ChIP experiments. To analyze the presence of an individual histone modifier in a single condition, peak detection methods provide a method to binarize the data [Y. Zhang et al., 2008], which reflects the fact that the modification is either present or absent. However, since the all-or-nothing decision is based on a rigid threshold, for borderline cases even a small difference in sequencing coverage may result in a completely different outcome. Furthermore, the analyzed sample is a mixture of various cells that potentially have different patterns of histone modifications, such that the peak-based approach might be an oversimplification. Chromatin segmentation tools analyze several histone modification tracks in parallel to categorize the genome in regions with similar combinations of histone modifications, either following discretization using peak detection [Ernst and Kellis, 2010], or by modeling the read densities [Mammana and Chung, 2015]. By assigning regulatory mechanisms to the histone modification classes, these approaches attempt to decrypt the histone code.

Several research questions require comparison of histone modifications between different conditions. A common approach for this task is based on peak detection for the individual conditions and subsequent determination of overlapping and distinct peaks between the sets [Heinz et al., 2010], or model based statistical tests in the peak regions [Ross-Innes et al., 2012]. Despite its popularity, this procedure harbors several potential problems. In addition to the simplified binary perspective resulting from peak calling, the shape and height of the enrichment signals is not properly represented in the peaks, which gives rise to both false positive and negative results. To avoid these pitfalls, approaches to detect differential methylation from enrichment based experiments, which are based on statistical modeling of read counts within genome-wide windows, have also been applied in the context of differential ChIP-seq analysis [Seumois et al., 2014].

3.2 MODELING READ ENRICHMENT: STATISTICAL PRINCIPLES

3.2.1 *Quantification of DNA Fragment Density*

The primary signal of enrichment based sequencing experiments is the DNA fragment density at a genomic position. Each fragment is represented by a sequencing read spanning a fraction of the actual fragment. The fragments can be mapped to a genomic position by aligning the read sequence with the reference genome.

A common approach to parameterize the genome-wide fragment density is to segment the reference sequence into windows of equal size, and to count the sequencing fragments per window. Each window is considered as an independent feature for which enrichment is quantified. Hence, the window size restricts the resolution of the analysis. On the other hand, for a given sequencing depth, smaller window size results in fewer reads per window, and therefore less predictive power at each window. Thus, for the choice of the window size, both the resolution of the experiment, determined by the DNA fragment size, and the sequencing depth should be considered.

For each window, the number of fragments serves as a measurement for the local fragment enrichment. To unambiguously assign the fragments to genomic windows, each fragment is counted for the window that is overlapping the center position of the fragment. For paired end sequencing, the genomic position of the sequenced fragments can be obtained directly. For single end, the end position can be approximated by extending the read to the average fragment length that has been estimated by gel electrophoresis. Due to inhomogeneous preference in the PCR steps, some fragments are overrepresented in the sample, and thus sequenced multiple times, leading to pile-ups in the alignment. To prevent a bias for the corresponding regions, reads with exactly the same positions are replaced by one representative.

3.2.2 *Between-Sample-Normalization*

In order to account for the different sequencing depth of different MeDIP seq libraries and thereby to obtain values for the fragment density which are commensurable between samples, the read counts must be normalized. To this end, the read counts are scaled by a sample specific factor reflecting the different sequencing depths. This step is similar to the normalization in other types of quantitative sequencing analysis, in particular RNA-seq. For RNA-seq, 3 different scaling methods have been suggested:

TOTAL COUNTS (TC) The simplest approach is to divide the reads by the total number of reads. For convenience, these values are typically multiplied by 10^6 to obtain reads per million (rpm). This approach, however, does not consider the different distributions of count values. For RNA-seq experiments, it has been observed that a large fraction of the reads originate from a small number of highly expressed genes. These genes would thus dominate the normalization factor, and, if differentially expressed, distort the results [M. D. Robinson and Oshlack, 2010]. For MeDIP-seq reads, differences in the experimental conditions can lead to similar effects.

UPPER QUANTILE (UC) The median provides a commonly used measure that is more robust to extreme values compared to the total read counts. However, for both RNA-seq and MeDIP-seq, typically the distribution of counts is skewed, and more than half of the features have no or only few reads. For RNA-seq, it has been suggested the reads be divided by the upper (75%) quantile [Bullard et al., 2010]. In contrast to RNA-seq, where this measure commonly provides reasonable and robust scaling factors, the UQ method is problematic for genome-wide enrichment experiments. Here, the number of features (genomic windows) is in the order of the number of reads. Especially for widespread marks such as DNA methylation, a large part of the genome is enriched, and the number of reads per window is too low to provide a reasonable scaling factor.

TRIMMED MEAN OF M-VALUES (TMM) This approach is based on the assumption that for most of the features the abundance does not change between samples [M. D. Robinson and Oshlack, 2010]. The idea is to scale the libraries, such that log ratios (m-values) are centered around 0. Since the variance of the count values is proportional to the mean, m-values of features with higher read counts have lower variance. To account for this dependency, the m-values are weighted by the inverse of the approximated asymptotic variance. In order to minimize the effect of differently abundant features and outliers, the extreme values are trimmed off. The TMM scaling factor is proportional to the weighted trimmed mean of m-values.

LIMITATIONS In the case of highly irregular read density distributions between samples, simple scaling may be insufficient for inter-sample normalization. This issue can be illustrated by M vs A plots, where the log ratios are plotted against the average log counts. Horizontal asymmetry indicates the need for more involved normalization methods in order to make samples commensurable. Quantile normalization is a method for unifying distributions that was first proposed in the context of expression micro-arrays [Bolstad et al., 2003]. In a first step a common reference distribution is generated from the rank-wise averages across samples. Next, each feature is assigned the value from the reference distribution that corresponds to its rank in the original distribution. While this method is in general applicable to quantitative sequencing count data, it may have implications on subsequent modeling, especially if sequencing depth or library complexity differs widely between samples.

3.2.3 *Within-Sample-Normalization*

If parts of the genome are more abundant in the primary sample, these regions will also be overrepresented in the enriched fraction of the sample, leading to a bias for the respective regions. To account for this effect, read densities can be scaled proportional to the relative abundance of the region by window specific scaling factors.

VARYING NUMBER OF CHROMOSOMES One obvious source of differently abundant genomic regions is the different number of sex chromosomes in males and females. While female individuals have two copies of the X chromosome, male individuals have one copy of both X and Y chromosomes. Other conditions leading to differences in chromosome copy number are trisomies, such as Down syndrome (Trisomy 21) or Klinefelter syndrome (XXY).

COPY NUMBER VARIATIONS (CNV) Frequent features of cancer cells are variations of DNA copy number, which are segmental duplications and deletions, affecting genomic regions from one kilobase to an entire chromosome. Proportional to level of the CNV, MeDIP read densities increase or decrease (Figure 5, left). Applying a region specific scaling factor fully accounts for this effect (Figure 5, right).

3.2.4 *Modeling Read Counts*

In order to statistically assess the fragment density, the number of reads per window can be described by statistical models. These models are the foundation of statistical inference from the observed data, such as statistical hypothesis testing and statistical estimators. In statistical models, the generation of data y (here the read counts per window) is characterized by a random variable Y, following a probability distribution. Here, I present common probability distributions that are applicable to the discrete count data from sequencing experiments.

3.2.4.1 *Poisson Distribution*

The Poisson process assumes that events occur at constant rate and independently of the previous events. The distribution expresses the probability of observing a given number of events within a fixed time interval. Assuming the pool from which DNA

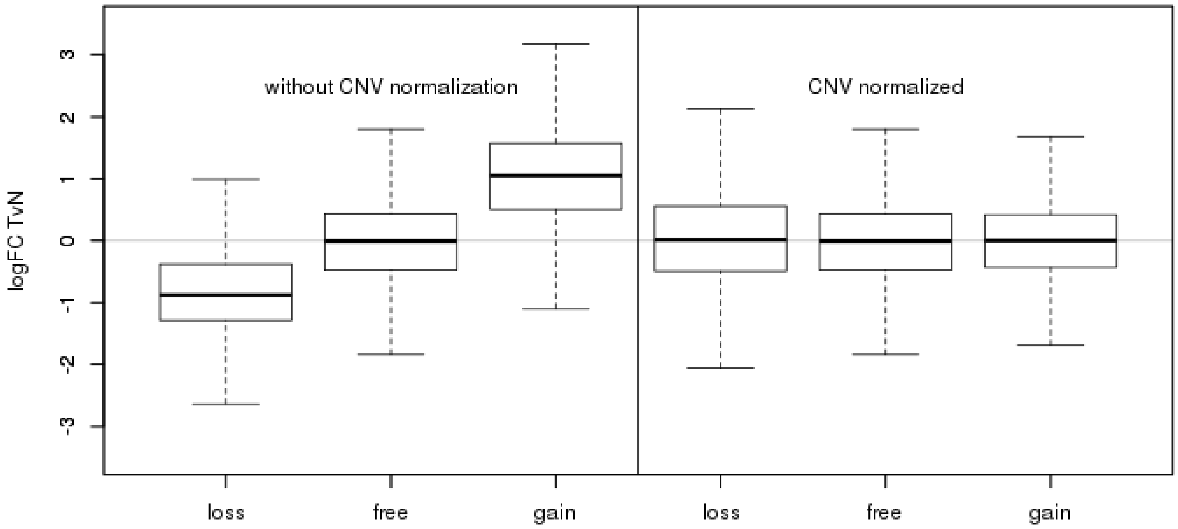


Figure 5: MeDIP read enrichment in regions featuring loss of copy number, CNV free regions and gain of copy number, before and after CNV normalization. Before accounting for CNV, the local read density is reduced in deleted regions and increased in amplified regions (left). Scaling by the estimated copy number results in balanced read densities for all CNV groups (right).

fragments are drawn for sequencing is infinitely large, then the number of reads y within a genomic region for a specific sample can be modeled as a Poisson process. The probability mass function of the Poisson distribution depends on a single rate parameter λ.

$$Y \sim Pois(\lambda)$$

$$Pr(Y = y|\lambda) = f_{Pois}(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad y \in \mathbb{N} \tag{1}$$

The expected value and the variance of the distribution correspond to the rate parameter λ. This distribution is popular due to its versatility and simplicity. It is adequate for modeling the probability of observing a given number of sequencing reads in a specific window for an individual sample. Examples for the probability mass for different rates are depicted in Figure 6A.

3.2.4.2 *Negative Binomial Distribution*

In contrast to the model for an individual sample, statistical modeling of a group mean must also consider the biological variation within the group. This additional variation leads to overdispersion, which cannot be covered by the Poisson distribution, where the variance is equal to the mean. An elegant approach to introduce this additional source of variation is to model the read counts by a Poisson distribution, where the rate parameter is scaled by a factor ϕ, which itself is the realization of a Gamma distributed random variable Φ with positive parameters $\alpha = \beta = \theta$. The expected value of the Gamma distribution $E[\Phi]$ is $\frac{\alpha}{\beta} = 1$, such that the scaling does not effect the mean. The variance $Var[\Phi]$, which models the variability between samples, is $\frac{\alpha}{\beta^2} = \frac{1}{\theta}$. Therefore, θ is called the dispersion parameter. Figure 6B shows the probability density of Φ for varying values for θ. It can be shown that this Poisson-Gamma-Mixture distribution is equivalent to a Negative Binomial distribution (see Appendix A.1).

$$Y \sim Pois(\lambda = \mu * \phi)$$

$$\Phi \sim \gamma(\alpha = \theta, \beta = \theta)$$

$$\implies Y \sim NB(r = \theta, p = \frac{\mu}{\mu + \theta}) \tag{2}$$

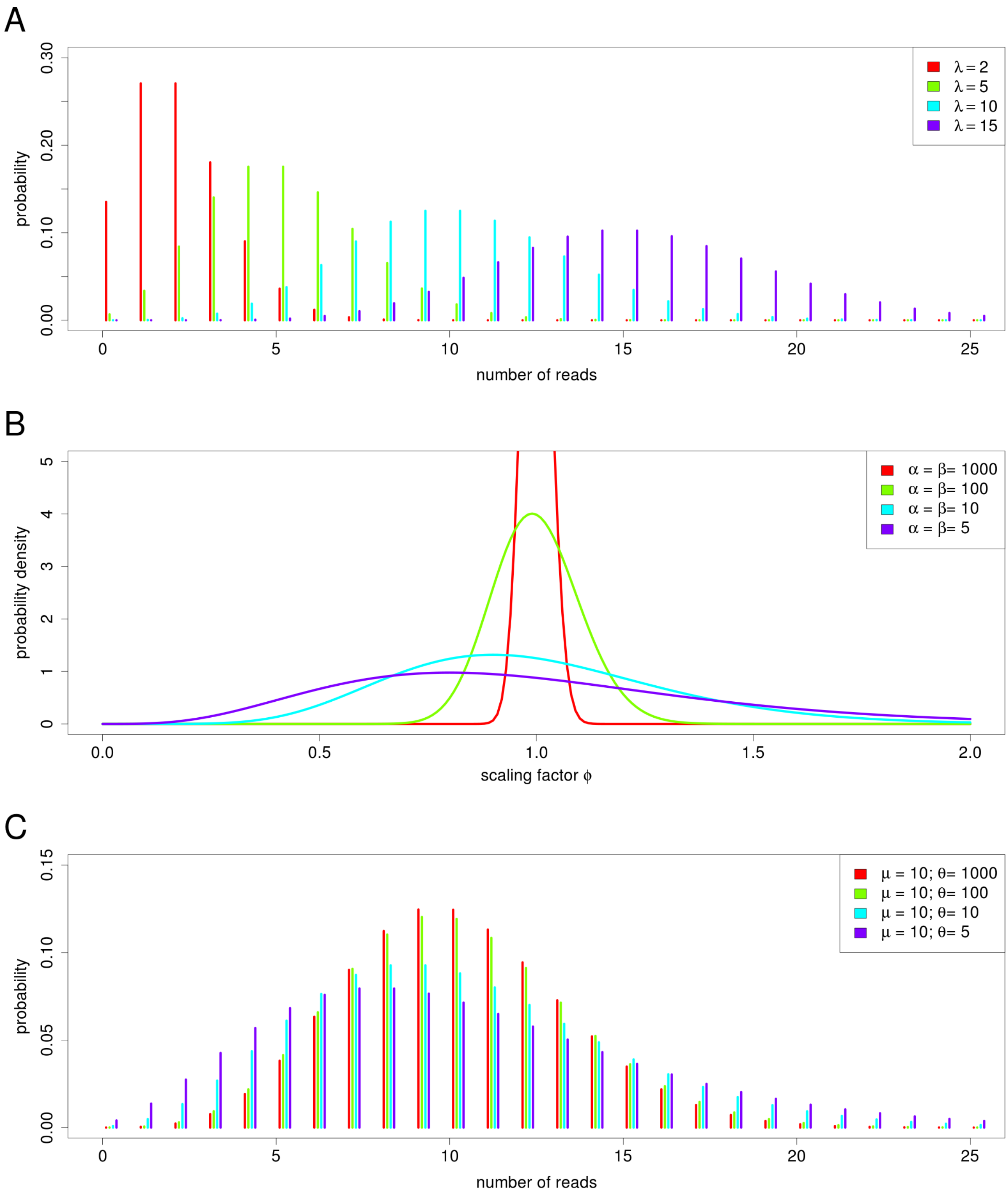


Figure 6: Probability distributions to model number of reads per window. (A) Poisson distribution for different λ. (B) Gamma distribution modeling the variability between samples. (C) Negative Binomial distribution with fixed mean and varying dispersion parameter.

The Negative Binomial distribution has the following probability mass function:

$$Pr(Y = y|\mu, \theta) = f_{NB}(r = \theta, p = \frac{\mu}{\mu + \theta}) = \frac{\Gamma(y + r)}{\Gamma(r)y!} p^y (1 - p)^r \tag{3}$$

with mean parameter $\mu = E[Y]$, dispersion parameter θ. $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, \mathrm{d}t$ is the gamma function. In contrast to the Poisson distribution, the overall variance of the Negative Binomial distribution exceeds the mean: $Var(Y) = \mu + \mu^2 * \frac{1}{\theta}$. The influence of the overdispersion parameter on the probability mass is depicted in Figure 6C .

3.2.5 *Bayesian Inference*

In Bayesian inference, the probability of a hypothesis is expressed in terms of the posterior probability distribution. According to Bayes' theorem, the posterior probability for the hypothesis H, given evidence E from data, is the consequence of three factors:

$$Pr(H|E) = \frac{Pr(E|H) * Pr(H)}{Pr(E)} \tag{4}$$

- The *likelihood function* $Pr(E|H)$ is derived from a statistical model for E, interpreted as a function in H.

- The *prior probability* of the hypothesis $Pr(H)$ reflects the probability of the hypothesis *prior to* observation of the current evidence E.

- The *model evidence* $Pr(E)$ is the likelihood of the data. It is computed by marginalizing over all possible hypotheses of the model.

The posterior probability provides the complete description of the model for the hypothesis. However, in order to evaluate several hypotheses and summarize the models, it is often helpful to condense the information of the posterior by providing point estimators, possibly complimented by credibility intervals.

BAYESIAN POINT ESTIMATORS summarize the information from the posterior by providing a specific estimate for the hypothesis. Three concepts for point estimators are common:

- The *maximum posterior* reflects the most credible hypothesis.

- The *posterior mean* is the expected value of the hypothesis. This estimator minimizes the mean squared error.

- The *posterior median* has equal probability mass above and below the estimate. It minimizes the absolute error.

The most appropriate point estimator depends on the posterior and the specific situation, as well as the computational complexity, which may vary greatly between the estimators, depending on the posterior. In general, mean and median have defined loss functions and better represent the entire spectrum of hypotheses, while the maximum is often easier to compute.

BAYESIAN CREDIBILITY INTERVALS provide a range of hypotheses that includes $100\% - \alpha$ of the credibility from the posterior. For a typical value of $\alpha = 5\%$, the interval can be interpreted as the range of credible hypotheses. The width of the interval therefore reflects the uncertainty of a point estimator. There are two concepts used to derive a credible interval.

- The *equal tail interval* (ETI) is the range, where the probability of H being below the interval as well as the probability of H being above the interval is in either case $\frac{\alpha}{2}$.

- The *highest density interval* (HDI) involves the $1 - \alpha$ most credible values for H. For unimodal posterior distributions, it is the narrowest credible interval.

Generally, the choice of the credibility interval depends on the posterior, as well as the point estimator. For symmetric posterior distributions, both intervals are equivalent. However, for highly skewed posteriors, the mean and maximum might not be contained in the ETI. Since the maximum is the most credible value for H, its exclusion would contradict the interpretation as credibility interval. On the other hand, for multimodal distributions, the HDI is not coherent, which might raise practical and conceptual issues.

3.2.6 *Statistical Significance Testing*

The evidence for differential methylation levels between groups of samples at genomic regions can be assessed by statistical hypothesis testing. The test considers the probability of the observed data, assuming the mean methylation level is the same between groups (H_0 hypothesis). There are two general classes of statistical hypothesis tests, parametric and non-parametric tests. Parametric tests are based on probability distributions underlying the observations, and make assumptions on the parameters of this distribution. In contrast, non-parametric tests are solely based on statistics independent of the distribution's parameters, such as order statistics. The probability of obtaining the observed data, assuming H_0, is called the p-value. If the p-value is below the critical significance level α, H_0 is considered to be unlikely and is rejected. In this case, the alternative hypothesis $H_1 \neg H_0$, stating that the groups have different means, can be accepted.

The probability of falsely rejecting a true H_0 is called type-1 error. On the other hand, the probability of correctly rejecting a false H_0 is called the power of the statistical test. In general, parametric tests have higher power compared to non-parametric tests. However, violations of the assumptions on the distribution made for parametric tests lead to decreased power and increased type-1 error. Therefore, non-parametric tests have broader applicability. Furthermore, the underlying statistical model is often simpler for non-parametric tests, and hence easier to compute.

3.2.7 *Multiple Testing and Independent Filtering*

When applying multiple statistical tests, the type-1 error accumulates. Techniques controlling for multiple testing, such as the false discovery rate (FDR) [Benjamini and Hochberg, 1995], adjust the p-value and thereby generally reduce the detection power. For example, segmenting the human genome in 250 base regions results in about 12 million regions. If the statistical test is applied on all regions, the smallest p-value must be below 10^{-9} in order to be significant at FDR of 10%. Therefore, the statistical power of the test should be as high as possible.

In addition, a strategy to increase the detection power in high throughput experiments is to decrease the number of tests by independent filtering [Bourgon et al., 2010]. To this end, regions are selected based on a criterion that is correlated with the power of the test, but is independent of the test statistic under H_0. The filtering threshold for this criterion can be adjusted such that the number of significant regions at a predefined FDR is maximized (Figure 7A). For sequencing derived count data, a filter based on the sum of the read counts over all samples has been suggested [Love et al., 2014]. If H_0 is true, i.e. there is no difference in the enrichment between groups, the actual coverage does not affect the distribution of the test statistic. However, if there is a difference in the enrichment, statistical power increases with coverage (Figure 7B).

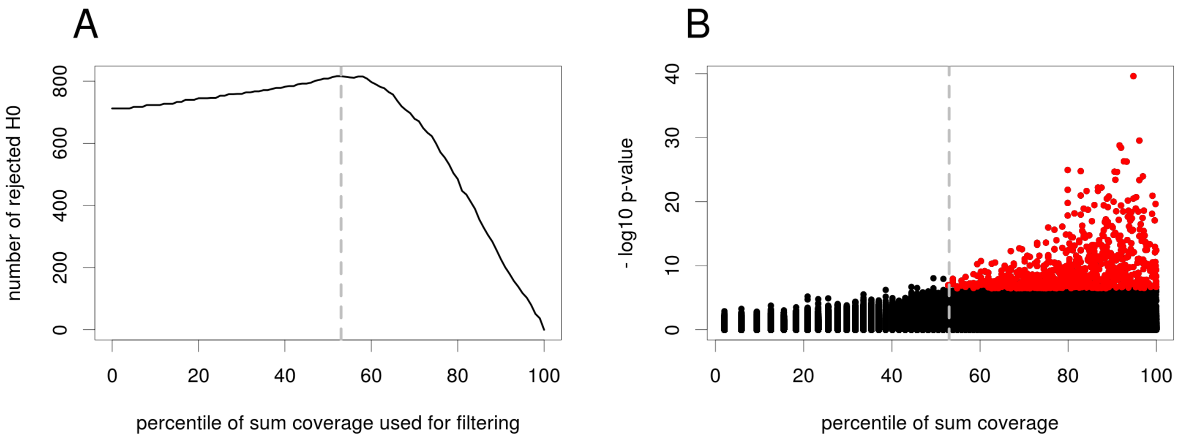


Figure 7: Independent filter for multiple testing. (A) The filter threshold is selected such that the number of rejected H_0 is maximal. For this example, the maximum for an FDR of 5% is at the 52 percentile, corresponding to 35 reads over all samples, indicated as gray dashed line. (B) As the statistical power to correctly reject H_0 increases with the read coverage, filtering allows maximizing the rejections at 5% FDR (red dots) while missing only few regions at similar significance level.

3.3 BENCHMARK DATASET

To test the specific methods for the analysis of enrichment based methylation experiments I used a benchmark dataset, consisting of samples from five human non small cell lung cancer (NSCLC) tissue that had been transplanted onto xenograft mice (patient derived xenograft models, PDX) as well as from normal lung tissue adjacent to the tumor of the same patients [Lienhard et al., 2016]. Including two cases of squamous cell lung cancer (SQC), one case of lung adenocarcinoma (ADC) and two cases of pleomorphic lung carcinoma (PLC), the cohort covers a variety of histological subtypes of NSCLC (Table 1). For these samples, we assayed genome-wide methylation using MeDIP-seq, as well as targeted bisulfite sequencing, using Methyl-seq. Additionally, parts of the libraries were sequenced prior to the enrichment step at low coverage (input sequencing). Further, we investigated the transcriptome of the samples using RNA sequencing.

This dataset is valuable in three different ways for the development of a rigorous method. First, the dataset offers the opportunity to analyze the characteristics of the MeDIP enrichment by comparing the MeDIP read density to absolute methylation levels obtained by bisulfite sequencing. From these observations, I deduce a statistical model of the enrichment that is used to estimate methylation levels from MeDIP enrich-

Sample	sex	molecular subtype
Patient 1	M	SQC
Patient 2	F	PLC
Patient 3	F	ADC
Patient 4	M	PLC
Patient 5	M	SQC

Table 1: Samples of the NSCLC dataset

ment signals. Second, since the dataset provides two independent DNA methylation assays – MeDIP-seq and Methyl-seq – for the same samples of the patient cohort, it allows for extensive benchmarking and methods comparison: In addition to assessing the accuracy of methylation estimates from MeDIP-seq for the 10 independent samples, this dataset allows individual differences between tumor and normal samples to be assessed and the alternative methods of analysis to be evaluated against each other according to how well they quantify these differences. Third, transcriptome analysis of the same samples allows investigating gene regulatory effects of DNA methylation.

Furthermore, I test the performance of the methods using an independent, second dataset, containing MDB-seq experiments as well as BS based methylation levels for the IMR90 cell line [Riebler et al., 2014]. This dataset is used to assess the robustness of the analysis methods with respect to the experimental protocol.

4

ESTIMATING ABSOLUTE METHYLATION VALUES

The read density, scaled by the sample specific and local normalization factors, provides an indirect signal of the methylation level at the genomic region. This signal is a relative measure, and thus allows the differential methylation levels between two samples to be assessed. However, since the number of reads is primarily dependent on the density of methylated cytosines and the enrichment efficiency, regions with different CpG densities cannot be compared directly. Furthermore, many use cases presuppose absolute methylation levels, such as assessing whether a specific region is methylated or unmethylated, charting whole genome methylation landscapes, or comparing the measurements with bisulfite based assays. The CpG dependence has been addressed by scaling the read density with a normalization factor that is coupled to the CpG density. Yet it is unclear how the resulting score relates to the fraction of methylated cytosines. Therefore, I developed an alternative method to estimate absolute methylation levels from the read counts. The estimation is based on a statistical model that takes the sample specific dependence of the enrichment on CpG density into account. As the method transforms read counts to the interval $[0, 1]$, the estimates can be interpreted as the fraction of methylated cytosines within the window and directly compared to BS based assays.

4.1 COUPLING FACTOR SCALING

A simple way to account for the dependence of the local CpG density on the enrichment is to introduce a CpG density dependent normalization factor. To this end, the dependency of fragment enrichment and CpG density is approximated by a linear function. Scaling by this CpG coupled factor results in an absolute methylation score [Down et al., 2008; Chavez et al., 2010]. However, the linear approximation fits poorly for higher

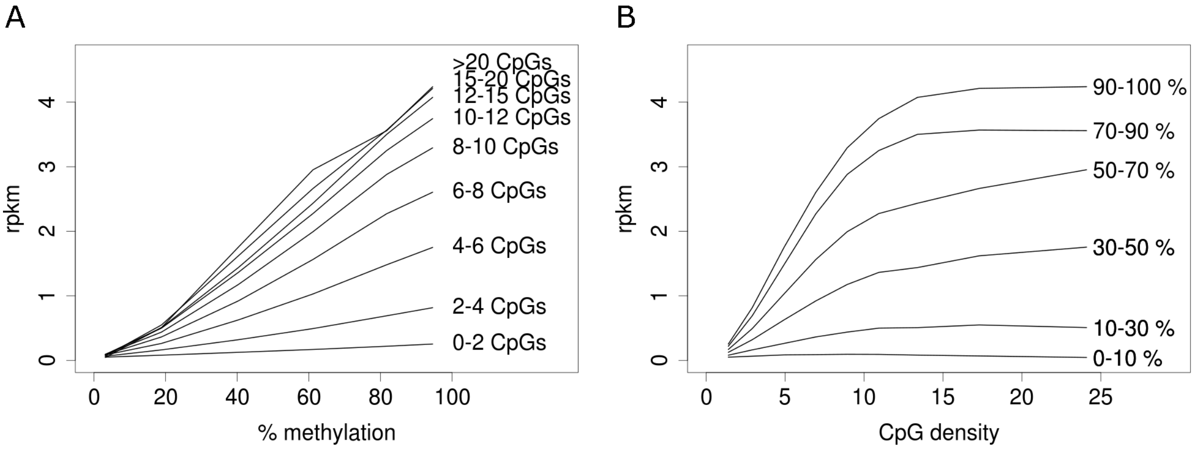


Figure 8: Observed enrichment characteristics of MeDIP fragments. (A) Linear dependency of enrichment and methylation level for groups with fixed CpG density. (B) Non-linear dependency of enrichment and CpG density for groups with fixed methylation level.

CpG densities, indicating that more sophisticated estimation between CpG density and enrichment is required. Furthermore, the range of these coupling factor scaled values is unbounded, and therefore, the absolute methylation score cannot be interpreted as the level of methylation. Hence, it is not directly and intuitively interpretable: assessing whether a certain genomic region is predominantly methylated or not requires the introduction of an arbitrary threshold, which, due to the artificial nature of the score, cannot be motivated biologically. Furthermore, such a methylation score cannot directly be compared to methylation levels derived from BS based experiments, which correspond to the fraction of methylated cytosines.

4.2 RELATION OF METHYLATION LEVEL AND ENRICHMENT

By comparing MeDIP-seq read densities with absolute methylation levels derived from BS-seq on the same samples, I observed the following enrichment characteristics. The MeDIP enrichment signal is dependent on the number of methylated cytosines within the fragment, which is limited by the number of CpGs. By grouping genomic windows according to similar CpG density, I observed a linear relation between absolute methylation, as measured by BS-seq, and the average normalized MeDIP-seq read coverage (Figure 8A).

On the other hand, for a fixed level of absolute methylation, I observed an increase in MeDIP enrichment from lower to medium CpG density that becomes saturated at higher levels of CpG density (Figure 8B). Further, I observed that regions lacking DNA methylation as well as regions lacking CpG dinucleotides are covered by an offset of reads. This read offset corresponds to unmethylated fragments remaining from the input although they were not bound by the antibody. If not taken into account, the "background reads" may lead to distortion of the enrichment signal, especially at regions with low CpG density or low methylation levels.

4.3 MODELING THE READ COVERAGE

The number of reads for a single sample can be adequately modeled with the Poisson distribution (see Section 3.2.4.1). In order to capture the linear impact of the methylation level on the enrichment, I model the rate parameter λ of the distribution linearly in the methylation level $ML = \beta$.

$$Y \sim Pois(\lambda = o + \beta * c) \tag{5}$$

The offset o is the expected number of reads of a region without any enrichment ("background reads"), corresponding to the situation where the region is completely unmethylated. The enrichment is modeled by the product of absolute methylation level β, which is between 0 and 1, and the expected maximal enrichment c for the region. This value reflects the sample specific relation of MeDIP enrichment and CpG density, and can be interpreted as the expected gain of reads when the region is fully methylated. This model yields the following probability density function for the number of reads Y in a region with specific methylation level β:

$$Pr(Y = y|ML = \beta) = \frac{(o + \beta * c)^y * exp(-o - \beta * c)}{y!} \text{ where } y \in \mathbb{N} \tag{6}$$

4.4 ESTIMATING LOCAL CpG DENSITY

The number of methylated cytosines within the fragment determines the affinity of the antibody and is therefore crucial for the MeDIP enrichment signal. Assuming DNA methylation occurs in the CpG context only, the number of methylated cytosines is limited by the number of CpGs in the fragment. In order to estimate the average number of CpGs per fragment within a genomic window wd of length l_{wd}, I assume that a fragment may be centered at each genomic position with equal probability. Next, if we further assume fixed fragment length l_f, each CpG is contained in exactly l_f potential fragments. As fragments are assigned uniquely to the window containing its center position, each window is covered by l_{wd} potential fragments. In this case, $\widehat{\rho}_{CpG}$ is the expected number of CpGs per fragment for a specific window:

$$\widehat{\rho}_{CpG}(wd) = \sum_{P \in CpG} {}^{n(P,wd)}/_{l_{wd}} \tag{7}$$

where $n(P, wd)$ is the number of potential fragments that are centered within window wd and overlap genomic position P. Note that, with this definition, CpGs in the ${}^{l_f}/_2$ neighborhood of window wd also have an impact on the CpG density of that window. With increasing distance of *i* from wd, the impact of *i* on the CpG density of wd decreases, as fewer fragments overlapping *i* are centered within wd (Figure 9A, red line).

However, in practice, fragment length is variable, and CpGs further away than ${}^{l_f}/_2$ may also influence the enrichment of fragments assigned to the window. Therefore, I model the fragment length by a Gaussian distribution: $l_f \sim \mathcal{N}(\mu, \sigma)$ (Figure 9B). In this case, $n(P, wd)$ becomes the sum over all fragments centered in wd, weighted by the probability that P is contained in the fragment. For practical reasons, the tails of the distribution are cut at $\mu \pm 3 * \sigma$. Figure 9C shows the impact of a CpG on the CpG density of a window depending on the position of the CpG relative to the window. For paired end sequencing, mean fragment size and standard deviation can be derived from the alignment. For single end reads, the fragment size can be set to the target size of the library preparation ±10%.

4.5 ESTIMATING ABUNDANCE OF BACKGROUND READS

The efficiency of the MeDIP enrichment step is highly variable and can range from 25 to >100 fold [Taiwo et al., 2012]. As a result, the fraction of "original" input fragments within the MeDIP enriched sample typically varies between 2% and 40%. This fraction of input fragments is the cause of the observation of methylation independent "background reads". In order to correctly compensate for the background reads, the sample specific offset parameter must be estimated.

For this purpose, I make use of regions that lack CpG dinucleotides, and which should therefore show no MeDIP enrichment. The average number of fragments cov-

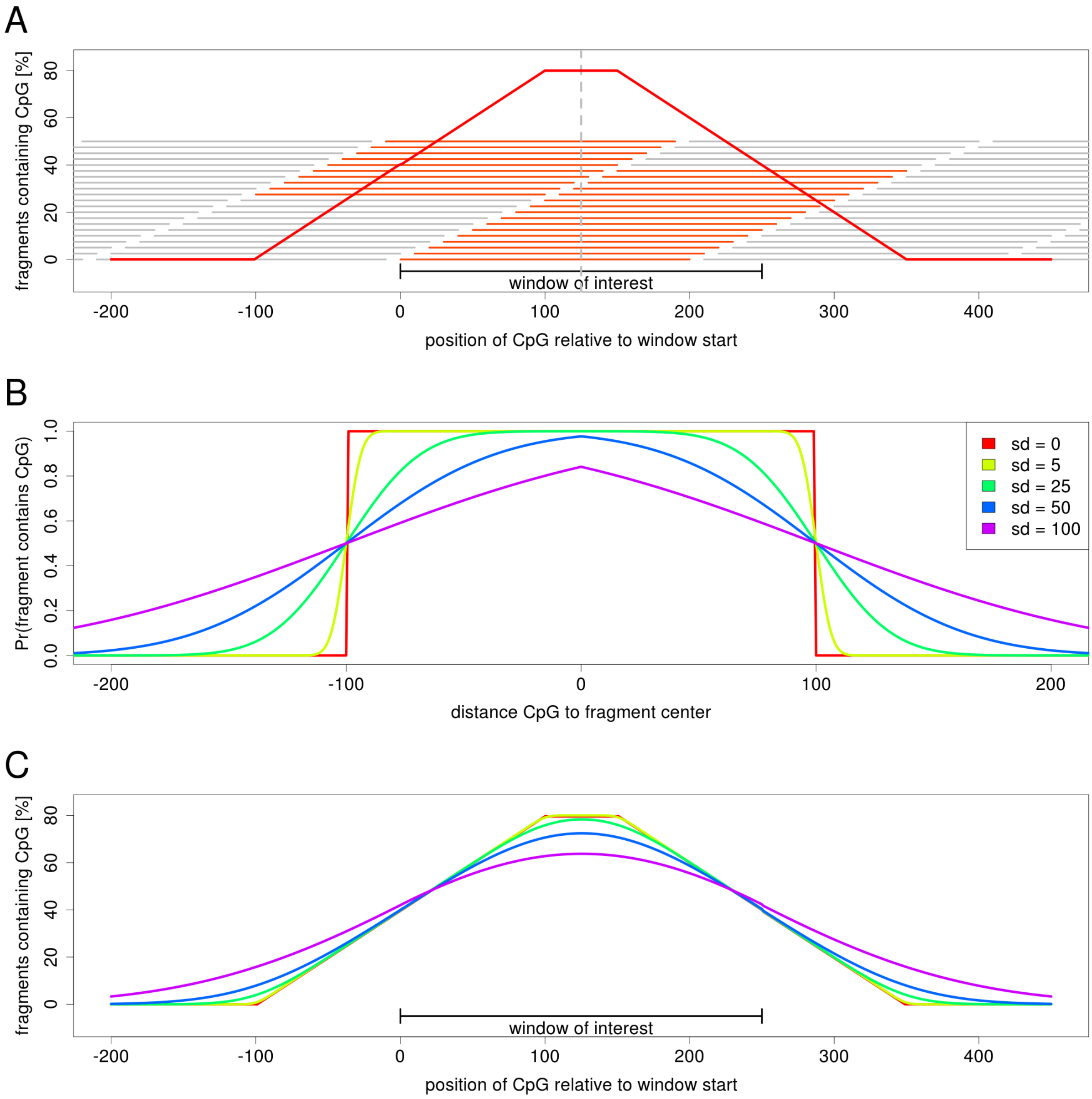


Figure 9: CpG density estimation. (A) Fraction of fragments assigned to genomic window containing CpG, assuming fixed fragment size of 200 bases. Horizontal bars represent potential sequencing fragments, orange bars are assigned to the window of interest; every 10^{th} fragment is depicted. A CpG at the center position of the window is contained in 200 of the 250 potential fragments of the window (80%). (B) Probability that a CpG is contained in a fragment with expected length 200, dependent on the distance to the fragment center, for different standard derivations. (C) Fraction of fragments assigned to genomic window containing CpG, assuming normally distributed fragment length with mean of 200 bases and different standard derivations.

ering these CpG absent windows relative to the overall average number of fragments provides an estimate for the fraction of background reads.

4.6 ASSESSING SAMPLE SPECIFIC ENRICHMENT

As observed for the MeDIP seq data, fragment enrichment in the model is linear regarding the methylation level. The slope of the enrichment is defined by the enrichment for completely methylated fragments, which depends on the CpG density. Due to different experimental conditions and sample compositions, these enrichment characteristics can vary between samples.

To assess the sample specific dependency of CpG density and enrichment, the model relies on knowledge about the methylation status for a subset of genomic regions. For these regions, the observed read densities are scaled according to the known methylation levels, resulting in the estimated enrichment given full methylation. The observed enrichment will be used in a following step to calibrate the model by inferring the general enrichment characteristics of the sample. To this end, the regions are grouped into bins of similar CpG density. For each bin, the expected maximal enrichment is estimated by averaging the scaled observed read densities. In order to allow subsequent interpolation of the enrichment to the complete spectrum of CpG densities in the genome, the regions used for this calibration must span a broad range of CpG densities. As highly methylated regions have the best signal to noise ratio, these regions are the most informative for estimating enrichment characteristics. Here, I describe three different strategies to estimate the sample specific enrichment, reflecting three different levels of prior knowledge:

"SAMPLE SPECIFIC CALIBRATION" This strategy requires additional bisulfite based calibration experiments of the same samples. These experiments provide absolute methylation levels for a subset of genomic regions, which are used to calibrate sample specific enrichment. In order to avoid noise in the calibration, these experiments are filtered for average to highly methylated regions with moderate to low variation over the samples.

"SAMPLE TYPE CALIBRATION" For this strategy, enrichment is calibrated based on regions that are consistently highly methylated in a large number of comparable samples. On account of the consistency of methylation levels within the comparable samples, it is reasonable to assume that corresponding regions are similarly methylated in the samples of interest as well, and thus can be used to calibrate the MeDIP enrichment profiles.

"BLIND CALIBRATION" This approach is based on the inverse relationship of methylation and CpG density in vertebrate methylomes. Commonly, regions with low CpG density are highly methylated, whereas methylation decreases with higher CpG density levels [Deaton and Bird, 2011b]. This correlation provides a rough estimate for the average methylation levels of windows in this range of CpG density that is used analogously to the previous calibration strategies.

For each group of genomic windows with similar CpG density, all strategies allow estimation of the mean enrichment for fully methylated regions (Figure 10B, black line). The precision of these observed group-wise enrichment factors depends on the validity of the assumptions involved in the calibration strategy, but also the number of observations for each group. Therefore, the expected precision of the observed enrichment varies for different levels of CpG density.

4.7 FITTING ENRICHMENT PROFILES

In the next step, the observed enrichment profile is smoothed and interpolated to the complete range of CpG density levels in the genome. At the same time, the enrichment estimates for groups with less observations is stabilized by borrowing strength from neighboring groups. To account for the saturation at higher levels of CpG density (Figure 8), the observed enrichment estimates are approximated by the sigmoidal function S:

$$S(x) = \frac{x}{\sqrt{1+x^2}} \tag{8}$$

This function is scaled and shifted with sample specific parameters x_1, x_2 and x_3 (Figure 10A), to match the relation of the CpG density ρ_{CpG} and the observed enrichment characteristics (Figure 10B):

$$c = cf(\rho_{CpG}) = (S(\frac{\rho_{CpG}}{x_3} - x_1) - S(-x_1)) * \frac{x_2}{1 - S(-x_1)} \tag{9}$$

The parameters are optimized to minimize the squared deviance from the observed group-wise sample specific enrichment estimates, weighted by the precision $(1/SEM)$ of the estimates. For the analyzed MeDIP samples, the function $cf(CpG)$ is capable of

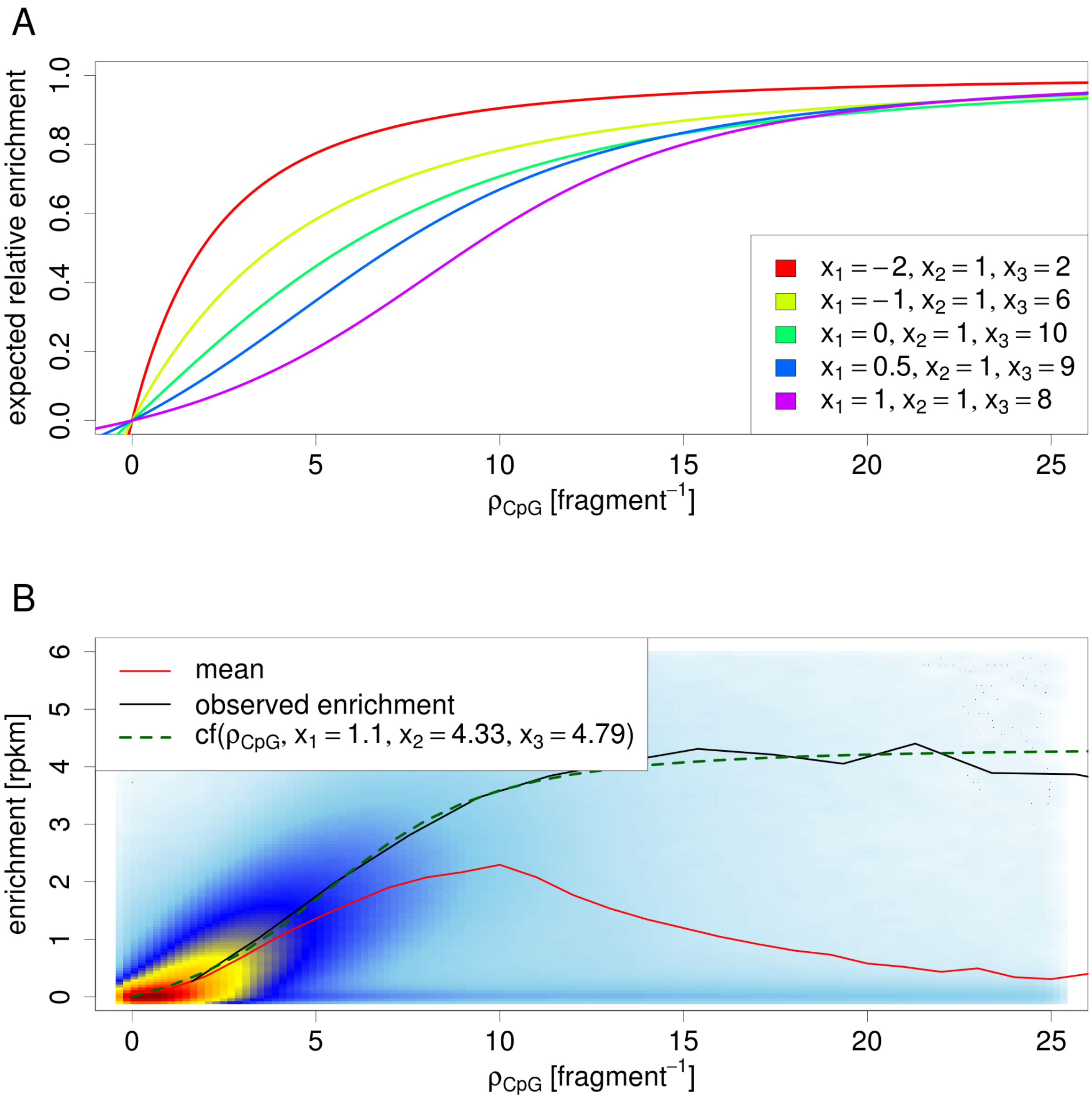


Figure 10: Sigmoidal function to model enrichment characteristics. (A) The three parameters allow the function to be fitted to the observed enrichment characteristics. x_1 controls the slope at the origin relative to the maximum slope. x_2 corresponds to the maximal enrichment. x_3 stretches the function horizontally. (B) Heat color representation of observed fragment enrichment and CpG density for one specific sample. The red line is the mean enrichment over all fragments, the black line the observed enrichment of fully methylated regions, and the dashed green line the fitted sigmoidal function cf.

fitting the observed enrichment characteristics, and it thus provides a good approximation of the observed dependency of maximal enrichment and CpG density levels.

4.8 STATISTICAL MODEL OF THE ABSOLUTE METHYLATION LEVEL

The Poisson model (5) describes the distribution of the read coverage y in genomic regions where methylation levels are known. In order to model the methylation level *ML* given the read coverage y, I apply Bayes' theorem to derive a Bayesian posterior distribution for the methylation level, given the number of reads y.

$$Pr(ML|Y) = \frac{Pr(Y|ML) * Pr(ML)}{Pr(Y)} \tag{10}$$

The prior distribution $Pr(ML)$ models probability of methylation levels without evidence from the enrichment experiment. In order to assign equal probability to all possible methylation levels, the uniform distribution in the interval $[0,1]$ is selected as a non-informative prior.

$$Pr(ML = \beta) = \begin{cases} 1 & \text{if } \beta \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The probability of observing y reads under the Poisson model (6) is derived by marginalizing over ML:

$$\begin{aligned} Pr(Y = y) &= \int_0^1 Pr(Y = y|ML = \beta)\,\mathrm{d}\beta \\ &= \frac{1}{y!}\int_0^1 (o + \beta * c)^y * exp(-o - \beta * c)\,\mathrm{d}\beta \quad \big| t = o + \beta * c \\ &= \frac{1}{c * y!}\int_{t=o}^{t=o+c} t^y * exp(-t)\,\mathrm{d}t \\ &= \frac{1}{c * y!}\gamma(y+1, o+c) - \gamma(y+1, o) \end{aligned} \tag{12}$$

where $\gamma(a,x) = \int_0^x t^{a-1} exp(-t)\,\mathrm{d}t$ is the lower incomplete Gamma function.

Plugging the likelihood (6), the prior (11) and the probability of y (12) into Bayes' theorem (10), yields the posterior distribution, which describes the probability density of the methylation level $f_{ML}(\beta)$:

$$\begin{aligned} Pr(ML = \beta|y) &= \frac{Pr(y|ML = \beta) * Pr(ML = \beta)}{Pr(Y = y)} \\ &= \begin{cases} \frac{(o+\beta*c)^{y}*c*exp(-o-\beta*c)}{\gamma(y+1,o+c)-\gamma(y+1,o)} & \text{if } \beta \in [0,1] \\ 0 & \text{otherwise} \end{cases} \qquad (13) \\ &= f_{ML}(\beta; y, c, o) \end{aligned}$$

Figure 11 B and C show examples for the posterior probability density for different sets of parameters.

It can be shown that the posterior distribution $f_{ML}(\beta; y, c, o)$ is equivalent to a truncated Erlang distribution (see Appendix A.3). This equivalence allows an alternative interpretation of the posterior distribution: Primarily, the Erlang distribution models the time x to the occurance of the k$^{\text{th}}$ event of a Poisson process with rate λ. Analogously, the posterior distribution (13) models the enrichment relative to the maximal enrichment c that explains the observed number of reads y. This enrichment is restricted to the range between the background (o) and the maximal enrichment $c + o$. Since enrichment is modeled relative to c, the Erlang distribution is truncated to the interval $[\frac{o}{c}, \frac{o}{c} + 1]$. Furthermore, the probability density, the cumulative distribution and the quantile function of this distribution are implemented in standard statistical libraries and will be used in the following to derive the Bayesian estimators.

4.9 BAYESIAN ESTIMATORS FOR THE ABSOLUTE METHYLATION LEVEL

The maximum a posteriori (MAP) estimator is given by:

$$MAP = max_{\beta}(f_{ML}(\beta, y, o, c)) = \begin{cases} 0 & \text{if } \frac{y-o}{c} < 0 \\ 1 & \text{if } \frac{y-o}{c} > 1 \\ \frac{y-o}{c} & \text{otherwise} \end{cases} \qquad (14)$$

The mean estimator of the methylation level can be derived by integrating over the posterior distribution:

$$\begin{aligned}
\hat{\beta}(y) &= E[Pr(ML|Y=y)] \\
&= \int_0^1 \beta * Pr(\beta|y)\,\mathrm{d}\beta \\
&= \frac{\int_0^1 \beta c(o+\beta c)^y exp(-o-\beta c)\,\mathrm{d}\beta}{\gamma(y+1,o+c)-\gamma(y+1,o))} \\
&= \frac{\int_0^1 (o+\beta c)^{y+1} exp(-o-\beta c) - o(o+\beta c)^y exp(-o-\beta c)\,\mathrm{d}\beta}{\gamma(y+1,o+c)-\gamma(y+1,o))} | t = o+\beta * c \\
&= \frac{\frac{1}{c}\int_{t=o}^{t=o+c} t^{y+1} * exp(-t) - o * t^y * exp(-t)\,\mathrm{d}t}{\gamma(y+1,o+c)-\gamma(y+1,o))} \\
&= \frac{\gamma(y+2,o+c)-\gamma(y+2,o)-o\gamma(y+1,o+c)+o\gamma(y+1,o)}{c\gamma(y+1,o+c)-c\gamma(y+1,o))} \qquad (15)
\end{aligned}$$

For large values of c and y, this representation becomes numerically unstable, since $\gamma(a,x)$ becomes very large, and a small relative error in $\gamma(a,x)$ leads to large relative error in $\hat{\beta}(y)$. To avoid this, the posterior mean can be expressed in terms of the cumulative distribution function of the Erlang distribution $F_E(x;\lambda,k) = \frac{\gamma(k,\lambda x)}{(k-1)!}$ with rate $\lambda = 1$, for which a stable implementation is available:

$$\begin{aligned}
\hat{\beta}(y) &= \frac{\gamma(y+2,o+c)-\gamma(y+2,o)-o\gamma(y+1,o+c)+o\gamma(y+1,o)}{c\gamma(y+1,o+c)-c\gamma(y+1,o))} \\
&= \frac{\begin{array}{l}(y+1)!(F_E(o+c;1,y+2)-F_E(o;1,y+2))- \\ \qquad oy!(F_E(o+c;1,y+1)-F_E(o;1,y+1))\end{array}}{cy!(F_E(o+c;1,k=y+1)-F_E(o;1,y+1))} \\
&= \frac{\begin{array}{l}(y+1)(F_E(o+c;1,y+2)-F_E(o;1,y+2))- \\ \qquad o(F_E(o+c;1,y+1)-F_E(o;1,y+1))\end{array}}{c(F_E(o+c;1,y+1)-F_E(o;1,y+1)} \qquad (16)
\end{aligned}$$

The median and the equal tails credibility interval are given by the quantile function $F_{ML}^{-1}(p;y,c,o)$ of the posterior. The quantile function is the inverse of the CDF

$F_{ML}(q; y, c, o)$, which is derived in Appendix A.2. In order to derive the quantile $q = F_{ML}^{-1}(p; y, c, o)$, the equation $F_{ML}(q; y, o, c) = p$ is solved for q. This can be done using the quantile function of the Erlang distribution $F_E^{-1}(p; \lambda, k)$:

$$\begin{aligned} F_{ML}(q; y, o, c) &= p \\ \frac{\gamma(y+1, o+\beta c) - \gamma(y+1, o)}{\gamma(y+1, o+c) - \gamma(y+1, o)} &= p \\ \frac{F_E(q; \lambda = 1, k = y+1) - F_E(o; \lambda = 1, k = y+1)}{F_E(o+c; \lambda = 1, k = y+1) - F_E(o; \lambda = 1, k = y+1)} &= p \\ F_E^{-1}((p * (F_E(o+c; 1, y+1) - F_E(o; 1, y+1)) + F_E(o; 1, y+1); 1, y+1)) &= q \end{aligned} \tag{17}$$

The interval containing $1 - \alpha$ of the highest posterior density (HDI) cannot be derived with a closed formula. If the probability density is lower at the $1 - \alpha$ quantile compared to 0%, or at the α quantile compared to 100%, the HDI is defined by $[0, F_{ML}^{-1}(p = 1 - \alpha)]$ or $[F_{ML}^{-1}(p = \alpha), 1]$, respectively. Otherwise, the lower and upper boundary of the interval must have equal probability density. For this case the HDI can be derived by equating the density function, using the bisection method (Algorithm 1). Figure 11 shows examples of the point estimators and the credible intervals for the methylation level.

While mean and median point estimators appear to be practically equivalent over the complete range of methylation levels, the MAP estimator agrees only at intermediate levels, deviating for methylation levels close to 0 or 1. In these cases, most of the posterior probability is on one side of the mode, and influences posterior mean and median, but not the mode. Practically, the median and the mode have two advantages over the MAP as estimators for the methylation level:

- Mode and mean are dependent on the probability of all possible explanations for the observed data, while MAP reflects only the most likely one.

- In contrast to MAP, mode and median estimators are strictly monotonically increasing: higher read counts lead to higher methylation estimates.

HDI and ETI both provide an interval containing $1 - \alpha$ of the credibility of the methylation level. In contrast to HDI, which must be approximated iteratively, the ETI can be computed directly using the quantile function of the posterior. However, especially for cases where most of the posterior probability is close to the extreme methylation levels 0

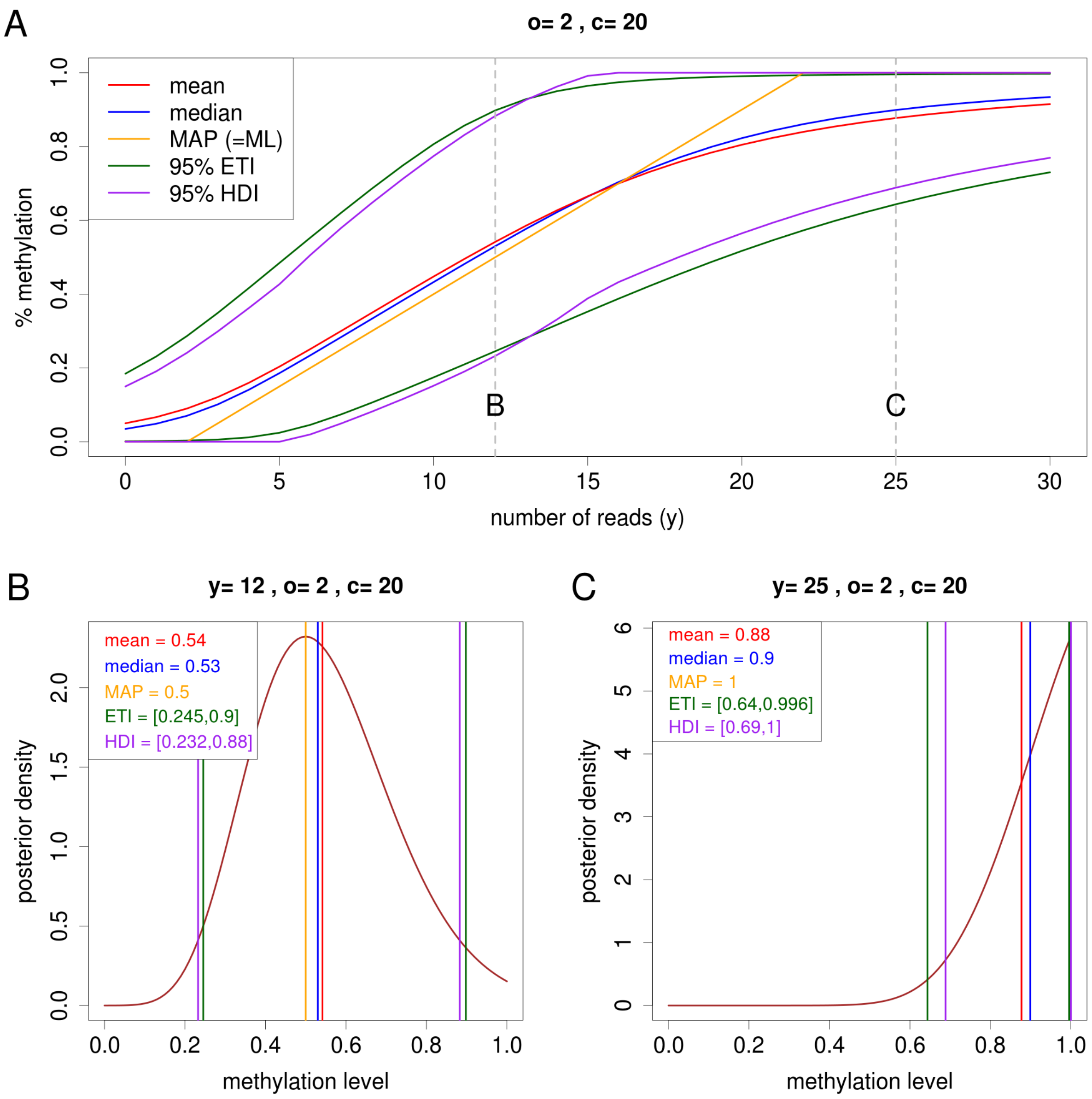


Figure 11: Comparison of Bayesian estimators. (A) Influence of observed number of reads on estimators, for a region with expected value of o=2 background reads and additional c=20 reads maximal enrichment. (B) Posterior density (brown line) and estimators for the same regions, assuming observation of 12 reads and (C) 25 reads respectively, as indicated in A. Note that for the second example C the maximum posterior is not contained in the ETI.

Algorithm 1 Highest density interval

```
function HDI($\alpha, y, c, o$)
    $lq \leftarrow F_{ML}^{-1}(p = \alpha, y, c, o)$                                ▷ check boundary cases
    $uq \leftarrow F_{ML}^{-1}(p = 1 - \alpha, y, c, o)$
    if $f_{ML}(0, y, c, o) > f_{ML}(uq, y, c, o)$ then
        return $(0, uq)$
    else if $f_{ML}(1, y, c, o) > f_{ML}(lq, y, c, o)$ then
        return $(lq, 1)$
    end if
    $step \leftarrow \alpha/2$                          ▷ Equate density function using bisection
    $lb \leftarrow step$
    while $f_{ML}(lb, y, c, o) - f_{ML}(lb + 1 - \alpha, y, c, o) < \epsilon$ do
        $step \leftarrow step/2$
        if $f_{ML}(lb, y, c, o) > f_{ML}(lb + 1 - \alpha, y, c, o)$ then
            $lb \leftarrow lb - step$
        else
            $lb \leftarrow lb + step$
        end if
    end while
    return $F_{ML}^{-1}(p = lb, y, c, o), F_{ML}^{-1}(p = lb + 1 - \alpha, y, c, o))$
end function
```

or 1, the ETI may not include values with the highest probability density (MAP), contradicting its interpretation as the interval of highest credibility. For this reason, the HDI provides a more intuitive description of the most credible values for the methylation level.

4.10 ASSESSMENT OF METHYLATION LEVEL ESTIMATION METHODS

The task of estimating absolute levels of methylation of individual samples from enrichment experiments has recently been addressed by two alternative methods: BayMeth [Riebler et al., 2014] and MeSiC [Xiao et al., 2015]. In the following section, I will briefly outline these methods, and their differences to the posterior mean estimator described above, to which I will refer to as QSEA for this comparison. Using the NSCLC and the IMR90 datasets, I will assess the performance of the methods by comparing the estimated methylation levels for the individual samples to BS derived methylation levels. For the NSCLC data, I will additionally analyze the ability of the different methods to capture individual differences between tumor and normal tissues.

4.10.1 *Alternative Approaches*

BAYMETH Similarly to QSEA, BayMeth uses Bayesian point estimators from a Poisson model to estimate methylation levels. However, the approach deviates in three main aspects: First, while the fraction of background fragments is explicitly estimated and considered in QSEA by the offset parameter o, BayMeth assumes that all fragments are due to enrichment. Second, BayMeth use informative priors, namely a Dirac-Beta-Dirac mixture, with point mass on 0 and 1. Finally, in order to estimate the parameters for the prior and the maximal enrichment rate, BayMeth groups the windows by CpG density and individually estimates the parameters for each group independent on the other groups, using Empirical Bayes method. To support parameter estimation, the authors of BayMeth suggest running an additional experiment on a sample that has been fully methylated with SssI treatment. Then, for each CpG density class, the prior and the rate parameters are fitted to maximize the likelihood of the observed number of reads in the primary and the SssI treated enrichment experiment. Alternatively, the method can be calibrated without the additional SssI treated experiment by fitting the parameters based on the enrichment experiment only.

MESIC MeSiC estimates methylation levels based on Random Forrest Regression (RFR) models. Each model predicts the expression for a set of genomic regions, defined by sequence features and functional elements. In particular, the authors consider CpG islands related annotation, gene related features and repetitive elements. For each of the genomic elements, the model is trained to fit methylation level estimates based on MeDIP-seq to BS-seq values, using data for H1 cell line. These models are considered general and used to estimate methylation levels for other samples with the same parameter set. Therefore, MeSiC neither offers options to correct for CNV influences, nor to provide information for the calibration of the method. Despite the technical limitations of the enrichment approach, the authors claim to estimate methylation at base resolution, and report methylation levels for individual CpGs. MeSiC is implemented as an online tool; the authors provide a web-site allowing read counts per genomic window to be uploaded in a specific format. The analysis is performed on the web server, and the user is informed when the results are available for download.

4.10.2 *Parameter Calibration*

BayMeth optionally allows the model parameters to be optimized using an additional calibration experiment with the same sample that has been fully methylated by SssI treatment. This calibration experiment is available for the IMR90 dataset, and the method performance can be assessed in both situations: "SssI calibration", using the SssI experiment, and "blind calibration", without using additional data. For the NSCLC benchmark dataset the calibration experiment is not available, and the evaluation is based on the "blind calibration" strategy only.

MeSiC does not allow to provide additional data to fit the model parameters, but the model has been trained on a dataset provided by the authors. The authors claim that these parameters are general such that MeSiC does not require parameter fitting.

As described above in Section 4.6, I suggest three different strategies to fit the enrichment profiles of the model, reflecting three different levels of prior knowledge:

- For "sample specific calibration", I directly use bisulfite derived methylation levels of the same sample to train the model. For the NSCLC dataset, I select between 146,455 and 184,099 genomic windows that are at least 50% methylated in the corresponding BS experiment, and at least 70% methylated in at least half of the samples. Accordingly, I selected 135,619 genomic windows that are at least 70% methylated in the IMR90 dataset.

- To demonstrate the strategy of "sample type calibration" for the NSCLC dataset, I used methylation values from microarray measurements of 54 adenocarcinoma samples and 32 adjacent normal tissue samples [TCGA Consortium, 2014], as well as 49 squamous cell lung cancer samples and 37 adjacent normal tissue samples [TCGA Consortium, 2012]. From these cohorts, I identified 18,587 genomic windows with average methylation levels > 0.9 over all samples, covering the full range of CpG density levels. Due to the lack of equivalent calibration data for the IMR90 dataset, I did not apply this calibration strategy on the IMR90 dataset.

- According to the "blind calibration" strategy, I assume that regions with low CpG density are 80% methylated on average and that with increasing CpG density, methylation decreases linearly to 25% for the mean CpG density of CpG islands (CGIs). These values have been applied for both the IMR90 and NSCLC datasets.

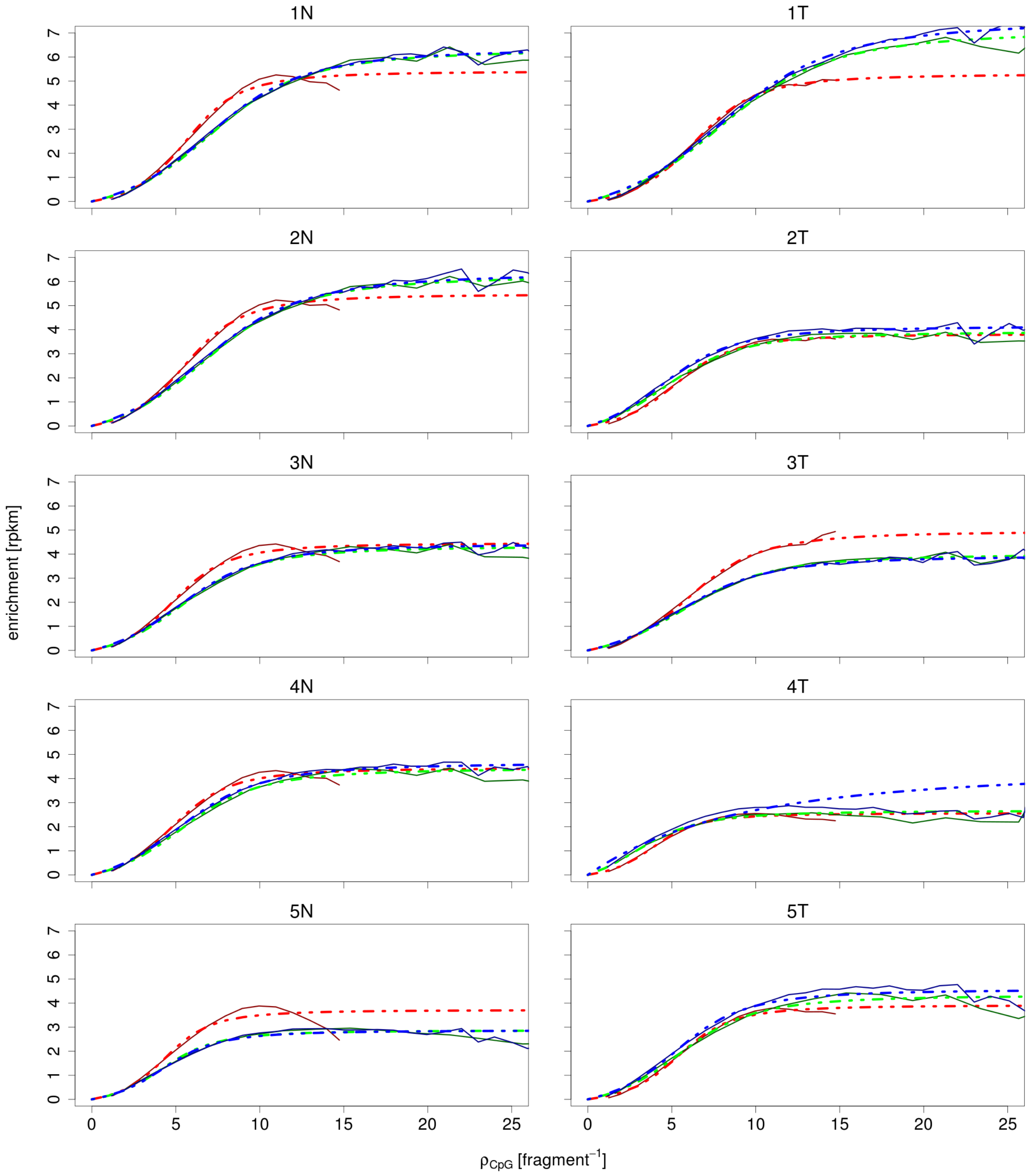


Figure 12: Enrichment characteristics for the 5 tumor (T) and normal (N) sample pairs of the NSCLC dataset. Solid lines represent observed enrichment, and dashed lines fitted sigmoidal function for blind calibration, sample type calibration, and sample specific calibration in red, green and blue, respectively

While the enrichment profiles differ between the samples, all three calibration strategies lead to similar results for a given sample (Figure 12). The similarity of the estimated enrichment profiles indicates the robustness of the estimation with regard to the different strategies.

4.10.3 *Performance Comparison*

IMR90 DATASET In order to assess the influence of the parameter estimation strategies of the different approaches, I compared the MeDIP-seq methylation level estimates with bisulfite measurements for different ranges of CpG density. While QSEA, both with sample specific bisulfite calibration (BS calibration) and blind calibration provides unbiased estimates for all ranges of CpG density, BayMeth generally overestimates intermediate methylation levels, especially for low CpG levels (Figure 13A). Surprisingly, this effect is more evident in the SssI calibrated estimates, leading to better performance of the blind estimate for this dataset. MeSiC overestimates methylation levels for genomic regions with low CpG density and underestimates methylation at high levels of CpG densities. Genome-wide, Spearman correlations of QSEA methylation estimates with 450k are high for both "sample specific calibration" (0.819) and "blind calibration" (0.805). BayMeth results in a correlation of 0.786 with "blind calibration" and 0.655 with "SssI calibration", reflecting the observation for lower CpG ranges. Methylation estimates of the MeSiC RFR model compared to 450k results in a correlation of 0.594 (Figure 13B).

NSCLC DATASET In line with the results from the IMR 90 dataset, QSEA performs comparably well for all three calibration configurations, as expected for the similar enrichment profiles. Genome-wide, QSEA estimates for all three calibration strategies are highly correlated for tumor and normal samples of all patients, resulting in Spearman correlation coefficients between 0.75 and 0.84. Correlation of BayMeth estimates with blind calibration is 0.64 and MeSiC estimates 0.38 on average (Figure 13B). Estimates for tumor tissue samples perform slightly better with QSEA and BayMeth, compared to the corresponding normal sample. For MeSiC, methylation levels of tumor tissue samples perform worse compared to normal tissue.

In contrast to QSEA, both alternative methods incorporate the distribution of methylation levels for different CpG levels, either indirectly, with the sequence features in MeSiC, or directly, with the informative prior in BayMeth, which is fitted for different levels of CpG density. This incorporation leads to the CpG dependent bias that can be

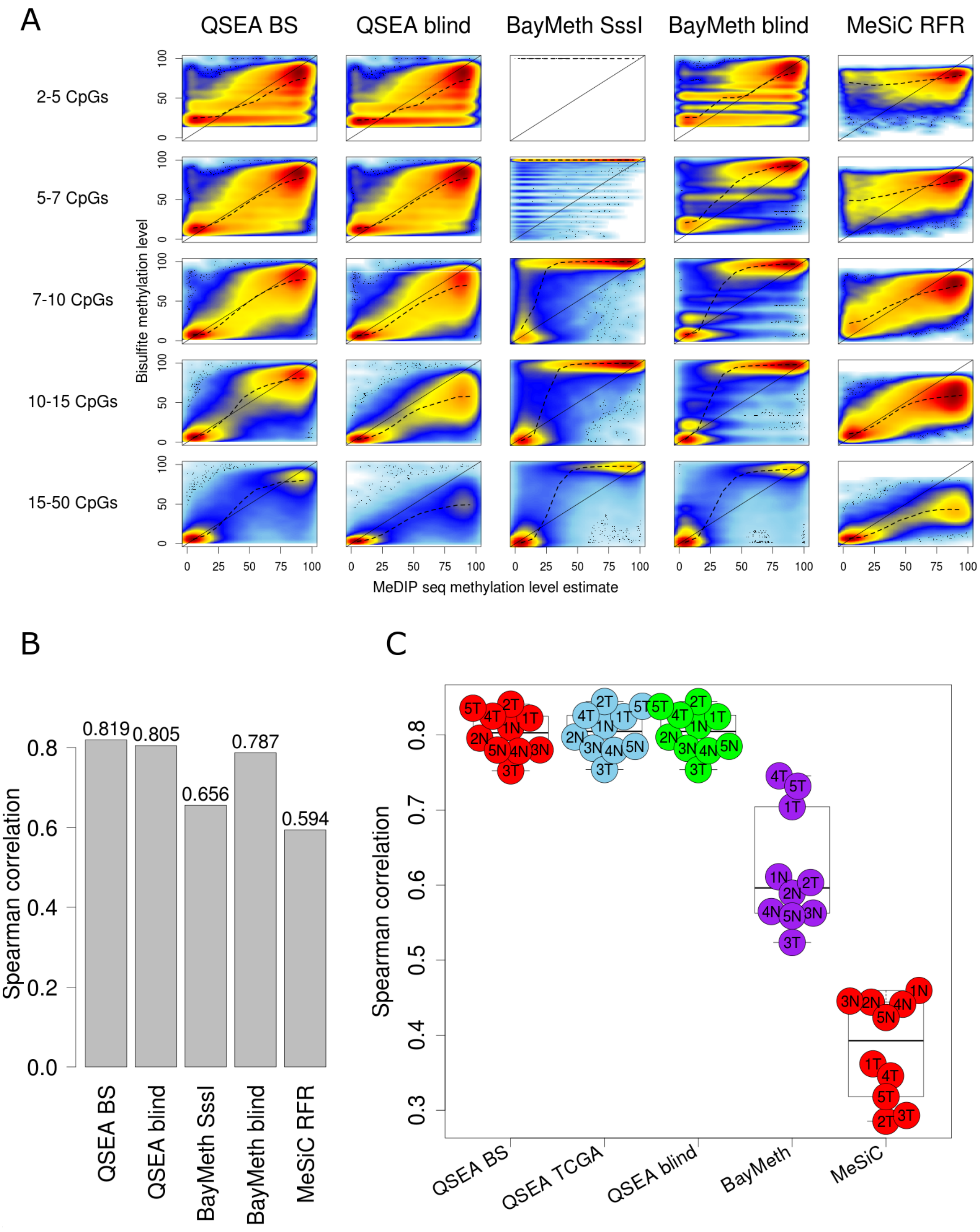


Figure 13: Comparison of methylation level estimation methods: (A) Comparison of bisulfite and MeDIP methylation level estimates for IMR90 dataset, for different ranges of CpG density. (B) Correlation of genome-wide MeDIP absolute methylation level estimates and bisulfite measurements for IMR90 dataset and (C) NSCLC dataset

observed for these methods. For example, genomic regions with low CpG density feature high methylation levels. This prior information leads to high methylation estimates for these regions, even for regions with intermediate enrichment.

METHYLATION DIFFERENCES In addition to assessing the accuracy of the methylation estimates, the NSCLC dataset provides the opportunity to analyze how well the methods correctly quantify methylation differences between pairs of samples, which is essential for comparative methylation analysis, in particular for the functional interpretation of the effect of the methylation differences. In order to analyze the ability of the different methods to capture individual differences between tumor and normal tissues, I calculated Spearman correlation coefficients between MeDIP-seq and BS-seq tumor-normal methylation differences for each patient. On average this correlation is 0.71 for "blind calibration" and 0.73 for "TCGA calibration" and "BS calibration" modes, 0.44 for BayMeth "blind calibration", and 0.02 for MeSiC. For comparison, the pairwise correlation between the BS tumor-normal differences of different patients is 0.51 on average. Based on this correlation analysis, I performed hierarchical clustering. For all calibration modes, QSEA estimates tightly cluster with the corresponding BS values, while for the other methods the sample relationships cannot be recovered (Figure 14). This implies that the differences between BS sequencing and the QSEA MeDIP estimates are minor compared to the differences between the tumor patients. In contrast, the alternative methods have higher variability between MeDIP and bisulfite estimates and are thus not appropriate for quantifying the differences in methylation levels.

4.11 CONCLUSION

In this chapter, I developed the QSEA method for transformation of relative methylation enrichment to absolute levels of methylation. The method is based on a statistical model that takes the CpG dependent enrichment characteristic as well as the abundance of unspecific background reads into account. From the model, I derived the posterior distribution, given the observed number of sequencing reads, and Bayesian point estimators as well as credibility intervals for the methylation level. The parameters of the model are biologically interpretable and can be derived from the data. To this end, I suggested 3 calibration strategies, reflecting different levels of prior knowledge.

I assessed the performance of my approach, as well as alternative methods by comparing the methylation level estimates with bisulfite validation data of the same samples. For the estimation of methylation levels for individual samples, QSEA and BayMeth

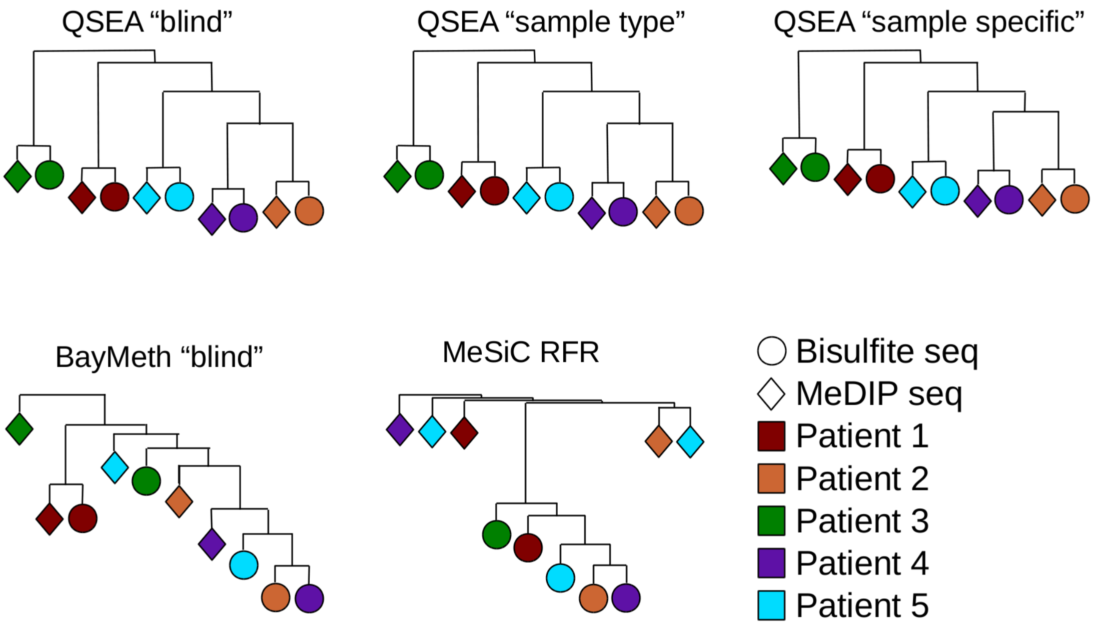


Figure 14: Performance of methylation level transformation methods with respect to differences between samples for NSCLC dataset. For all calibration strategies, QSEA estimates cluster with respective bisulfite values, while BayMeth and MeSiC estimates do not recapitulate the bisulfite pattern.

	IMR90	NSCLC	Δ NSCLC
QSEA sample specific	**0.819**	0.8	0.71
QSEA sample type	NA	**0.81**	**0.73**
QSEA blind	0.805	0.81	0.73
BayMeth SssI	0.656	NA	NA
BayMeth blind	0.787	0.64	0.44
MeSiC RFR	0.594	0.38	0.02

Table 2: Performance of methylation level estimation approaches, assessed by correlation coefficient of methylation level estimates to bisulfite based methylation levels for IMR 90 and NSCLC datasets, as well as correlation of estimated tumor-normal differences for NSCLC dataset (Δ NSCLC). Correlations for NSCLC dataset are averages over all samples. For each comparison, the best performing method is highlighted in bold.

performed comparably well. In contrast, MeSiC failed to recapitulate bisulfite methylation levels, especially for the tumor samples. One possible explanation for the poor performance might be overfitting of the predefined parameters to the cell line used for training. These parameters may fail to generalize to the irregular alterations of the tumors.

Quantifying the difference of methylation levels between different conditions, tissues or cell types is crucial to assessing the biological effect of the differences. To this end, I assessed the ability of the methods to reliably estimate tumor-normal methylation differences. While QSEA methylation differences are in accordance with bisulfite results, differences of estimates by the alternative methods have high deviance to bisulfite differences. This deviance exceeds the variability between the patients, disqualifying the alternative methods for the quantification of methylation level differences. Therefore, QSEA is the appropriate method for studies assessing methylation level differences in development and disease.

However, methylation estimates depend on the validity of the assumptions made for a given calibration strategy. In particular, for the blind calibration, these assumptions have to be checked carefully. Even though the assumptions worked equally well for the two benchmark datasets, samples with different methylation level distributions, such as from embryonic stem cells during epigenetic reprogramming, or from different species, may require that the strategy be modified.

5

DIFFERENTIAL METHYLATION ANALYSIS

The moderate cost of enrichment based sequencing methylation experiments makes the approach widely applicable for the analysis of large sets of samples. Detecting differences in DNA methylation profiles between groups of samples allows the functional impact of epigenetics to be investigated in a variety of settings: first, detecting distinct methylation profiles in different types of tissues helps in finding cell type specific function of DNA methylation [Ziller et al., 2013]. Furthermore, aberrant methylation in disease animal models or human disease samples compared to healthy control samples points toward the epigenetic basis of disease related regulatory malfunctions [Weber et al., 2005b]. Finally, differential methylation between treated and untreated samples allows to assess the epigenetic effect of chemicals, drugs and environmental conditions [Heyn and Esteller, 2012]. These analyses require efficient and reliable statistical tests.

Given two groups of samples Y_1 and Y_2, with n_1 and n_2 samples, the statistical test assesses the significance of a difference δ in the mean methylation levels of the two groups $\bar{\beta}_{Y1}$ and $\bar{\beta}_{Y2}$. For genome-wide analysis, the statistical tests are applied to tens of millions of regions. Calculation of the test statistics for all these regions must be performed in a reasonable amount of time. Therefore, it is necessary the statistical tests be based on simple models and efficient algorithms to estimate the parameters of the model.

5.1 WILCOXON'S RANK SUM TEST

Wilconxon's rank sum test is a non-parametric test based on the assumption that groups of observations Y_1 and Y_2 follow statistical distributions that differ only by a shift δ in the mean parameter of the distribution. Under the general assumption that enrichment

rank order	W	$p - value$
xxxxxyyyyy	0	0.0079
xxxxyxyyyy	1	0.0159
xxxyxxyyyy	2	0.0317
xxxxyyxyyy	2	0.0317

Table 3: Significant cases for Wilcoxon's rank sum test for two groups of samples with size 5

is a monotone function of the methylation level, the test can be applied to absolute methylation level estimates, as well as directly to the normalized read counts, since the test statistic is independent of the distributions of Y_1 and Y_2. Under the null hypothesis $H_0 : \delta = 0$, the order of the pooled sample is independent of the grouping. The test statistic W is the sum of ranks of the first group Y_1 minus the smallest possible rank sum for the n samples of Y_1:

$$W_{n_1,n_2} = \sum_{i=1}^{n_1} R(y_{1,i}) - \frac{n_1(n_1+1)}{2} \tag{18}$$

For small samples n_1 and $n_2 < 20$ the exact distribution of W under H_0 is derived based on the random order of the pooled sample. For larger sample size, the distribution of the test statistic is approximated by a normal distribution $W_{n_1,n_2} \sim \mathcal{N}(\mu, \sigma)$ with parameters $\mu = \frac{n*m}{2}$ and $\sigma = \sqrt{\frac{n_1 n_2 *(n_1+n_2+1)}{12}}$.

The Wilcoxon test has a simple test statistic which can be computed efficiently. However, especially in the situation of low sample size, the test has limited power. For group size of 5, such as in the NSCLC dataset, the critical value of the test statistic W at a significance level $\alpha = 0.05$ is 2, which leaves only 4 distinct orders with p-value < 0.05 (Table 3). With the minimal p-value of 0.0079 for a test with 2 times 5 samples, the statistical power is insufficient to survive multiple testing correction: in order to be significant at 10% FDR, the number of tests considered can be at most 12. Nevertheless, the application of this test in the context of differential methylation analysis has been proposed to avoid making assumptions on the distribution of the read counts or methylation levels, and thereby to minimize the computational complexity [Chavez et al., 2010; Grimm et al., 2013; Chavez, 2011]. The authors circumvent multiple testing correction and apply additional filters on the mean count and logFC in order to prevent false positive results.

5.2 GENERALIZED LINEAR MODEL LIKELIHOOD RATIO TEST

According to the model presented in Section 3.2.4.2, the read counts of individual samples within the group follow a negative binomial distribution with $\mu_i = o_i + c_i * \bar{\beta}_x$, and dispersion parameter ϕ, reflecting the variance within the groups. By fitting Generalized Linear Models (GLMs), maximum likelihood parameters of the negative binomial distributions are determined for both groups simultaneously. Next, the likelihood ratio to a null model, which does not distinguish between the groups, serves as test statistic to assess the significance of the difference of group means.

Due to its flexibility and efficiency this approach is widely used for differential analysis of quantitative sequencing data, mainly in the context of transcriptome analysis with RNA-seq [M. D. Robinson et al., 2010; Anders and Huber, 2010].

5.2.1 *Parameter Fitting with GLM*

GLMs model the mean parameter of a distribution μ_i dependent on a linear predictor η_i. η and μ are associated by a a smooth and invertible link function g. For the purpose of differential enrichment, g describes η as the logarithm of sample mean μ_i relative to the expected maximal enrichment c_i (see Section 4.7). Thus, the exponential of the linear predictor can be interpreted as an estimate of the group mean methylation levels $\bar{\beta}_Y$ that neglects the offset of background reads. The linear predictor for sample i is a linear combination of m explanatory variables $x_{i,j}$ scaled by the coefficients b_j.

$$y_i \sim NB(\mu_i, \phi) \tag{19}$$

$$log(\frac{\mu_i}{c_i}) = \eta_i \tag{20}$$

$$\eta_i = b_1 * x_{1,i} + ... + b_m * x_{i,m} \tag{21}$$

$$\begin{pmatrix} \eta_1 \\ ... \\ \eta_n \end{pmatrix} = \begin{pmatrix} x_{11} & ... & x_{1m} \\ & ... & \\ x_{n1} & ... & x_{nm} \end{pmatrix} \begin{pmatrix} b_1 \\ ... \\ b_m \end{pmatrix} = Xb \tag{22}$$

The matrix of explanatory variables X is called the design matrix and specifies the relation of the samples to the coefficients.

Applying the iteratively re-weighted least squares algorithm (IRLS) provides the optimal coefficients b, which yield the mean parameter μ of the negative binomial distribution that maximizes the likelihood for the observed data. At each step of the IRLS algorithm, the dispersion parameter ϕ is re-estimated by equating the model deviance to the residual degrees of freedom [Venables and Ripley, 2013]:

$$2\sum(y_i * log(\frac{\mu_i}{y_i}) - \sum(y_i + \phi) * log(\frac{\mu_i + \phi}{y_i + \phi}) = n - m \tag{23}$$

5.2.2 *Model Structure*

The dependency of the samples on the coefficients for the model M_1 is encoded in the design matrix X. In order to use GLMs to fit the mean parameters for two groups, the design matrix has two columns to assign the samples to groups: the first column corresponds to the intercept and is 1 for all samples. The second column is 0 for each sample from the first group and 1 for each sample from second group. In this model, the linear predictor for samples of the first group is $\eta = b_1$, and for samples of the second group $\eta = b_1 + b_2$. Thus, b_1 corresponds to the log enrichment of the first group, and b_2 the log fold change of enrichment of first and second group. Since this model corresponds to the alternative hypothesis of the statistical test H_1, stating a difference between the groups, it is called alternative model M_1. The residuals of M_1 have $(n_1 + n_2) - 2$ degrees of freedom. Figure 15 depicts the structure of this model.

However, the framework is not restricted to modeling two groups. By defining additional explanatory variables, more complex models can be constructed. For example, let Y_1 and Y_2 be two sets of tissue samples from the same donors, before and after treatment. If the methylation varies between individuals, it might be reasonable to introduce a patient specific offset on η. In this case, the design matrix would be extended by one column for each donor, whose component values would be one if the sample belonged to the corresponding donor, and zero otherwise. This model structure corresponds to a paired experimental design. Analogously, the approach can be extended to compensate for experimental differences between the samples that would otherwise bias the results. All model parameters are estimated simultaneously.

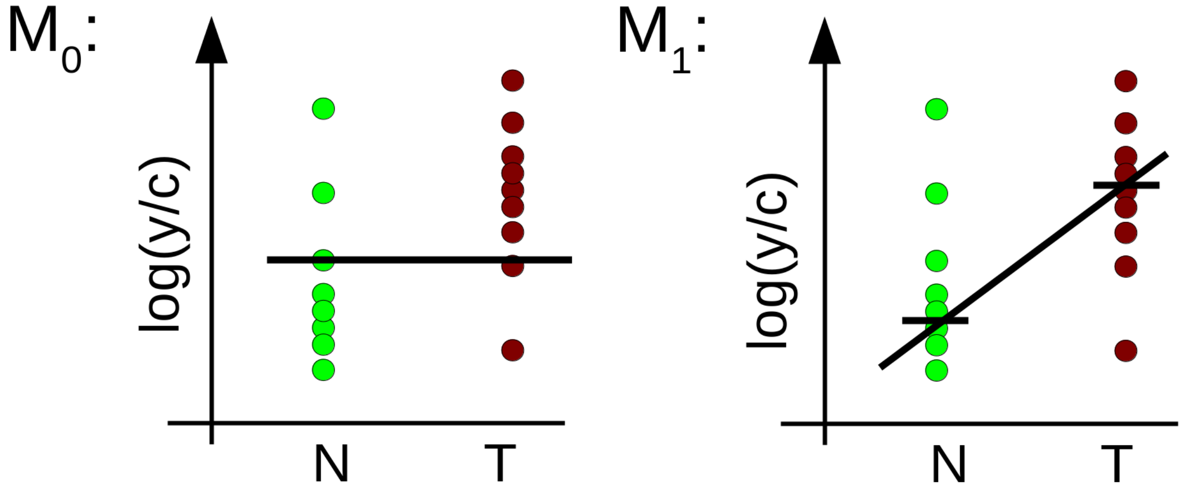


Figure 15: Example of model structure of GLM. For M_0, the linear predictor is only dependent on the intercept, while M_1 involves a second explanatory variable modeling the group difference. For both models, the model coefficients are fit to optimize the likelihood of the linear predictor (black line), assuming negative binomial distribution for the read counts y.

5.2.3 *Null Model*

H_0 states there is no difference between groups; hence, for the two variable model with intercept and group coefficients, the second coefficient b_2 equals zero. According to this hypothesis, a null model M_0 is fitted, where the linear predictor does not depend on the group coefficient b_2. Therefore, the design matrix X for M_0 is reduced by the second column, and only the intercept coefficient b_1 remains. M_0 is called a nested model in M_1, as M_1 can be transformed into M_0 by constraining the coefficient b_2 to zero. Since M_0 has one parameter fewer than M_1, the residual degrees of freedom of the null model $rdf(M_0)$ exceed the residual degrees of freedom of M_1 by one. As an example, the null model corresponding to the model with intercept and group assignment is depicted in Figure 15.

5.2.4 *Statistical Test*

For significance testing, the likelihood of the complete model M_1 is compared to the nested null-model M_0. If there is no difference between group mean methylation levels, the group assignment is irrelevant for the model, and the likelihood ratio (LR) of the full model and the reduced model can be approximated by a χ^2-distribution with $rdf(M_0) - rdf(M_1)$ degrees of freedom.

$$-2\log\frac{\mathcal{L}(M_0)}{\mathcal{L}(M_1)} \sim \chi^2_{\Delta rdf} \tag{24}$$

Hence, if this test statistic exceeds the $(1-\alpha)$-quantile of the χ^2-distribution, H_0 can be rejected, and the difference of group means is significant at level α.

5.3 COMPARISON OF STATISTICAL TESTS

In this section, I assess the suitability of the presented statistical approaches for detection of DMRs. Since the set of true DMRs is not known, I base the benchmark on differential regions detected by bisulfite sequencing as the gold standard. The targeted bisulfite experiment of the NSCLC dataset covers 109,224 genomic 250 base regions with at least 100 reads on average, about 1% of the genome. Of these, 1,692 regions have gain and 1,020 loss of methylation in tumor compared to normal tissue at 1% FDR, according to a beta-binomial test [Ziller et al., 2013]. For the purpose of this benchmark these regions are considered truly differentially methylated in order to assess the power of the tests. However, Methyl-seq targets do not represent the whole genome, but enrich known regulatory regions of the genome, such as CpG islands. Therefore, the ratio of hyper and hypomethylated regions is not comparable to the results from the genome-wide enrichment based approaches.

Next, I detected tumor-normal DMRs from MeDIP-seq using the two methods presented above, the WRS test and the GLM-LR test, and compared the results to regions found differentially methylated by targeted bisulfite sequencing. Using the GLM-LR test, 83,453 out of the 12,382,699 genome-wide regions are detected as differentially methylated at FDR $<$ 1%. Of those, 11,129 are hypermethylated and 72,324 are hypomethylated in the tumors. As discussed above, the power of the WRS test is not sufficient to correct for multiple testing. In accordance to the practice in previous studies [Grimm et al., 2013], regions with p-value < 0.01, $|\text{logFC}| > |log(4/3)|$ and mean normalized read coverage of > 0.25 rpm in at least one of the groups are considered differentially methylated. These criteria yield 134,589 DMRs, of which 13,846 regions show gain and 120,743 regions loss of methylation. Respectively for gain and loss of methylation, 55.6% and 62.4% of the regions detected by GLM-LR approach were also detected by the WRS test (Figure 16A and B).

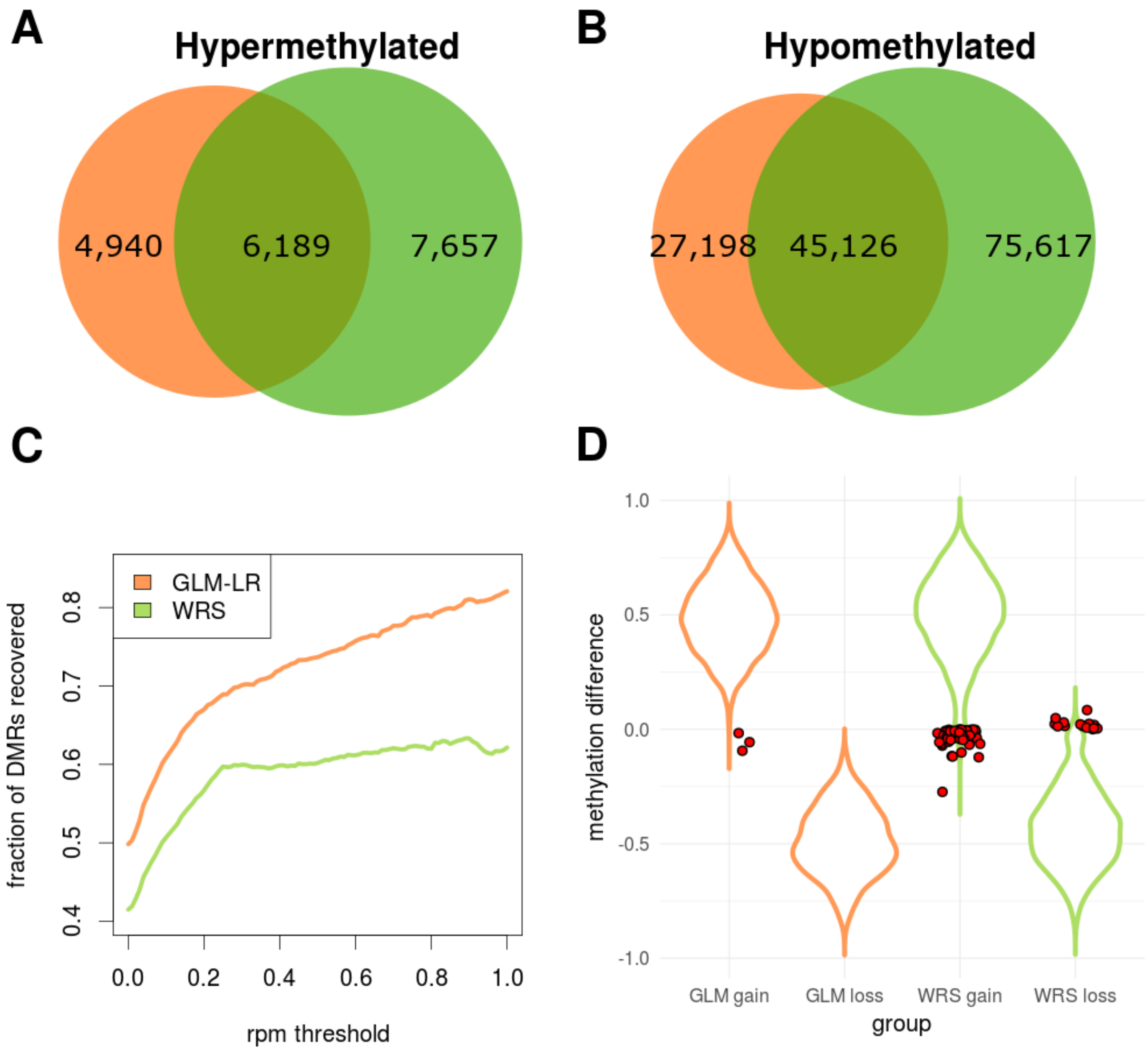


Figure 16: Comparison of generalized linear model likelihood ratio test (GLM-LR, orange) and Wilcoxon Rank Sum test (WRS, green). Overlap of hypermethylated (A) and hypomethylated (B) regions, detected by the two approaches. (C) Fraction of recovered true positive DMRs defined by targeted BS-seq, depending on MeDIP-seq mean read coverage. (D) Distribution of methylation level differences estimated from BS-seq, for the DMRs found by the two approaches. Red dots indicate DMRs with opposing bisulfite estimates.

Overall, GLM-LR retrieved 49.8%, and WRS 41.4% of those true positive DMRs. However, the power of the test is highly dependent on the average read coverage within the region: the higher the mean coverage within the region, the higher the fraction of retrieved true DMRs (Figure 16 C). For regions covered by at least 0.6 rpm, the fraction of recovered true DMRs is 75.7% and 61.2% for GLM-LR and WRS tests respectively. Since the enrichment is dependent on the CpG density, this correlation indicates that the power of enrichment based approaches is generally higher in CpG rich regions, such as CpG islands. Even though GLM-LR detected 38% fewer DMRs compared to WRS, the fraction of recovered true DMRs is higher at all coverage levels, indicating the greater power of the approach.

In order to assess the type-1 error of the tests, I evaluate methylation level differences within the DMRs estimated from the bisulfite sequencing. From the regions detected by GLM-LR, only 3 with gain, and none with loss of methylation have opposing difference as estimated by the bisulfite experiment. In contrast, for WRS, 51 regions with gain and 16 regions with loss of methylation are contradicted by the bisulfite estimates. These numbers indicate a higher rate of false positive results for WRS, compared to GLM-LR (Figure 16 D).

5.4 CONCLUSION

The comparison of the two statistical tests indicated both greater power and lower false positive rate for the GLM-LR approach, compared to the WRS test. In contrast to WRS, GLM-LR provides sufficient statistical power to control the rate of false discovery. Accordingly, the decision can be based solely on the statistical evidence, and does not depend on arbitrary thresholds. Despite being statistically more complex compared to the non-parametric WRS test, GLM-LR can be computed efficiently using the IRLS algorithm, allowing its application to millions of genomic regions. Additionally, GLM-LR provides a more flexible framework, which facilitates modeling of different experimental designs and allows technical influences such as batch effects to be modeled explicitly.

In summary, GLM-LR is superior to WRS for the detection of DMRs from methylation enrichment based sequencing experiments. This conclusion is also supported by the results from [Lienhard et al., 2014], where the properties of the two statistical approaches have been compared for a MeDIP-seq dataset of mouse intestinal samples (discussed in Supplementary Text 1 of the publication).

6

IMPLEMENTATION

The methods described in the previous chapters are implemented as two R packages, MEDIPS and QSEA, which are available within the Bioconductor repository. The Bioconductor repository ensures the installation of the packages to be simple, and provides access to a great variety of resources for genomic reference sequences and annotations. While QSEA has more general applicability and implements all methods presented within this thesis, MEDIPS has some complementary functionality, and a broad user basis. Therefore, both packages are actively maintained and developed.

6.1 MEDIPS PACKAGE

MEDIPS is a bioconductor package for the analysis of enrichment based methylation data, initially described in [Chavez et al., 2010]. The focus of the functionality of the package is on quality control metrics and the detection of differentially methylated regions (DMRs). With about 2,000 downloads from distinct IPs in 2016, MEDIPS is among the 20% most widely used packages within Bioconductor.

Since the methods for detecting differential enrichment can in principle be applied to other enrichment based experiments, such as ChIP-seq, the functions are implemented generically. For example, [Lienhard and Chavez, 2016] describes the application of MEDIPS for the detection of cell type specific histone mark H3K4me2 from ChIP-seq experiments.

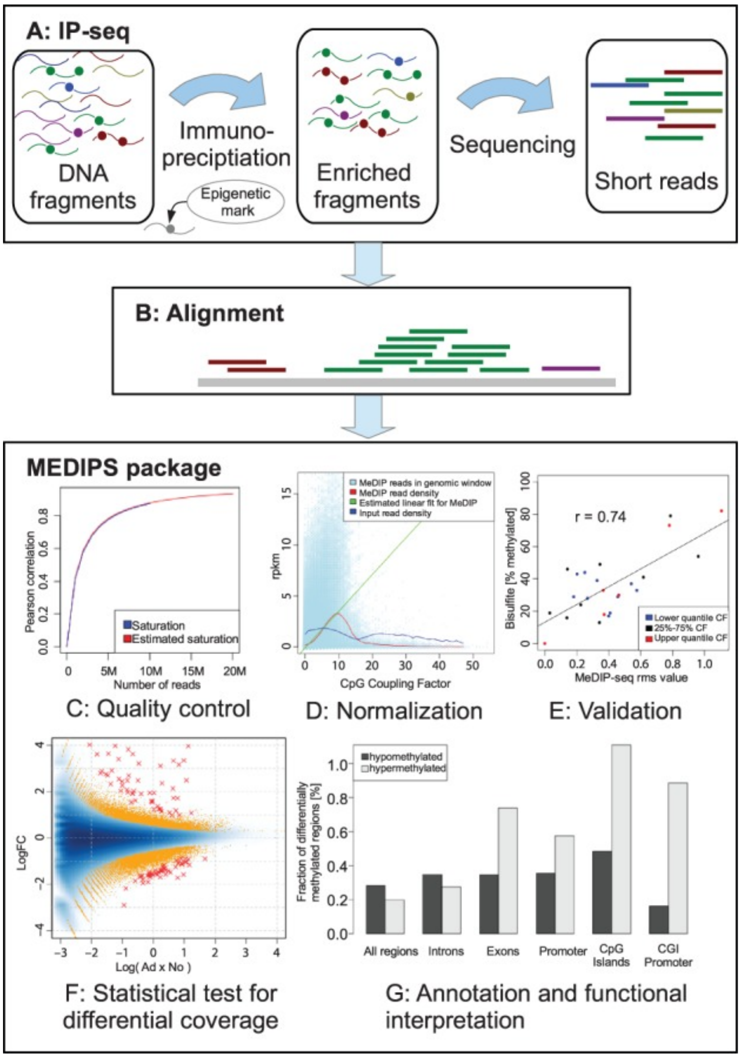


Figure 17: Workflow and functionality of the MEDIPS analysis pipeline.

6.1.1 *Functionality of MEDIPS*

Figure 17 depicts the MeDIP-seq work-flow, as well as the main components of the current version of the MEDIPS package. MEDIPS provides the following functionality:

PREPROCESSING The computational analysis starts with the importing of alignment files into MEDIPS which counts sequencing fragments within genome-wide tiling windows. In order to estimate local CpG density, sequence information from the reference genome is imported using the BSgenome bioconductor package [Pagès, 2016], currently providing reference genome sequences for 24 different species.

QUALITY CONTROL MEDIPS provides two methods for quality control of MeDIP-seq data: first, saturation analysis aims to assess whether the depth of sequencing is sufficient by estimating the saturation of the coverage profile (Figure 17E). This is done by assessing the correlation of downsampled subsets of the read counts. Second, the efficiency of methylation enrichment is assessed by calculating the depth of coverage at genomic CpG sites.

NORMALIZATION In MEDIPS, between sample normalization is achieved by scaling the read counts by a sample-specific factor proportional to the total read counts, and, in versions since 2014 by upper quantile or TMM scaling factors. However, the design, the data structure, and the functions do not allow local scaling factors to be applied, hence MEDIPS cannot consider CNVs in normalization.

ABSOLUTE METHYLATION VALUES MEDIPS provides a scaling CpG density normalization to retrieve an absolute methylation score (AMS). While this normalization improves the correlation to bisulfite sequencing data (Figure 17E), MEDIPS does not actually transform the enrichment values to absolute methylation levels. AMS values are thus not directly comparable to bisulfite sequencing and not interpretable as % methylation.

DIFFERENTIAL METHYLATION ANALYSIS Initially, two statistical tests were implemented for the detection of DMRs: the t-test as well as the non-parametric Wilcoxon rank sum test. To address the issues with the statistical test described in Section 5.1, I extended the functionality of the package and implemented the statistical test based on the likelihood ratio of nested GLMs in an update (Figure 17F), [Lienhard et al., 2014]. The implementation is based on the methods implemented in the edgeR bioconductor package, which applies the same methods for the analysis of differentially expressed genes from RNA-seq data.

ANNOTATION To assist in the functional interpretation of genomic regions, MEDIPS provides gene based annotation features by accessing the ENSEMBL database.

6.2 QSEA PACKAGE

The implementation of region specific normalization factors to account for variations in copy number is not compatible with the MEDIPS data structure. Therefore, I developed a new analysis package, QSEA, as the successor of MEDIPS. QSEA stands for "quantitative sequencing enrichment analysis". It implements CNV normalization, as well as transformation to absolute levels of methylation. These novel features extend the field of applications of the analysis package: In tumor samples, a large fraction of the genome is affected by structural variations; accordingly, normalizing for CNV is essential. Transformation to absolute methylation levels greatly enhances the interpretability of the data, since it facilitates:

- the quantification of effect size and biological interpretation of the differences,
- the comparison to BS data, and
- the comparison of methylation levels in different regions of a genome or across different genomes.

In addition to the functional extensions, QSEA implements several improvements on different steps of the analysis, enhancing the usability of the package. The individual analysis steps of the QSEA workflow are depicted in Figure 18.

6.2.1 *Preprocessing*

In QSEA, the description of the samples can be imported from a sample table containing all information on files and properties of the samples. Similarly to MEDIPS, the reference genome is divided into windows, and the number of fragments mapping to each genomic window is counted. Scanning the alignment files and counting reads within genome-wide windows is computationally expensive and consumes a considerable fraction of the total run time of the analysis. For this reason, the preprocessing step has been parallelized in QSEA, allowing multicore processors to be utilized.

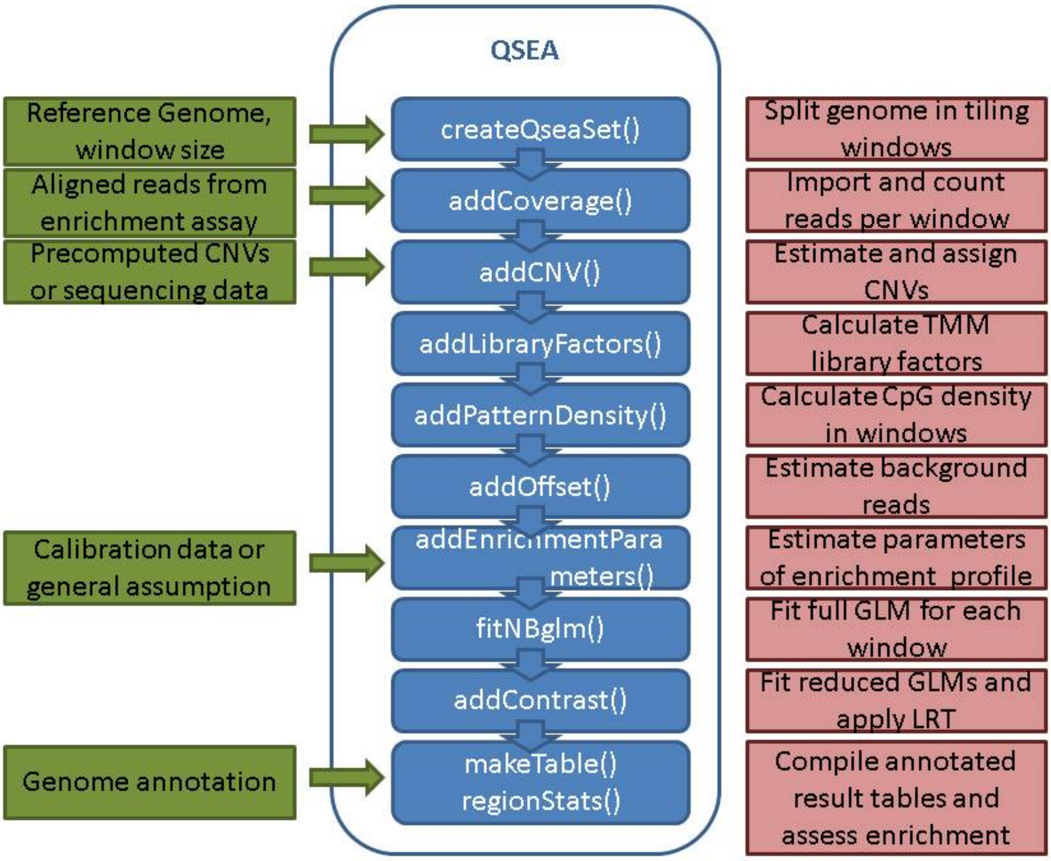

QSEA
Reference Genome, window size
Aligned reads from enrichment assay
Precomputed CNVs or sequencing data
Calibration data or general assumption
Genome annotation
createQseaSet()
addCoverage()
addCNV()
addLibraryFactors()
addPatternDensity()
addOffset()
addEnrichmentParameters()
fitNBglm()
addContrast()
makeTable()
regionStats()
Split genome in tiling windows
Import and count reads per window
Estimate and assign CNVs
Calculate TMM library factors
Calculate CpG density in windows
Estimate background reads
Estimate parameters of enrichment profile
Fit full GLM for each window
Fit reduced GLMs and apply LRT
Compile annotated result tables and assess enrichment

Figure 18: Overview of functionality and work-flow of the QSEA analysis package. Green boxes represent data input, functions implemented in QSEA are depicted in blue, and red boxes describe the respective analysis step performed within these functions. Modified from [Lienhard et al., 2016]

6.2.2 *Normalization*

In addition to the sample specific scaling normalization factors, which are also available in MEDIPS, the QSEA package implements an extended data structure, enabling the specification of region specific normalization factors. This allows explicit modeling of CNVs, which is important for the analysis of tumor samples. To facilitate the usage of this novel feature, the functionality is supplemented by methods for estimating CNVs from sequencing data. Alternatively, CNVs can be estimated by an dedicated experiment, for example using genotyping arrays [Carter, 2007], whole-genome sequencing [Alkan et al., 2009] or a combined strategy. In this way, the resolution and fine mapping of the CNVs breakpoints can be optimized.

CNV ESTIMATION QSEA incorporates a method for estimating CNVs from whole genome sequencing which segments the genome into regions with similar read density using the HMM algorithm [Lai et al., 2016]. In the case where sequencing of the input library is available, this method can be directly applied to estimate CNVs.
For the other case, where whole-genome sequencing is not available for the samples under analysis, I developed and implemented a strategy to apply the same method also on methylation enriched sequencing data [Lienhard et al., 2016]. To prevent distortion by the enriched reads, only fragments without CpG dinucleotides are considered. These fragments are completely unmethylated and thus unaffected by the MeDIP enrichment step. About 10% of the fragments of a typical MeDIP library do not contain any CpGs and can be used to estimate CNVs.
To review this strategy, I applied the method on MeDIP-seq as well as low coverage whole genome sequencing (input-seq) for the 5 tumor and corresponding normal tissue samples from the NSCLC dataset. I found a striking consistency of CNVs detected using the two different types of sequencing data: The two approaches agree for 85% to 98% of the genome, and Spearman correlation of CNV profiles is between 0.89 and 0.97 for the individual samples. Figure 19 provides an overview of genome-wide CNV profiles for the tumor samples, estimated from the two different data sources. The high similarity indicates that both methods are equally reliable.

6.2.3 *Absolute Methylation Values*

The central novelty of QSEA is the transformation of enrichment signals to absolute levels of methylation, as described in the introduction of Section 4. Besides the CpG density of the region, this transformation depends on two parameters that need to be estimated from the data: first, background read abundance, and second, CpG-dependent read enrichment characteristics. The estimation of these parameters is implemented following the concepts described in chapter 4. Briefly, CpG density is estimated as the number of dinucleotides per fragment, assuming uniformly distributed read coverage and normally distributed fragment length. To estimate the abundance of background fragments, QSEA makes use of the average read coverage in CpG free regions. The CpG density dependent enrichment is observed from regions where methylation levels are known or can be estimated roughly. The enrichment profiles are then smoothed and extrapolated by fitting a sigmoidal function to the observations.

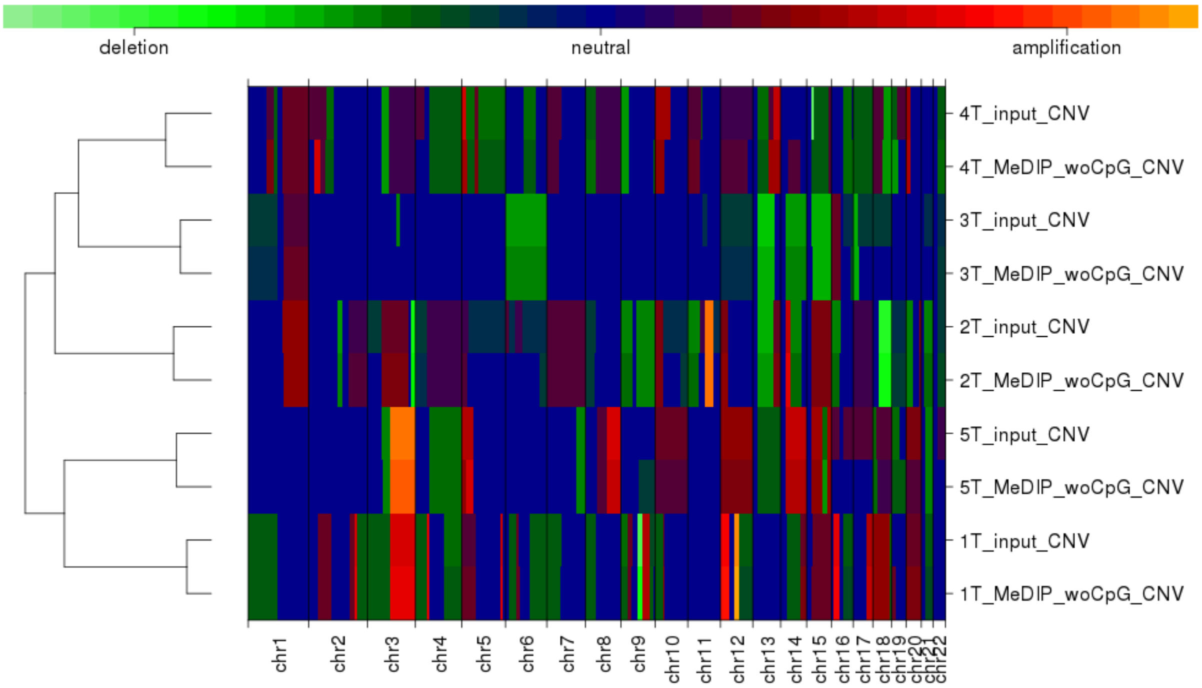


Figure 19: Genome-wide overview of CNVs for the 5 NSCLC patient samples. CNVs estimated from input and MeDIP-seq are highly similar for each sample.

6.2.4 *Quality control*

To assess the quality of the data, QSEA focuses on the coherence of estimated model parameters and assists in judging whether the data is suitable for addressing the research hypothesis. The approaches are complementary to the saturation and coverage analysis implemented in MEDIPS, which primarily assess the adequate depth of sequencing. In QSEA, insufficient sequencing depth is not directly detected, but results in broader credibility intervals of the methylation level estimates.

ENRICHMENT AND BACKGROUND PARAMETERS The parameters of the model used to estimate absolute methylation levels help in assessing the efficiency of the MeDIP enrichment step of the protocol and thus provide a useful measure of quality for the MeDIP-seq experiment. Insufficient enrichment of methylated fragments results in a high fraction of background reads and poor estimates of enrichment characteristics. Depending on the quality of the antibody and the DNA, the fraction of background reads typically varies between 5% and 40% of the library. Correspondingly, the observed enrichment is expected to be dependent on the CpG density, following the sigmoidal function. More background reads require higher depth of sequencing in order to obtain precise estimates for the methylation levels. Estimated background fractions above 60% are indicative of deficient enrichment, and such samples should be discarded. Poor fit of the enrichment characteristic may indicate violation of calibration assumptions and/or experimental bias. Figures 20 A to D depict the background fraction and enrichment characteristics for high and poor quality MeDIP-seq samples.

EXPLORATORY ANALYSIS QSEA provides exploratory data analysis methods, allowing the assessment of whether the research hypothesis is reflected by the estimated methylation profiles. This hypothesis may be the correlation of the methylation profiles and primary influences, such as the treatment or disease under study. Furthermore, exploratory data analysis assists in identifying secondary influences, such as age, sex, environmental or experimental factors. The relationship between samples can be depicted in two dimensions by principle component analysis (PCA) on the methylation estimates. Since methylation levels are stable for the major part of the genome, the PCA is focused on the most variable regions only. This selection allows the most prominent alterations between the samples to be depicted, while avoiding the noise emerging from the unchanged part. Comparing the first principle components to experimental conditions helps in identifying primary and secondary influences on the overall methylation levels, and relating the effect size to the variability within homogeneous groups. Furthermore, this approach helps to detect outliers which may indicate experimental errors and may require special consideration or exclusion. Figure 20 E depicts the relationship of methylation profiles for the NSCLC samples, indicating strong alteration between tumor and normal samples, but also a big variability within the tumor samples compared to normal samples.

6.2.5 *Differential Methylation Analysis*

The GLM implementation of the edgeR package, which is used within MEDIPS, is not ideally suited for the application on genome-wide windows for two reasons: first, the GLM functionality within edgeR is based on a package-specific data structure, and the transformation of the data to this format requires additional computational resources. Second, the edgeR package was designed to estimate parameters for roughly 30.000 genes, rather than tens of millions of genomic windows. Thus, run-time and memory efficiency were not primary goals in the development of edgeR, and for convenience information is often stored redundantly. For these reasons, I optimized the run-time

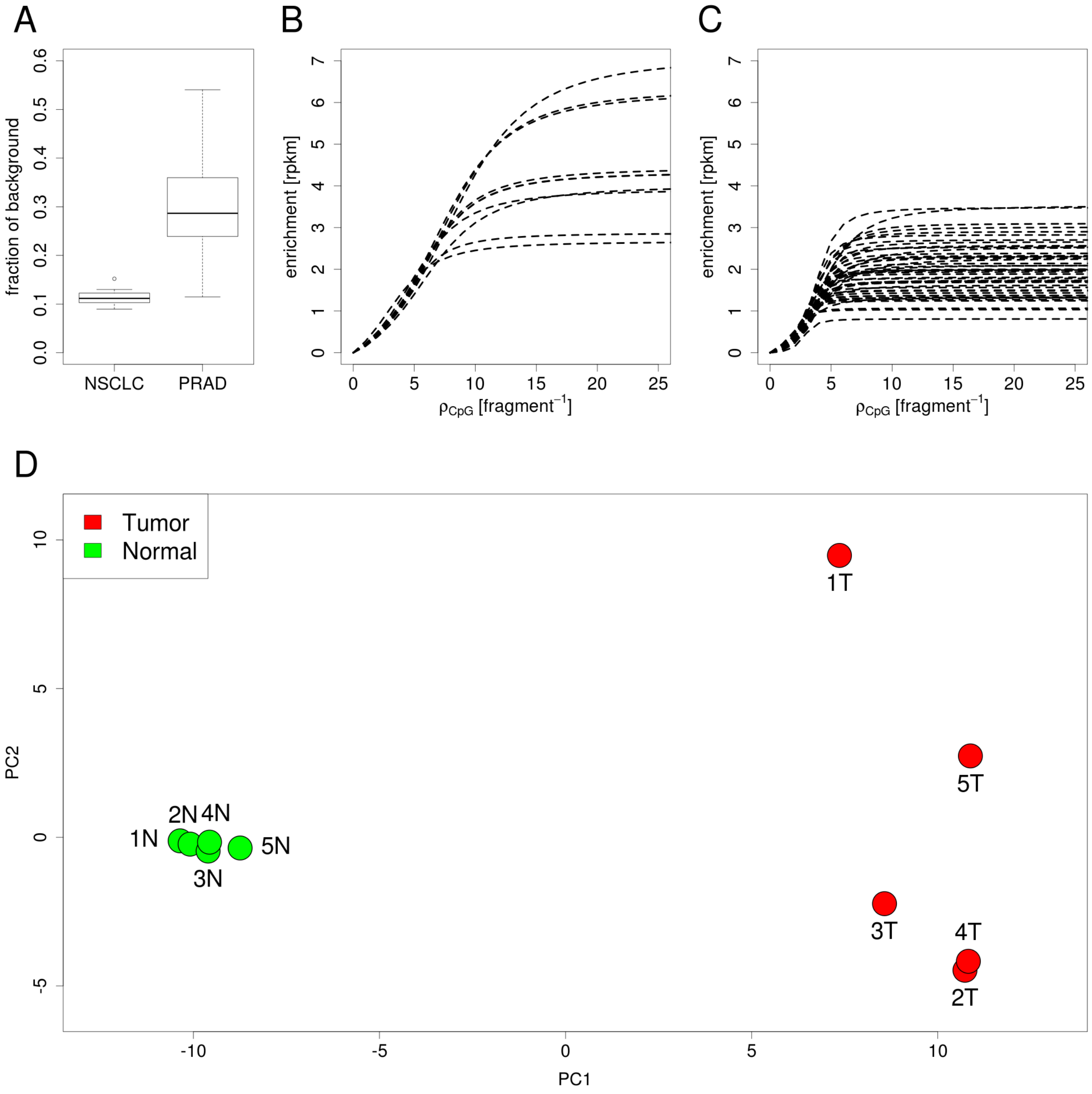


Figure 20: Quality control metrics in QSEA for the high quality NSCLC dataset, and for an unpublished prostata adenocarcinoma MeDIP-seq dataset (PRAD) of questionable quality. (A) Fraction of sequencing fragments originating from background reads is around 10 % for NSCLC, but up to 55% for PRAD dataset. (B) Enrichment characteristics of individual samples for NSCLC dataset. (C) For PRAD dataset, reduced MeDIP enrichment is expressed in degraded enrichment characteristics. (E) Principle component analyisis for NCLC samples separates tumor from normal samples, and shows variability within tumor samples.

efficiency and memory consumption of the iteratively re-weighted least squares algorithm to fit the GLMs within QSEA differential enrichment analysis.

6.2.6 *Assessing Regulatory Effects of DNA Methylation*

In order to approach the functional interpretation of identified DMRs, these regions can be associated with known genomic features. For example, the UCSC table browser [Karolchik et al., 2004] provides several types of genomic features for different reference genomes, which are suitable for such an analysis. Gene based annotation, such as RefSeq [Pruitt et al., 2007] can be used to associate regions with transcriptional start sites (TSS) and to compare gene body and intergenic regions. Sequence based features, such as model based CpG islands, shores and shelves and predicted transcription factor binding sites, mark regions with potential regulatory influence and can be used to focus the analysis on previously known effects. Furthermore, annotations based on experimental data in several cell lines is available form the ENCODE [ENCODE Consortium, 2012]. For the human genome, this resource provides experimentally validated cell type specific as well as generic binding sites for 161 transcription factors. QSEA provides two approaches to infer functional implications of DNA methylation using genomic annotation:

- By associating regions with specific features to neighboring genes, the regulatory effect of DNA methylation can be inferred and validated.
- Enrichment of classes of genomic features may indicate functional mechanisms responsible for the epigenetic alterations, as well as help in inferring regulatory effects of the altered methylation at the regions.

GENE ASSOCIATION Pursuant to the well described relation of CpG Island promoter hypermethylation and gene silencing, integrating DNA methylation patterns near transcriptional start sites (TSS) with gene expression information allows the regulatory influence of DNA methylation to be inferred. The main difficulty for this approach is matching a genomic region to a gene on which the region potentially has regulatory influence. The most common strategies are either to assign each genomic region to the next TSS, or to assign all regions within a certain distance (e.g. $\pm 1kb$) of the TSS. The regions considered can be further restricted to CpG islands, known TFBS or heterochromatic regions that have been identified before. QSEA can use annotation from several sources to annotate genomic regions of potential interest and provides functions to create output tables for these regions, containing all relevant information. In any case, these approaches assign multiple genomic regions to one gene, of which potentially only one or few truly have an impact on the expression of the gene. To enable a one to one matching of genomic region and gene, QSEA provides the option to select the window with the most significant difference from all regions that have been assigned to a gene. Resulting gene lists can then be subject to further downstream analysis, such as gene set enrichment analysis (GSEA, [Subramanian et al., 2005]), overrepresentation analysis, or induced network analysis [Herwig et al., 2016]. Such downstream analysis serves to assign functional mechanisms to the differentially methylated regions.

ENRICHMENT ANALYSIS Typically only a minor fraction of DMRs can be assigned to a TSS. The function of DNA methylation at regions outside the CpG Islands at gene promoters is still elusive, and they are covered to a lesser extent by targeted approaches such as 450k human methylation microarrays or Methyl-seq. In order to exploit the full potential of genome-wide methylation information from enrichment based methylation assays, QSEA facilitates inference of functional mechanisms by providing methods to analyze the enrichment of annotated features within DMRs. For example, if binding sites of specific transcription factors are particularly affected by differential methylation, these factors might directly be involved in the alteration of the methylation patterns. Alternatively, transcription factor binding may be methylation dependent, and a factor specific regulatory mechanism may be responsible for alteration of methylation levels.

6.2.7 *Runtime and Memory Performance*

The complete QSEA analysis of 10 human MeDIP-seq samples from the NSCLC dataset, including CNV analysis using low coverage input sequencing takes 95 minutes on a single core computer and allocates a maximum of 14 GB main memory. A large part of the run time is required for processing the alignment files: import of MeDIP-seq alignment files and counting of reads overlapping genome-wide 250 base windows takes 37 minutes, and CNV analysis including the import of low coverage input alignment files takes 11 minutes. The analysis of CpG density of the human genome takes 21 minutes. Calculation of the remaining normalization parameters, including calculation of effective library size, estimation of offset reads and analysis of MeDIP enrichment takes about 2 minutes. The detection of differentially methylated regions takes 13 minutes to fit the full model and estimate the dispersion for genome-wide windows, and 12 minutes to fit the nested null model and test the contrast.

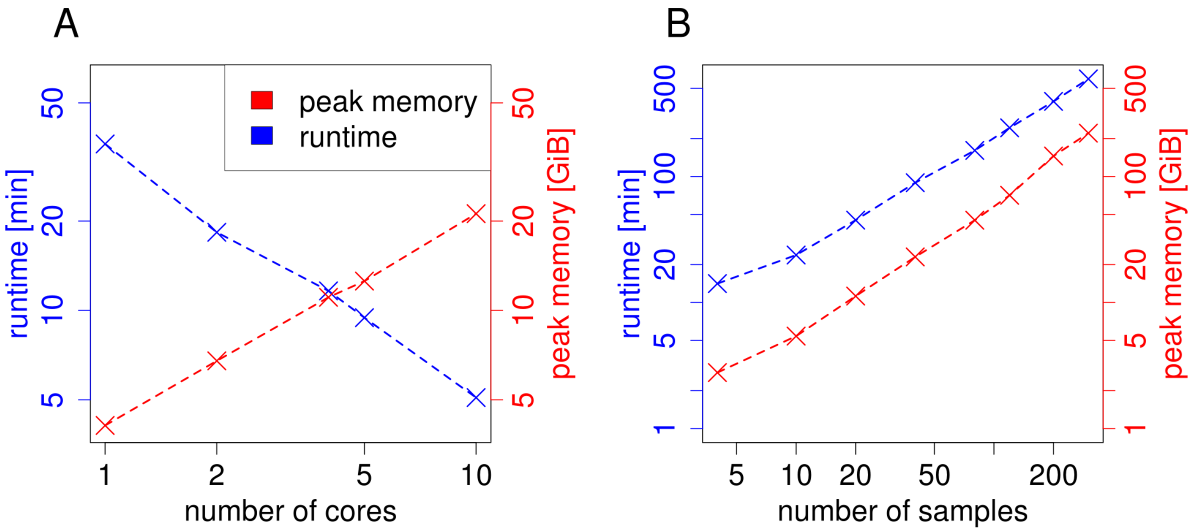


Figure 21: Run-time and memory requirements for QSEA analysis. (A) Required computational resources for the import of alignment files for 10 samples in QSEA with parallel processing, and (B) for the remaining QSEA analysis, including calibration of the enrichment model and differential analysis, for different numbers of samples, using one core.

QSEA supports parallel scanning of alignment files on multicore computers, which reduces run time for this step to 5 minutes on 10 cores (Figure 21A). Importantly, once these computational steps are performed, QSEA provides functions to retrieve all information for any regions of interest, for example, regions defined by genome annotations or by differential methylation. Normalized values and methylation estimates for those regions are computed on request, rather than precomputed for all genome-wide windows and stored separately. This approach allows both efficient usage of memory, as well as fast and flexible access to results of interest. For example, it takes about one minute to compile a table for all 105,426 genome-wide DMRs of the NSCLC dataset, containing the raw read counts, normalized coverage, and estimated methylation values including the credibility interval for the estimates and additional comprehensive annotation.

6.3 CONCLUSION

The two biocoductor packages, QSEA and MEDIPS, provide a comprehensive collection of methods for the analysis of enrichment based methylation experiments. The packages are easy to use, especially for non-expert computer users. Integration in Bioconductor facilitates the installation process and ensures the coherency of package versions and environment. All functions are documented, including descriptions of all parameters and output values, and executable examples demonstrating their usage. Furthermore, detailed tutorials demonstrate the application of the packages for practically relevant use cases. The methods are implemented efficiently, by making use of optimized bioconductor packages e.g. for processing and counting of sequencing reads. In QSEA, parallelization of the most time consuming steps further increases the efficiency of the work-flow. The packages are capable of processing large datasets, with several hundreds of samples (Figure 21B). In general, the implemented methods are versatile and can be applied to multiple protocols. For example, differential enrichment analysis also finds application in other genome-wide enrichment assays, such as ChIP-seq. The main improvement of QSEA over MEDIPS is the implementation of Bayesian estimators for the methylation level. This step is presupposed by many use cases, such as assessing whether a specific region is methylated or unmethylated, comparing with bisulfite based assays, and charting whole genome methylation landscapes. Furthermore, normalization for CNVs enables the application of the methods to samples with structural variations, such as tumor samples. Although QSEA has advanced functionality compared to MEDIPS, the two packages also have complementary aspects: while quality metrics in MEDIPS primarily assess the sufficiency of sequencing depth, QSEA is focused on the coherence of estimated model parameters, and the data’s suitability for addressing the research hypothesis.

7

APPLICATIONS

Both packages, QSEA and MEDIPS, have been intensively used to analyze methylome data. In this chapter, I present four recent studies, in which I contributed to the analysis, and which are of relevance for the development of the methods presented. For each study, I outline the experimental and computational methods, highlight the main results and conclusions, and expand on the impact of the study on this thesis.

[Grimm et al., 2013] detected altered methylation patterns in colon cancer mouse models using enrichment based methylation analysis, and thereby identified epigenetic pathways involved in tumorigenesis. To this aim, biological replicates of each sample were produced, allowing within group variability for the detection of DMRs to be considered.

[Lienhard et al., 2016] served as a pilot study to assess the application of enrichment based sequencing in a clinical study. To this aim, DNA methylation was profiled by MeDIP-seq as well as bisulfite-seq for the same samples. This dataset provided the basis for the development of the transformation to absolute methylation levels, as described in this thesis.

[Grasse et al., in preparation] avails the results of the pilot study and extends the cohort to assess the influence of DNA methylation on chemotherapy treatment response in lung cancer. With this analysis, we identified a potential epigenetic biomarker to predict carboplatin response.

[Seumois et al., 2014] used histone ChIP-seq to detect memory specific enhancers in the immune cells of asthmatic donors. The study showcases the versatility of the approach by providing an example of the use of the methods on ChIP-seq data.

7.1 DNA-METHYLOME ANALYSIS OF MOUSE INTESTINAL ADENOMA

In this study, we investigated the impact of DNA methylation on early tumor development using a colon cancer mouse model [Grimm et al., 2013]. We found targets of Polycomb Repressive Complex II (PRC2) to be enriched among regions with aberrant methylation in intestinal adenoma. Further, we identified a core set of DMRs that is conserved between early adenoma in mouse and advanced human colon cancer, indicating functional importance of these regions. Since aberrant methylation at these sites was detected in early stages of tumor development, they may serve as candidates for the selection of specific clinical epigenetic biomarkers for early cancer diagnosis.

7.1.1 *Introduction*

Histone modifying factor *Polycomp repressive complex II* (PRC2) interacts with DNA methyltransferases (DNMTs) to initiate long-term silencing of genes via de-novo methylation. Accordingly, deregulated PRC2 activity is associated with aberrant epigenetic patterns of tumor cells, resulting in the silencing of tumor suppressor genes. Intestinal tumors are often caused by genetic loss of the tumor suppressor APC, leading to hyperactivation of Wnt/beta-catenin signaling. Subsequently in tumor development, further tumor suppressors, like Cdkn2a, Dkk1, Sfrp, or Hic1, are epigenetically silenced. In contrast to well characterized genetic mutations, time and mode of occurrence of these epigenetic alterations during tumorigenisis is largely unknown. APC deficient mice (APC^{MIN}) develop multiple intestinal adenoma and thus resemble the early stage of the human colon cancer phenotype. We use MeDIP-seq of APC^{MIN} mice to characterize epigenetic alterations in early tumor development, resulting in a comprehensive catalog of differentially methylated regions in the adenoma.

7.1.2 *Methods*

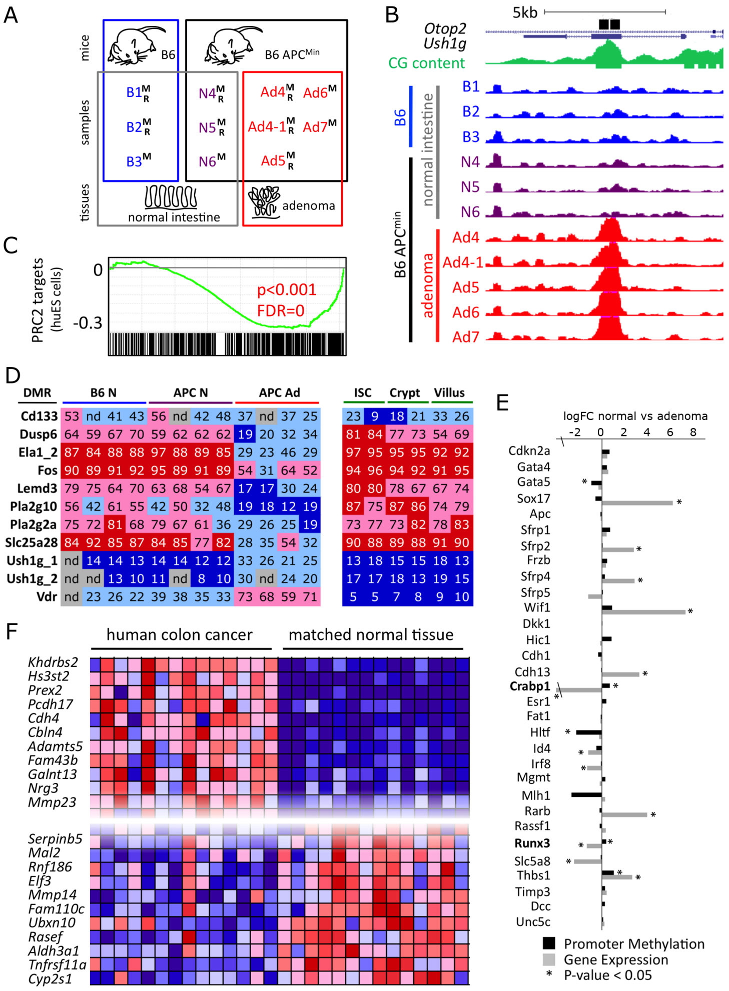
Using MeDIP-seq, we assessed the methylome of 5 adenoma and 3 normal intestinal epithelia from APC^{MIN} mice as well as 3 intestinal epithelia from wildtype mice, as controls (Figure 22A). For genome-wide 500 base windows, MeDIP-seq analysis was performed with the MEDIPS package. DMRs were detected by Wilcoxons Rank Sum Test, p-value < 0.01, and additional thresholds on the ratio (> 1.33) and minimum average abundance of normalized read counts (rpm > 0.25). To assess the regulatory influence of altered DNA methylation on gene expression, transcriptome analysis by RNA-seq was performed on the same samples. Differentially expressed genes were detected using the edgeR package, with FDR $\leq 0.1\%$. To infer the underlying functional mechanisms of promoter hypermethylation, gene set enrichment analysis (GSEA) was performed on targets of different epigenetic modifiers .

7.1.3 *Results*

We identified 5,135 hypermethylated regions and 8,845 hypomethylated regions in APC Min adenoma compared to the normal tissues. DMRs were found highly enriched in CpG islands, promoter and exonic regions. Many of the genes with differentially methylated promoters are known to be regulated by DNA methylation, such as Ush1g (Figure 22B). Accordingly, cluster analysis of methylation patterns in these regions clearly separated normal intestine from adenoma samples. 21 randomly selected DMRs were all validated by bisulfite pyrosequencing measurements in samples from the same animals. Additionally, 18 out of 21 DMRs were validated in samples from different animals, while three regions showed variability between the mice. In order to exclude the possibility that adenoma specific methylation patterns originate from progenitor cells, we compared methylation status of 11 DMRs at intestinal stem cells, as well as purified cells of the villus and crypt (Figure 22D). These specific cell types resemble the methylation status of the bulk normal cells, suggesting that DMRs form *de novo* in adenoma.

Using GSEA, we examined the enrichment of tumor suppressor genes (TSG), as well as genes known to be targeted by different epigenetic modifiers within the promoter DMRs. In contrast to TSG, targets of trithorax group complex and TET1, where no enrichment was found, PCR1/2 targets were enriched among genes with a hypermethylated promoter in adenoma (Figure 22C). Accordingly, we found several PRC2 components overexpressed in adenoma, and increased H3K27me3 mark at five out of six hypermethylated promoters, indicating local PRC2 activity.

We looked for a genome-wide anti-correlation of promoter methylation and gene expression, pursuant to the common model, according to which hypermethylation at promoter regions leads to silencing of the gene. Although co-occurrence of hypermethylation and down expression was slightly enriched, the majority of genes did not show a clear correlation of methylation and expression (Figure 22E). Although we observed general up-regulation of Wnt target genes and down-regulation of differentiation signatures in adenoma, there was no consistent trend of differential promoter or gene body methylation for these groups of genes.



DMR	B6 N				APC N				APC Ad				ISC		Crypt		Villus	
Cd133	53	nd	41	43	56	nd	42	48	37	nd	37	25	23	9	18	21	33	26
Dusp6	64	59	67	70	59	62	62	62	19	20	32	34	81	84	77	73	54	69
Ela1_2	87	84	88	88	97	88	89	85	29	23	46	29	97	95	95	95	92	92
Fos	90	89	91	92	95	89	91	89	54	31	64	52	94	96	94	92	91	95
Lemd3	79	75	64	70	63	68	54	67	17	17	30	24	80	80	78	67	67	74
Pla2g10	56	42	61	55	42	50	32	48	19	18	12	19	87	75	87	86	74	79
Pla2g2a	75	72	81	68	79	67	61	36	29	26	25	19	73	77	73	82	78	83
Slc25a28	84	92	85	87	84	85	77	82	28	35	54	32	90	88	89	88	91	90
Ush1g_1	nd	14	14	13	14	14	12	12	33	26	21	25	13	18	15	15	18	13
Ush1g_2	nd	nd	13	10	11	nd	8	10	30	nd	24	20	17	17	18	13	19	15
Vdr	nd	23	26	22	39	38	35	33	73	68	59	71	5	5	7	8	9	10

Figure 22: DNA-methylome analysis of mouse intestinal adenoma. (A) Summary of tissue samples used for genome-wide analyses. (B) Visualization of the adenoma-hypermethylated region in Ush1g. Black bars indicate regions confirmed by bisulfite-based validation. (C) Hypermethylated promoters enrich PRC2 targets. Genes are sorted by normal vs adenoma promoter methylation, green line depicts enrichment score, black bars indicate target genes, p-value and FDR are based on GSEA. (D) Color-coded table of CpG methylation analyses of 11 DMRs in primary tissues and purified cell types, using bisulfite pyrosequencing. Dark blue, $< 20\%$ methylation; light blue, $20-50\%$ methylation, light red, $50-80\%$ methylation; bright red, $> 80\%$ methylation. (E) Differential gene expression and promoter methylation for 31 selected epigenetically regulated tumor suppressor genes. Only two of the examples, *Crabp1* and *Runx3* are hypermethylated and down-regulated. (F) Promoter methylation of human colon cancer samples for genes orthologous to hyper- (top) and hypomethylated (bottom) promoters in APC^{MIN}.

Modified from [Grimm et al., 2013].

We then asked to what extent the detected methylation signature of mouse intestinal adenoma also applies to advanced human colon cancer. To this end, we compared the methylation patterns detected with promoter methylation at orthologous genes in 14 human colon cancer patients. Strikingly, we found a core set of adenoma specific methylation patterns conserved in human colon cancer, including many genes previously suggested as cancer biomarkers (Figure 22F). This indicates that the core epigenetic signature is established early during tumor formation and retained during progression to carcinoma.

7.1.4 *Conclusions*

Our results demonstrate preferential DNA hypermethylation at PRC2 target sites in adenoma, supporting the model that DNA methylation is guided by PRC2 activity via H3K27me3. Comparison of adenoma specific methylation profiles with intestinal stem cells and purified crypt and villus cells suggest that these DMRs form *de novo* during adenoma development. However, in contrast to advanced human colon cancer we did not find enrichment of TSG within hypermethylated promoters. This finding suggests that promoter hypermethylation in tumors is a stochastic process, and that TSG are selected during tumor evolution. For adenoma, the general regulatory paradigm, which states that promoter methylation leads to gene silencing, holds true for a minor fraction of genes only. Further studies are required for a deeper understanding of the functional effects of altered methylation patterns. The identified core set of DMRs, conserved between early mouse adenoma and advanced human colon carcinoma, are promising candidates for diagnostic biomarkers for early stage intestinal cancer.

7.1.5 *Relevance for the Thesis*

As one of the first projects investigating replicates of genome-wide epigenetic profiles for samples in different conditions, the design of this study allowed us to consider the biological and technical variability in the detection of DMRs. To this aim, we developed an *ad hoc* approach based on the non-parametric Wilcoxon rank sum test in combination with additional criteria; with this approach positive results based on weak evidence or small effect size could be avoided." The limitations of this approach, which are discussed in Section 5.1, later motivated the adoption of the GLM-LM based statistical test for application on MeDIP-seq data, which was implemented in an update of the MEDIPS package [Lienhard et al., 2014]. Furthermore, the study demonstrates the inference of functional mechanisms from genome-wide epigenomic profiling by enrichment analysis.

7.2 QSEA – MODELING GENOME-WIDE DNA METHYLATION ENRICHMENT

For clinical studies, enrichment based experiments provide genome-wide DNA methylation measurements at reasonable costs. This application depends on statistical methods for detecting differentially methylated regions (DMRs) between groups of samples. Furthermore, in order to provide full interpretability, relative DNA methylation enrichment measurements require transformation to absolute methylation levels. In this pilot study, we demonstrate the applicability of our novel analysis package, QSEA, for clinical studies. To this end, we profile DNA methylation of tumor and normal tissue from non-small cell lung cancer patients, and apply QSEA for quantification of methylation levels and detection and functional interpretation of DMRs. Methylation differences are strongly correlated with BS-seq measurements, and detected DMRs can be confirmed by the literature as well as experimental validation.

7.2.1 *Introduction*

Beside detection of DMRs, common questions in the analysis of DNA methylation profiles presuppose absolute methylation levels, in particular in clinical studies. For example, absolute levels are required to assess whether a specific region is methylated or not, which is crucial for inferring the biological effect of DNA methylation. Comparing methylation at different genomic loci within or across genomes and charting whole genome methylation landscapes also depends on absolute measurements. Furthermore, transformation of methylation enrichment values is required for experimental validation, since bisulfite based experiments report absolute methylation levels. Previous methods for this transformation have been developed for *in vitro* samples only, and their ability to reliably quantify differences between pairs of *in vivo* samples, and thus their applicability to clinical studies, has not yet been demonstrated. In this project, we demonstrated the application of our novel workflow, QSEA, to clinical studies by detecting and quantifying aberrant methylation on pairs of tumor and adjacent normal tissue from five non-small cell lung cancer (NSCLC) patients. We assessed the functional background of aberrant methylation in tumors and monitored its effect on gene expression regulation.

7.2.2 *Methods*

DNA methylation of tumor samples and adjacent normal tissue of 5 NSCLC patients was profiled by MeDIP-seq, as well as Methyl-seq targeted bisulfite-sequencing. MeDIP-seq was analyzed with QSEA in genomic 250 base windows, using TCGA LUAD and LUSC cohorts to calibrate enrichment characteristics. Methyl-seq was analyzed with Bismark, and methylation levels in 250 base windows were averaged and compared to MeDIP results. Gene expression was assessed by RNA-seq for the same samples and analyzed using DESeq2. To confirm the effects of detected promoter hypermethylation on gene expression, three lung cancer cell lines (H1299, H1650, HCC827) were demethylated by applying four concentrations of decitabine, an inhibitor of DNA methyltransferase, as well as DMSO as control. Subsequently, gene expression was measured by qRT-PCR.

7.2.3 *Results*

We identified 11,098 regions with gain and 94,328 regions with loss of methylation in tumors compared to normal tissue. A minor fraction of those, 1,306 and 250 for gain and loss of methylation respectively, were located in CpG island promoters. Among the genes with methylation gains, we found 107 known tumor suppressor genes, many of which have already been described in the context of lung cancer. Furthermore, we observed a very strong correlation (0.87) between QSEA tumor-normal methylation differences and Methyl-seq differences (Figure 23A), confirming the reliability of the QSEA methylation quantification and validating the identified DMRs.

We selected 757 genes with altered DNA methylation at CpG island promoters that are expressed in at least two samples. According to the expected regulatory effect of DNA methylation at CGI promoters, 330 of these genes showed anti-correlated expression and promoter methylation ($\rho < -0.5$, Figure 23B). To confirm the causative silencing effect of promoter hypermethylation *in vitro*, we cleared DNA methylation in NSCLC cell lines with treatment of decitabine at different concentrations, and monitored changes in gene expression for seven hypermethylated genes. All seven genes showed increased expression after reversal of promoter regulation, confirming the regulation of these genes by promoter methylation (Figure 23C).

A major fraction of identified DMRs lies outside promoter regions, emphasizing the benefit of whole genome methylation profiling methods over targeted approaches. To

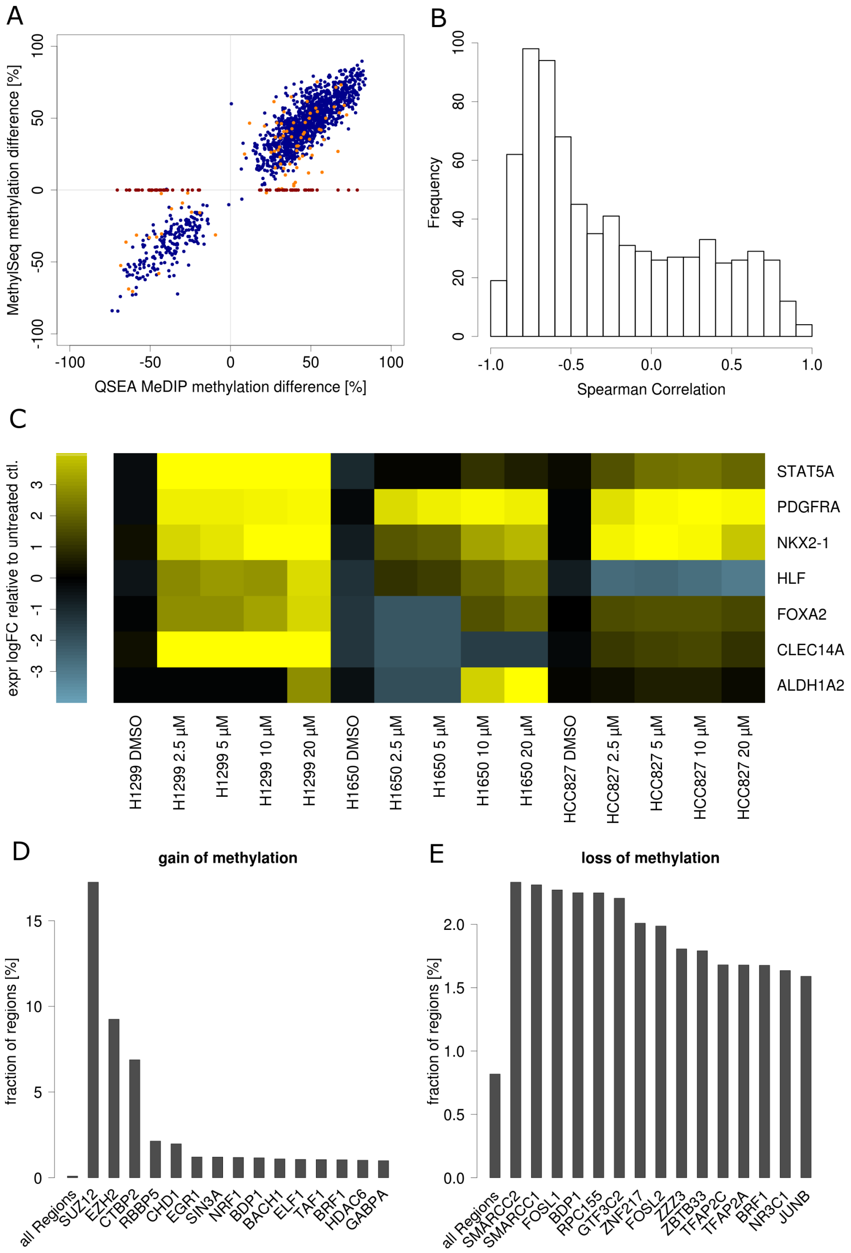


Figure 23: Genome-wide DNA methylation alterations in NSCLC. (A) Scatterplot of methylation differences quantified by QSEA and Methyl-seq. Blue: directly covered by Methyl-seq; orange: neighborhood covered by Methyl-seq; red: not covered by Methyl-seq. (B) Histogram of correlation between CpG island promoter methylation and gene expression. (C) Gene expression in four NSCLC cell lines after de-methylation validation experiment. (D) Sixteen most enriched transcription factor binding sites for hypermethylated regions and (E) hypomethylated regions respectively. Modified from [Lienhard et al., 2016].

exploit the full potential of genome-wide measurements, we used enrichment analysis to infer functional mechanisms other than CGI promoter hypermethylation. Based on the experimentally defined TFBS from ENCODE, we identified factors of polycomb repressive complex 2 (PRC2) to be more than 100 fold enriched for hypermethylation in NSCLC (Figure 23D). This finding is in line with the functional impact of PRC2 in tumor development, also observed in [Grimm et al., 2013]. Globally, hypomethylation is predominant in tumors, but less enriched in annotated regions. Among the top hypermethylation enriched binding sites are histone modifiers SMARCC1, SMARCC2 and SMARCB1, which show 2- to 3-fold enrichment. This enrichment suggests DNA methylation plays a role in chromatin remodeling by the SWI/SNF complex, a mechanism known to be involved in carcinogenesis (Figure 23E).

7.2.4 *Conclusions*

We performed a comprehensive methylome analysis of cancer samples from MeDIP-seq experiments. Detected differentially methylated regions were confirmed with bisulfite-sequencing. These regions disturb gene expression *in vitro* and thus have the potential to directly influence the cancer phenotype. By analyzing the enrichment of TFBS in tumor specific DMRs, we identified specific functional epigenetic mechanisms involved in lung cancer. These results demonstrate the potential of MeDIP-seq, analyzed with the QSEA methodology, for methylation analysis in a clinical context.

7.2.5 *Relevance for the Thesis*

The samples of this pilot study constitute the benchmark dataset, which was used in Section 4.2 to analyze the relationship of enrichment and methylation level and to construct the model on which the transformation is based. Furthermore, this dataset was used in Section 4.10 and Section 5.3 to assess the different methods' performance.

7.3 PREDICTING THERAPY RESISTANCE IN NSCLC BY EPIGENOMIC PROFILING

Understanding the epigenetic characteristics of drug resistance in cancer treatment is key to the selection of appropriate and optimal therapies for individual patients. In

[Grasse et al., in preparation], we use xenograft models of non-small cell lung cancer (NSCLC) to detect genomic regions, in which DNA methylation levels are correlated to carboplatin response.

7.3.1 *Introduction*

Non-small cell lung cancer (NSCLC) accounts for the largest fraction of cancer-related deaths worldwide. A major problem in lung cancer treatment is the development of resistance to standard chemotherapy, occurring in about 50% of the cases. Thus, biomarkers indicating the response to standard and alternative therapies have great potential to improve patient outcomes and reduce side effects. In contrast to primary tumors, cancer model systems allow several treatments to be tested in a controlled environment. However, predictive marks discovered in cell line derived lung cancer models perform poorly in primary tumors. Patient derived xenografts (PDX) provide a model system that resembles the features of primary tumors more closely. For this model, human tumor cells are transplanted into immune deficient mice that do not reject human cells. In this study, we analyzed genome-wide methylation profiles of 22 PDX models, to detect resistance mechanisms of 7 chemotherapies. Focusing on carboplatin, the standard therapy for NSCLC, we identified candidate regions where methylation level is correlated with therapy response and validated the predictive power with an independent patient cohort.

7.3.2 *Methods*

Using MeDIP-seq, we analyzed whole genome DNA methylation of 22 PDX and corresponding normal human lung tissue from the same patients, as well as primary tumor tissue from 6 of the patients. Additionally, we used Illumina gene expression microarrays for transcriptome profiling of PDX and normal lung samples. From each PDX model, six replicates were respectively treated with carboplatin, gemcitabine, paclitaxel, cetuximab, erlotinib, bevacizumab and etoposide, and drug response was assessed by average tumor volume relative to untreated control samples. Figure 24A provides an overview of the samples from this cohort. MeDIP-seq analysis was performed with the QSEA package, including CNV normalization, estimation of absolute methylation levels, and detection of DMRs between xenograft and normal as well as between drug responders and non-responders. CNVs were identified based on the MeDIP-seq reads, as described in Section 6.2.2. Parameters for absolute methylation level estimates were calibrated using publicly available methylation profiles of NSCLC samples from TCGA. Tumor specific DMRs, as well as treatment response dependent differentially methylated regions (trDMRs), were identified using the GLM-LR approach. For trDMRs, the linear predictor of the GLM was modeled as dependent on the log scaled relative tumor volume. Potential biomarker regions for carboplatin, where DNA methylation correlates with treatment response, were validated by methylation specific PCR in primary tumor tissues of an independent patient cohort.

7.3.3 *Results*

We observed that aberrant DNA methylation in PDX compared to normal tissue correlates with the alterations observed in primary NSCLC from TCGA (Figure 24B). Furthermore, we found large blocs corresponding to lamina associated domains hypermethylated in PDX, which is a characteristic of solid human tumors (Figure 24C). These findings confirm that the model preserves specific epigenetic tumor profiles during the encraftment process, and thus adequately models the tumor methylome.

To assess the impact of epigenetic mechanisms in treatment resistance, we focused on carboplatin, as platin based drugs are the standard therapy for NSCLC. In order to identify epigenetic biomarker candidates predicting treatment response, we correlated genomic methylation levels in PDX with relative tumor volume after treatment. This approach revealed 2,510 regions with characteristic methylation profiles for carboplatin treatment resistance (trDMRs). Carboplatin trDMRs are enriched for specific transcription factors, indicating distinct resistance mechanisms: while NFE2 binding sites are most preferentially hypermethylated in carboplatin resistant as opposed to carboplatin sensitive tumors, binding sites of PRC2 component SUZ12 are most enriched at hypomethylated regions (Figure 24D). A homolog of NFE2 as well as PRC2 has been previously linked to platin response, supporting the functional relevance of the identified mechanisms.

We further evaluated the functional impact of trDMRs by integrating DNA methylation and transcriptome information. For 547 genes, we observed an inverse correlation of gene expression and promoter methylation. Ingenuity pathway analysis revealed an enrichment of signaling pathways known to be involved in platin resistance, such as MYC-mediated apoptosis signaling, STAT3 and HER-2 signaling, Glutathione-mediated detoxification, ERK/MAPK signaling and BMP signaling. This enrichment confirms epigenetic regulation of key carboplatin resistance mechanisms, and thereby the relevance of identified trDMRs.

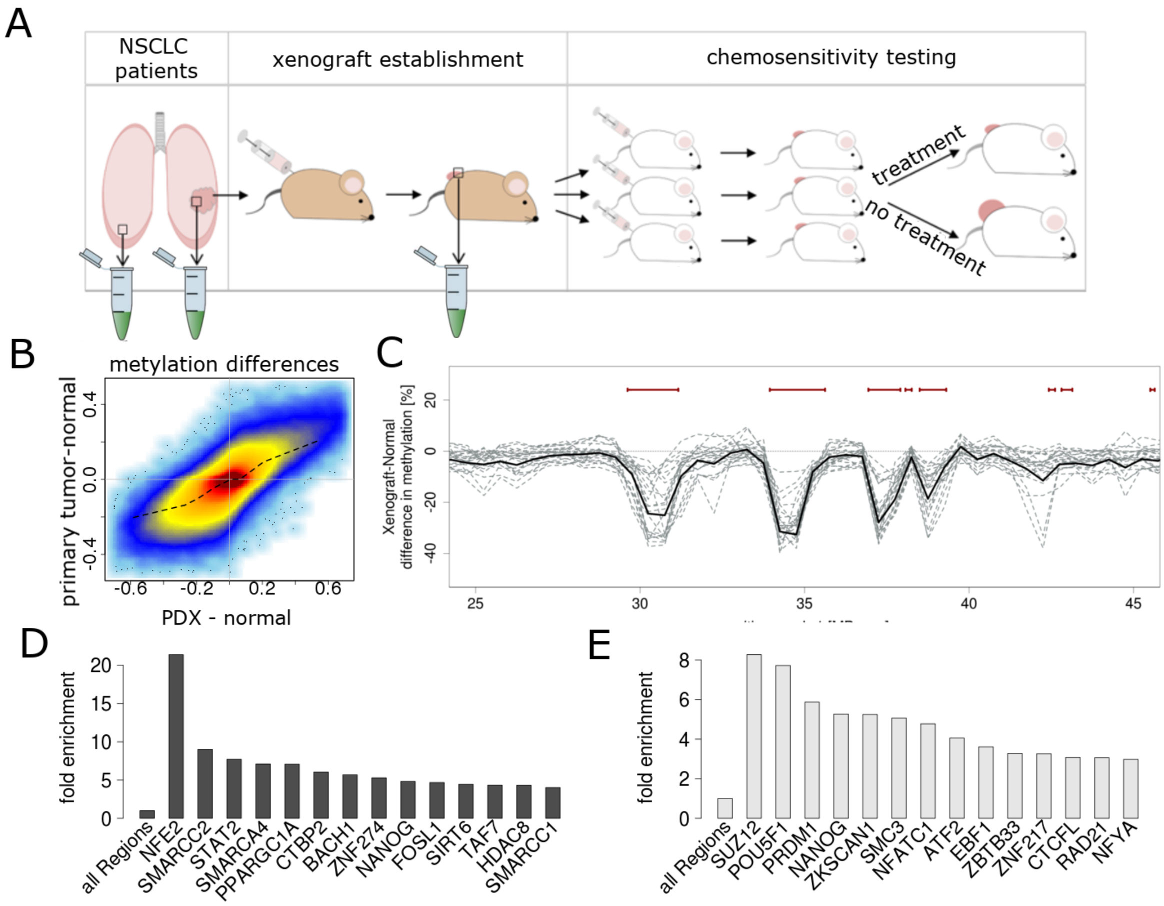


Figure 24: Predicting therapy resistance in NSCLC by epigenomic profiling. (A) Summary of samples used for genome-wide analyses and treatment response assessment. Tubes indicate sample extraction for methylation and gene expression analysis. (B) Methylation differences between PDX and normal generally conform with differences between primary tumor and normal. (C) Section of chr1, where PDX feature large hypomethylated blocks (LHB), corresponding to lamina associated domains (red bars). Dashed lines are smoothed tumor normal methylation differences for individual patients, the solid line is the mean difference over all patients. (D) TFBS enriching for caboplatin trDMRs, featuring gain and (E) loss of methylation in resistant tumors respectively. Modified from [Grasse et al., in preparation].

Based on methylation level differences between resistant and sensitive samples, as well as evidence on functionality taken from the literature, we selected 13 of the 2,510 trDMRs as candidate biomarkers. For these regions, we tested the biomarker potential by retrospectively predicting carboplatin response based on local methylation levels in an independent patient cohort and comparing the prediction to the observed treatment outcome. Among the tested regions, we found a candidate region to be highly predictive for carboplatin response. This candidate region is currently under patent consideration, and therefore undisclosed within this thesis.

7.3.4 *Conclusion*

We demonstrated that PDXs provide an appropriate model for studying epigenetic alterations in cancer. Focusing on carboplatin, we identified genomic regions where DNA methylation potentially influences treatment response mechanisms. In an independent patient cohort, we validated the predictive potential of the methylation levels at particular regions with respect to observed clinical outcomes. Analogously to carboplatin, trDMRs for the other screened drugs may be an indicator of expected patient response to alternative treatment.

7.3.5 *Relevance for the Thesis*

The experimental design of this project motivated the development of the transformation of relative methylation enrichment to absolute methylation levels, as presented in this thesis. The transformation also allows the difference in methylation level to be quantified, thereby allowing the biological effect of differential methylation to be assessed. Therefore, this quantification is crucial for the interpretation of treatment response specific DMRs with respect to potential biological effects. Furthermore, the application of the GLM approach with continuous explanatory variables (modeling the influence of treatment response on methylation) rather than binary dummy variables (e.g. encoding group assignment) demonstrates the flexibility of the approach. In general, the study provides an example of how the "quantitative sequencing enrichment analysis" approach can be applied in a clinical setting.

7.4 EPIGENOMIC ANALYSIS OF IMMUNE CELLS FROM ASTHMA PATIENTS

Although MEDIPS and QSEA were developed for the enrichment of DNA methylation, the methods implemented in the two packages also apply to other enrichment experiments, such as ChIP-seq. For example, in [Seumois et al., 2014] the MEDIPS was used to detect cell type specific H3K4me2 marks in naïve and differentiated T-cells from blood samples of healthy and asthmatic donors. We thereby identified a catalog of enhancer elements specifically regulating memory differentiation of CD4$^+$ T cell subtypes T_H1 and T_H2. Strikingly, we found single nucleotide variants associated with asthma susceptibility enriched within T_H2 specific enhancers, indicating an interplay between genetic and epigenetic mechanisms involved in the manifestation of the disease.

7.4.1 *Introduction*

Acquisition of immunological memory involves naive T and B cells specializing to become differentiated cells that recognize specific pathogens. This specialization is controlled by epigenetic mechanisms. Abnormal memory responses of this adaptive immune system can lead to autoimmune diseases, such as asthma. Asthma features airway inflammation, mediated by excessive immune response to inhaled pollen and other allergens. The prevalence of asthma is increasing globally, and most patients depend on long term medication, as today there is no curative therapy available. A molecular characteristic of asthma is the increased differentiation of T cell subtype T_H2. Understanding the epigenetic mechanisms underlying the differentiation of memory cells will help in tackling immune system mediated diseases. Epigenetic profiling in primary human T cells is constrained by the low number of cells that can be purified, for example from blood samples of donors. Therefore, we chose to profile cell type specific cis regulatory elements based on a single histone modification, H3K4me2, which features both active and poised enhancers.

7.4.2 *Methods*

Naïve T cells and differentiated CD4$^+$ subtypes T_H1 and T_H2 were isolated from blood samples of 12 asthmatic and 12 healthy donors, based on the expression of surface receptors CD4 and CCR4. Samples were profiled for H3K4me2 using a modified ChIP-seq protocol to account for small amounts of DNA from as few as 10^4 cells. In order to identify cell type and disease specific regulatory elements, we applied the GLM-LR approach implemented in MEDIPS to detect regions that were differentially enriched by the H2K4me2 ChIP, based on genome-wide 500 nucleotide windows. Additionally, transcription was profiled for all samples by RNA-seq, and analyzed for differential expression using DESeq. Interaction networks of potentially regulated genes were analyzed using the induced network approach of CPDB.

7.4.3 *Results*

In order to assess the functional impact of genomic regions identified by the microscaled H3K4me2 ChIP, we compared the enrichment profiles at subtype specific receptor genes. We found increased enrichment of known enhancer elements of the CCL5 and the CCR4 gene in T_H1 and T_H2 cells respectively, as expected from the known expression pattern of the surface receptor proteins (Figure 25A). Hence, despite the low amount of input DNA, H3K4me2 ChIP-seq effectively revealed cis regulatory DNA elements in the primary human T cells.

Based on this premise, we aimed at identifying further enhancer elements that are specific to CD4+ T cell differentiation. To this end, we compared genome-wide H3K4me2 enrichment profiles of T_H1, T_H2 and naïve T cells. We found 71,640 regions, accounting for about 1% of the genome, differentially enriched in at least one of the pairwise comparisons (Figure 25B). Genes in proximity to these putative memory specific enhancer and promoter elements are enriched for biological processes involved in adaptive immune responses, such as regulation of lymphocyte and leukocyte activation. Additionally, many of these genes show concordant changes in gene expression, further confirming the regulatory impact of the elements. Analysis of the interaction network of genes that gained promoter H3K4me2 in T_H2 cells revealed three genes, MYC, E2F2, and E2F4, as potential master regulators of T_H2 growth and survival (Figure 25C).

In order to identify potential co-factors of T-cell differentiation, we analyzed the enrichment of TFBS, co-occurring at cell type specific enhancers. In addition to the binding sites of lineage defining factors GATA-3 and T-BET, which were respectively enriched in T_H1 and T_H2 as expected, this analysis revealed distinct transcriptional co-factors. Most prominently, the antioxidant response elements binding factor NFE2 was enriched in T_H2 specific enhancer sites. NFE2 has previously also been suggested as a driver for T_H2 differentiation in mice. The potential role of NRF2 in human T_H2 driven disease is of special interest and requires further investigation, since the factor is stimulated by synthetic antioxidants which are commonly used as food preservatives.

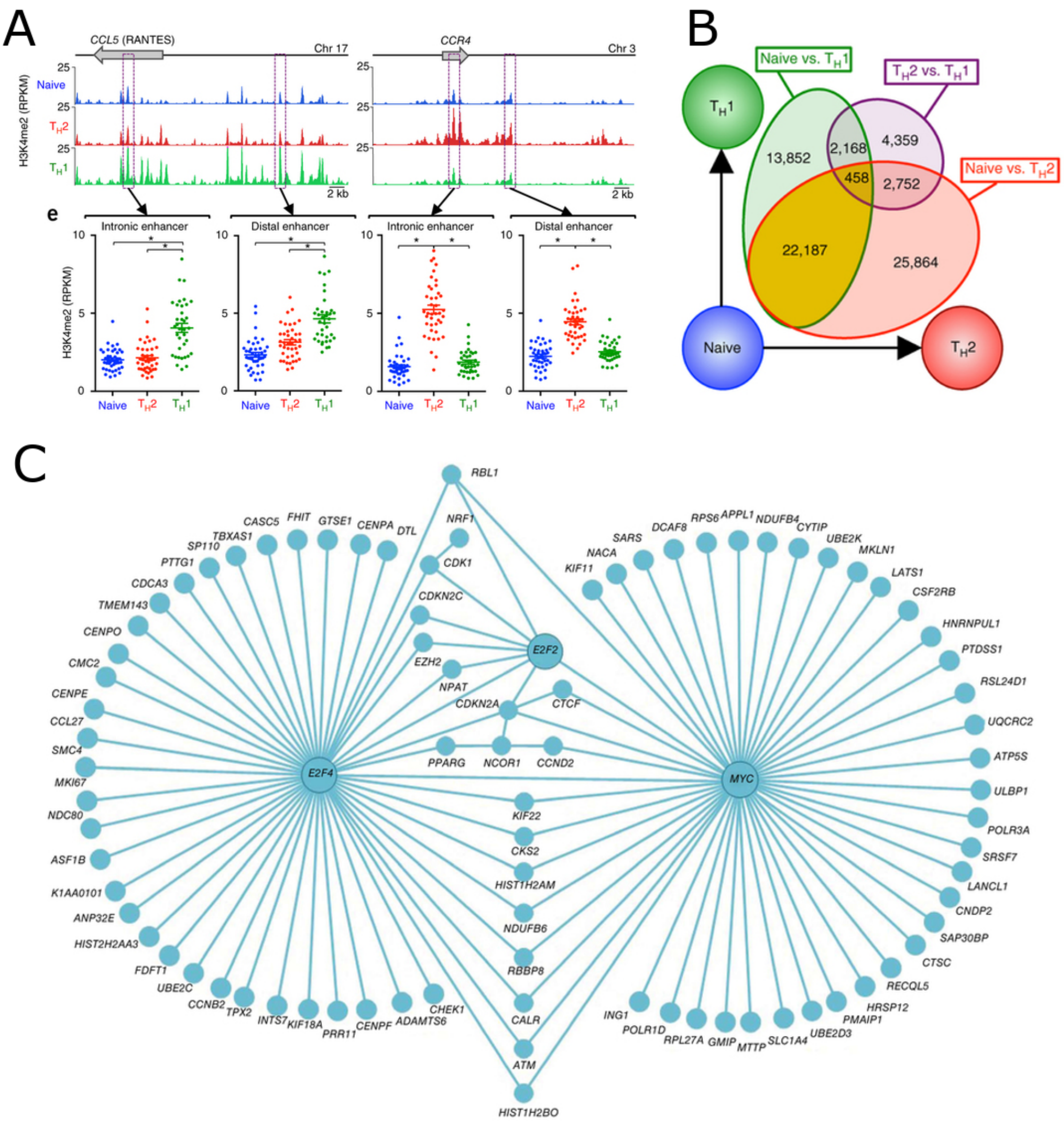


Figure 25: Epigenomic analysis of immune cells from asthma patients. (A) ChIP-seq reveals cell-type specific H3K4me2 enrichment patterns for enhancer elements of CLL5 and CCR4. Tracks in the upper panels are merged over 24 donors, while dots in lower panels represent individual samples (including assay replicates). (B) Overlap among the differential enriched regions identified for each pairwise comparison. (C) Induced gene-regulatory network analysis of genes associated with T_H2 specific enhancer elements reveals MYC, E2F2, E2F4 as key upstream regulators of this group of genes. Modified from [Seumois et al., 2014].

We further investigated the influence of genetic susceptibility of asthma on T_H2 specific enhancer sites. SNPs that are associated with asthma risk are enriched in T_H2 cell enhancers, which implies that the SNPs may modulate the activity of enhancers, and thereby shape the pathological gene expression patterns observed in disease.

Finally, the study revealed a set of 200 enhancer regions in T_H2 cells of asthma patients, featuring aberrant H3K4me2 patterns compared to the same cell type in healthy donors. These asthma associated enhancers also enrich TFBS involved in T cell differentiation, such as GATA3, TBX21 and RUNX3, highlighting their functional role. Many of the genes associated with these enhancers, including CCR5 binding chemokines CCL3L1, CCL3L3 and CCL4L2, are involved in chemokine and toll-like receptor signaling pathways, indicating a potential role in pathogenesis.

7.4.4 *Conclusion*

The study provides a genome-wide catalog of putative enhancer regions associated with specific T cell lineages in human *in vivo*. This catalog enabled us to identify E2F2, E2F4 and MYC as key master regulators of T_H2 memory differentiation. Consistent with the increased number of T_H2 memory cells in asthmatic patients, enhancer regions specific for this lineage enrich genetic variants associated with asthma risk, indicating a pathogenic role of deregulated T_H2 differentiation in asthma. By comparing enhancer elements active in the T_H2 cells of asthmatic patients to the same cell type in healthy controls, the study revealed a list of asthma associated enhancers. Further studies will clarify the role of these enhancers and the regulated genes in the genesis of asthma.

7.4.5 *Relevance for the Thesis*

The classic approach of ChIP-seq enrichment analysis is based on the detection of peaks to identify genomic regions featuring the mark of interest. Based on this approach, it has been suggested ChIP enrichment be compared between different conditions by analyzing the overlap of peaks detected within the groups individually. However, this procedure proved to be inappropriate for several reasons: first, peak detection inevitably relies on rigid thresholds, to binarize the ChIP-seq enrichment, which is an oversimplification in many cases. This is especially problematic for differential analysis, and may, for example, lead to a situation in which a peak is just above the detection threshold in one group of samples and just below in another. This region would appear as differentially enriched, although in fact the enrichment differs only slightly between groups. Second, neither for peak calling nor comparison is variation within groups typically considered, which may lead to false positive results in variable regions. Third, for peak calling, the shape of the peaks is typically not taken into account, and distinct peaks may get merged to a single peak. Therefore, two overlapping peaks in different conditions may in fact span distinct genomic regions. In contrast, the GLM-LR test explicitly models the read counts of individual samples within groups, and thereby the size of differences between groups and variability within them. By sectioning the genome in fixed regions, the approach ensures the correspondence of compared regions. For these reasons, the GLM-LR test, implemented in MEDIPS, has been applied to detect differential ChIP-seq enrichment between groups of samples, which demonstrates the versatility of the approach. A detailed protocol of this use case, including a discussion of alternative approaches for detecting differential ChIP-seq has been published in [Lienhard and Chavez, 2016]. Additionally, the study showcases the efficiency of the implementation, allowing an integrated analysis of more than 100 samples including replicates to be processed.

8

DISCUSSION

Epigenetic mechanisms cooperatively control the differentiation and maintenance of cellular identity. In mammals, these mechanisms include methylation of cytosines in CpG context, which contributes to gene regulation and genomic stability. During replication, the methylation pattern can be copied to the nascent DNA strand, and thus inherited by daughter cells. DNA methylation is essential for the establishment and maintenance of important regulatory processes like X-inactivation and genomic imprinting. It therefore plays important roles in development and disease, and a deep understanding of the underlying regulatory machinery is of great biological and clinical relevance.

Based on high throughput sequencing, two types of assays measuring DNA methylation have been established: bisulfite-sequencing and enrichment of methylated DNA fragments followed by sequencing. While whole genome bisulfite-sequencing directly provides genome-wide absolute methylation levels at single base resolution, it requires deep sequencing, which makes the approach costly. Therefore, for large sets of samples, enrichment based methods provide an attractive alternative to measure genome-wide DNA methylation. This approach requires substantially less sequencing depth and is thus more cost effective. However, sequencing read density provides a relative methylation signal that is also dependent on the number of CpGs per fragment and the enrichment characteristics. The comparison of different genomic regions within and across samples and the derivation of absolute methylation levels require specific normalization and further transformation.

Within this thesis, I developed the *quantitative sequencing enrichment analysis* (QSEA) work-flow for the normalization and subsequent transformation of methylation enriched sequencing read density to absolute methylation level. The normalization procedure accounts for the influence of alterations in DNA copy number, which makes the method applicable to cancer samples and cancer derived cell lines. The transformation to absolute methylation levels is based on a statistical model of the read counts; it incorporates both unspecific background reads and the relation of sample specific enrichment characteristics and CpG density.

I presented different calibration strategies for the parameters, reflecting different levels of prior knowledge of the samples and making the approach flexible and versatile. The model can be calibrated based on additional calibration experiments, sample specific prior knowledge or general assumptions. In practice, additional independent experiments are the exception, and mostly used for validation purposes. Therefore, the model is commonly calibrated using bisulfite based methylation profiles of comparable samples. For many cancer types, these profiles are available from the ICGC and TCGA consortia, and for an increasing number of cell lines, methylation profiles can be obtained from public repositories. However, for specific tissues or models, where this information is not available, the calibration depends on general assumptions regarding the average methylation level in relation to the CpG density. Since the methylation level may vary between species and different developmental stages, there is no general recommendation; rather the assumptions must be carefully matched to the samples under investigation.

For the analyzed samples, QSEA methylation levels retrieved by all three calibration strategies were highly correlated to bisulfite based methylation levels. I compared the results with two recently published alternative approaches, MeSiC [Xiao et al., 2015] and BayMeth [Riebler et al., 2014]. Like QSEA, BayMeth uses a Bayesian approach to estimate absolute methylation levels, but differs regarding the statistical model and parameter estimation methods. MeSiC estimate absolute methylation levels using random forest regression based on predefined annotated genomic features. The comparison showed that QSEA outperforms the alternative methods with respect to accuracy of methylation levels for individual samples. The improvement becomes even more evident when assessing the difference of methylation levels between samples. Tumor-normal methylation level differences estimated by QSEA are in accordance with bisulfite-sequencing results. In contrast, the estimates of methylation level yielded by the alternative methods deviate more from the bisulfite methylation differences in the same sample pair compared to the variability between the pairs. Since the deviance of the alternative methods exceeds the sample variability, these methods are not appropriate for quantifying methylation level differences between samples. Exact quantification of methylation level difference between different conditions, tissues or cell types is required to assess the biological effect of the differences. Therefore, QSEA transformation is appropriate for methylation analysis in the context of developmental or clinical studies.

Moreover, I adapted and compared two different statistical approaches for the detection of differentially methylated regions: the non-parametric Wilcoxon Rank Sum test (WRS) and the likelihood ratio test of nested generalized linear models (GLM-LR). While the WRS test statistic is computationally less complex, the comparison indicates greater power and a lower false positive rate for GLM-LR. Furthermore, GLM-LR is more flexible, and enables the modeling different experimental designs and correction for technical biases. For these reasons, the GLM-LR approach is superior to the WRS for the detection of genome-wide differentially methylated regions.

All methods described in this thesis are implemented in two R/bioconductor packages MEDIPS and the successor QSEA. The implementation of the methods is efficient and flexible, optionally allowing incorporation of prior knowledge and additional data. All functions comprise detailed documentation and executable examples. For both packages, application is demonstrated by tutorials containing practically relevant use cases. The main novelties of QSEA compared to the predecessor MEDIPS are the transformation to absolute methylation levels, and the CNV normalization. These features enhance the interpretability and extend the applicability of the methods to cancer samples. While both packages contain overlapping functionality, some features of MEDIPS, such as quality control metrics, are complementary to QSEA. Because of this and the popularity of MEDIPS, both packages are actively maintained and developed.

The popularity of the packages is reflected in the bioconductor download statistics as well as the citation count of the papers presenting the methods. Together, the packages are downloaded by more than 200 distinct IPs per month, and regular requests via the bioconductor user and support forum document active usage. This has made MEDIPS/QSEA the standard work-flow for analyzing enrichment based methylation experiments, with, in total, 181 citations[1] of the three papers [Chavez et al., 2010; Lienhard et al., 2014; Lienhard et al., 2016].

In order to consider copy number variation (CNV) of cancer samples, the normalization method depends on accurate estimations of local DNA copy number, as well as the position of the breakpoints. Commonly, estimation methods depend on distinct experiments like genotyping arrays or whole-genome-sequencing. I investigated the possibility of detecting CNVs directly from MeDIP-seq reads, ignoring all fragments that contain CpG dinucleotides. The abundance of CpG-less fragments is independent of the local methylation level, since cytosines outside the CpG context are unmethylated in mammals and thus not enriched by MeDIP. I found CNV estimated from MeDIP-seq correspond well to CNVs estimated from whole-genome-sequencing at similar cover-

[1] according to google scholar, 3/1/2017

age. This strategy means that no additional CNV experiment needs to be conducted, which further increases the efficiency of the MeDIP approach.

After normalization and transformation, methylation levels from enrichment based methylation assays are comparable to bisulfite results. However, while bisulfite-sequencing provides methylation levels for individual CpGs, the resolution of enrichment based approaches is restricted by the length of the sequencing fragments, which are typically around 250 bases. Since the methylation state of neighboring CpGs is highly correlated ($r = 0.94$ to 0.95) [Hodges et al., 2009], 250 base window averages are typically sufficient to describe local methylation levels. Therefore, merging windows with similar methylation level the genome to be segmentedin blocks with consistently methylated regions. However, if the boundary between high and low methylated regions falls close to the center of a window, this window will spuriously appear as intermediately methylated. Functional division of the genome, for example irregular sized windows that consider CpG island boundaries, or flexible segmentation based on HMM may reduce this effect.

Mapping DNA fragments to a genomic region by aligning sequencing reads to a reference genome is based on two assumptions: first, the reads of a fragment can be uniquely aligned to the corresponding region, and second, the sample genome is sufficiently similar to the reference, such that differences do not disturb the alignment. By masking repetitive regions of the genome, the analysis is restricted to regions where the first assumption holds true. However, this excludes investigation of DNA methylation for repetitive elements, such as micro-satellites, long terminal repeats or short and long interspersed nuclear elements. In any case, due to the variable copy number of these elements, which affect the measurement, enrichment based experiments are not particularly eligible to assess methylation levels within repeats. In contrast, bisulfite-sequencing is not disturbed by varying DNA copy number, and therefore allows a limited analysis of methylation of repetitive elements. Although the sequence similarity within one class of repetitive elements does not permit distinguishing individual elements, the overall methylation of repeat classes can be assessed by aligning bisulfite-seq reads with consensus sequences of the repeat classes, and summarizing all CpGs within one class [Kang et al., 2015]. Violations of the second assumption might be particularly relevant when comparing samples with genomes similar to the reference to samples with genomes dissimilar to the reference. For example, comparing two different mouse strains, one of which corresponds to the reference, the other of which does not, might yield false DMRs. Regions with sequence variants in the second strain may have reduced mappability and hence, incorrectly appear as unmethylated. This issue can be resolved using different reference sequences for the two samples, and mapping corresponding regions in the two references. However, corresponding regions of the different genomes may have different size and CpG density, further complicating the analysis.

The GLM-LR test to detect DMRs implemented in MEDIPS and QSEA can be applied to other kinds of enrichment based assays, such as histone ChIP-seq, to detect differences between groups of samples. This use case has been demonstrated to detect cell type specific enhancer elements, using ChIP-seq of H3K4me2 [Seumois et al., 2014]. In principle, the transformation to absolute levels could also be applied to histone ChIP-seq. Estimating the fraction of cells featuring the modification would help in assessing the biological effect and facilitate the comparison of different genomic regions within and across samples. For MeDIP, several methylated CpGs within one fragment can potentially be bound by the antibody, resulting in dependence of the enrichment on CpG density. In contrast, for histone ChIP-seq, the DNA fragment typically contains only one potential epitope for the antibody, since the fragment contains only one histone complex. Therefore, the model for the transformation would need modification to the specific enrichment characteristics of histone ChIP-seq. Furthermore, the development of calibration strategies to adjust these enrichment characteristics to the data would require further investigation.

The analysis of specific oxidative forms of cytosine methylation, in particular 5-hydroxy-methyl-cytosine, is of interest, since it has been proposed that they are independent epigenetic marks with distinct regulatory functionality. However, distinguishing oxidative forms of cytosine methylation is experimentally laborious. For example, bisulfite-sequencing cannot distinguish 5-methyl-cytosine (5mC) from 5-hydroxy-methyl-cytosine (5hmC) and is insensitive to other oxidative forms of methylated cytosines. In contrast, enrichment based methylation assays are specific to 5mC, and recently, the development of antibodies specific to 5hmC has allowed the adoption of the MeDIP-seq protocol to detect 5hmC (hMeDIP) [Nestor and Meehan, 2014]. In principle, the QSEA analysis methods can be directly applied to sequencing reads from hMeDIP experiments. However, intrinsic enrichment characteristics of the 5hmC antibody and decisive differences in the genome-wide distribution of the mark may require modification of parameter definitions, calibration strategies and estimation methods for the model.

Taken together, the methods presented in this thesis allow comprehensive analysis of genome-wide methylation profiles based on enrichment of methylated DNA fragments. Resulting absolute methylation levels are comparable to bisulfite derived methylation levels, allowing the interpretation of the biological effect of methylation and the comparison of different regions across the genome. The GLM-LR approach enables detection of genome-wide differentially methylated parts of the genome that are characteristic for the condition under study. The methods are implemented generically and can also be applied to other kinds of experiments, such as to inferring the fraction of cells featuring local 5-hydroxy-methylation or histone modification. The broad range of possible applications of the methods, the user friendly implementation in the two bioconductor packages MEDIPS and QSEA, and the interest of the user community in these packages promise further utilization and sustained impact of the methods developed here.

BIBLIOGRAPHY

Alkan, C., J. M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J. O. Kitzman, C. Baker, M. Malig, O. Mutlu, et al. (2009). "Personalized copy number and segmental duplication maps using next-generation sequencing." In: *Nature genetics* 41.10, pp. 1061–1067.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." In: *Genome biology* 11.10, p. 1.

Beerli, R. R., D. J. Segal, B. Dreier, and C. F. Barbas (1998). "Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks." In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14628–14633.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." English. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, ISSN: 00359246.

Berger, S. L. (2007). "The complex language of chromatin regulation during transcription." In: *Nature* 447.7143, pp. 407–412.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, et al. (2006). "A bivalent chromatin structure marks key developmental genes in embryonic stem cells." In: *Cell* 125.2, pp. 315–326.

Bodi, Z., S. Zhong, S. Mehra, J. Song, H. Li, N. Graham, S. May, and R. G. Fray (2012). "Adenosine methylation in Arabidopsis mRNA is associated with the 3′ end and reduced levels cause developmental defects." In: *Frontiers in plant science* 3, p. 48.

Bolstad, B. M., R. A. Irizarry, M. Åstrand, and T. P. Speed (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." In: *Bioinformatics* 19.2, pp. 185–193.

Booth, M. J., T. W. Ost, D. Beraldi, N. M. Bell, M. R. Branco, W. Reik, and S. Balasubramanian (2013). "Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine." In: *Nature protocols* 8.10, pp. 1841–1851.

Bourgon, R., R. Gentleman, and W. Huber (2010). “Independent filtering increases detection power for high-throughput experiments.” In: *Proceedings of the National Academy of Sciences* 107.21, pp. 9546–9551.

Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit (2010). “Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.” In: *BMC bioinformatics* 11.1, p. 94.

Cantone, I. and A. G. Fisher (2013). “Epigenetic programming and reprogramming during development.” In: *Nature structural & molecular biology* 20.3, pp. 282–289.

Capuano, F., M. MuÌˆlleder, R. Kok, H. J. Blom, and M. Ralser (2014). “Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species.” In: *Analytical chemistry* 86.8, pp. 3697–3702.

Carter, N. P. (2007). “Methods and strategies for analyzing copy number variation using DNA microarrays.” In: *Nature genetics* 39, S16–S21.

Chavez, L., J. Jozefczuk, C. Grimm, J. Dietrich, B. Timmermann, H. Lehrach, R. Herwig, and J. Adjaye (2010). “Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage.” In: *Genome Res.* 20, pp. 1441–1450.

Chavez, L. (2011). “Multivariate statistical analysis of epigenetic regulation with application to the analysis of human embryonic stem cells.” Doctoral dissertation. Freie Universität Berlin.

Chen, B.-F. and W.-Y. Chan (2014). “The de novo DNA methyltransferase DNMT3A in development and cancer.” In: *Epigenetics* 9.5, pp. 669–677.

Deaton, A. M. and A. Bird (2011a). “CpG islands and the regulation of transcription.” In: *Genes & development* 25.10, pp. 1010–1022.

Deaton, A. M. and A. Bird (2011b). “CpG islands and the regulation of transcription.” In: *Genes & development* 25.10, pp. 1010–1022.

Deng, J., R. Shoemaker, B. Xie, A. Gore, E. M. LeProust, J. Antosiewicz-Bourget, D. Egli, N. Maherali, I.-H. Park, J. Yu, et al. (2009). “Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming.” In: *Nature biotechnology* 27.4, pp. 353–360.

Down, T. A., V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Graef, N. Johnson, J. Herrero, E. M. Tomazou, et al. (2008). “A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis.” In: *Nature biotechnology* 26.7, pp. 779–785.

ENCODE Consortium et al. (2012). “An integrated encyclopedia of DNA elements in the human genome.” In: *Nature* 489.7414, pp. 57–74.

Ernst, J. and M. Kellis (2010). “Discovery and characterization of chromatin states for systematic annotation of the human genome.” In: *Nature biotechnology* 28.8, pp. 817–825.

Falkenberg, K. J. and R. W. Johnstone (2014). “Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders.” In: *Nature reviews Drug discovery* 13.9, pp. 673–691.

Grasse, S., M. Lienhard, S. Frese, M. Kerick, C. Grimm, J. Rolff, M. Becker, F. Dreher, U. Schirmer, S. Börno, A. Ramisch, G. Leschber, B. Timmermann, M. Odenthal, C. Grohé, H. Lüders, R. Büttner, I. Fichtner, H. Sültmann, H. Lehrach, R. Herwig, and M. R. Schweiger (in preparation). “Genome-wide DNA methylation profiles for the prediction of therapy resistance in NSCLC.” In:

Greer, E. L., M. A. Blanco, L. Gu, E. Sendinc, J. Liu, D. Aristizábal-Corrales, C.-H. Hsu, L. Aravind, C. He, and Y. Shi (2015). “DNA methylation on N 6-adenine in C. elegans.” In: *Cell* 161.4, pp. 868–878.

Grimm, C., L. Chavez, M. Vilardell, A. L. Farrall, S. Tierling, J. W. Böhm, P. Grote, M. Lienhard, J. Dietrich, B. Timmermann, et al. (2013). “DNA–methylome analysis of mouse intestinal adenoma identifies a tumour-specific signature that is partly conserved in human colon cancer.” In: *PLoS Genet* 9.2, e1003250.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass (2010). “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.” In: *Molecular cell* 38.4, pp. 576–589.

Herman, J. G., A. Merlo, L. Mao, R. G. Lapidus, J.-P. J. Issa, N. E. Davidson, D. Sidransky, and S. B. Baylin (1995). “Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers.” In: *Cancer research* 55.20, pp. 4525–4530.

Herwig, R., C. Hardt, M. Lienhard, and A. Kamburov (2016). “Analyzing and interpreting genome data at the network level with ConsensusPathDB.” In: *Nature Protocols* 11.10, pp. 1889–1907.

Heyn, H. and M. Esteller (2012). “DNA methylation profiling in the clinic: applications and challenges.” In: *Nature Reviews Genetics* 13.10, pp. 679–692.

Hodges, E., A. D. Smith, J. Kendall, Z. Xuan, K. Ravi, M. Rooks, M. Q. Zhang, K. Ye, A. Bhattacharjee, L. Brizuela, et al. (2009). “High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing." In: *Genome research* 19.9, pp. 1593–1605.

Illingworth, R. S., U. Gruenewald-Schneider, S. Webb, A. R. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews, and A. P. Bird (2010). "Orphan CpG islands identify numerous conserved promoters in the mammalian genome." In: *PLoS Genet* 6.9, e1001134.

Ito, S., L. Shen, Q. Dai, S. C. Wu, L. B. Collins, J. A. Swenberg, C. He, and Y. Zhang (2011). "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine." In: *Science* 333.6047, pp. 1300–1303.

Jones, P. L., G. C. J. Veenstra, P. A. Wade, D. Vermaak, S. U. Kass, N. Landsberger, J. Strouboulis, and A. P. Wolffe (1998). "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription." In: *Nature genetics* 19.2, pp. 187–191.

Kang, J., M. Lienhard, W. A. Pastor, A. Chawla, M. Novotny, A. Tsagaratou, R. S. Lasken, E. C. Thompson, M. A. Surani, S. B. Koralov, S. Kalantry, L. Chavez, and A. Rao (2015). "Simultaneous deletion of the methylcytosine oxidases Tet1 and Tet3 increases transcriptome variability in early embryogenesis." In: *Proc. Natl. Acad. Sci. U.S.A.* 112.31, E4236–4245.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent (2004). "The UCSC Table Browser data retrieval tool." In: *Nucleic acids research* 32.suppl 1, pp. D493–D496.

Kornberg, R. D. (1977). "Structure of chromatin." In: *Annual review of biochemistry* 46.1, pp. 931–954.

Kornblith, A. B., J. E. Herndon, L. R. Silverman, E. P. Demakos, R. Odchimar-Reissig, J. F. Holland, B. L. Powell, C. DeCastro, J. Ellerton, R. A. Larson, et al. (2002). "Impact of azacytidine on the quality of life of patients with myelodysplastic syndrome treated in a randomized phase III trial: a Cancer and Leukemia Group B study." In: *Journal of Clinical Oncology* 20.10, pp. 2441–2452.

Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, et al. (2015). "Integrative analysis of 111 reference human epigenomes." In: *Nature* 518.7539, pp. 317–330.

Lai, D., G. Ha, and S. Shah (2016). *HMMcopy: Copy number prediction with correction for GC and mappability bias for HTS data*. R package version 1.16.0.

Lienhard, M. and L. Chavez (2016). "Quantitative Comparison of Large-Scale DNA Enrichment Sequencing Data." In: *Statistical Genomics: Methods and Protocols*, pp. 191–208.

Lienhard, M., C. Grimm, M. Morkel, R. Herwig, and L. Chavez (2014). "MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments." In: *Bioinformatics* 30.2, pp. 284–286.

Lienhard, M., S. Grasse, J. Rolff, S. Frese, U. Schirmer, M. Becker, S. Börno, B. Timmermann, L. Chavez, H. Sültmann, et al. (2016). "QSEA – modelling of genome-wide DNA methylation from sequencing enrichment experiments." In: *Nucleic Acids Research*, gkw1193.

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, et al. (2009). "Human DNA methylomes at base resolution show widespread epigenomic differences." In: *nature* 462.7271, pp. 315–322.

Liu, X. S., H. Wu, X. Ji, Y. Stelzer, X. Wu, S. Czauderna, J. Shu, D. Dadon, R. A. Young, and R. Jaenisch (2016). "Editing DNA methylation in the mammalian genome." In: *Cell* 167.1, pp. 233–247.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." In: *Genome biology* 15.12, p. 1.

Mammana, A. and H.-R. Chung (2015). "Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome." In: *Genome biology* 16.1, p. 1.

Margueron, R. and D. Reinberg (2010). "Chromatin structure and the inheritance of epigenetic information." In: *Nature Reviews Genetics* 11.4, pp. 285–296.

Meissner, A., A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch (2005). "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." In: *Nucleic acids research* 33.18, pp. 5868–5877.

Nestor, C. E. and R. R. Meehan (2014). "Hydroxymethylated DNA immunoprecipitation (hmeDIP)." In: *Functional Analysis of DNA and Chromatin*, pp. 259–267.

Pagès, H. (2016). *BSgenome: Infrastructure for Biostrings-based genome data packages and support for efficient SNP representation*. R package version 1.42.0.

Patra, S. K., M. Deb, and A. Patra (2010). "Molecular marks for epigenetic identification of developmental and cancer stem cells." In: *Clinical epigenetics* 2.1, p. 27.

Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." In: *Nucleic acids research* 35.suppl 1, pp. D61–D65.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." In: *Cell* 159.7, pp. 1665–1680.

Riebler, A., M. Menigatti, J. Z. Song, A. L. Statham, C. Stirzaker, N. Mahmud, C. A. Mein, S. J. Clark, and M. D. Robinson (2014). "BayMeth: improved DNA methylation quantification for affinity capture sequencing data using a flexible Bayesian approach." In: *Genome biology* 15.2, p. 1.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics* 26, pp. 139–140.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." In: *Genome Biology* 11, R25.

Rossetto, D., A. W. Truman, S. J. Kron, and J. Côté (2010). "Epigenetic modifications in double-strand break DNA damage signaling and repair." In: *Clinical Cancer Research* 16.18, pp. 4543–4552.

Ross-Innes, C. S., R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, et al. (2012). "Differential oestrogen receptor binding is associated with clinical outcome in breast cancer." In: *Nature* 481.7381, pp. 389–393.

Russo, V. E., R. A. Martienssen, A. D. Riggs, et al. (1996). *Epigenetic mechanisms of gene regulation.* Cold Spring Harbor Laboratory Press.

Sado, T., M. H. Fenner, S.-S. Tan, P. Tam, T. Shioda, and E. Li (2000). "X inactivation in the mouse embryo deficient for Dnmt1: distinct effect of hypomethylation on imprinted and random X inactivation." In: *Developmental biology* 225.2, pp. 294–303.

Saxonov, S., P. Berg, and D. L. Brutlag (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." In: *Proceedings of the National Academy of Sciences* 103.5, pp. 1412–1417.

Serre, D., B. H. Lee, and A. H. Ting (2010). "MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome." In: *Nucleic acids research* 38.2, pp. 391–399.

Seumois, G., L. Chavez, A. Gerasimova, M. Lienhard, N. Omran, L. Kalinke, M. Vedanayagam, A. P. V. Ganesan, A. Chawla, R. Djukanović, et al. (2014). "Epigenomic analysis of primary human T cells reveals enhancers associated with T$_H$2 memory cell differentiation and asthma susceptibility." In: *Nature immunology* 15.8, pp. 777–788.

Sharma, S., T. K. Kelly, and P. A. Jones (2010). "Epigenetics in cancer." In: *Carcinogenesis* 31.1, pp. 27–36.

Song, C.-X., K. E. Szulwach, Y. Fu, Q. Dai, C. Yi, X. Li, Y. Li, C.-H. Chen, W. Zhang, X. Jian, et al. (2011). "Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine." In: *Nature biotechnology* 29.1, pp. 68–72.

Splinter, E., H. Heath, J. Kooren, R.-J. Palstra, P. Klous, F. Grosveld, N. Galjart, and W. de Laat (2006). "CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus." In: *Genes & development* 20.17, pp. 2349–2354.

Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, et al. (2011). "DNA-binding factors shape the mouse methylome at distal regulatory regions." In: *Nature*.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550.

Sved, J. and A. Bird (1990). "The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model." In: *Proceedings of the National Academy of Sciences* 87.12, pp. 4692–4696.

Tahiliani, M., K. P. Koh, Y. Shen, W. A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L. M. Iyer, D. R. Liu, L. Aravind, et al. (2009). "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1." In: *Science* 324.5929, pp. 930–935.

Taiwo, O., G. A. Wilson, T. Morris, S. Seisenberger, W. Reik, D. Pearce, S. Beck, and L. M. Butcher (2012). "Methylome analysis using MeDIP-seq with low DNA concentrations." In: *Nature protocols* 7.4, pp. 617–636.

TCGA Consortium (2012). "Comprehensive genomic characterization of squamous cell lung cancers." In: *Nature* 489.7417, pp. 519–525.

TCGA Consortium (2014). "Comprehensive molecular profiling of lung adenocarcinoma." In: *Nature* 511.7511, pp. 543–550.

Tetzner, R., F. Model, G. Weiss, M. Schuster, J. Distler, K. V. Steiger, R. Grützmann, C. Pilarsky, J. K. Habermann, P. R. Fleshner, et al. (2009). "Circulating methylated SEPT9 DNA in plasma is a biomarker for colorectal cancer." In: *Clinical chemistry* 55.7, pp. 1337–1346.

Timp, W., H. C. Bravo, O. G. McDonald, M. Goggins, C. Umbricht, M. Zeiger, A. P. Feinberg, and R. A. Irizarry (2014). "Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors." In: *Genome medicine* 6.8, p. 1.

Tsankova, N., W. Renthal, A. Kumar, and E. J. Nestler (2007). "Epigenetic regulation in psychiatric disorders." In: *Nature Reviews Neuroscience* 8.5, pp. 355–367.

Turner, B. M. (2000). "Histone acetylation and an epigenetic code." In: *Bioessays* 22.9, pp. 836–845.

Venables, W. N. and B. D. Ripley (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Verkerk, A. J., M. Pieretti, J. S. Sutcliffe, Y.-H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, F. Zhang, et al. (1991). "Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome." In: *Cell* 65.5, pp. 905–914.

Vilain, A., N. Vogt, B. Dutrillaux, and B. Malfoy (1999). "DNA methylation and chromosome instability in breast cancer cell lines." In: *FEBS letters* 460.2, pp. 231–234.

Waddington, C. (1957). *The strategy of the genes*.

Wang, H., M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, et al. (2012). "Widespread plasticity in CTCF occupancy linked to DNA methylation." In: *Genome research* 22.9, pp. 1680–1688.

Weber, M., J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schuebeler (2005a). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." In: *Nature genetics* 37.8, pp. 853–862.

Weber, M., J. J. Davies, D. Wittig, E. J. Oakeley, M. Haase, W. L. Lam, and D. Schuebeler (2005b). "Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells." In: *Nature genetics* 37.8, pp. 853–862.

Weisenberger, D., D. Van Den Berg, F. Pan, B. Berman, and P. Laird (2008). "Comprehensive DNA methylation analysis on the Illumina Infinium assay platform." In: *Illumina, San Diego*.

Xiao, Y., F. Yu, L. Pang, H. Zhao, L. Liu, G. Zhang, T. Liu, H. Zhang, H. Fan, Y. Zhang, et al. (2015). "MeSiC: A Model-Based Method for Estimating 5 mC Levels at Single-CpG Resolution from MeDIP-seq." In: *Scientific reports* 5.

Zhang, F., L. Cong, S. Lodato, S. Kosuri, G. M. Church, and P. Arlotta (2011). “Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription.” In: *Nature biotechnology* 29.2, pp. 149–153.

Zhang, T., S. Cooper, and N. Brockdorff (2015). “The interplay of histone modifications–writers that read.” In: *EMBO reports* 16.11, pp. 1467–1481.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, et al. (2008). “Model-based analysis of ChIP-Seq (MACS).” In: *Genome biology* 9.9, p. 1.

Ziller, M. J., K. D. Hansen, A. Meissner, and M. J. Aryee (2015). “Coverage recommendations for methylation analysis by whole-genome bisulfite sequencing.” In: *Nature methods* 12.3, pp. 230–232.

Ziller, M. J., H. Gu, F. Müller, J. Donaghey, L. T.-Y. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, et al. (2013). “Charting a dynamic DNA methylation landscape of the human genome.” In: *Nature* 500.7463, pp. 477–481.

Appendices

A

MATHEMATICAL PROOFS

A.1 GAMMA POISSON MIXTURE CORRESPONDS TO NEGATIVE BINOMIAL

Let Y be Poisson distributed with rate around a mean μ, scaled by a factor ϕ, which is drawn from a gamma distributed random variable Φ with equal shape and rate parameters $\alpha = \beta = \theta$:

$$Y \sim Pois(\lambda = \mu * \phi)$$

$$\Phi \sim \gamma(\alpha = \theta, \beta = \theta)$$

$$f_{Pois}(y; \lambda) = \frac{\lambda^y exp(-\lambda)}{y!} \tag{25}$$

$$f_\gamma(\phi; \alpha, \beta) = \frac{\beta^\alpha \phi^{\alpha-1} exp(-\phi\beta)}{\Gamma(\alpha)} \tag{26}$$

Note that the scaling does not affect the expected value of Y:

$$\begin{aligned} E[Y] &= E[\mu * \phi] \\ &= \mu * E[\Phi] \\ &= \mu * \frac{\theta}{\theta} \\ &= \mu \end{aligned} \tag{27}$$

The variance of the scaling factor is $var(\Phi) = \frac{\alpha}{\beta^2} = \frac{1}{\theta}$. Marginalizing the joint probability over the scaling factor θ yields f_Y, the probability density function of Y:

$$\begin{aligned}
f_Y(y;\mu,\theta) &= \int_0^\infty f_{Pois}(y;\phi\mu) * f_\gamma(\phi;\theta,\theta)\,\mathrm{d}\phi \\
&= \int_0^\infty \frac{(\mu\phi)^y exp(-\mu\phi)}{y!} * \frac{\theta^\theta \phi^{\theta-1} exp(-\phi\theta)}{\Gamma(\theta)}\,\mathrm{d}\phi \\
&= \frac{\theta^\theta \mu^y}{\Gamma(\theta) y!} \int_0^\infty \phi^y exp(-\mu\phi) * \phi^{\theta-1} exp(-\phi\theta)\,\mathrm{d}\phi \\
&= \frac{\theta^\theta \mu^y}{\Gamma(\theta) y!} \int_0^\infty \phi^{y+\theta-1} exp(-\phi(\mu+\theta))\,\mathrm{d}\theta \\
&= \frac{\theta^\theta \mu^y}{\Gamma(\theta) y!} \int_0^\infty \left(\frac{(\mu+\theta)\phi}{\mu+\theta}\right)^{y+\theta-1} exp(-\phi(\mu+\theta))\,\mathrm{d}\phi \\
&= \frac{\theta^\theta \mu^y}{\Gamma(\theta) y!} \frac{1}{(\mu+\theta)^{y+\theta-1}} \int_0^\infty (\phi(\mu+\theta))^{y+\theta-1} exp(-\phi(\mu+\theta))\,\mathrm{d}\phi
\end{aligned} \tag{28}$$

Now we substitute ϕ with $g(x) = \frac{x}{\mu+\theta}$. Applying the substitution rule

$$\int_a^b f(g(x)) g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} f(\phi)\,\mathrm{d}\phi \tag{29}$$

from right to left yields the gamma function for the integral of $f(g(x))$:

$$\begin{aligned}
f_Y(y;\mu,\theta) &= \frac{\theta^\theta \mu^y}{\Gamma(\theta) y!} \frac{1}{(\mu+\theta)^{y+\theta-1}} \int_0^\infty x^{y+\theta-1} exp(-x) * \frac{1}{\mu+\theta}\,\mathrm{d}x \qquad (30) \\
&= \frac{\Gamma(y+\theta)}{\Gamma(\theta) y!} \left(\frac{\mu}{\mu+\theta}\right)^y \left(\frac{\theta}{\mu+\theta}\right)^\theta
\end{aligned}$$

Thus, f_Y is equivalent to the probability density function of the Negative Binomial distribution f_{NB} with parameters $r = \theta$ and $p = \frac{\mu}{\mu+\theta}$:

$$f_{NB}(y; r, p) = \frac{\Gamma(y+r)}{\Gamma(r)y!} p^y (1-p)^r \tag{31}$$

$$\implies Y \sim NB\left(r = \theta, p = \frac{\mu}{\mu+\theta}\right) \quad \square \tag{32}$$

A.2 DERIVATION OF METHYLATION LEVEL POSTERIOR CDF

The cumulative distribution function of the methylation level posterior can be derived as follows.

$$\begin{aligned} F_{ML}(q; y, c, o) &= Pr(\beta < q | y, c, o) \\ &= \int_0^q f_{ML}(\beta; y, c, o)\,\mathrm{d}\beta \\ &= \int_0^q \frac{(o+\beta*c)^y * c * exp(-o-\beta*c)}{\gamma(y+1, o+c) - \gamma(y+1, o)}\,\mathrm{d}\beta \\ &= \frac{\int_0^q (o+\beta*c)^y * exp(-o-\beta*c) * c\,\mathrm{d}\beta}{\gamma(y+1, o+c) - \gamma(y+1, o)} \end{aligned} \tag{33}$$

Substituting $g(\beta) = o + \beta * c$ with t by applying the substitution rule (29) from left to right yields the CDF.

$$\begin{aligned} F_{ML}(q; y, c, o) &= \frac{\int_o^{o+q*c} t^y * exp(-t)\,\mathrm{d}t}{\gamma(y+1, o+c) - \gamma(y+1, o)} \\ &= \frac{\gamma(y+1, o+qc) - \gamma(y+1, o)}{\gamma(y+1, o+c) - \gamma(y+1, o)} \end{aligned} \tag{34}$$

A.3 EQUIVALENCE OF METHYLATION LEVEL POSTERIOR AND TRUNCATED ERLANG

For $x >= 0$, positive rate $\lambda > 0$ and the shape $k \in \mathbb{N}$ the Erlang distribution is described by the density function $f(x)$ and the cumulative distribution function $F(q)$:

$$f_E(x; \lambda, k) = \begin{cases} \lambda^k x^{k-1} e^{-\lambda x} \frac{1}{(k-1)!} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{35}$$

$$F_E(q; \lambda, k) = \begin{cases} \frac{\gamma(k, \lambda q)}{(k-1)!} & \text{if } q \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{36}$$

In order to truncate the probability density $f(x)$ to an interval $[x_1, x_2]$, the function is scaled such that the integral equals to one.

$$f_{tr}(x) = \begin{cases} \frac{f(x)}{F(x_2) - F(x_1)} & \text{if } x \in [x_1, x_2] \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

Thus, the Erlang distribution, with rate $\lambda = c$ and shape $k = y + 1$, truncated to the interval $[\frac{o}{c}, \frac{o}{c} + 1]$ has the following density function:

$$f_{trE}(x; \lambda = c, k = y + 1) = \begin{cases} \frac{c^{y+1} x^y e^{-xc}}{\gamma(y+1, (\frac{o}{c}+1) * x) - \gamma(y+1, \frac{o}{c} * x)} & \text{if } x \in [\frac{o}{c}, \frac{o}{c} + 1] \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

Substituting x with $\beta + \frac{o}{c}$ results in the posterior distribution function of the methylation level (13).

$$f_{trE}(\beta; y, c, o) = \begin{cases} \frac{(o + \beta * c)^y * c * exp(-o - \beta * c)}{\gamma(y+1, o+c) - \gamma(y+1, o)} & \text{if } \beta \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{39}$$

$$= f_{ML}(\beta; y, c, o) \quad \square \tag{40}$$

CURRICULUM VITAE

For reasons of data protection, the Curriculum vitae is not published in the online version

PUBLICATIONS

Grasse, S., **M. Lienhard**, S. Frese, M. Kerick, C. Grimm, J. Rolff, M. Becker, F. Dreher, U. Schirmer, S. Börno, A. Ramisch, G. Leschber, B. Timmermann, M. Odenthal, C. Grohé, H. Lüders, R. Büttner, I. Fichtner, H. Sültmann, H. Lehrach, R. Herwig, and M. R. Schweiger (in preparation). “Genome-wide DNA methylation profiles for the prediction of therapy resistance in NSCLC.” In:

Herwig, R., C. Hardt, **M. Lienhard**, and A. Kamburov (2016). “Analyzing and interpreting genome data at the network level with ConsensusPathDB.” In: *Nature Protocols* 11.10, pp. 1889–1907.

Lienhard, M. and L. Chavez (2016). “Quantitative Comparison of Large-Scale DNA Enrichment Sequencing Data.” In: *Statistical Genomics: Methods and Protocols*, pp. 191–208.

Lienhard, M., S. Grasse, J. Rolff, S. Frese, U. Schirmer, M. Becker, S. Börno, B. Timmermann, L. Chavez, H. Sültmann, et al. (2016). “QSEA – modelling of genome-wide DNA methylation from sequencing enrichment experiments.” In: *Nucleic Acids Research*, gkw1193.

Etchegaray, J. P., L. Chavez, Y. Huang, K. N. Ross, J. Choi, B. Martinez-Pastor, R. M. Walsh, C. A. Sommer, **M. Lienhard**, A. Gladden, S. Kugel, D. M. Silberman, S. Ramaswamy, G. Mostoslavsky, K. Hochedlinger, A. Goren, A. Rao, and R. Mostoslavsky (2015). “The histone deacetylase SIRT6 controls embryonic stem cell fate via TET-mediated production of 5-hydroxymethylcytosine.” In: *Nat. Cell Biol.* 17.5, pp. 545–557.

Kang, J., **M. Lienhard**, W. A. Pastor, A. Chawla, M. Novotny, A. Tsagaratou, R. S. Lasken, E. C. Thompson, M. A. Surani, S. B. Koralov, S. Kalantry, L. Chavez, and A. Rao (2015). “Simultaneous deletion of the methylcytosine oxidases Tet1 and Tet3 increases transcriptome variability in early embryogenesis.” In: *Proc. Natl. Acad. Sci. U.S.A.* 112.31, E4236–4245.

Lienhard, M., C. Grimm, M. Morkel, R. Herwig, and L. Chavez (2014). “MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments.” In: *Bioinformatics* 30.2, pp. 284–286.

Rasche, A., **M. Lienhard**, M. L. Yaspo, H. Lehrach, and R. Herwig (2014). "ARH-seq: identification of differential splicing in RNA-seq data." In: *Nucleic Acids Res.* 42.14, e110.

Seumois, G., L. Chavez, A. Gerasimova, **M. Lienhard**, N. Omran, L. Kalinke, M. Vedanayagam, A. P. V. Ganesan, A. Chawla, R. Djukanović, et al. (2014). "Epigenomic analysis of primary human T cells reveals enhancers associated with T_H2 memory cell differentiation and asthma susceptibility." In: *Nature immunology* 15.8, pp. 777–788.

Grimm, C., L. Chavez, M. Vilardell, A. L. Farrall, S. Tierling, J. W. Böhm, P. Grote, **M. Lienhard**, J. Dietrich, B. Timmermann, et al. (2013). "DNA–methylome analysis of mouse intestinal adenoma identifies a tumour-specific signature that is partly conserved in human colon cancer." In: *PLoS Genet* 9.2, e1003250.

Delft, J. van, S. Gaj, **M. Lienhard**, M. W. Albrecht, A. Kirpiy, K. Brauers, S. Claessen, D. Lizarraga, H. Lehrach, R. Herwig, and J. Kleinjans (2012). "RNA-Seq provides new insights in the transcriptome responses induced by the carcinogen benzo[a]pyrene." In: *Toxicol. Sci.* 130.2, pp. 427–439.

Rybak, J., A. Kuss, H. Lamecker, S. Zachow, H. C. Hege, **M. Lienhard**, J. Singer, K. Neubert, and R. Menzel (2010). "The Digital Bee Brain: Integrating and Managing Neurons in a Common 3D Reference System." In: *Front Syst Neurosci* 4.

SUMMARY

Enrichment of methylated DNA followed by sequencing offers a reasonable compromise between experimental cost and genomic coverage, allowing genome-wide DNA methylation to be assessed for large numbers of samples, which is a common requirement for clinical studies. However, the computational analysis of these experiments is complex, and depends on specific normalization and statistical approaches. Furthermore, quantification of the enrichment signals in terms of absolute levels of methylation requires specific transformation.

In this dissertation, I introduce specific computational methods for the individual steps of the analysis workflow. I assess the impact of sequencing library size, alterations in DNA copy number and CpG density on the local enrichment, and present a suitable normalization procedure. As the central part of the workflow, I developed a statistical model for the enrichment read counts, which is deployed in the Bayesian estimation of absolute levels of methylation. The model involves experimental parameters, such as sample specific enrichment characteristics. Accounting for different levels of prior knowledge, I suggest several calibration strategies for the model's parameters, which use either additional data or certain general assumptions. The transformation to absolute methylation levels greatly enhances interpretability and facilitates comparison with other methylation assays. By comparing the results with bisulfite sequencing validation data, I demonstrate the accuracy of the transformation, as well as the improvement over existing alternative methods. A common objective of methylome analysis is the detection of differentially methylated regions between groups of samples. I compare different statistical approaches for this task and discuss the inherent properties. I thereby identify likelihood ratio tests of nested generalized linear models to be well suited in terms of reliability and efficiency. The methods are implemented in two different R/bioconductor packages, MEDIPS and QSEA, which are easy to use and provide comprehensive functionality for the analysis of enrichment based experiments. All functions are documented and demonstrated by runnable examples, as well as detailed tutorials for specific practically relevant use cases. By presenting four representative studies published in peer-reviewed journals, I demonstrate the applicability and the versatility of the introduced methods. Taken together, this dissertation provides new computational methods for the analysis of enrichment based methylation experiments; these methods enhance the interpretability and reliability of the results from these experiments.

ZUSAMMENFASSUNG

Hochdurchsatzsequenzierung von angereicherter methylierter DNS erlaubt genomweite Methylierungsmessung zu relativ günstigen Kosten, wodurch die Analyse von zahlreichen Proben, zum Beispiel für klinische Studien, ermöglicht wird. Die computergestützte statistische Auswertung dieser Experimente ist jedoch komplex, und bedarf spezieller Normalisierungsmethoden und Schätzverfahren. In dieser Dissertation stelle ich spezifische computergestützte Methoden für die einzelnen Analyseschritte der Auswertung vor. Ich untersuche den Einfluss von Sequenziertiefe, Amplifikationen oder Deletationen der DNS, sowie der Häufigkeit von CpGs auf die Anreicherung der entsprechenden genomischen Region, und führe ein geeignetes Normalisierungsverfahren ein. Als zentralen Analyseschritt rekonstruiere ich das absolute Methylierungsniveau aus der relativen Anreicherung mittels Bayes'schen Schätzern. Hierfür habe ich ein statistisches Modell der angereicherten sequenzierten DNS-Fragmente entwickelt. Abhängig vom Vorwissen über die Proben schlage ich verschiedene Kalibrierungsstrategien für die probenspezifischen Anreicherungsparameter des Modells vor, basierend auf zusätzlichen Daten oder allgemeinen Annahmen. Die Umwandlung in absolute Methylierungswerte erhöht die Interpretierbarkeit erheblich und erleichtert den Vergleich mit anderen Methylierungsexperimenten. Durch Vergleich der Ergebnisse mit Bisulfit-Sequenzierung Validierungsdaten zeige ich die Schätzgenauigkeit des Verfahrens sowie die Verbesserung gegenüber bestehender alternativer Methoden. Ein häufiges Ziel der Methylomanalyse ist der Nachweis von differentiell methylierten Regionen zwischen Probengruppen. Ich vergleiche verschiedene statistische Ansätze für diesen Schritt und zeige diesbezüglich die Eignung von Likelihood-Quotienten-Tests geschachtelter generalisierter linearer Modelle hinsichtlich Zuverlässigkeit und Effizienz. Die vorgestellten Methoden sind in zwei R / Bioconductor-Paketen implementiert, MEDIPS und QSEA. Die Pakete sind einfach zu bedienen bieten umfassende Funktionalität. Alle Funktionen sind dokumentiert und werden mittels ausführbarer Beispiele, sowie ausführlichen Tutorials zu spezifischen praktisch relevanten Anwendungsfällen veranschaulicht. Vier vorgestellte repräsentative Studien, welche in wissenschaftlichen Fachzeitschriften veröffentlicht wurden, demonstrieren die praktische Anwendbarkeit und die Vielseitigkeit der eingeführten Methoden. Zusammengefasst bietet diese Dissertation neue computergestützte Methoden zur Analyse anreicherungsbasierter Methylierungsexperimente, welche sowohl die Interpretierbarkeit als auch die Zuverlässigkeit der Ergebnisse solcher Experimente erhöhen.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Date

Signature

Matthias Lienhard