The role of Dnmts and Tets in shaping the DNA methylation landscape of mouse embryonic stem cells

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

von

Peter Pascal Giehr

Universität des Saarlandes

Saarbrücken
September 2019

Datum des Kolloquiums: 13.09.2019

Dekan: Prof. Dr. Guido Kickelbick

Berichterstatter: Prof. Dr. Jörn E. Walter

Prof. Dr. Verena Wolf

Prof. Dr. Alexander Meissner

Vorsitzender: Prof. Dr. Volkhard Helms

Akad. Mitarbeiter: Dr. Björn Becker

ACKNOWLEDGEMENTS

This work would not have been possible without the constant support from my colleagues, friends and family, to whom I would like to express my sincere gratitude.

First, I would like to deeply thank Prof. Dr. Jörn Walter, who accepted me as diploma student, and later as PhD candidate and who introduced me to the fascinating world of epigenetics. I highly value his trust in my scientific work and I thank him for giving me the freedom to follow my own scientific ideas. He involved me in cooperations and projects beyond my PhD project and provided me the opportunity to gather important experiences in different areas of scientific work. Regarding both, success and failure, I could always rely on his support.

I am also very grateful to Prof. Dr. Verena Wolf for reviewing my thesis, and furthermore, I want to thank her for many years of interesting and fruitful collaboration. I equally thank Alexander Lück and Charalampos Kyriakopoulos for introducing me to the world of Hidden Markov Modelling. Both of them hold a large share in the here presented projects. Without their excellent work in translating complex biological processes into complex stochastic models, this work would not have been possible.

Many thanks to my co-authors and cooperation partners, who essentially contributed to these studies. I am thankful to Dr. Gabriella Ficz, Prof. Dr. Ferdinand von Meyenn, Dr. Stefan Wallner, Dr. Fabian Müller and Prof. D.r Wolf Reik for valuable scientific discussions, providing biological samples and for bioinformatic support.

I thank the entire 'Saarbrücker epigenetics family' for the great time and their constant support. Besides scientific projects, I fondly remember unique experiences like a journey through prehistorical times, explorations of caves and wine cellars, as well as adventures on rails or on the water. Many thanks to my personal fitness coach Karl for numerous runs in the woods and on the sports ground. They have always provided oxygen and physical condition for the weekly challenges. I want to thank Sascha, Kostya, Jie and Pavlo for exciting table football games, surpassing 'Bundesliga'! I owe particular thanks to Nina, Nicole, Sarah, Sascha, Gilles and Kostya for always providing me support regarding technical and biological issues. I am equally thankful to Abdul, Karl, Pavlo and Gilles for computational assistance. Abdul and Karl are especially acknowledged for conducting a series of bioinformatic analyses in the course of the here presented studies. I wish to thank my fellow PhD students Anna, Annemarie, Fon and Kathrin, who have always been supportive. I am very thankful to Nicole, Beate, Christina and Jasmin. It is indisputable that, without them, the lab would fall into chaos.

My sincere thanks goes to my family. My parents Monika and Michael, as well as my grandparents Edith and Rudolf, who always supported me and encouraged me to stay curious and creative. At all times and situations, I could count on my siblings Anna Lena and Andreas.

I would finally like to thank Judith, who has always been at my side for the past two years, both private and professional. Especially in difficult times, she encouraged me with understanding, the right words and a loving smile.

SUMMARY

DNA methylation in mammals substantially contributes to regulation of gene expression and thus defines cell fate and identity. Generation and stable inheritance of cell specific methylation patterns are ensured by the activity of DNA methyltransferases (Dnmts), while removal of DNA methylation is often linked to oxidised cytosine forms and the activity of ten-eleven translocation di-oxygenases (Tets).

Within this cumulative work, the cooperation of Dnmts and Tets in sustaining, but also altering an existing methylome has been investigated. For this, new hairpin sequencing techniques for the simultaneous and strand specific detection of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) were designed: One local, sequence specific analysis which permits the generation of ultra deep sequencing data, as well as one genome wide hairpin sequencing approach, capturing about 4 million CpGs equally distributed across the genome. In addition, distinct novel hidden Markov models have been developed, which, based on hairpin sequencing data, conduct a series of stochastic analyses. The first model estimates the enzyme efficiency of Dnmts in form of maintenance and *de novo* activity, as well as the efficiency of hydroxylation by Tets and also the strand specific distribution of 5mC and 5hmC. Furthermore, a second model determines the impact of neighbouring CpG methylation on the activity of Dnmts. The pipelines were applied to WT, as well as Dnmt and Tet knockout mouse embryonic stem cells under distinct environmental conditions.

Altogether, the here presented studies demonstrate that Dnmts and Tets do not act mutually exclusive on particular CpGs, but clearly in an opposed manner. In other words, methylated regions display high methylation efficiency, while unmethylated domains exhibit reduced methylation efficiency, but at the same time high hydroxylation efficiencies. Furthermore, in contrast to previous observations, the here presented data suggest a notable reduction of the demethylation rate in the absence of Tet enzymes, as well as an ectopic increase in both maintenance and *de novo* methylation. Finally, investigation of spatial methylation patterns reveals that the activity of Dnmt3a and 3b at a given CpG position is affected by the methylation state of its 5' neighbouring CpG. In summary, this work provides new strategies for the investigation of dynamic DNA methylation and demethylation processes and, furthermore, insight into the underlying mechanisms of how Dnmts and Tets balance the patterns and levels of DNA methylation.

ZUSAMMENFASSUNG

DNA-Methylierung in Säugern trägt maßgeblich zur Regulierung der Gen-Expression bei und definiert somit auch das Schicksal und die Identität von Zellen. Die Generierung und stabile Weitervererbung von zellspezifischen Methylierungsmustern werden durch die Aktivität von DNA-Methyltransferasen (Dnmts) gewährleistet, wohingegen der Abbau von DNA-Methylierung oft auf oxidative Cytosinformen und die Aktivität der *Ten-Eleven Tranlocation* Enzyme (Tets) zurückgeführt wird.

Im Rahmen dieser kumulativen Arbeit wurde die Zusammenarbeit von Dnmts und Tets in Bezug auf die Erhaltung, aber auch den Umbau eines bestehenden Methyloms untersucht. Zu diesem Zweck wurden neue *Hairpin*-Sequenzierungs-Techniken, die der gleichzeitigen und strangspezifischen Detektion von 5-Methylcytosin (5mC) und 5-Hydroxymethylcytosin (5hmC) dienen, etabliert: Eine lokale, sequenzspezifische Analyse, die die Generierung von Ultra-Tiefensequenzierungs-Daten erlaubt, sowie eine genomweite Hairpin-Sequenzierungs-Analyse, die rund vier Millionen CpGs, welche gleichmäßig über das gesamte Genom verteilt sind, erfasst. Zusätzlich wurden einzelne neuartige *hidden Markov Modelle* entwickelt, die, basierend auf den *Hairpin*-Sequenzierungs-Daten, eine Reihe von stochastischen Analysen durchführen. Das erste Modell kalkuliert die Enzym-Effizienz der Dnmts als *Maintenance*- oder *de novo*-Aktivität, die Hydroxylierungs- Effizienz der Tet-Enzyme, sowie die strangspezifische Ver- teilung von 5mC und 5hmC. Des Weiteren bestimmt ein zweites Modell den Einfluss von benachbarter CpG-Methylierung auf die Aktivität von Dnmts. Diese Herangehensweisen wurden auf embryonale Wildtyp-, aber auch Dnmt und Tet *Knockout* Stammzelllinien aus der Maus angewandt, welche unterschiedlichen Kultivierungsbedingungen ausgesetzt wurden.

Zusammenfassend zeigen die hier aufgeführten Studien, dass sich die Enzymaktivitäten von Dnmts und Tets an bestimmten CpG-Positionen nicht ausschließen, aber klar gegenläufig verhalten. Methylierte Regionen weisen starke Methylierungseffizienzen auf, wohingegen unmethylierte Bereiche niedrige Methylierungs-, dafür jedoch hohe Hydroxylierungs-Effizienzen aufweisen. Entgegen vorhergehender Beobachtungen zeigen die hier präsentierten Daten in Abwesenheit der Tet Enzyme eine deutliche Reduktion der Demethylierungsrate, sowie einen Anstieg von sowohl ektopischer *Maintenance*-, als auch *de novo*-Methylierung. Darüber hinaus zeigt die Untersuchung von benachbarten CpGs, dass die Enzym-Aktivität von Dnmt3a und 3b durch den Methylierungs-Zustand des jeweiligen benachbarten 5'-CpG's beeinflusst wird. Zusammenfassend beinhaltet die vorliegende

Arbeit neue Strategien zur Untersuchung von dynamischen DNA-Methylierungs- und Demethylierungs-Prozessen und gibt darüber hinaus Einblicke in die Mechanismen, mit denen Dnmts und Tets die DNA- Methylierungsmuster beeinflussen und im Gleichgewicht halten.

ABSTRACT

DNA methylation is an important epigenetic mark, which is set and maintained by DNA methyltransferases (Dnmts) and removed via passive or active mechanisms involving Ten eleven translocation enzyme (Tet) mediated oxidation. Stable cell type specific methylation patterns can only be achieved if methylation and demethylation events are in balance. Yet, the genome wide regulation of Dnmt and Tet activity is still not fully understood.

The present studies use novel hairpin sequencing techniques coupled with oxidative bisulfite sequencing, which permits the simultaneous and strand specific detection of 5-methylcytosine and 5-hydroxymethylcytosine. Application of HMM models then facilitates the estimation of enzyme efficiencies for Dnmts and Tets. Furthermore, spatial modelling of hairpin bisulfite data allows the investigation of how Dnmts interpret pre-existing methylation patterns.

Taken together, the results of the presented studies show that methylation and hydroxylation are antagonistic, but not mutual exclusive events. In this context, the data shows that Tet efficiency is highest at open and accessible chromatin. Furthermore, the absence of Tets leads to a considerable misregulation of Dnmts, resulting in an increase in both maintenance and *de novo* methylation efficiency. Lastly, the spatial analysis of methylation patterns reveals that the *de novo* methyltransferases Dnmt3a and 3b depend in their activity on pre-existing neighbouring CpG methylation.

KURZFASSUNG

DNA Methylierung is eine epigenetische Modifikation, welche durch DNA Methyltransferasen (Dnmts) gesetzt und beibehalten wird. Entfernt wird DNA Methylierung durch aktive oder passive Mechanismen welche die Oxidation von DNA Methylierung durch *Ten-Eleven Translocation* Enzyme (Tets) involviert. Stabile, Zelltyp-spezifische Methylierungsmuster können nur erreicht werden, wenn Methylierungs- und Demethylierungsvorgänge im Gleichgewicht sind. Dennoch ist die genomweite Regulation von Dnmts und Tets nicht vollständig geklärt.

Die hier gezeigten Studien verwenden neue *Hairpin*-Sequenzierungs-Verfahren, gekoppelt mit oxidativer Bisulfit-Sequenzierung, was eine simultane und strangspezifische Analyse von 5-Methylcytosin und 5-Hydroxymethylcytosin erlaubt. Die Anwendung von *hidden Markov Modellen* erlaubt im Anschluss die Berechnung von Enzymeffizienzen für Dnmts und Tets. Darüber hinaus erlaubt eine räumliche Modellierung von Methylierungsmustern die Untersuchung, wie Dnmts bereits bestehende Methylierung interpretieren.

Die Ergebnisse zeigen, dass Methylierung und Hydroxylierung antagonistische, aber keinesfalls sich ausschließende Ereignisse sind. Dabei zeigen Tets ihre stärkste Aktivität an offenem und zugänglichem Chromatin. Zudem führt der Verlust von Tets zu einer deutlichen Missregulation von Dnmts, welche sich durch eine Zunahme der *Maintenance* und *de novo*-Methylierungseffizienz äußert. Schließlich zeigt die räumliche Modellierung, dass die *de novo*-Methyltransferasen bei ihrer Aktivität abhängig von bereits bestehender DNA Methylierung sind.

CONTENTS

1. BACKGROUND

The construction of a multicellular organism from a single cell is a complex and astounding process. Sperm and oocyte fuse to one primordial cell, the zygote, which after a vast number of subsequent cell divisions gives rise to hundreds of distinct cell types. The genetic information, the blueprint for the construction is provided in form of a nucleotide sequence, the deoxyribonucleic acid (DNA). The DNA itself is organised in functional groups, genes, which yield the design for proteins and functional ribonucleic acids (RNAs), the basic components of each cell. Since all cells originate from the same cell, they all share the same DNA sequence. However, each cell has its individual function and executes its own distinct and highly specific program. Such cell type specific gene expression is realised by epigenetic mechanisms.

1.1 Epigenetics

The prefix *epi-* ($\epsilon\pi\iota$) originates from Greek and means ‘on top’ or ‘over’ and refers to epigenetic mechanisms as a second layer of information alongside the DNA sequence. Epigenetic information does not alter the genomic sequence of the DNA, but will also be inherited from one cell generation to the next. These epigenetic mechanisms include small regulatory RNAs, post translational modifications of histones, as well as covalent DNA modifications.

1.2 Histones and Histone Modifications

Eukaryotic DNA does not simply exist as promiscuous linear molecules, instead, it is organised in a tightly controlled structure, the chromatin [1]. The basic component of chromatin is the ‘nucleosome’ which in turn consists of 145 to 147 base pairs (bp) of DNA wrapped around a protein octamer complex. Each octamer is constructed by small basic globular proteins, the histones. In mammals, a ‘typical’ nucleosome contains pairs of the canonical histones H2A, H2B, H3 and H4, respectively [2, 3]. Histones themselves are built by a highly organised globular part forming the octamers, as well as a sparsely structured N-terminal tail. On average, nulceosomes occur every 200 ±40bp [4]. Nevertheless, based on their actual density one distinguishes between transcriptional active, loose ‘euchromatin’ and the inactive, dense ‘heterochromatin’.

The chromatin state is defined by post-translational modifications (PTMs) of histones. Various types of PTMs are known, including acetylation, methylation, phosphorylation, ubiquitination, sumoylation, ADP ribosylation, deamination, as well as the more recently discovered propionylation and butyrylation [5]. Until today, the most studied modifications are PTMs of the loosely organised histone tails, even though PTMs of the histone central domain are not less versatile. PTMs can change the chromatin structure in two ways. First, based on their chemical properties, they will increase or decrease either the interaction with DNA and/or other nucleosomes [6, 7]. Secondly, PTMs are able to recruit transcription factors and chromatin remodeller which de-condense or condense chromatin resulting in activation or de-activation of transcription [8, 9]. Trimethylation of lysine 4 and lysine 36 of the histone H3 (H3K4me3 and H3K36me3) for example causes transcriptional activation, while trimethylation of lysine 9 and lysine 27 (H3K9me3 and H3K27me3) results in transcriptional repression [10, 11, 12].

However, in mammals, a stable transcriptional control can only be achieved in cooperation with DNA methylation, which will be the main focus of the presented cumulative work. Several studies show that DNA methylation can influence PTMs of histones and *vice versa* [13, 14, 15, 16].

1.3 Regulatory RNAs

RNA is considered the first molecule of life, both, in form of genetic information and in the context of catalytic reactivity [17]. However, a long pursued dogma in biology pictured RNA as a transport molecule, shipping the genetic information from DNA to the protein synthesising ribosome [18, 19]. During the last two decades, a vast amount of RNA classes have been identified which display distinct characteristics and involvement in almost all biological processes.

Growing evidence also suggests the participation of RNAs in epigenetic regulation of gene transcription [20, 21]. Several small RNAs, such as siRNAs (short interfering RNAs), piRNAs (PIWI associated RNAs) or tiRNAs (transcription initiation RNAs), but also long non-coding RNAs (lncRNAs) have been suggested to regulate epigenetic processes [20, 22, 23, 24]. Thus, small RNAs interact with Polycomb group proteins and are involved in transcriptional silencing [25]. lncRNAs, for example, take part in X-chromosome silencing and parental imprinting in mammals [26]. In addition, lncRNAs have been found to interact with certain histone types, the Polycomb repressive complex and also DNA methyltransferases [27, 28, 29].

1.4 DNA Methylation

The concept of DNA methylation is shared by many organisms, from bacteria over plants and insects up to mammals, concerning several modified bases such as N6-methyladenine, N4-methylcytosine (in bacteria) or C5-methylcytosine (5mC). Irrespectively of the modified base, DNA methylation is generated by a conserved mechanism. Specific enzymes, the DNA methyltransferases (Dnmts), catalyse the transfer of a methyl group from S-adenosylmethionine to their target base. In bacteria, DNA methylation cooperates with endo-nucleases as a protective mechanism against invading viral DNA, whereas eukaryotes use methylated bases to silence retro-viral DNA already integrated into their genome and also as a mechanism for gene regulation.

1.4.1 DNA Methylation in Mammals

The importance of DNA methylation in mammals has been shown in mice, where the absence of methylation marks leads to a developmental arrest in the early embryo [30, 31]. Basically, DNA methylation is essential in two ways. First, non-coding sequences such as repetitive-, as well as transposable elements, which represent the majority of the genome, are silenced to ensure genome integrity [32]. Secondly, DNA methylation regulates the transcription of genes according to cell type specific methylation patterns. The most abundant form of DNA methylation in mammals is 5-methylcytosine (5mC), which predominantly occurs in a 5'-CpG-3' (Cytosine-phosphatidyl-Guanine) di-nucleotide context [33, 34]. In somatic cells, between 70% to 80% of all CpGs are methylated [35, 36]. However, studies indicate that CpGs tend to be methylated according to their frequency in appearance [37]. Low CpG density leads to high methylation level and *vice versa*. Interestingly, CpGs are underrepresented in the mammalian genome [37]. Studies showed that highly methylated sequences include repetitive elements, satellite DNA, intergenic DNA and exon sequences. However, across the genome there are regions with high CpG density, so called CpG islands (CGIs) [38]. These CGIs are about 1kb in length and unmethylated in most cell types. The majority of CGIs can be assigned to gene promoters. Nevertheless, a smaller number of CGIs can also be found in intra- or intergenic regions.

1.4.2 DNA Methyltransferases

CpG methylation in mammals is set and maintained by a family of C5 DNA methyltransferases (Dnmts) which all share a highly conserved C-terminal domain [40]. In fact, after identification of Dnmt1, other Dnmts were identified by sequence homology search, which in total revealed the existence of five conserved Dnmts, namely Dnmt1, Dnmt2, Dnmt3a, Dnmt3b and Dnmt3l [41, 42].

Dnmt1, 3a and 3b represent the canonical Dnmts with catalytic activity, responsible for the setting and maintenance of stable methylation patterns [41, 31]. In contrast, Dnmt2

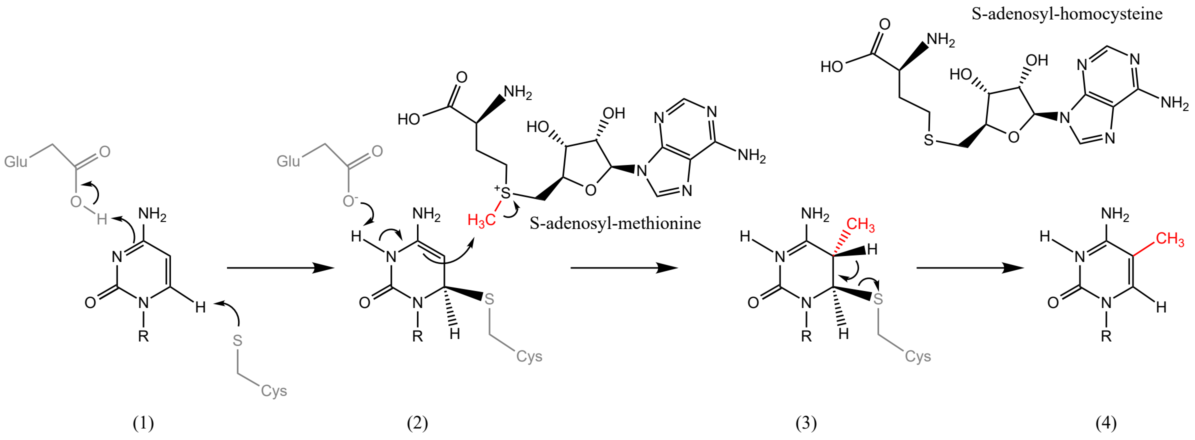


Fig. 1.1: Methylation mechanism of cytosine by Dnmts according to Lyko, 2018 [39]. (1) Nucleophilic attack on the C6 position of the cytosine ring by a conserved cysteine residue facilitated by a similarly conserved glutamic acid residue. (2) Transfer of a methyl group from S-adenosyl-methionine to C5. (3) Deprotonation of C5. (4) Generated 5-methylcytosine.

and Dnmt3l do not posses any catalytic activity against DNA [43, 44]. Later studies revealed Dnmt2 as RNA methyltransferase, which methylated the position 38 in tRNA, whereas Dnmt3l functions as a cofactor for Dnmt3a and 3b [45, 46, 47, 32, 48]. Recently, Dnmt3c has been identified as a sixth member of the cytosine-5-methyltransferase (C5-MTase) family. However, Dnmt3c occurs only in mouse with a low catalytic activity, male germ cells [49].

All canonical Dnmts methylate DNA *via* the same mechanism (Figure 1.1). Initially, the targeted cytosine base is rotated into the catalytic pocket using base-flipping. Subsequently, a conserved cysteine residue mediates a nucleophilic attack on the C6 position followed by the transfer of the methyl group from S-adenosylmethionine (SAM) to the C5 atom in the cytosine ring (Figure 1.1).

With the exception of Dnmt2, all Dnmts contain a regulatory, N-terminal domain. These domains are quite distinct in size and their composition of sub-domains, which strongly influences the properties of the individual enzyme. Based on their structure and biological function, the canonical Dnmts are categorised into the *de novo* methyltransferases Dnmt3a and 3b, and the maintenance methyltransferase Dnmt1 (Figure 1.3).

Maintenance Methyltransferase Dnmt1. The main function of Dnmt1 is the truthful inheritance of methylation patterns across replication and cell division (Figure 1.2). Experiments *in vitro* revealed that Dnmt1 displays a much higher affinity towards DNA which is only methylated on one DNA strand (hemimethylated) [50]. Depending on the conditions, the activity is up to 50 times higher compared to completely unmethylated DNA. At the same time, Dnmt1 is characterised by a high processivity, meaning that it methylates CpGs in a consecutive manner [50]. Through interaction with PCNA and Uhrf1, Dnmt1 is closely coupled with the replication machinery and Uhrf1 further enhances the affinity of

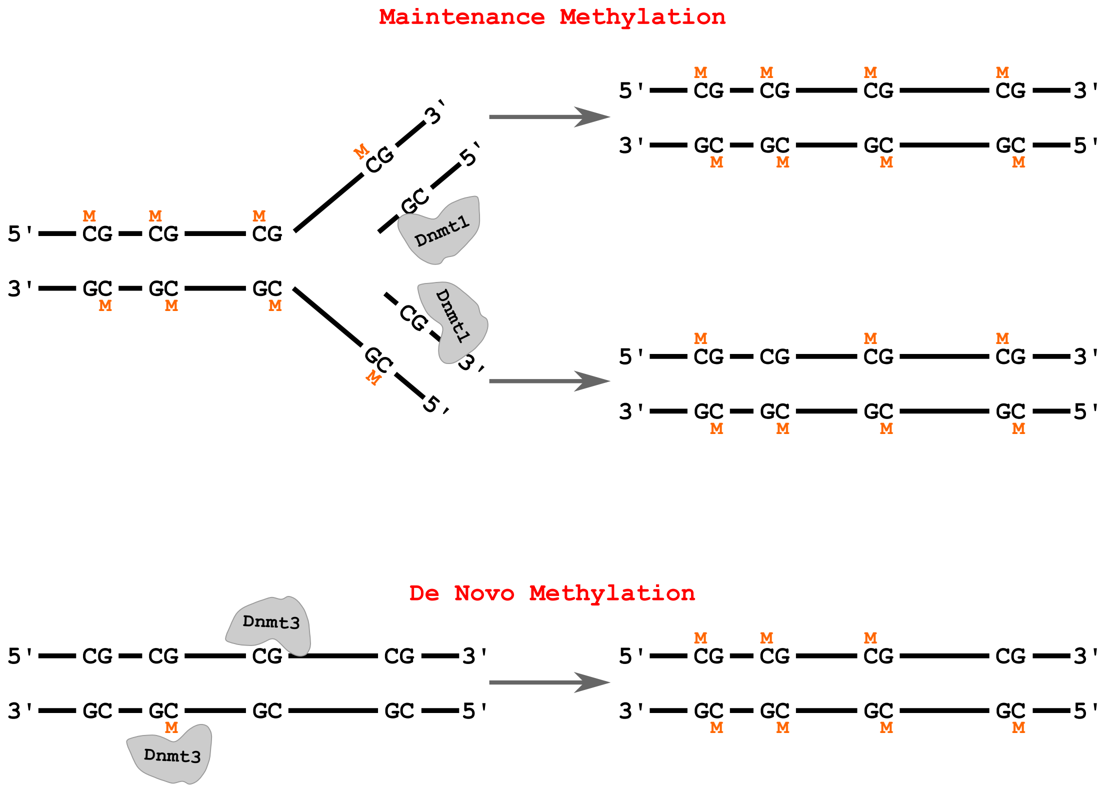

Maintenance Methylation
5' — CG — CG — CG
3' — GC — GC — GC
M
CG 3'
GC 5'
Dnmt1
CG 3'
GC 5'
Dnmt1
5' — CG — CG — CG — CG — 3'
3' — GC — GC — GC — GC — 5'
De Novo Methylation
Dnmt3
5' — CG — CG — CG — CG — 3'
3' — GC — GC — GC — GC — 5'
Dnmt3
5' — CG — CG — CG — CG — 3'
3' — GC — GC — GC — GC — 5'

Fig. 1.2: Schematic display of maintenance and *de novo* methylation. After replication, Dnmt1 detects hemimethylated CpGs and restores the methylation pattern on the newly synthesised DNA strand, i.e. maintenance methylation. Dnmt3a and 3b methylate DNA independently of methylation status or replication and can create new methylation pattern, i.e. *de novo*.

Dnmt1 to hemimethylated CpGs [51, 52, 53, 54, 55]. After replication the Uhrf1/Dnmt1 complex determines the methylation status of the parental DNA strand and transfers the information with high precision to the newly synthesised DNA strand. The literature provides different information about the accuracy of Dnmt1. With a precision of 95% to 96%, Vilkaitis *et al.* grant Dnmt1 a relatively high error rate, whereas Goyal *et al.* determines a vast fidelity of Dnmt1 with 99.7% accuracy [56, 57]. Knock-Out (KO) of Dnmt1 results in an almost complete loss of DNA methylation in ES cells and an early embryonic lethality [30, 58].

Dnmt1 was the first Dnmt described in mammals and represents the largest member of the MTase family (Figure 1.3). The complex N-terminal part of Dnmt1 can be divided into multiple conserved sub-domains which facilitate protein interactions and modulates DNA binding, as well as methylation activity. Amongst others, Dnmt1 contains a nuclear localisation signal (NLS) as well as a PCNA interacting motif which guides Dnmt1 into the nucleus and to replication foci, respectively. The Dnmt1-associated protein 1 (DMAP1) binding domain facilitates, as the name suggests, the interaction with the transcriptional repressor DMAP1, but also binding to HDAC2 and contributes to the stable localisation of Dnmt1 to replication foci [59]. However, the actual recruitment to the replication fork is accomplished by the replication foci targeting sequence (RFTS) [60]. In addition, Dnmt1 contains a CXXC, a conserved zinc-finger domain. CXXC binds unmethylated CpG rich DNA and in the case of Dnmt1, binding of the CXXC domain onto the DNA positions an auto-inhibitory linker region between DNA and catalytical domain which prevents the methylation of unmethylated CpGs [61]. However, other investigations indicate that this mechanism does not apply for Dnmt1 protein [62]. Close to the C-terminal catalytic domain, there are two bromo-adjacent homology (BAH) domains, but their function remains elusive.

To some extent, methylation of cytosine can also be found in a nonCpG i.e. CpH (H = A, T, C) context. Particular abundance of nonCpG methylation has been described in oocytes, ES cells and also neurons [63, 64, 65, 66]. NonCpG methylation is generated by the *de novo* methylation and initially has been considered a side product of strong Dnmt3a and 3b activity [63, 67]. Yet, increasing evidence suggest also a functional role of nonCpG methylation in gene regulation [67, 68, 69, 70]. Thus, the presence of nonCpG methylation is involved in transcriptional silencing by recruiting repressing factors such as MeCP2 or REST [68, 69]. A schematic representation summarising location and relative sizes of the functional domains of Dnmts is given in Figure 1.3.

De novo *Methyltransferases Dnmt3a and Dnmt3b.* During embryonic development the inner cell mass gives rise to numerous cell types and subsequently forming a multicellular organism. ES cells are epigenetically ‘naive’, which is reflected by almost complete absence of DNA methylation. The new methylation patterns needed for creation of somatic cell types are set by Dnmt3a and 3b, also referred to as *de novo* methyltransfearses. In contrast to Dnmt1, these enzymes exhibit no preferences for hemi- or unmethylted substrates *in vitro* and are consequently able to generate entirely new methylation patterns (Figure 1.2) [31, 71]. Dnmt3a and 3b deficient ESCs and embryos failed to *de novo* methylate repetitive elements and retroviral sequences. In addition, it has been demonstrated that the presence of Dnmt3a is necessary for the proper methylation of imprinted genes [72].

Both proteins are considerably smaller compared to Dnmt1 and contain distinct regulatory domains (Figure 1.3). The PWWP domain resembles a conserved ‘proline-trypthophan - trypthophan-proline’ motif facilitating chromatin association by specific interaction with histone 3 trimethylated at lysine 36 (H3K36me3) and is essential for targeting major satellite repeats in pericentric heterochromatin [73, 74, 75]. The ADD or ATRX domain allows protein interaction such as binding to histone deacetylase 1 (HDAC1). Among each other, Dnmt3a and 3b share a highly conserved catalytical domain with 85% sequence homology.

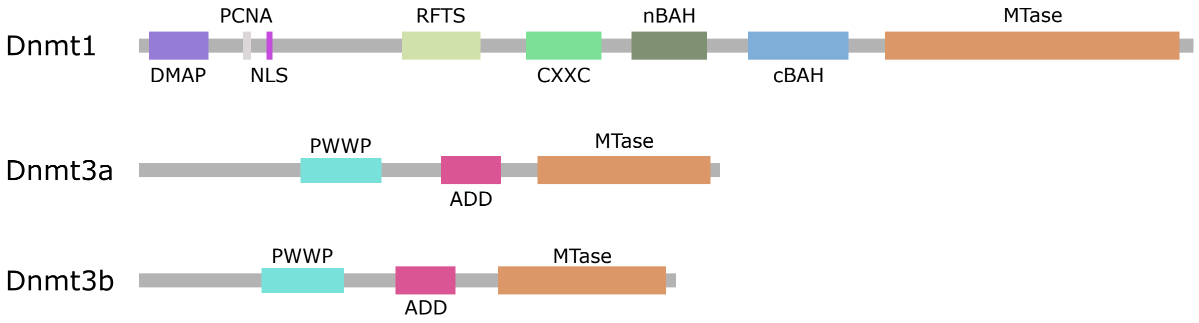


Fig. 1.3: Schematic protein domain representation of the canonical mammalian DNA methyltransferases Dnmt1 (1620 amino acids), Dnmt3a (908 aa) and Dnmt3b (859 aa). DMAP = DMAP1 binding domain, PCNA = PCNA binding domain, NLS = nuclear localisation sequence, RFTS = replication foci targeting sequence, CXXC = conserved zinc finger DNA binding domain, nBAH = N-terminal bromo-adjacent homology domain, cBAH = c-terminal bromo-adjacent homology domain, MTase = catalytic DNA methyl transferase domain, PWWP = Proline-Tryptophan-Tryptophan-Proline domain, ADD = ATRX-Dnmt3-Dnmt3L domain.

1.5 *Detection of 5-Methylcytosine*

Over the years, several techniques have been developed to measure levels and distribution of 5mC. Methylation sensitive restriction enzymes coupled with quantitative real-time PCR (qPCR), allow to determine the methylation level at certain CpGs [76, 77, 78, 79]. 5mC specific antibodies are used in staining or enrichment based approaches, which are either linked with qPCR or sequencing, revealing a relative abundance of 5mC across the genome [80, 81, 82].

However, the gold standard in 5mC detection is bisulfite sequencing (BS) [83]. Bisulfite treatment of DNA basically converts the illegible methylation signal into a readable genomic sequence and permits quantification of 5mC at single base resolution. Under aqueous and mild acidic conditions (around pH 5) bisulfite anions react with the C6 position of unmethylated cytosine forming cytosinesulphonate, whereas 5mC remains unaffected (Figure 1.4). Subsequently, the molecule undergoes hydrolysis and deamination resulting in uracilsulphonate. Desulphonation under alkaline conditions eventually generates uracil (U).

As a result, subsequent PCR and sequencing results in a thymine (T) ‘signal’ for unmethylated cytosines, whereas 5mC presents itself as cytosine. Alignment with the genomic sequence of a reference sequence then allows quantification and pinpointing of methylated cytosines. Nowadays, modern ‘next generation’ sequencing techniques facilitate high throughput applications of BS and furthermore provide ultra deep sequencing depth or even genome wide detection of 5mC.

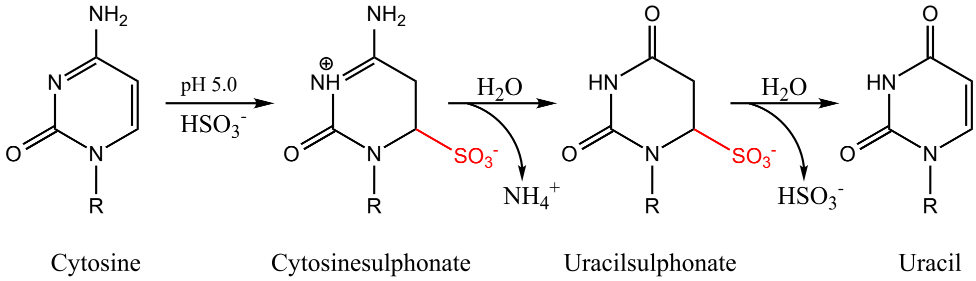


Fig. 1.4: Bisulfite conversion of unmethylated cytosine. Under mild acidic conditions, cytosine reacts with bisulfite anions forming cytosinesulphonate. Subsequent hydrolysis and deamination result in the generation of uracilsulphonate. Lastly, desulfonation under alkaline conditions results in the formation of uracil.

1.6 Hairpin Bisulfite Sequenicng

Hairpin bisulfite sequencing (HBPS) presents the experimental basis of the here published studies. This section will give an overview about the methods concept while Chapter 2 will discuss HPBS in detail. HPBS overcomes limitations of conventional BS approaches and was developed by Laird *et al.* to study methylation fidelity or rather developmental methylation changes at particular loci [84, 85].

During bisulfite treatment, successful conversion of C to U requires the denaturation of both complementary DNA strands. As a consequence, subsequent sequencing recovers only the information of one DNA strand. In HPBS, genomic DNA is first subjected to enzymatic digestion using endonucleases. Next, a short hairpin oligonulceotide is attached to each end of the DNA fragment which physically connects upper and lower strand. Following bisulfite treatment, PCR and sequencing, the methylation patterns from both complementary DNA strand of one individual chromosome i.e. DNA molecule, can be derived. In other words, HPBS detects whether a given CpG dyad is methylated at only one DNA strand (hemimethylated), at both strands (fully methylated) or completely unmethylated. Recently, a first genome wide hairpin approach has been developed by Zhao *et al.* [86] who investigated the methylation fidelity in mouse ES cells.

1.7 Tet Enzymes and Oxidation of 5mC

Ten-eleven translocation (Tet) enzymes were identified as iron-II and oxoglutarate dependent di-oxygenases, which oxidise 5mC to 5-hydroxymethyl cytosine (5hmC) [87, 88]. The enzymes are named after the ten-eleven translocation, t(10;11)(q22;q23), sometimes found in patients bearing a rare case of acute myeloid and lymphocytic leukaemia, in which the first family member, Tet1, was first described [89, 90]. Indication for enzymatic activity of Tet enzymes arose from comparisons with J binding proteins from *Trypanosoma brucei* which are involved in the generation of base ‘J’ by oxidizing the T to 5-hydroxyuracil. Computational screens identified a large family of JBP homologue

DNA modifying enzymes including the metazoan Tets [87, 91]. During evolution, the mammalian Tet precursor gene went through triplication, giving rise to three catalytic active Tet family members, namely Tet1, Tet2 and Tet3 (Figure 1.5).

All Tet enzymes share a catalytic C-terminal domain which comprises amino acid motifs for the binding of iron-II (Fe^{2+}) ions, as well as oxoglutarate (α-ketoglutaric acid) [91, 88]. The catalytic domain is constructed by a double-stranded β-helix fold and contains a cystein (Cys) rich region at its terminal part which is limited to metazoan Tet-JBP families. Nine conserved Cys coordinate zinc ions (Zn^{2+}) and putatively contribute to DNA binding. Positioning of essential cofactors (Fe^{2+}) and 2-oxogluterate within the catalytical pocket is facilitated by two conserved histidines (His), one of which is present in a His-Xaa-Arg/Glu motif , and an likewiseconserved Argenine (Arg), respectively [92, 93].

In addition, Tet1 and Tet3 contain a N-terminal located CXXC domain which inherently facilitates DNA binding and is shared by many other DNA interacting proteins.

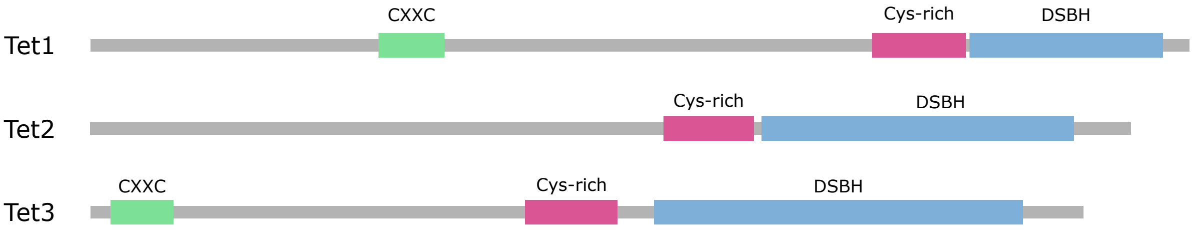


Fig. 1.5: Schematic protein domains representation of Ten-eleven translocation di-oxigenase enzymes Tet1 (2007 aa), Tet2 (1912 aa) and Tet3 (1668 aa). CXXC = conserved zinc finger DNA binding domain, Cys-rich = cysteine, DSBH = double-stranded β-helix, Cys-rich and DSBH form the catalytic domain.

Publications suggest that the CXXC domain of Tet1 displays substantial differences compared to CXXC domains of other proteins. It has been proposed that the distinct differences actually prevent Tet1 binding to DNA, whereas other studies suggest unique features allowing the binding of unmodified, methylated and hydroxymethylated DNA [94, 95, 96]. Moreover, Tet1 naturally locate at CGIs. Yet, mutation in the CXXC domain abolishes recruitment to CGIs, demonstrating the role of the CXXC domain in the context of Tet regulation.

Tet2 underwent evolutionary chromosomal inversion and as a result, became separated from its CXXC domain [97]. Now, the CXXC domain is encoded by the gene Idax, located in close proximity to Tet2. IDAX was shown to interact and recruit Tet2 to DNA and consequently, Tet2 exhibits distinct binding patterns depending on the presence or absence of IDAX [98]. Interestingly, recruitment of Tet2 by IDAX to DNA results in caspase-dependent degradation of the protein complex [98].

In the case of Tet3, the characteristics of its CXXC domain directs the enzyme mainly to unmethylated, CpG rich regions [99]. Additionally, in the absence of the CXXC domain, Tet3 displays an increased activity, suggesting a negative regulation by the CXXC motif.

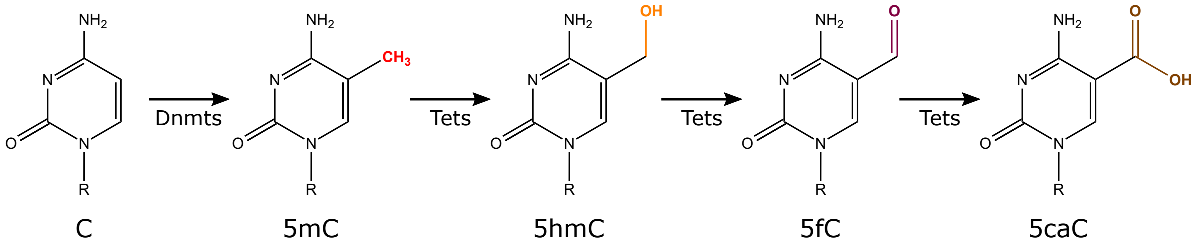


Fig. 1.6: Stepwise modification of cytosine by Dnmts and Tets to 5mC (5-methylcytosine), 5hmC (5-hydroxymethylcytosine), 5fC (5-formylcytosine) and 5caC (5-carboxylcytosine).

Generally, all three Tet enzymes were shown to catalyse the oxidation of 5mC to 5hmC [88]. Pursuing studies then revealed that Tet enzymes subsequently oxidise 5hmC in a consecutive manner to 5fC and eventually 5caC [100]. Figure 1.6 displays the set of cytosine forms and the DNA modifying enzymes responsible for their generation.

Mammalian Tets are broadly expressed throughout the organism but nevertheless, display distinct expression profiles. Tet1 for example represents the main oxygenase in mouse ES cells and primordial germ cells (PGCs) [88, 101, 102, 103], while Tet2 and Tet3 are expressed in hematopoietic stem cells, somatic cell types and neuronal cells, respectively [104, 105]. Moreover, Tet3 is the sole family member present in oocytes and early single cell zygotes [106, 107]. In accordance to the cell type specific expression profiles of Tets, oxidative cytosine forms (oxCs) exist in various cell types, though far less abundant compared to 5mC. Furthermore, whereas the overall 5mC levels are quite constant throughout somatic tissues (4.3% of all Cs), amounts of oxCs are considerable variable and cell type specific [108, 109]. The highest levels of 5hmC are found in brain cells and ES cells with 0.3% to 0.7% of all Cs [88, 108].

1.8 DNA Demethylation

Even though DNA methylation is considered a rather stable epigenetic mark, developmental stages, as well as cellular differentiation, require the erasure of 5mC from the DNA. Such demethylation events can occur genome wide but also in a local, sequence specific manner. Global loss of 5mC, for example, can be found after fertilisation where both, maternal and paternal nuclei undergo massive genome wide demethylation, building the basis for the formation of pluripotent embryonic stem cells [110, 111]. Similarly, maturation of PGCs in post-implantation embryos requires the genome wide erasure of parental specific methylation imprints. Local demethylation are manifold and practically can be found during any cellular differentiation process [112, 113, 114]. Promoters and enhancers for example become demethylated prior to or as a consequence of gene activation to ensure stable expression of developmental and cell type specific genes.

Based on the underlying mechanism, demethylation can be divided into passive and active demethylation. While passive methylation is the result of replication dependent dilution of 5mC, active loss of DNA methylation involves enzymatic removal of the modified cytosine or even direct elimination of the methyl group itself. In either case, growing evidence indicates an involvement of Tets and oxidative cytosine species in the controlled removal of 5mC [107, 106, 115]. Figure 1.7 summarises several discussed active and passive demethylation pathways.

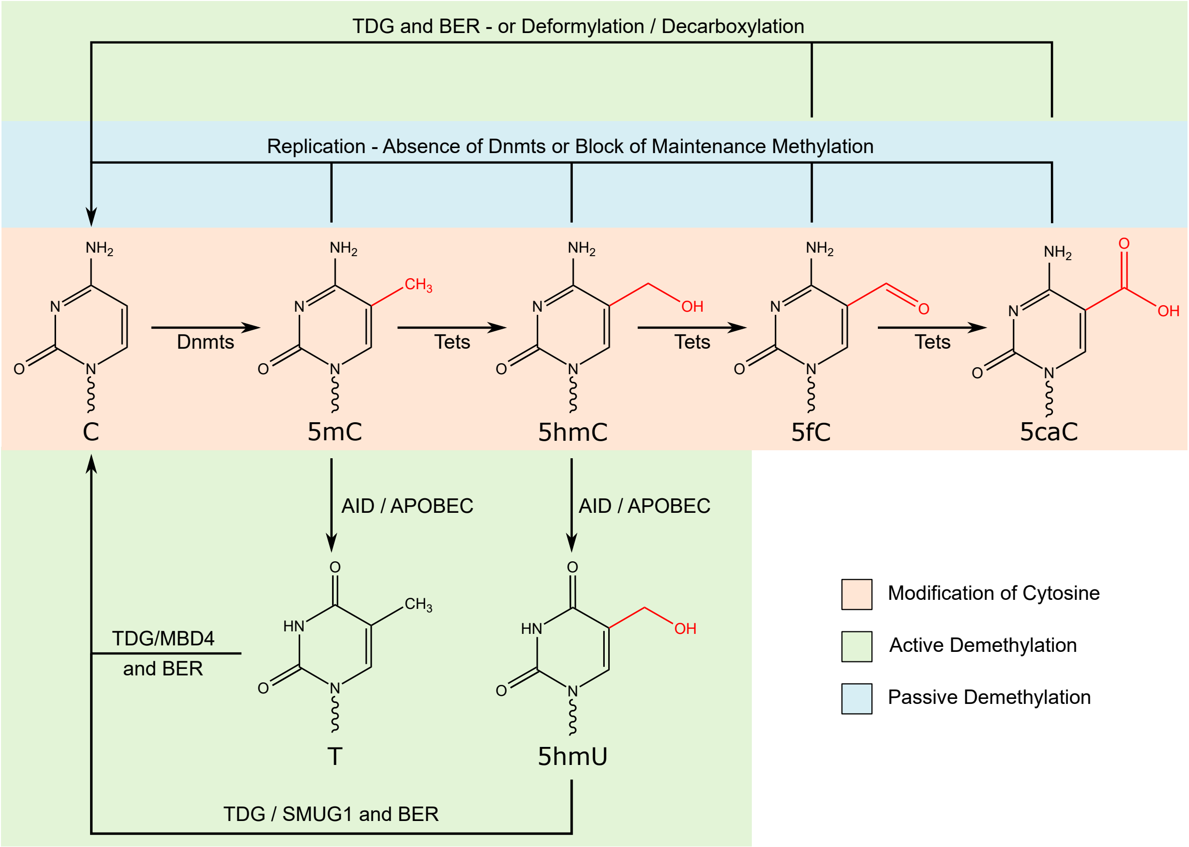


Fig. 1.7: Graphical display of active and passive demethylation pathways; DNA modification events (methylation and oxidation) are highlighted in orange, passive demethylation events in blue and active demethylation events in green.

1.8.1 Passive Demethylation

During passive demethylation, 5mC becomes depleted from the genome by successive rounds of DNA replication and the concurrent absence of maintenance methylation activity. The absence of maintenance can be the result either absence of Dnmts or functional blockage of maintenance methylation ativity itself. Previous studies demonstrate that in oocyte and early zygote, distinct Dnmt1 isoforms are actively retained from entering the nucleus, subsequently preventing effective preservation of DNA methylation [116, 117, 118, 119]. Moreover, 5hmC has been considered to block maintenance methylation by inhibition of the Dnmt1-Uhrf1 complex. Even though it is controversially discussed whether Uhrf1 can recognise 5hmC, *in vitro* studies unanimously describe a strongly reduced Dnmt1 activity in the presence of hemihydroxylated DNA [120, 121]. One study indicates that not only 5hmC, but also 5fC and 5caC influence the activity of Dnmts even at neighbouring CpGs [122].

1.8.2 Active Demethylation

Several mechanisms have been described in which enzymatic removal of modified cytosines will restore unmethylated CpGs. In plants, particular DNA glycosylases recognise 5mC directly and catalyse its removal from the DNA [123, 124]. However, in mammals, so far no orthologs have been identified. Yet, some *in vitro* studies suggest that in the absence of SAM, Dnmts might act as de-hydroxylases or de-carboxylases, converting 5hmC into unmodifed cytosine [125, 126, 127]. Similarly, experiments with ES cell lysate guessed the presence of 5caC specific decarboxylases, but a liable enzyme could not been identified [128].

A more likely mechanism involves the deamination of 5mC or 5hmC by members of the AID/APOBEC deaminase family. Indications come from experiments with over-expression of AID/APOBECs. In neuronal cells, their enrichment leads to a reduction of 5hmC and a simultaneous artificial expression of Tet1 causes accumulation of the 5hmC deamination product 5hmU [115]. Subsequently, thymine DNA glycosylase (TDG) or single-strand selective monofunctional uracil glycosylase 1 (SMUG1) which show a high activity against U:G and T:G mismatches, will cleave the base from the DNA before unmodified cytosine is restored by the base excision repair machinery (BER) [129]. However, APOBEC deaminases prefer single stranded substrates which might reduce the biological relevance of this mechanism. Furthermore, a recent publication demonstrates that, *in vitro*, one member of the AID family strongly discriminates against oxCs, including 5hmC [130].

One of the most accepted active demethylation pathways includes the activity of TDG. While the intrinsic substrates of the enzyme are T:G and U:G mismatches, it also recognises 5fC:G and 5caC:G pairs in double stranded DNA [131, 132]. In fact, the activity towards 5fC and 5caC is even higher compared the its 'natural' targets [132]. TDG deficiency in mouse ES cells causes a strong increase in 5fC and 5caC levels while a combined over expression of Tet1 and TDG leads to almost complete depletion of both bases[131, 133]). Furthermore, TDG is the only member of uracil DNA glycosylases which displays an activity against 5fC/5caC and the only glycosylase known to be essential for early embryogenesis [134, 129].

1.9 Detection of Oxidised Cytosine Forms

In order to decipher the role of oxC forms, several methods for their detection have been developed. Comparable to the analysis of 5mC, detection strategies for oxCs differ in sensitivity and resolution, ranging from the global measurement of oxC levels by liquid chromatography or mass spectrometry to base resolution application using next generation sequencing [100, 135, 136]. In most cases, the detection of oxC forms using sequencing is coupled to BS treatment. However, as indicated in Figure 1.8, 5hmC, 5fC and 5caC all display individual chemical properties which influence their conversion during incubation with bisulfite. While 5hmC remains resistant to BS treatment, 5fC and 5caC both undergo conversion to 5-formyluracil (5fU) and 5-carboxyluracil (5caU), respectively [137, 138, 139]. Hence, after PCR and sequencing, 5hmC will present itself as C, whereas 5fC and 5caC will be detected as T. Consequently, a clear separation between 5mC and 5hmC, as well as C, 5fC and 5caC is not possible. Novel sequencing techniques overcome this limitation by applying additional chemical or enzymatic treatments, which alter the chemical properties of oxCs. In the following, the most common approaches will be shortly presented.

1.9.1 Sequencing of 5hmC

Oxidative Bisulfite Sequencing: In order to separate the collective C signal from 5mC and 5hmC after classical bisulfite sequencing, Booth *et al.* developed oxidative bisulfite sequencing (oxBS) [140, 141]. The method uses potassium perruthenate ($KRuO_4$) to selectively oxidise 5hmC to 5fC prior to BS treatment. As a result, only 5mC will later be detected as C. Parallel application of BS and oxBS of the same sample followed by data comparison will then permit the estimation 5hmC amount and location. A schematic overview about BS and oxBS conversion is given in Figure 1.8.

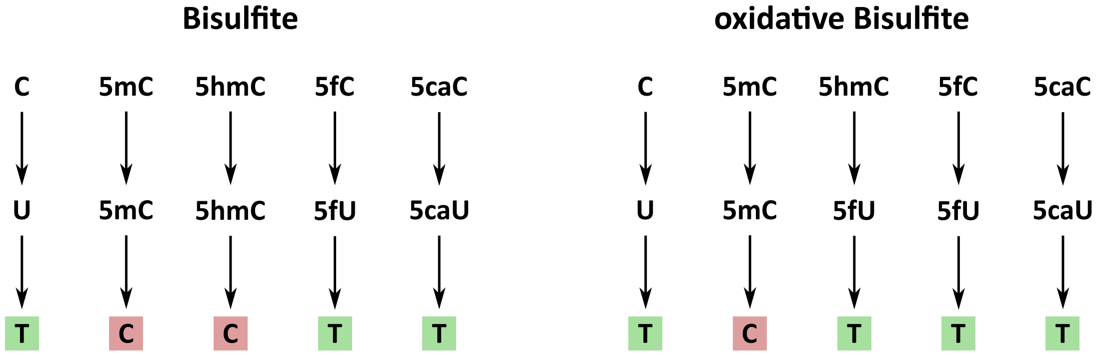


Fig. 1.8: Conversion of cytosine and modified cytosines during BS and oxBS treatment. C, 5fC, as well as 5caC are converted during bisulfite treatment, while 5mC and 5hmC remain unchanged. In oxBS, 5hmC becomes first oxidised to 5fC and will later be converted to U, consequently only 5mC is detectable as C during sequencing.

Tet Assisted Bisulfite Sequencing: In parallel, Yu *et al.* developed Tet assisted bisulfite sequencing (TAB-Seq) [142]. Within this pipeline, 5hmC is initially glycosylated by T4-β-glucosyltransferase (βGT). Next, the DNA is subjected to Tet oxidation, which will convert all 5mC to 5fC or 5caC, while glycosylated 5hmC remains resistant to Tet activity.

Eventually, after bisulfite treatment, PCR and sequencing the only 5hmC will be marked as C while all other cytosine forms will appear as T.

APOBEC-Coupled Epigenetic Sequencing: Recently, a bisulfite-free localisation method for 5hmC has been developed, APOBEC-coupled epigenetic sequencing (ACE-Seq). In ACE-Seq, 5hmC is again shielded by glycosylation before the DNA is subjected to denaturation and enzymatic treatment with APOBEC (A3A). Except for 5hmC, all other cytosine forms will be deaminated by A3A and thus, converted to T. During sequencing, the presence of 5hmC in the genome will be indicated by a cytosine signal [143].

1.9.2 Sequencing of 5fC and 5caC

The low abundance of 5fC and 5caC makes an accurate detection of both cytosine forms particular challenging. Nevertheless, several techniques have been developed, which allow their detection.

CAB Sequencing: In chemically assisted bisulfite (CAB), 5fC (fCAB) or 5caC (caCAB) are selectively labelled using O-ethylhydroxylamine ($EtONH_2$) or 1-ethyl-3-[3-dimethyl-aminopropyl]- carbodiimide hydrochloride (EDC), respectively [135, 144, 145]. Once labelled, both bases become resistant to downstream deamination during bisulfite treatment. Localisation of 5fC or 5caC is then deduced by subtracting the traditional BS signal from the fCAB/caCAB sequencing readout.

MAB Sequencing: Methylation or M.SssI assisted bisulfite sequencing (MAB-Seq) detects the collective signal of 5fC and 5caC [138, 136]. Prior to the incubation with bisulfite, the DNA is subjected to M.SssI catalysed methylation reaction, which converts all C in a CpG context to 5mC. Once more, 5mC remains unaffected during bisulfite treatment and any detection of T (unmethylated CpGs) after sequencing corresponds to 5fC or 5caC.

1.10 Novel Sequencing Techniques

New possibilities for the detection of oxCs, or rather, DNA modification in general, emerge from novel sequencing techniques such as nanopore and single molecule real-time (SMRT) sequencing [146, 147, 148]. Both techniques sequence native single DNA molecules without the need of prior amplification and are also able to detect modified DNA bases and furthermore, allow long read lengths of more than 20kb.

Nanopore Sequencing: In nanopore sequencing, single-stranded DNA slips through a small pore within a electroconductive membrane. When passing through the pore, the

DNA will alter the ionic current, which is applied to the membrane. Thus, base composition and chemical modifications e.g. 5mC/5hmC will result in unique changes of the applied current, which can be used to identify DNA sequence and also DNA modifications.

SMRT Sequencing SMRT sequencing uses sequencing-by-synthesis. For this, the DNA is first fragmented and processed into double stranded dumbbell-like structures. Next, sequencing primer and polymerase are annealed before single molecules are loaded into nanoscale observation chambers. The polymerases will then incorporate fluorescently labelled nucleotides and the omitted signal will be recorded in real-time. Real-time detection allows to determine the duration between nucleotide incorporations. In this context, epigenetic modifications will cause a delay, which is unique for distinct modification types and hence, allows the characterisation of DNA modifications.

1.11 *Embryonic Stem Cells - Cultivation and Epigenetic Constitution*

After fusion of sperm and oocyte, subsequent cell division eventually gives rise to the structure of the blastocyst which contains the inner cell mass comprised of cells. ES cells resemble the basic building block for each arising multicellular organism. Applying the appropriate conditions, ES cells can be isolated and cultivated under a proliferating state *in vitro*, in such way that they retain their pluripotent capacity [149, 150]. In epigenetics, ES cells are widely used for the investigation of mechanistic processes related to the early embryonic development. In the present thesis, distinct mouse ES cell lines are used to investigate the role of DNA modifying enzymes. Therefore, the following section will provide an overview of common culture conditions for ES cells and the corresponding epigenetic characteristics.

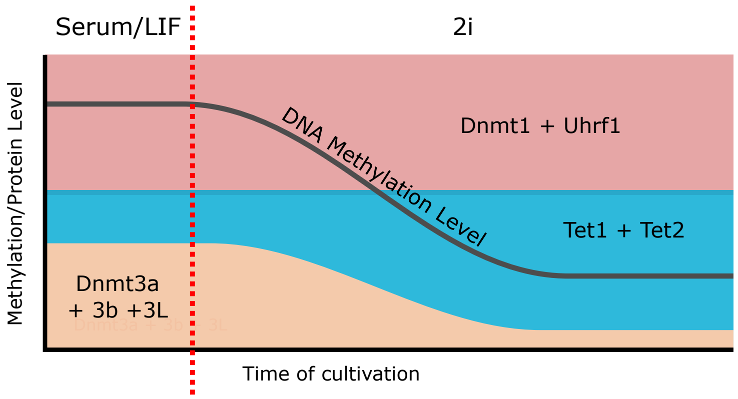


Fig. 1.9: Schematic display of DNA demethylation and expression of Dnmts (Dnmt3a/3b = orange, Dnmt1/Uhrf1 = red) and Tets (blue) during ES cell cultivation on Serum/LIF and subsequent transfer into 2i medium; red line indicates the transition from Serum/LIF to 2i medium.

Traditionally, ES cells were cultured on feeder cells, mitotically inactivated mouse fibroblasts (MFs), together with fetal calf serum (FCS), still a commonly used strategy. FCS and the excrete additives from MFs promote ES cell proliferation and sustain the undifferentiated phenotype [151, 152]. A more recent protocol uses a combination of the leukaemia inhibiting factor (LIF) and FCS for feeder cell free cultivation of ES cells on gelatin coated plates and has emerged as the most commonly used cultivation system [153, 154]. While the presence of LIF suppresses differentiation of ES cells, the multifactorial composition of FCS releases also some pro-differentiating signals [155]. Thus, these cells are considered as developmentally 'primed' ES cells. Epigenetically, these cells are described by a hypermethylated phenotype as a result of strong expression of the *de novo* methyltransferases Dnmt3a and 3b, as well as their co-factor Dnmt3l [156, 157]. However, the cells also display a considerable level of Tet1 and Tet2. A recent publication suggests that this combination of Dnmt and Tet expression leads to a oscillation methylation at particular genomic regions [158]. Likely as a consequence, primed ES cells show a mosaic expression of pluripotency genes such as Nanog [159]. ES cells from the inner cell mass posses a much lower level of DNA methylation, which shows that the hypermethylation is a side effect of the applied cultivation protocol.

A novel protocol for the cultivation of ES cells avoids the use of FCS and instead applies two inhibitors (2i medium), PD0325901 and CHIR99021, which target the mitogen-activated protein kinase (Mek) and the glycogen synthase kinase 3 (Gsk3) [160]. Inhibition of both kinases shields the cells from differentiation signals and introduces a naive stem cell state. Compared to Serum/LIF cultivated cells, naive ES cells show considerably lower expression of Dnmt3a and 3b on transcriptional, as well as on the protein level [156]. At least on the transcriptional level, this is also true for Dnmt3l. Moreover, a recent publication suggests reduced targeting of the maintenance complex Uhrf1 and Dnmt1 due to reduction in H3K9me3 [161]Consequently, 2i cultivated ES cells display a much lower genome wide methylation level, which matches more closely the one of ES cells from inner cell mass and are furthermore defined by a much more uniform expression of pluripotency factors [156, 157, 161].

ES cells can be transferred from Serum/LIF to 2i medium and *vice versa*. After transfer, the cells will adapt their gene expression and methylation level based on the environmental conditions. Figure 1.9 provides a schematic overview of the epigenetic changes during the adjustment of ES cells towards 2i medium after long term cultivation under Serum/LIF conditions. The Serum-to-2i shift is widely accepted as a model system to study DNA demethylation and is also used in parts of the here presented studies [156, 157, 161].

1.12 Hidden Markov Models

Markov models represent probabilistic models describing a sequence of stochastic events, in which the probability of a future event only depends on the previous realised states [162]. Figure 1.10-A provides a simple example with two distinct states. Over time, a series of events is imaginable, in which states might change with the probability (m_X or m_Y), or remain unchanged (1-m_X or 1-m_Y). However, in a hidden Markov model (HMM), states which are aimed to be described cannot be directly observed, i.e. remain hidden. Instead, these 'hidden states' are indirectly determined by using events that are observable (observable states) and are connected to the underlying hidden states (Figure 1.10-B).

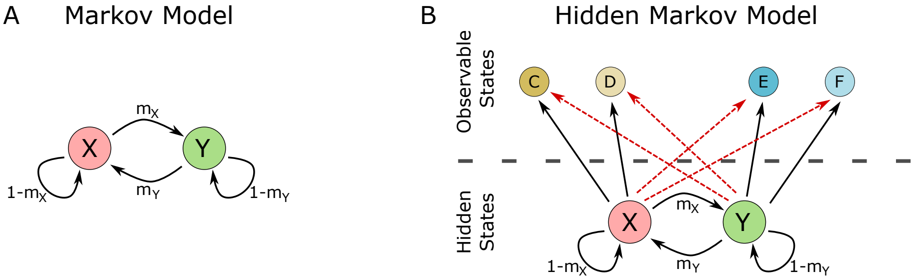


Fig. 1.10: (A) Example of a simple Markov model with two observable states and (B) hidden Markov model with four observable (C, D, E, F) and two hidden states (X, Y). Arrows indicate possible transitions between the distinct states, m_X = probability to move from state X to Y, 1-m_X = probability to remain within state X, m_Y = probability to move from state Y to X, 1-m_Y = probability to remain within state Y. Arrows connecting hidden and observable states represent emission probabilities.

HMMs are used in many research areas and have also emerged as a powerful tool in computational biology. Application of HMM to biological problems includes for example gene prediction, pairwise and multiple alignment, prediction of protein secondary structures, as well as annotation of non-coding RNA [163, 164, 165, 166, 167]. In the present studies, HMMs are used to describe the evolution of DNA methylation patterns over time. The hidden states in this particular case correspond to the 'real' methylation state of a given CpG position (e.g.: CpG or 5mCpG), while the observable states are given by sequencing of bisulfite converted DNA (TpG or CpG) Chapter 3, Figure 3.4. Moreover, the transition probability between the hidden states are, amongst others, determined by methylation events such as *de novo* or maintenance methylation. A detailed description of the underlying HMMs used in presented studies can be found in Chapter 3 and Chapter 6.

1.13 Aim

DNA methylation can be robustly inherited across many cell generations. Nevertheless, removal of cytosine methylation during early development and cell differentiation is key to the formation of pluripotent ES cells, as well as somatic cell types. While DNA methylation is established and maintained by Dnmts, many proposed demethylation pathways are defined by the involvement of Tets and oxidised cytosine forms such as 5hmC. However, many of the underlying mechanisms of DNA demethylation, as well as a potential reciprocal influence of Dnmts and Tets,still remain elusive. The here summarised studies aim to address mechanisms involved in the preservation and removal of DNA methylation.

One focus is placed on the contribution of 5hmC and Tet enzymes during genome wide DNA methylation in mouse ES cells. Hence, this work investigates on the one hand, if or rather to what extent the presence of 5hmC might facilitate a replication dependent loss of 5mC by blocking maintenance methylation activity. Secondly, by comparing WT and Tet KO systems, a possible direct impact of Tets on Dnmt methylation activity is examined. The second aspect of this work focuses on impact of existing methylation pattern on Dnmt activity. More precisely, the study seeks to answer the question of how the methylation status of neighbouring CpGs affect Dnmt activity at a given CpG position.

In order to address these questions, novel sequencing techniques have been extended and newly developed, which allow to determine the strand specific distribution of 5mC, as well as 5hmC from individual DNA molecules. This includes local deep sequencing applications, but also genome wide strand specific sequencing. The obtained data was then subjected to stochastic hidden Markov models, which have been developed by the department of Modelling and Simulation from the Saarland Univiersity, to determine enzymatic efficiency of Dnmts and Tets.

Bibliography

[1] GY Hadlaczky, M Went, and NR Ringertz. Direct evidence for the non-random localization of mammalian chromosomes in the interphase nucleus. *Experimental cell research*, 167(1):1–15, 1986.

[2] Graham A Bentley, Anita Lewit-Bentley, John T Finch, Alberto D Podjarny, and M Roth. Crystal structure of the nucleosome core particle at 16 å resolution. *Journal of molecular biology*, 176(1):55–75, 1984.

[3] Karolin Luger, Armin W Mäder, Robin K Richmond, David F Sargent, and Timothy J Richmond. Crystal structure of the nucleosome core particle at 2.8 å resolution. *Nature*, 389(6648):251, 1997.

[4] Roger D Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–871, 1974.

[5] Tony Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[6] Michael Shogren-Knaak, Haruhiko Ishii, Jian-Min Sun, Michael J Pazin, James R Davie, and Craig L Peterson. Histone h4-k16 acetylation controls chromatin structure and protein interactions. *Science*, 311(5762):844–847, 2006.

[7] Asifa Akhtar and Peter B Becker. Activation of transcription through histone h4 acetylation by mof, an acetyltransferase essential for dosage compensation in drosophila. *Molecular cell*, 5(2):367–375, 2000.

[8] Joanna Wysocka, Tomek Swigut, Hua Xiao, Thomas A Milne, So Yeon Kwon, Joe Landry, Monika Kauer, Alan J Tackett, Brian T Chait, Paul Badenhorst, et al. A phd finger of nurf couples histone h3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098):86, 2006.

[9] M Vettese-Dadey, PA Grant, TR Hebbes, C Crane-Robinson, CD Allis, and JL Workman. Acetylation of histone h4 plays a primary role in enhancing transcription factor binding to nucleosomal dna in vitro. *The EMBO journal*, 15(10):2508–2518, 1996.

[10] Bassem Al-Sady, Hiten D Madhani, and Geeta J Narlikar. Division of labor between the chromodomains of hp1 and suv39 methylase enables coordination of heterochromatin spread. *Molecular cell*, 51(1):80–91, 2013.

[11] Heribert Talasz, Herbert H Lindner, Bettina Sarg, and Wilfried Helliger. Histone h4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. *Journal of Biological Chemistry*, 280(46):38814–38822, 2005.

[12] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.

[13] Peter L Jones, Gert C Jan Veenstra, Paul A Wade, Danielle Vermaak, Stefan U Kass, Nicoletta Landsberger, John Strouboulis, and Alan P Wolffe. Methylated dna and mecp2 recruit histone deacetylase to repress transcription. *Nature genetics*, 19(2):187, 1998.

[14] Bernhard Lehnertz, Yoshihide Ueda, Alwin AHA Derijck, Ulrich Braunschweig, Laura Perez-Burgos, Stefan Kubicek, Taiping Chen, En Li, Thomas Jenuwein, and Antoine HFM Peters. Suv39h-mediated histone h3 lysine 9 methylation directs dna methylation to major satellite repeats at pericentric heterochromatin. *Current Biology*, 13(14):1192–1200, 2003.

[15] Silvina Epsztejn-Litman, Nirit Feldman, Monther Abu-Remaileh, Yoel Shufaro, Ariela Gerson, Jun Ueda, Rachel Deplus, François Fuks, Yoichi Shinkai, Howard Cedar, et al. De novo dna methylation promoted by g9a prevents reprogramming of embryonically silenced genes. *Nature structural & molecular biology*, 15(11):1176, 2008.

[16] Cindy Yen Okitsu and Chih-Lin Hsieh. Dna methylation dictates histone h3k4 methylation. *Molecular and cellular biology*, 27(7):2746–2757, 2007.

[17] Gilbert Walter. Origin of life: the rna world. *Nature*, 319(20):618, 1986.

[18] George W Beadle and Edward L Tatum. Genetic control of biochemical reactions in neurospora. *proceedings of the National Academy of Sciences*, 27(11):499–506, 1941.

[19] Sydney Brenner, François Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, 1961.

[20] Kevin V Morris, Simon W-L Chan, Steven E Jacobsen, and David J Looney. Small interfering rna-induced transcriptional gene silencing in human cells. *Science*, 305(5688):1289–1292, 2004.

[21] Daniel H Kim, Pål Sætrom, Ola Snøve, and John J Rossi. Microrna-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences*, 2008.

[22] Alexei Aravin, Dimos Gaidatzis, Sébastien Pfeffer, Mariana Lagos-Quintana, Pablo Landgraf, Nicola Iovino, Patricia Morris, Michael J Brownstein, Satomi Kuramochi-Miyagawa, Toru Nakano, et al. A novel class of small rnas bind to mili protein in mouse testes. *Nature*, 442(7099):203, 2006.

[23] Angélique Girard, Ravi Sachidanandam, Gregory J Hannon, and Michelle A Carmell. A germline-specific class of small rnas binds mammalian piwi proteins. *Nature*, 442(7099):199, 2006.

[24] Ryan J Taft, Cas Simons, Satu Nahkuri, Harald Oey, Darren J Korbie, Timothy R Mercer, Jeff Holst, William Ritchie, Justin JL Wong, John EJ Rasko, et al. Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nature structural & molecular biology*, 17(8):1030, 2010.

[25] Charlotte Grimaud, Frédéric Bantignies, Manika Pal-Bhadra, Pallavi Ghana, Utpal Bhadra, and Giacomo Cavalli. Rnai components are required for nuclear clustering of polycomb group response elements. *Cell*, 124(5):957–971, 2006.

[26] Neil Brockdorff, Alan Ashworth, Graham F Kay, Veronica M McCabe, Dominic P Norris, Penny J Cooper, Sally Swift, and Sohaila Rastan. The product of the mouse xist gene is a 15 kb inactive x-specific transcript containing no conserved orf and located in the nucleus. *Cell*, 71(3):515–526, 1992.

[27] Marcel E Dinger, Paulo P Amaral, Tim R Mercer, Ken C Pang, Stephen J Bruce, Brooke B Gardiner, Marjan E Askarian-Amiri, Kelin Ru, Giulia Soldà, Cas Simons, et al. Long noncoding rnas in mouse embryonic stem cell pluripotency and differentiation. *Genome research*, pages gr–078378, 2008.

[28] Takashi Nagano, Jennifer A Mitchell, Lionel A Sanz, Florian M Pauler, Anne C Ferguson-Smith, Robert Feil, and Peter Fraser. The air noncoding rna epigenetically silences transcription by targeting g9a to chromatin. *Science*, 322(5908):1717–1720, 2008.

[29] Faizaan Mohammad, Tanmoy Mondal, Natalia Guseva, Gaurav Kumar Pandey, and Chandrasekhar Kanduri. Kcnq1ot1 noncoding rna mediates transcriptional gene silencing by interacting with dnmt1. *Development*, pages dev–048181, 2010.

[30] En Li, Timothy H Bestor, and Rudolf Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

[31] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[32] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96, 2004.

[33] Suhua Feng, Shawn J Cokus, Xiaoyu Zhang, Pao-Yang Chen, Magnolia Bostick, Mary G Goll, Jonathan Hetzel, Jayati Jain, Steven H Strauss, Marnie E Halpern, et al. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences*, 107(19):8689–8694, 2010.

[34] Assaf Zemach, Ivy E McDaniel, Pedro Silva, and Daniel Zilberman. Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, 328(5980):916–919, 2010.

[35] Melanie Ehrlich, Miguel A Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C Kuo, Roy A McCune, and Charles Gehrke. Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research*, 10(8):2709–2721, 1982.

[36] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.

[37] Adrian P Bird. Dna methylation and the frequency of cpg in animal dna. *Nucleic acids research*, 8(7):1499–1504, 1980.

[38] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando J Miller, and Donald Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40(1):91–99, 1985.

[39] Frank Lyko. The dna methyltransferase family: a versatile toolkit for epigenetic regulation. *Nature Reviews Genetics*, 19(2):81, 2018.

[40] János Pósfai, Ashok S Bhagwat, György Pósfai, and Richard J Roberts. Predictive motifs derived from cytosine methyltransferases. *Nucleic acids research*, 17(7):2421–2435, 1989.

[41] Timothy H Bestor and Vernon M Ingram. Two dna methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with dna. *Proceedings of the National Academy of Sciences*, 80(18):5559–5563, 1983.

[42] Timothy Bestor, Andrew Laudano, Robert Mattaliano, and Vernon Ingram. Cloning and sequencing of a cdna encoding dna methyltransferase of mouse cells: the carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of molecular biology*, 203(4):971–983, 1988.

[43] Masaki Okano, Shaoping Xie, and En Li. Dnmt2 is not required for de novo and maintenance methylation of viral dna in embryonic stem cells. *Nucleic acids research*, 26(11):2536–2540, 1998.

[44] Ulla Aapola, Kazunori Shibuya, Hamish S Scott, Juha Ollila, Mauno Vihinen, Maarit Heino, Ai Shintani, Kazuhiko Kawasaki, Shinsei Minoshima, Kai Krohn, et al. Isolation and initial characterization of a novel zinc finger gene, dnmt3l, on 21q22. 3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics*, 65(3):293–298, 2000.

[45] Mary Grace Goll, Finn Kirpekar, Keith A Maggert, Jeffrey A Yoder, Chih-Lin Hsieh, Xiaoyu Zhang, Kent G Golic, Steven E Jacobsen, and Timothy H Bestor. Methylation of trnaasp by the dna methyltransferase homolog dnmt2. *Science*, 311(5759):395–398, 2006.

[46] Matthias Schaefer, Tim Pollex, Katharina Hanna, Francesca Tuorto, Madeleine Meusburger, Mark Helm, and Frank Lyko. Rna methylation by dnmt2 protects transfer rnas against stress-induced cleavage. *Genes & development*, 24(15):1590–1595, 2010.

[47] Déborah Bourc'his, Guo-Liang Xu, Chyuan-Sheng Lin, Brooke Bollman, and Timothy H Bestor. Dnmt3l and the establishment of maternal genomic imprints. *Science*, 294(5551):2536–2539, 2001.

[48] Da Jia, Renata Z Jurkowska, Xing Zhang, Albert Jeltsch, and Xiaodong Cheng. Structure of dnmt3a bound to dnmt3l suggests a model for de novo dna methylation. *Nature*, 449(7159):248, 2007.

[49] Joan Barau, Aurélie Teissandier, Natasha Zamudio, Stéphanie Roy, Valérie Nalesso, Yann Hérault, Florian Guillou, and Déborah Bourc'his. The dna methyltransferase dnmt3c protects male germ cells from transposon activity. *Science*, 354(6314):909–912, 2016.

[50] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 2004.

[51] Tetsuo Iida, Isao Suetake, Shoji Tajima, Hiroshi Morioka, Satoshi Ohta, Chikashi Obuse, and Toshiki Tsurimoto. Pcna clamp facilitates action of dna cytosine methyltransferase 1 on hemimethylated dna. *Genes to cells*, 7(10):997–1007, 2002.

[52] Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.

[53] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908, 2007.

[54] Kyohei Arita, Mariko Ariyoshi, Hidehito Tochio, Yusuke Nakamura, and Masahiro Shirakawa. Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. *Nature*, 455(7214):818, 2008.

[55] Chengmin Qian, Side Li, Jean Jakoncic, Lei Zeng, Martin J Walsh, and Ming-Ming Zhou. Structure and hemimethylated cpg binding of the sra domain from human uhrf1. *Journal of Biological Chemistry*, 2008.

[56] Giedrius Vilkaitis, Isao Suetake, Saulius Klimašauskas, and Shoji Tajima. Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase. *Journal of Biological Chemistry*, 280(1):64–72, 2005.

[57] Rachna Goyal, Richard Reinhardt, and Albert Jeltsch. Accuracy of dna methylation pattern preservation by the dnmt1 methyltransferase. *Nucleic acids research*, 34(4):1182–1188, 2006.

[58] Jing Liao, Rahul Karnik, Hongcang Gu, Michael J Ziller, Kendell Clement, Alexander M Tsankov, Veronika Akopian, Casey A Gifford, Julie Donaghey, Christina Galonska, et al. Targeted disruption of dnmt1, dnmt3a and dnmt3b in human embryonic stem cells. *Nature genetics*, 47(5):469, 2015.

[59] Michael R Rountree, Kurtis E Bachman, and Stephen B Baylin. Dnmt1 binds hdac2 and a new co-repressor, dmap1, to form a complex at replication foci. *Nature genetics*, 25(3):269, 2000.

[60] Heinrich Leonhardt, Andrea W Page, Heinz-Ulrich Weier, and Timothy H Bestor. A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei. *Cell*, 71(5):865–873, 1992.

[61] Jikui Song, Olga Rechkoblit, Timothy H Bestor, and Dinshaw J Patel. Structure of dnmt1-dna complex reveals a role for autoinhibition in maintenance dna methylation. *Science*, 331(6020):1036–1040, 2011.

[62] Pavel Bashtrykov, Gytis Jankevicius, Anita Smarandache, Renata Z Jurkowska, Sergey Ragozin, and Albert Jeltsch. Specificity of dnmt1 for methylation of hemimethylated cpg sites resides in its catalytic domain. *Chemistry & biology*, 19(5):572–578, 2012.

[63] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, 2000.

[64] Shin-ichi Tomizawa, Hisato Kobayashi, Toshiaki Watanabe, Simon Andrews, Kenichiro Hata, Gavin Kelsey, and Hiroyuki Sasaki. Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-cpg methylation in oocytes. *Development*, pages dev–061416, 2011.

[65] Ryan Lister, Eran A Mukamel, Joseph R Nery, Mark Urich, Clare A Puddifoot, Nicholas D Johnson, Jacinta Lucero, Yun Huang, Andrew J Dwork, Matthew D Schultz, et al. Global epigenomic reconfiguration during mammalian brain development. *Science*, 341(6146):1237905, 2013.

[66] Kenjiro Shirane, Hidehiro Toh, Hisato Kobayashi, Fumihito Miura, Hatsune Chiba, Takashi Ito, Tomohiro Kono, and Hiroyuki Sasaki. Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-cpg methylation and role of dna methyltransferases. *PLoS genetics*, 9(4):e1003439, 2013.

[67] Romain Barrès, Megan E Osler, Jie Yan, Anna Rune, Tomas Fritz, Kenneth Caidahl, Anna Krook, and Juleen R Zierath. Non-cpg methylation of the pgc-1α promoter through dnmt3b controls mitochondrial density. *Cell metabolism*, 10(3):189–198, 2009.

[68] Lin Chen, Kaifu Chen, Laura A Lavery, Steven Andrew Baker, Chad A Shaw, Wei Li, and Huda Y Zoghbi. Mecp2 binds to non-cg methylated dna as neurons mature, influencing transcription and the timing of onset for rett syndrome. *Proceedings of the National Academy of Sciences*, page 201505909, 2015.

[69] Donghong Zhang, Bingruo Wu, Ping Wang, Yidong Wang, Pengfei Lu, Tamilla Nechiporuk, Thomas Floss, John M Greally, Deyou Zheng, and Bin Zhou. Non-

cpg methylation by dnmt3b facilitates rest binding and gene silencing in developing mouse hearts. *Nucleic acids research*, 45(6):3102–3115, 2016.

[70] Christopher L Keown, Joel B Berletch, Rosa Castanon, Joseph R Nery, Christine M Disteche, Joseph R Ecker, and Eran A Mukamel. Allele-specific non-cg dna methylation marks domains of active chromatin in female mouse brain. *Proceedings of the National Academy of Sciences*, 114(14):E2882–E2890, 2017.

[71] Taiping Chen, Yoshihide Ueda, Jonathan E Dodge, Zhenjuan Wang, and En Li. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by dnmt3a and dnmt3b. *Molecular and cellular biology*, 23(16):5594–5605, 2003.

[72] Masahiro Kaneda, Masaki Okano, Kenichiro Hata, Takashi Sado, Naomi Tsujimoto, En Li, and Hiroyuki Sasaki. Essential role for de novo dna methyltransferase dnmt3a in paternal and maternal imprinting. *nature*, 429(6994):900, 2004.

[73] Taiping Chen, Naomi Tsujimoto, and En Li. The pwwp domain of dnmt3a and dnmt3b is required for directing dna methylation to the major satellite repeats at pericentric heterochromatin. *Molecular and cellular biology*, 24(20):9048–9058, 2004.

[74] Marco Morselli, William A Pastor, Barbara Montanini, Kevin Nee, Roberto Ferrari, Kai Fu, Giancarlo Bonora, Liudmilla Rubbi, Amander T Clark, Simone Ottonello, et al. In vivo targeting of de novo dna methylation by histone modifications in yeast and mouse. *Elife*, 4:e06205, 2015.

[75] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520(7546):243, 2015.

[76] Zhenggang Xiong and Peter W Laird. Cobra: a sensitive and quantitative dna methylation assay. *Nucleic acids research*, 25(12):2532–2534, 1997.

[77] James G Herman, Jeremy R Graff, SBDN Myöhänen, Barry D Nelkin, and Stephen B Baylin. Methylation-specific pcr: a novel pcr assay for methylation status of cpg islands. *Proceedings of the national academy of sciences*, 93(18):9821–9826, 1996.

[78] Christopher C Oakes, Sophie La Salle, Bernard Robaire, and Jacquetta M Trasler. Evaluation of a quantitative dna methylation analysis technique using methylation-sensitive/dependent restriction enzymes and real-time pcr. *Epigenetics*, 1(3):146–152, 2006.

[79] Ko Hashimoto, Shoichi Kokubun, Eiji Itoi, and Helmtrud I Roach. Improved quantification of dna methylation using methylation-sensitive restriction enzymes and real-time pcr. *Epigenetics*, 2(2):86–91, 2007.

[80] Chhaya W Achwal and H Sharat Chandra. A sensitive immunochemical method for detecting 5mc in dna fragments. *FEBS letters*, 150(2):469–472, 1982.

[81] Filipe V Jacinto, Esteban Ballestar, and Manel Esteller. Methyl-dna immunoprecipitation (medip): hunting down the dna methylome. *Biotechniques*, 44(1):35–43, 2008.

[82] Oluwatosin Taiwo, Gareth A Wilson, Tiffany Morris, Stefanie Seisenberger, Wolf Reik, Daniel Pearce, Stephan Beck, and Lee M Butcher. Methylome analysis using medip-seq with low dna concentrations. *Nature protocols*, 7(4):617, 2012.

[83] Marianne Frommer, Louise E McDonald, Douglas S Millar, Christina M Collis, Fujiko Watt, Geoffrey W Grigg, Peter L Molloy, and Cheryl L Paul. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual dna strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831, 1992.

[84] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[85] Brooks E Miner, Reinhard J Stöger, Alice F Burden, Charles D Laird, and R Scott Hansen. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite pcr. *Nucleic acids research*, 32(17):e135–e135, 2004.

[86] Lei Zhao, Ming-an Sun, Zejuan Li, Xue Bai, Miao Yu, Min Wang, Liji Liang, Xiaojian Shao, Stephen Arnovitz, Qianfei Wang, et al. The dynamics of dna methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome research*, pages gr–163147, 2014.

[87] Lakshminarayan M Iyer, Mamta Tahiliani, Anjana Rao, and L Aravind. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell cycle*, 8(11):1698–1710, 2009.

[88] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[89] Ryoichi Ono, Tomohiko Taki, Takeshi Taketani, Masafumi Taniwaki, Hajime Kobayashi, and Yasuhide Hayashi. Lcx, leukemia-associated protein with a cxxc domain, is fused to mll in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). *Cancer research*, 62(14):4075–4080, 2002.

[90] RB Lorsbach, J Moore, S Mathew, SC Raimondi, ST Mukatira, and JR Downing. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia*, 17(3):637, 2003.

[91] Lakshminarayan M Iyer, Vivek Anantharaman, Maxim Y Wolf, and L Aravind. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *International journal for parasitology*, 38(1):1–31, 2008.

[92] Christoph Loenarz and Christopher J Schofield. Expanding chemical biology of 2-oxoglutarate oxygenases. *Nature chemical biology*, 4(3):152, 2008.

[93] Christoph Loenarz and Christopher J Schofield. Physiological and biochemical aspects of hydroxylations and demethylations catalyzed by human 2-oxoglutarate oxygenases. *Trends in biochemical sciences*, 36(1):7–18, 2011.

[94] Carina Frauer, Andrea Rottach, Daniela Meilinger, Sebastian Bultmann, Karin Fellinger, Stefan Hasenöder, Mengxi Wang, Weihua Qin, Johannes Söding, Fabio Spada, et al. Different binding properties and function of cxxc zinc finger domains in dnmt1 and tet1. *PloS one*, 6(2):e16627, 2011.

[95] Haikuo Zhang, Xin Zhang, Erin Clark, Michelle Mulcahey, Stephen Huang, and Yujiang Geno Shi. Tet1 is a dna-binding protein that modulates dna methylation and gene transcription via hydroxylation of 5-methylcytosine. *Cell research*, 20(12):1390, 2010.

[96] Yufei Xu, Feizhen Wu, Li Tan, Lingchun Kong, Lijun Xiong, Jie Deng, Andrew J Barbera, Lijuan Zheng, Haikuo Zhang, Stephen Huang, et al. Genome-wide regulation of 5hmc, 5mc, and gene expression by tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell*, 42(4):451–464, 2011.

[97] Lakshminarayan M Iyer, Saraswathi Abhiman, and L Aravind. Natural history of eukaryotic dna methylation systems. In *Progress in molecular biology and translational science*, volume 101, pages 25–104. Elsevier, 2011.

[98] Myunggon Ko, Jungeun An, Hozefa S Bandukwala, Lukas Chavez, Tarmo Äijö, William A Pastor, Matthew F Segal, Huiming Li, Kian Peng Koh, Harri Lähdesmäki, et al. Modulation of tet2 expression and 5-methylcytosine oxidation by the cxxc domain protein idax. *Nature*, 497(7447):122, 2013.

[99] Yufei Xu, Chao Xu, Akiko Kato, Wolfram Tempel, Jose Garcia Abreu, Chuanbing Bian, Yeguang Hu, Di Hu, Bin Zhao, Tanja Cerovina, et al. Tet3 cxxc domain and dioxygenase activity cooperatively regulate key genes for xenopus eye and neural development. *Cell*, 151(6):1200–1213, 2012.

[100] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[101] Petra Hajkova, Sean J Jeffries, Caroline Lee, Nigel Miller, Stephen P Jackson, and M Azim Surani. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science*, 329(5987):78–82, 2010.

[102] Shinsuke Ito, Ana C D'alessio, Olena V Taranova, Kwonho Hong, Lawrence C Sowers, and Yi Zhang. Role of tet proteins in 5mc to 5hmc conversion, es-cell self-renewal and inner cell mass specification. *nature*, 466(7310):1129, 2010.

[103] Shinpei Yamaguchi, Kwonho Hong, Rui Liu, Azusa Inoue, Li Shen, Kun Zhang, and Yi Zhang. Dynamics of 5-methylcytosine and 5-hydroxymethylcytosine during germ cell reprogramming. *Cell research*, 23(3):329, 2013.

[104] Myunggon Ko, Hozefa S Bandukwala, Jungeun An, Edward D Lamperti, Elizabeth C Thompson, Ryan Hastie, Angeliki Tsangaratou, Klaus Rajewsky, Sergei B Koralov, and Anjana Rao. Ten-eleven-translocation 2 (tet2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. *Proceedings of the National Academy of Sciences*, page 201112317, 2011.

[105] Huimei Yu, Yijing Su, Jaehoon Shin, Chun Zhong, Junjie U Guo, Yi-Lan Weng, Fuying Gao, Daniel H Geschwind, Giovanni Coppola, Guo-li Ming, et al. Tet3 regulates synaptic transmission and homeostatic plasticity via dna oxidation and repair. *Nature neuroscience*, 18(6):836, 2015.

[106] Tian-Peng Gu, Fan Guo, Hui Yang, Hai-Ping Wu, Gui-Fang Xu, Wei Liu, Zhi-Guo Xie, Linyu Shi, Xinyi He, Seung-gi Jin, et al. The role of tet3 dna dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366):606, 2011.

[107] Mark Wossidlo, Toshinobu Nakamura, Konstantin Lepikhov, C Joana Marques, Valeri Zakhartchenko, Michele Boiani, Julia Arand, Toru Nakano, Wolf Reik, and Jörn Walter. 5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications*, 2:241, 2011.

[108] Aleksandra Szwagierczak, Sebastian Bultmann, Christine S Schmidt, Fabio Spada, and Heinrich Leonhardt. Sensitive enzymatic quantification of 5-

hydroxymethylcytosine in genomic dna. *Nucleic acids research*, 38(19):e181–e181, 2010.

[109] Myunggon Ko, Yun Huang, Anna M Jankowska, Utz J Pape, Mamta Tahiliani, Hozefa S Bandukwala, Jungeun An, Edward D Lamperti, Kian Peng Koh, Rebecca Ganetzky, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant tet2. *Nature*, 468(7325):839, 2010.

[110] SARAH K Howlett and WOLF Reik. Methylation levels of maternal and paternal genomes during preimplantation development. *Development*, 113(1):119–127, 1991.

[111] J Oswald, S Engemann, N Lane, W Mayer, A Olek, R Fundele, W Dean, W Reik, and J Walter. Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478, 2000.

[112] Stefan Wallner, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, Daniela Beißer, Sven Rahmann, Andreas S Richter, Thomas Manke, Ulrike Bönisch, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics & chromatin*, 9(1):33, 2016.

[113] Aras Toker, Dirk Engelbert, Garima Garg, Julia K Polansky, Stefan Floess, Takahisa Miyao, Udo Baron, Sandra Düber, Robert Geffers, Pascal Giehr, et al. Active demethylation of the foxp3 locus leads to the generation of stable regulatory t cells within the thymus. *The Journal of Immunology*, page 1203473, 2013.

[114] Vincenzo Calvanese, Agustín F Fernández, Rocío G Urdinguio, Beatriz Suarez-Alvarez, Cristina Mangas, Vicente Pérez-García, Clara Bueno, Rosa Montes, Verónica Ramos-Mejía, Pablo Martínez-Camblor, et al. A promoter dna demethylation landscape of human hematopoietic differentiation. *Nucleic acids research*, 40(1):116–131, 2011.

[115] Junjie U Guo, Yijing Su, Chun Zhong, Guo-li Ming, and Hongjun Song. Hydroxylation of 5-methylcytosine by tet1 promotes active dna demethylation in the adult brain. *Cell*, 145(3):423–434, 2011.

[116] M Cristina Cardoso and Heinrich Leonhardt. Dna methyltransferase is actively retained in the cytoplasm during early development. *The Journal of cell biology*, 147(1):25–32, 1999.

[117] Sarayu Ratnam, Carmen Mertineit, Feng Ding, Carina Y Howell, Hugh J Clarke, Timothy H Bestor, J Richard Chaillet, and Jacquetta M Trasler. Dynamics of dnmt1 methyltransferase expression and intracellular localization during oogenesis and preimplantation development. *Developmental biology*, 245(2):304–314, 2002.

[118] Maik Grohmann, Fabio Spada, Lothar Schermelleh, Natalia Alenina, Michael Bader, M Cristina Cardoso, and Heinrich Leonhardt. Restricted mobility of dnmt1 in preimplantation embryos: implications for epigenetic reprogramming. *BMC developmental biology*, 5(1):18, 2005.

[119] Ryutaro Hirasawa, Hatsune Chiba, Masahiro Kaneda, Shoji Tajima, En Li, Rudolf Jaenisch, and Hiroyuki Sasaki. Maternal and zygotic dnmt1 are necessary and sufficient for the maintenance of dna methylation imprints during preimplantation development. *Genes & development*, 22(12):1607–1616, 2008.

[120] Victoria Valinluck, Hsin-Hao Tsai, Daniel K Rogstad, Artur Burdzy, Adrian Bird, and Lawrence C Sowers. Oxidative damage to methyl-cpg sequences inhibits the binding of the methyl-cpg binding domain (mbd) of methyl-cpg binding protein 2 (mecp2). *Nucleic acids research*, 32(14):4100–4108, 2004.

[121] Hideharu Hashimoto, Yiwei Liu, Anup K Upadhyay, Yanqi Chang, Shelley B Howerton, Paula M Vertino, Xing Zhang, and Xiaodong Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849, 2012.

[122] Debin Ji, Krystal Lin, Jikui Song, and Yinsheng Wang. Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1-and DNMT3a-mediated cytosine methylation. *Molecular BioSystems*, 10(7):1749–1752, 2014.

[123] Mary Gehring, Jin Hoe Huh, Tzung-Fu Hsieh, Jon Penterman, Yeonhee Choi, John J Harada, Robert B Goldberg, and Robert L Fischer. Demeter dna glycosylase establishes medea polycomb gene self-imprinting by allele-specific demethylation. *cell*, 124(3):495–506, 2006.

[124] Teresa Morales-Ruiz, Ana Pilar Ortega-Galisteo, María Isabel Ponferrada-Marín, María Isabel Martínez-Macías, Rafael R Ariza, and Teresa Roldán-Arjona. Demeter and repressor of silencing 1 encode 5-methylcytosine dna glycosylases. *Proceedings of the National Academy of Sciences*, 103(18):6853–6858, 2006.

[125] Chun-Chang Chen, Keh-Yang Wang, and Che-Kun James Shen. The mammalian de novo dna methyltransferases dnmt3a and dnmt3b are also dna 5-hydroxymethylcytosine dehydroxymethylases. *Journal of Biological Chemistry*, 287(40):33116–33121, 2012.

[126] Zita Liutkevičiūtė, Edita Kriukienė, Janina Ličytė, Milda Rudytė, Giedrė Urbanavičiūtė, and Saulius Klimašauskas. Direct decarboxylation of 5-carboxylcytosine by dna c5-methyltransferases. *Journal of the American Chemical Society*, 136(16):5884–5887, 2014.

[127] Zita Liutkevičiūtė, Gražvydas Lukinavičius, Viktoras Masevičius, Dalia Daujotytė, and Saulius Klimašauskas. Cytosine-5-methyltransferases add aldehydes to dna. *Nature chemical biology*, 5(6):400, 2009.

[128] Stefan Schiesser, Benjamin Hackner, Toni Pfaffeneder, Markus Müller, Christian Hagemeier, Matthias Truss, and Thomas Carell. Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angewandte Chemie International Edition*, 51(26):6516–6520, 2012.

[129] Salvatore Cortellino, Jinfei Xu, Mara Sannai, Robert Moore, Elena Caretti, Antonio Cigliano, Madeleine Le Coz, Karthik Devarajan, Andy Wessels, Dianne Soprano, et al. Thymine dna glycosylase is essential for active dna demethylation by linked deamination-base excision repair. *Cell*, 146(1):67–79, 2011.

[130] Christopher S Nabel, Huijue Jia, Yu Ye, Li Shen, Hana L Goldschmidt, James T Stivers, Yi Zhang, and Rahul M Kohli. Aid/apobec deaminases disfavor modified cytosines implicated in dna demethylation. *Nature chemical biology*, 8(9):751, 2012.

[131] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[132] Atanu Maiti and Alexander C Drohat. Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011.

[133] Li Shen, Hao Wu, Dinh Diep, Shinpei Yamaguchi, Ana C D'Alessio, Ho-Lim Fung, Kun Zhang, and Yi Zhang. Genome-wide analysis reveals tet-and tdg-dependent 5-methylcytosine oxidation dynamics. *Cell*, 153(3):692–706, 2013.

[134] Daniel Cortázar, Christophe Kunz, Jim Selfridge, Teresa Lettieri, Yusuke Saito, Eilidh MacDougall, Annika Wirz, David Schuermann, Angelika L Jacobs, Fredy Siegrist, et al. Embryonic lethal phenotype reveals a function of tdg in maintaining epigenetic stability. *Nature*, 470(7334):419, 2011.

[135] Chun-Xiao Song, Keith E Szulwach, Qing Dai, Ye Fu, Shi-Qing Mao, Li Lin, Craig Street, Yujing Li, Mickael Poidevin, Hao Wu, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153(3):678–691, 2013.

[136] Francesco Neri, Danny Incarnato, Anna Krepelova, Stefania Rapelli, Francesca Anselmi, Caterina Parlato, Claudio Medana, Federica Dal Bello, and Salvatore Oliviero. Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter dna methylation dynamics. *Cell reports*, 10(5):674–683, 2015.

[137] Yun Huang, William A Pastor, Yinghua Shen, Mamta Tahiliani, David R Liu, and Anjana Rao. The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PloS one*, 5(1):e8888, 2010.

[138] Hao Wu, Xiaoji Wu, Li Shen, and Yi Zhang. Single-base resolution analysis of active dna demethylation using methylase-assisted bisulfite sequencing. *Nature biotechnology*, 32(12):1231, 2014.

[139] Sascha Tierling, Beate Schmitt, and Jörn Walter. Comprehensive evaluation of commercial bisulfite-based dna methylation kits and development of an alternative protocol with improved conversion performance. *Genetics & epigenetics*, 10:1179237X18766097, 2018.

[140] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, 2012.

[141] Michael J Booth, Tobias WB Ost, Dario Beraldi, Neil M Bell, Miguel R Branco, Wolf Reik, and Shankar Balasubramanian. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature protocols*, 8(10):1841, 2013.

[142] Miao Yu, Gary C Hon, Keith E Szulwach, Chun-Xiao Song, Peng Jin, Bing Ren, and Chuan He. Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nature protocols*, 7(12):2159, 2012.

[143] Emily K Schutsky, Jamie E DeNizio, Peng Hu, Monica Yun Liu, Christopher S Nabel, Emily B Fabyanic, Young Hwang, Frederic D Bushman, Hao Wu, and Rahul M Kohli. Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a dna deaminase. *Nature biotechnology*, 36(11):1083, 2018.

[144] Xingyu Lu, Chun-Xiao Song, Keith Szulwach, Zhipeng Wang, Payton Weidenbacher, Peng Jin, and Chuan He. Chemical modification-assisted bisulfite sequencing (cab-seq) for 5-carboxylcytosine detection in dna. *Journal of the American Chemical Society*, 135(25):9315–9317, 2013.

[145] Xingyu Lu, Dali Han, Boxuan Simen Zhao, Chun-Xiao Song, Li-Sheng Zhang, Louis C Doré, and Chuan He. Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide dna demethylation dynamics. *Cell research*, 25(3):386, 2015.

[146] Jiwook Shim, Gwendolyn I Humphreys, Bala Murali Venkatesan, Jan Marie Munz, Xueqing Zou, Chaitanya Sathe, Klaus Schulten, Farhad Kosari, Ann M Nardulli, George Vasmatzis, et al. Detection and quantification of methylation in dna using solid-state nanopores. *Scientific reports*, 3:1389, 2013.

[147] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[148] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of dna methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6):461, 2010.

[149] Gail R Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638, 1981.

[150] Martin J Evans and Matthew H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *nature*, 292(5819):154, 1981.

[151] GL Meng, NI Zur Nieden, SY Liu, JT Cormier, MS Kallos, and DE Rancourt. Properties of murine embryonic stem cells maintained on human foreskin fibroblasts without lif. *Molecular Reproduction and Development: Incorporating Gamete Research*, 75(4):614–622, 2008.

[152] Livia Eiselleova, Iveta Peterkova, Jakub Neradil, Iva Slaninova, Ales Hampl, and Petr Dvorak. Comparative study of mouse and human feeder cells for human embryonic stem cells. *International Journal of Developmental Biology*, 52(4):353–363, 2004.

[153] R Lindsay Williams, Douglas J Hilton, Shirley Pease, Tracy A Willson, Colin L Stewart, David P Gearing, Erwin F Wagner, Donald Metcalf, Nicos A Nicola, and Nicholas M Gough. Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature*, 336(6200):684, 1988.

[154] Austin G Smith, John K Heath, Deborah D Donaldson, Gordon G Wong, J Moreau, Mark Stahl, and David Rogers. Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. *Nature*, 336(6200):688, 1988.

[155] Jun Cheng, Amalia Dutra, Aya Takesono, Lisa Garrett-Beal, and Pamela L Schwartzberg. Improved generation of c57bl/6j mouse embryonic stem cells in a defined serum-free media. *genesis*, 39(2):100–104, 2004.

[156] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[157] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[158] Steffen Rulands, Heather J Lee, Stephen J Clark, Christof Angermueller, Sebastien A Smallwood, Felix Krueger, Hisham Mohammed, Wendy Dean, Jennifer Nichols, Peter Rugg-Gunn, et al. Genome-scale oscillations in dna methylation during exit from pluripotency. *bioRxiv*, page 338822, 2018.

[159] Ian Chambers, Jose Silva, Douglas Colby, Jennifer Nichols, Bianca Nijmeijer, Morag Robertson, Jan Vrana, Ken Jones, Lars Grotewold, and Austin Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230, 2007.

[160] Qi-Long Ying, Jason Wray, Jennifer Nichols, Laura Batlle-Morera, Bradley Doble, James Woodgett, Philip Cohen, and Austin Smith. The ground state of embryonic stem cell self-renewal. *Nature*, 453(7194):519, 2008.

[161] Ferdinand von Meyenn, Mario Iurlaro, Ehsan Habibi, Ning Qing Liu, Ali Salehzadeh-Yazdi, Fátima Santos, Edoardo Petrini, Inês Milagre, Miao Yu, Zhenqing Xie, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861, 2016.

[162] AA Markov. Theory of algorithms [translated by jacques j. schorr-kon and pst staff] imprint moscow, academy of sciences of the ussr, 1954 [jerusalem, israel program for scientific translations, 1961; available from office of technical services, united states department of commerce] added tp in russian translation of works of the mathematical institute, academy of sciences of the ussr, v. 42. *Original title: Teoriya algorifmov.[QA248. M2943 Dartmouth College library. US Dept. of Commerce, Office of Technical Services, number OTS 60-51085]*, 1954.

[163] Kasper Munch and Anders Krogh. Automatic generation of gene finders for eukaryotic species. *BMC bioinformatics*, 7(1):263, 2006.

[164] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge university press, 1998.

[165] Lior Pachter, Marina Alexandersson, and Simon Cawley. Applications of generalized pair hidden markov models to alignment and gene finding problems. *Journal of Computational Biology*, 9(2):389–399, 2002.

[166] Kyoung-Jae Won, Thomas Hamelryck, Adam Prügel-Bennett, and Anders Krogh. An evolutionary method for learning hmm structure: prediction of protein secondary structure. *BMC bioinformatics*, 8(1):357, 2007.

[167] Zasha Weinberg and Walter L Ruzzo. Sequence-based heuristics for faster annotation of non-coding rna families. *Bioinformatics*, 22(1):35–39, 2005.

2. HAIRPIN BISULFITE SEUQENCING: METHYLATION ANALYSIS OF BOTH COMPLEMENTARY DNA STRANDS OF ONE INDIVIDUAL CHROMOSOME

The content of Chapter 2 has been published as:

Giehr P.[†], Walter J.[†] (2018) Hairpin Bisulfite Sequencing: Synchronous Methylation Analysis on Complementary DNA Strands of Individual Chromosomes. In: Tost J. (eds) DNA Methylation Protocols. Methods in Molecular Biology, vol 1708. Humana Press, New York, NY.

AUTHOR CONTRIBUTIONS

Pascal Giehr: Authoring of the manuscript including the generation of all figures and tables.

Prof. Dr. Jörn Walter: Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Abstract

The accurate and quantitative detection of 5-methylcytosine is of great importance in the field of epigenetics. The method of choice is usually bisulfite sequencing because of the high resolution and the possibility to combine it with next generation sequencing. Nevertheless, also this method has its limitations. Following the bisulfite treatment DNA strands are no longer complementary such that in a subsequent PCR amplification the DNA methylation patterns information of only one of the two DNA strand is preserved. Several years ago Hairpin Bisulfite sequencing was developed as a method to obtain the pattern information of complementary DNA strands. The method requires fragmentation (usually by enzymatic cleavage) of genomic DNA followed by a covalent linking of both DNA strands through ligation of a short DNA hairpin oligonucleotide to both strands. The

[†] Department of Biological Sciences, Saarland University, D-66123 Saarbrücken, Germany

ligated covalently linked dsDNA products are then subjected to a conventional bisulfite treatment in which all unmodified cytosines are converted to uracil. During the treatment the DNA is denatured forming non-complementary ssDNA circles. These circles serve as a template for a locus specific PCR to amplify chromosomal patterns of the region of interest. As a result one ends up with a linearized product, which contains the methylation information of both complementary DNA strands.

2.1 Indroduction

Hairpin Bisulfite Sequencing (HBS) is a method to detect DNA methylation on both complementary DNA strands in individual DNA molecules [1]. HBS allows to discriminate if both strands are methylated or if a hemimethylation is present in only one of the two complementary DNA strands (upper or lower strand) or if both strands are symmetrically unmethylated. It also allows to discriminate a true non-CpG methylation from a natural polymorphic (mutated) site. A major advantage of hairpin bisulfite sequencing over conventional bisulfite sequencing is when one needs to detect the symmetry of DNA methylation patterns on both DNA strands i.e. when analyzing active demethylation, de novo methylation or maintenance methylation events during cell replication or stages of reprogramming [2, 3, 4].

For the use of the HBS method the following general steps should be considered. A standard HBS approach starts with the digestion of DNA by a defined restriction enzymes (usually 4 base cutter) that is not sensitive to DNA methylation, followed by a covalent linking of the DNA fragments (upper and lower DNA strand) to a short hairpin DNA oligonucleotide using conventional ligation (Figure 2.1). Restriction enzymes generating “sticky ends” should be preferred, since this will increase the efficiency of linker ligation. However, in our experience also the use of enzymes creating non-overhanging “blunt ends” is possible. The ligation is carried out using T4 DNA ligase. To ensure a high yield of DNA hairpin constructs the hairpin oligonucleotide is provided in excess to minimize the likelihood of re-ligation of DNA fragments. The overhang of the linker is designed to having complementary overhangs. For example, the restriction enzyme MspI will leave a 5’ CG overhang, accordantly the hairpin linker also has to have a 5’ “CG” overhang (Figure 2.2). Because of the enzymatic steps we recommend to use high quality (non-degraded) double stranded DNA (dsDNA). The circular constructs obtained after ligation are then subject to bisulfite treatment in which all cytosines are converted to uracils. The steric closeness (intertwined ssDNA rings) of the complementary DNA strands favors a quicker renaturation to dsDNA. Hence the bisulfite conversion of hairpin constructs is more challenging than that of normal DNA. To avoid a reannealing we recommend to use cycling bisulfite protocols with additional denaturation steps or alternatively higher incubation temperatures. After bisulfite treatment the DNA molecules are present in form

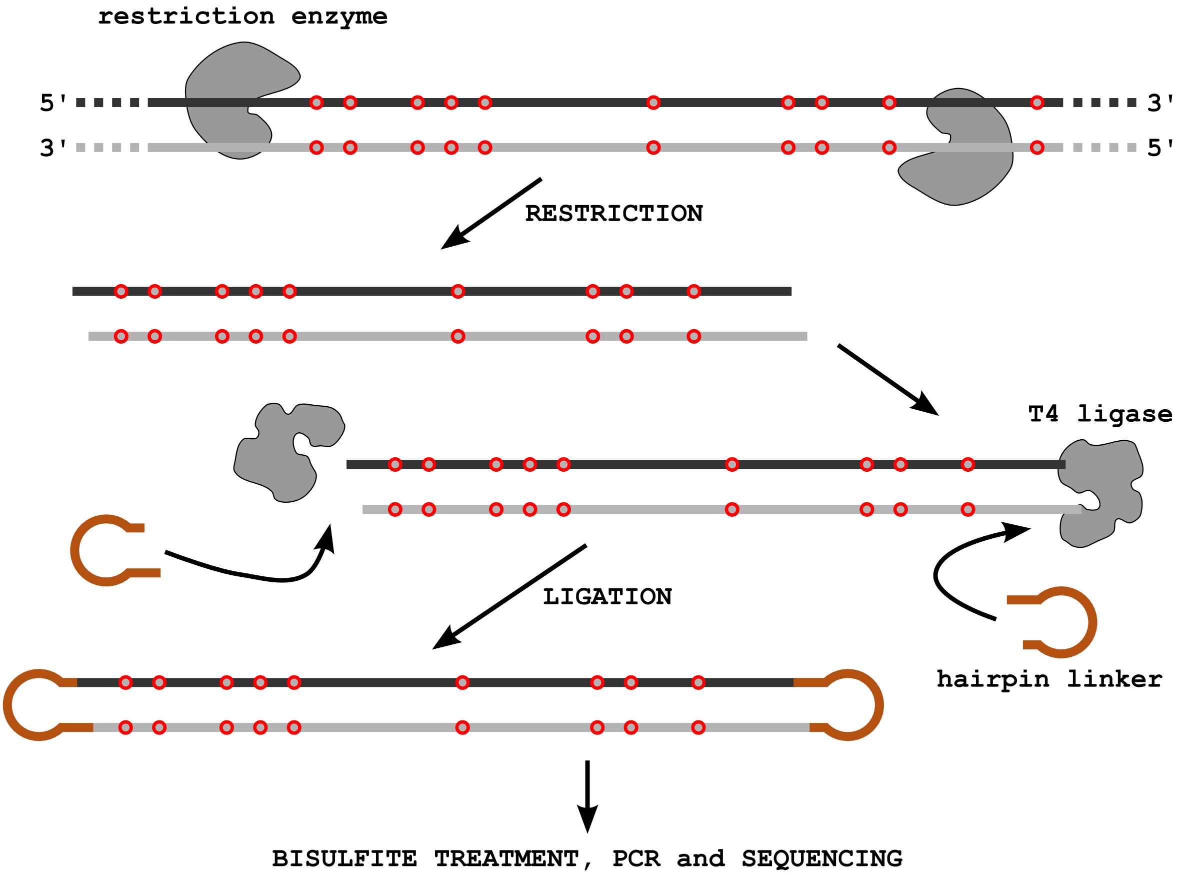


Fig. 2.1: Workflow of the Hairpin Bisulfite Sequencing protocol; genomic DNA is cut using a restriction enzyme, which is unaffected by DNA methylation. A complementary hairpin oligonucleotide is ligated to link upper and lower strand covalently together. The constructs are in the next step subject to bisulfite treatment resulting in single stranded circular DNA. After treatment the converted DNA serves as a template for a locus-specific PCR. PCR Products are then purified and sequenced. *Straight* and *dashed lines* indicate DNA strands, *red circles* illustrate CpG positions.

of single-stranded circular DNAs that contain uracil instead of unmethylated cytosines (Figure 2.2). They serve in the next step as a template in a site specific PCR to amplify the region of interest. Here it is essential to utilize a polymerase that accepts uracils as a template. The product of this PCR holds now the methylation information of the upper as well as the lower strand. The generated amplicons can be treated like “normal” PCR products and can be sequenced either by Sanger or next generation sequencing (NGS). We have successfully combined hairpin bisulfite sequencing with the Roche FLX pyrosequencing system and the Illumina MiSeq system.

For the subsequent analysis we use the trimmed and QC’ed FASTQ files and apply two bioinformatics tools develop in our lab. The first tool is the BiQ Analyzer `http://epigenetik.uni-saarland.de/de/software/` [5]. The program aligns the sequenced FASTQ reads to a reference HBS sequence (needs to be generated and provided). BiQ will provide an overview of CpG methylation and non-CpG methylation in the sequences. As outputs BiQ generates a tab separated table and different graphical represen-

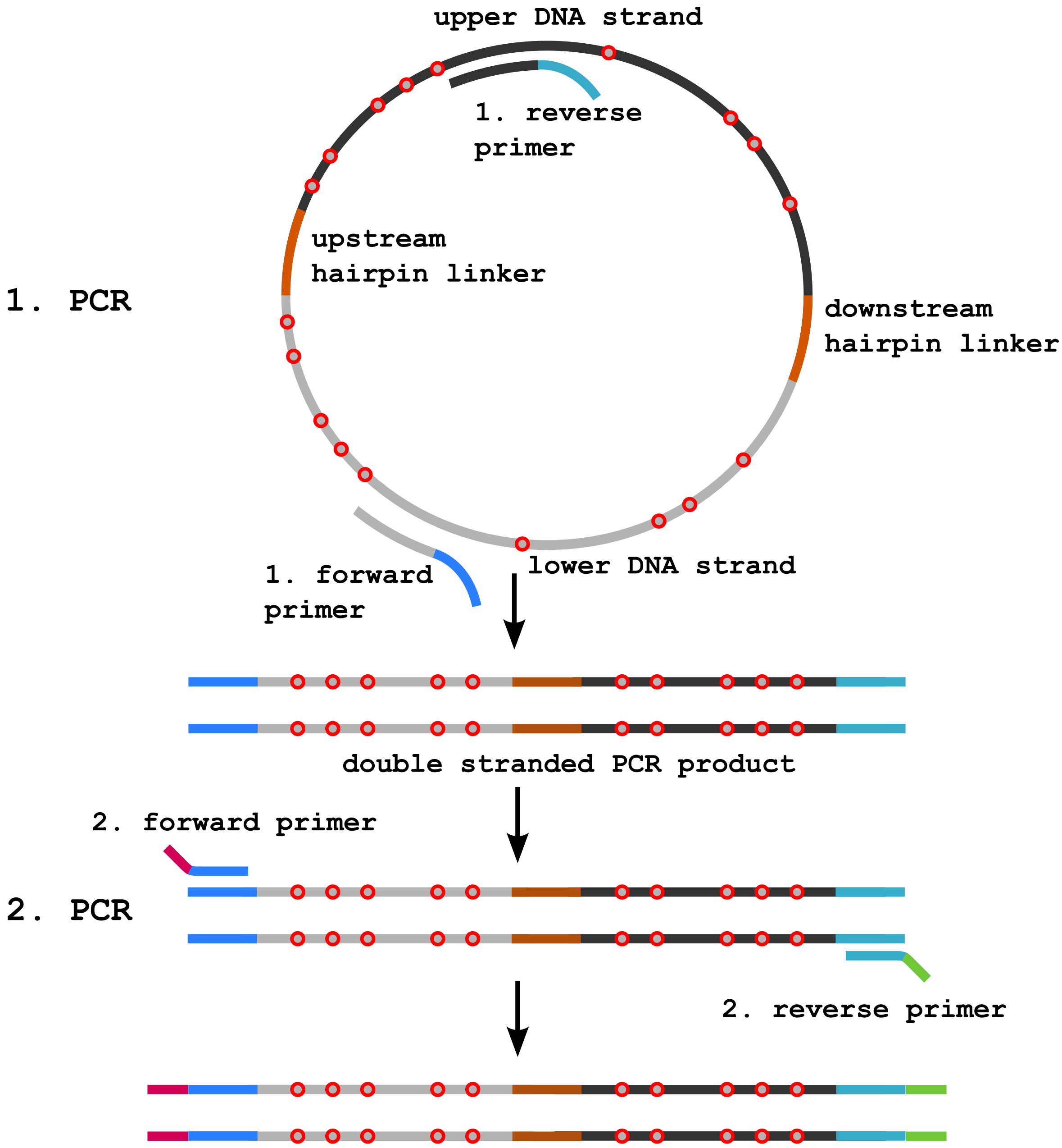


Fig. 2.2: Workflow after bisulfite treatment; bisulfite treated circular hairpin constructs are amplified in two consecutive PCRs; the fusion primers used in the first amplification step carry on the 5' end parts of the sequencing adapters which will become part of the PCR product. In the second PCR, the rest of the adapter sequence is introduced to the amplicon. *Lines* indicate DNA strands, *red circles* illustrate CpG positions.

tations, such a CpG methylation pattern map and quantitative pearl-necklace diagrams. The table output is then used in the next step by the Hairpinanalyzer script (available at `http://epigenetik.uni-saarland.de/de/software/`) to back-fold the single strand information into a double stranded format. A more detailed description of the workflow is given in the Methods section.

The double stranded hairpin bisulfite sequencing output now allows to detect and quantitate the symmetry of CpG and CpNpG DNA methylation patterns on both DNA strands of one individual chromosome and to unambiguously identify non-symmetrical cytosine methylation. We have applied the method to identify the massive occurrence of hemimethylated sites and general loss of methylation in Dnmt1KO ES-cells [2]. The hairpin bisulfite sequencing method provides a matched stranded information which directly shows the hemimethylated pattern formation. The use of a hairpin linker also offers additional technical advantages. Since the added linker contains unmodified cytosines, it is possible to directly calculate the true conversion rate obtained during bisulfite treatment and at the same time can estimate the true amount of methylated non-CpG positions (Figure 2.3). Further, in the loop sequences of the linker indices of variable nucleotides can be introduced creating a barcoding for each DNA molecule. This allows to identify duplicated sequences generated during PCR and to exclude them from further analysis.

Despite of the many advantages of the Hairpin Bisulfite Sequencing method HBS also has some experimental limitations. In a locus specific HBS analysis the obligate use of a suitable restriction enzyme (absence of recognitions site, distance to the analyzed region) can become a limitation. Moreover, the restriction enzymes must not be affected by DNA methylation, which would otherwise lead to a massive underrepresentation of methylated sites. Further, the size of the region that can be analyzed is limited. Based on the fact, that the PCR product contains the information of both upper and lower strands the product will be double the size of the genomic region. However, this is only a small disadvantage, since modern sequencing techniques allow to analyze sequences with a size over 500 base pairs (bp) in length (particularly on a FLX or MiSeq sequencer).

2.2 Materials

2.2.1 Experimental Design

(1) Identify suitable restriction sites close to the region of interest using NEBcutter (`http://tools.neb.com/NEBcutter2/`) or WatCut restrictions analysis (`http://watcut.uwaterloo.ca/template.php`).

(2) Design of the reference sequence (upper strand-linker-lower strand) for primer search and for the BiQ subsequent methylation analysis. The precise outline will be described in the method section.

(3) Primers for PCR are either designed “by eye” or with the help of online tools like Primer3Plus (`http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/`).

(4) Hairpin linkers are designed according to match the restriction enzyme ends and to containing indices/bar codes.

2.2.2 Sample Preparation

(1) Qubit® BR assay kit for measurement of the DNA concentration (Thermo Fisher Scientific).

(2) Restriction enzymes.

(3) T4 DNA Ligase (New England Biolabs).

(4) EZ DNA Methylation™ Kit; EZ DNA Methylation-Gold™ Kit (Zymo Research) or manual protocol.

(5) HOT FIREPol® DNA Polymerase (Solis BioDyne) or HotStarTaq DNA polymerase (Qiagen) or similar.

(6) Agencourt® AMPure® XP beads (Beckman Coulter).

(7) AveGene Gel purification kit.

(8) Primer and hairpin linker from a commercial primer supplier.

(9) 1xTE buffer; 10mM Tris HCL pH 7.4; 1mM EDTA pH8.0.

2.2.3 Sequencing and Data Analysis

(1) Genome Sequencer FLX 454 System (Roche) or MiSeq Desktop Sequencer (Illumina)

(2) For Methylation analysis BiQ Analyzer HT (`http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/`) was used followed by the use of Hairpinanalyzer. The python script of the hairpin analyzer can be received from `http://epigenetik.uni-saarland.de/de/software/`

2.3 Methods

2.3.1 Experimental Design

A proper planning and design of the experiment for Hairpin Bisulfite Sequencing includes three main steps.

Selecting Restriction Enzymes

The first step when designing the Hairpin Bisulfite experiment is the search for suitable restriction enzymes within the region of interest. When selecting enzymes a few things should be considered.

(1) First, the restriction enzyme should not be affected by DNA methylation to ensure that ummethylated, hemimethylated and fullmethylated regions are equally represented in the later results.

(2) The use of enzymes, which create sticky ends, should be preferred since the ligation will work more efficient. Nevertheless, the use of blunt end creating enzymes is possible.

(3) There are several online tools, which are suitable for the search of restriction enzymes for example "NEBcutter" or "WatCut". Both tools also provide the information on methylation sensitivity and sometimes even other modifications like 5hmC, 5fC and 5caC.

Hairpin Linker and Reference Sequence Design

The hairpin linker itself can be divided into three sections.

(1) The first part is variable and depends on the used restriction enzyme (Figure 2.2). For example MspI will create a 5'−CG overhang, therefore also the linker must contain a 5'−CG overhang including a free phosphate group to allow later ligation to the DNA.

(2) The second part always has the same sequence and facilitates the formation of the hairpin structure (Figure 2.2). The use of unmodified cytosine within this linker part allows later an exact und unbiased calculation of the conversion rate during the bisulfite treatment and permits a more accurate detection of nonCpG methylation.

(3) The last part is forming the loop structure of the linker. It contains a unique sequence that cannot form any double strand structures (Figure 2.2).

(4) As shown in Figure 2.2 the loop can obtain 6−8 variable nucleotide positions. This will allow to create an individual barcode for each DNA molecule and to exclude duplicates of the PCR in further computational analysis. Using eight of these positions it is theoretically possible to distinguish between 6,561 (3^8) sequences.

Reference Sequence and Primer Design

The next step in the experimental design is the construction of the reference sequence. This sequence is needed in order to design primers and also for later sequencing and methylation analysis. The reference sequence can easily be designed with any software that can handle text files.

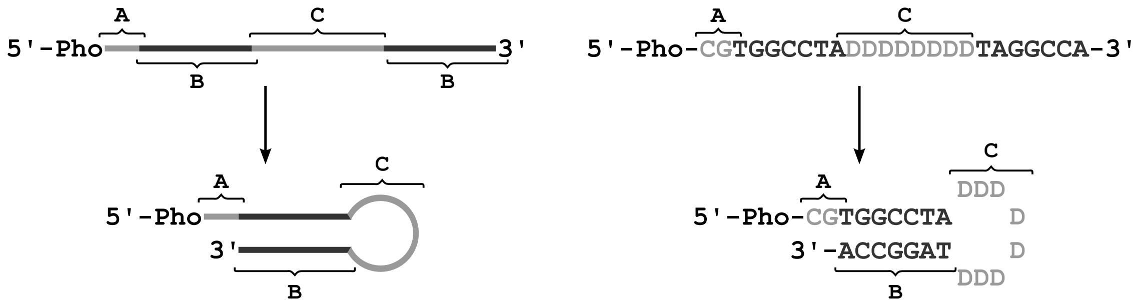


Fig. 2.3: Schematic illustration of the hairpin linker; A, variable, restriction enzyme dependent sequence; B, constant formation facilitating sequence; C, variable loop sequence; as an example the structure and sequence of a hairpin oligo for MspI restriction is shown in a denatured, single strand and annealed, folded state.

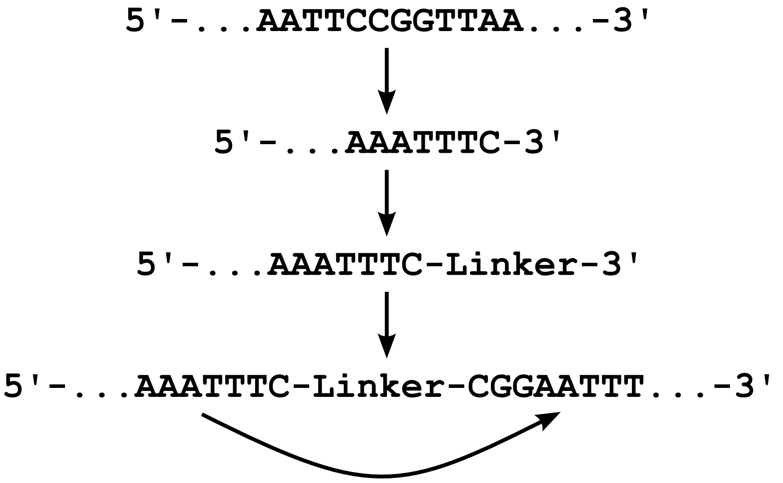


Fig. 2.4: Example of the design of a hairpin construct. The left part next to the restriction site is removed. The linker sequence is pasted followed by the reverse complement sequence of the right part of the DNA sequence.

(1) Download the genomic sequence of the region, mark the restriction site used, delete the sequence in front or after the restriction site, add the linker sequence and finally paste the reverse complement of the remaining sequence on the other site of the linker (Figure 2.3).

(1) Replace all cytosines outside a CpG context by T to obtain the bisulfite sequence of the hairpin construct.

(1) 3. The Primers can then be designed either manually or using software- or online tools **Note 1**.

2.3.2 Experimental Procedure

Restriction, ligation and bisulfite treatment are done in the same reaction tube without any purification steps in between which minimizes the loss of DNA. The described protocol below is the standard protocol used in our lab, however depending on the amount of DNA or the restriction enzyme used in the reaction an optimization or adjustments might be necessary. For a starting material, genomic DNA from phenol/chloroform or Kit extraction can be used. Due to the nature of the method it is important to work with intact dsDNA. Therefore the concentration of the DNA should be determined using the Qubit system that will only detect dsDNA.

Restriction

(1) Cleave 200 to 500 ng of genomic DNA with 20 units of a restriction enzyme in 1x Reaction buffer in a total reaction volume of 17µl.

(2) Incubate the reaction for at least 3 hours at the recommended temperature, followed by a heat inactivation **Note 2**.

Preparation of Hairpin Linker and Ligation

(1) Dissolve the oligonucleotide, which will later form the hairpin linker, in 1xTE, resulting in a 100 µM stock and store at -20°C.

(2) Before usage, form the oligo into the right structure. Heat 50 µl of the 100 µM solution to 98°C for 15 min followed by cool down using the slowest cooling rate of a thermocycler until 20°C is reached. In this form the hairpin linker is rather stable but should be stored at -20°C.

(3) Add 1 µl of a 100 µM hairpin linker solution to the reaction together with 200U T4 DNA ligase and 2 µL of 10mM ATP.

(4) Incubate the reaction for at least 3 h at 16°C. The ligation can also be performed overnight.

Bisulfite Treatment

As mentioned before, there are several bisulfite kits available which can be used for the conversion of hairpin constructs. Since hairpin DNA molecules tend to fold back rather fast and the conversion of cytosine can only occur only on single stranded DNA, we recommend a protocol with higher incubation temperature or additional denaturation steps. Kits successfully used in our lab are listed above in the material section. When using a homemade protocol, the conversion rate can be verified by looking into the conversion of cytosines included in the hairpin linker. A manual protocol used in our lab has been described previously [2]. In addition, the hairpin protocol is also extendable to oxidative bisulfite sequencing or other “chemical” forms of sequencing [6].

PCR

Because of the bisulfite treatment it is essential to use a polymerase that recognizes uracil in the template strand for the PCR. The best results in our lab were achieved using the HotFirePol from SolisBioDyne or HotStarTaq from Qiagen. We recommend a total reaction volume 30µl reaction and perform multiple PCRs in parallel to ensure a low number of duplicated reads. Pipette the reaction mixture according to Table 2.1 using either enzyme.

Tab. 2.1: Examples of PCR protocols to amplify Hairpin Bisulfite molecules

HotFirePol PCR	HotStartTaq PCR
3 µl 10x Reaction Buffer	3 µl 10x Reaction Buffer
3 µl 25 mM MgCl2	1.2 µl 25mM MgCl2
2.4 µl 10mM dNTPs	2.4 µl 10mM dNTPs
0.5 µl 166nM Forward Primer	0.5 µl 166nM Forward Primer
0.5 µl 166nM Reverse Primer	0.5 µl 166nM Revere Primer
0.5 µl HotFirePol DNA polymerase	0.3 µl HotStartTaq DNA Polymerase
Ad 30 µl ddH$_2$O	Ad 30 µl ddH$_2$O

Both enzymes share similar temperature characteristics during PCR. Purification of the PCR products is performed using 27µl AMPureXP Beads (0.9x). A typical cycler protocol for both enzymes is given in Table 2.2.

Tab. 2.2: Temperature profile of HotFire/HotStarTaq

95°C	15min	
95°C	1min	
50-62°C	1min	35-45x
72°C	1min	
72°C	5min	
4°C	hold	

Sequencing

Hairpin bisulfite is compatible with both Sanger and next generation sequencing (NGS). To prepare the samples for NGS one has to introduce adapter on each side of the amplicon which are compatible with the sequencing platform one once to use. These adapter bind to the sequencing platform and are the start point of the sequencing process. In addition each adapter carries a unique sequence ID which allows sequencing of multiple samples at the same time. The adapter sequence is introduced by the use of fusion primers in two consecutive PCRs (Figure 2.2).

In the first PCR, the primers consist 3' of the target specific sequence complementary to the bisulfite treated DNA and carry 5' the first part of the adapter sequence resulting in an amplification of the target sequence and the introduction of the first part of the adaptor (Table 2.1 and 2.2 for PCR conditions).

In the second PCR the primer will bind the adaptor part introduced during the first PCR step. These primers carry the sequence which later binds to the sequencing platform and in addition carry a sample specific sequence ID. Table 2.3 provides all primer and adapter sequences needed for sequencing on the Illumina MiSeq platform.

Tab. 2.3: Illumina adapter and primer Sequences; i5/i7 = index; grey = flow cell binding sequence; Oligonucleotide sequences © 2016 Illumina, Inc. All rights reserved.

1. PCR Forward	TCTTTCCCTACACGACGCTCTTCCGATCT -AmpliconSpecific
1. PCR Reverse	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT -AmpliconSpecific
2. PCR Forward (i5, 6bp index e.g.: CGTGAT)	AATGATACGGCGACCACCGAGATCTACAC [i5] ACACTCTTTCCCTACACGACGCTCTTCCGATCT
2. PCR Forward (i7, 6bp index e.g.: AAGCTA)	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC [i7] ATCTCGTATGCCGTCTTCTGCTTG

The second amplification can be performed as a multiplex PCR where several amplicons of distinct genomic regions can be prepared for sequencing at the same time. For this the concentration of each amplicons must be adjusted to 5nM and pooled into one reaction. Table 2.4 lists chemicals and cycling conditions for the second PCR amplification **Note 3**.

Tab. 2.4: List of chemicals and cycler condition for the second PCR

Enrichment PCR	Cycler conditions	
25.0µl 5nM amplicon pool		
5.0µl 10x Buffer HotStarTaq	95°C - 15min	
2.0µl 25mM MgCl2	95°C - 30sec	5 cylces
4.0µl 10mM dNTPs	60°C - 30sec	
2.5µl 10 µM Index primer	72°C - 30sec	
2.5µl 10 µM Universal primer	72°C - 5min	
0.6µl HotStarTaq	4°C - hold	
8.4µl ddH2O		

After incubation the reaction is again cleaned up using 55µl AMPureXP beads (1.1x) and adjusted to a 10nM concentration. The final amplicon library is prepared by pooling all enrichment PCRs into one reaction. Following the MiSeq preparation protocol from Illumina the library is diluted stepwise to a final concentration of 18pM.

2.3.3 Data Analysis

To obtain the information of the symmetric DNA methylation two separate steps are necessary. First, the methylation information has to be obtained from the sequenced PCR product. Second, the methylation information has to be translated back to the DNA double strand. For this we developed in our lab two different bio informatics tools.

Methylation Analysis

The methylation analysis is performed using the BiQ HT Analyzer [5]. The BiQ HT is a Java based program designed for locus-specific DNA methylation analysis of high-throughput bisulfite sequencing data **Note 4**. The program aligns the sequenced reads to a reference sequence. hereby comparing all positions where a cytosine is expected it detects the methylation status of the sequenced loci. Both reference sequence and sequencing data has to be provided in a FASTA file format. The program calculates different quality scores including alignment score, sequence identity, bisulfite conversion rate and number of missing sites. The different quality scores can also be used to filter the data, for example against low quality reads or low sequence identity. BiQHT will automatically use the default settings for filtering but each value can be adjusted manually by the user. Besides the analysis of CpGs, BiQ is also able to detect methylation in a non-CpG context (CpHpG; CpHpH). After analysis the data is presented and can be stored in different ways. A typical output consist of a tab stop separated table which includes all quality and methylation values and different types of methylation diagrams such as pattern maps and pearl-necklace diagram.

For the Hairpin analysis three independent analyzing steps with BiQ HT are necessary. The first step is the detection of CpG methylation. The reference sequence used in this step consists of the sequence of both DNA strands and the linker sequence in between.

(1) Replace in the linker sequence all cytosines by thymine because the linker will be analyzed later in an independent step.

(2) Depending on the type of loci, adjust the filter sequence identity. When analyzing repetitive elements for example a lower sequence identity (80%) should be chosen because of the variability of the sequence of those elements.

(2) Calculate the nonCpG methylation. Here it can be advantageous to replace all CpGs in the sequencing data by NpN to avoid confounding by possible mutations or sequencing failures, which create new CpG positions and lead to wrong estimations of nonCpG methylation. Again this is especially important when analyzing repetitive

elements, due to their sequence variability. Note that also all CpGs in the reference should be replaced by NpN to allow an accurate alignment of the sequenced reads.

(2) Estimate the conversion rate using the unconverted linker sequence as the reference sequence. Like in the non-CpG analysis the method to analyze Cs has to be chosen. When using wobble-position, within the linker loop, lower the sequence identity to 70 or even 60% otherwise most of the sequences will be filtered out.

For each of the three analysis steps a separate folder has to be created where the results are stored. The information in the different folders is then used to reconstruct the double strand information. In principle only the CpG folder is needed to reconstruct the double strand, but then conversion rate and non CpG position will not be analyzed (*see* **Note 5**).

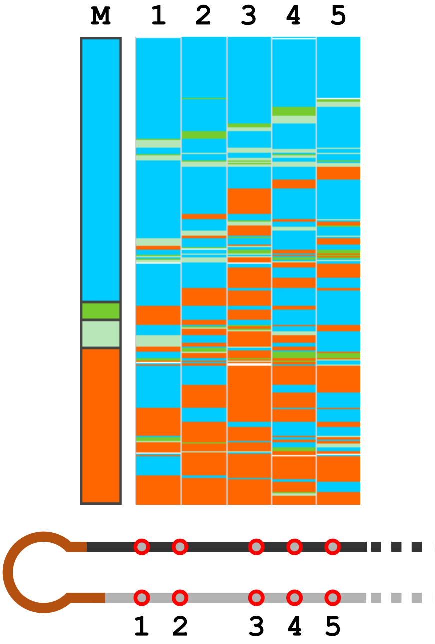


Fig. 2.5: Example of an methylation pattern map created by the Hairpinanalyzer; each column represents one CpG dyad (1-5) and one row a sequenced read; read = fully methylated CpG dyads; *light* and *dark* green = hemimethylated CpG dyads; *blue* = unmethylated CpG dyads; *white* bars indicate mutated or not analyzable CpGs; the bar on the left site shows a summary of the methylation over all CpG positions analyzed (M)

Reconstructing the Double Strand

The four folders containing the information about CpG methylation, non-CpG methylation, conversion rate and optionally SNPs are then used to reconstruct the double strand and calculating the amount of both strands methylated, only upper strand methylated, only lower strand methylated or both strands unmethylated.

For this purpose we have developed in our group a script, the so-called Hairpinanalyzer. The program is based on python and is able to translate the single strand information, created by the BQ HT, into a double strand output. The Hairpinanalyzer has no graphical interface and has to be run via command line and configuration of the source code. Like the BiQ the Hairpinanalzer creates an output in form of a tab separated table and a methylation pattern map. An example of a pattern map is given in Figure 2.5. Each column of the map represents a CpG position whereas each row indicates one sequenced read. The different methylation statistics, both strands methylated, left (of the linker) strand methylated, right strand methylated both strands methylated and mutated or not detectable. The colors can be chosen manually. The tab separated table contains information about number of reads, number of analysed CpGs, methylation status, mutations or sequencing errors as well as the information about SNPs and nonCpG methylation (*see* **Note 6**).

2.4 Notes

(1) Even though both strands are no longer perfectly complementary after bisulfite treatment due to the conversion of cytosine; there are still regions that are complementary to some extent. This makes the primer design sometimes difficult because of dimer formations. To avoid this, it is sometimes necessary to use relatively short primers. This on the other hand will result in relatively low annealing temperatures. In our experience primers with a size of 23 bases and an annealing temperature over 50°C are working fine. However, if possible a higher temperature should always be preferred, since it will increase the specificity of the PCR

(2) The restrictions conditions described in this protocol were suitable for all restriction enzymes used in our lab so far (BsaWI, DdeI, Eco47I, MspI, TaqI). However, depending on the amount of DNA and the used restriction enzyme it might be necessary to change the parameters to obtain optimal reaction conditions.

(3) When using next generation sequencing, the fusion primers of the PCR have to be adjusted to the sequencing system. In our lab we have successfully used FLX as well as Illumina systems for the sequencing of Hairpin Bisulfite amplicons.

(4) The BQ Analyzer HT can be downloaded from `http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de`. There is also a Java Web Start version available as well as detailed documentation.

(5) It is also possible to include a fourth folder that contains the information about SNPs and the barcode of the linker. For that a fourth analysis with BiQ has to be

performed. By repeating mainly the analysis for CpG and this time also using the option output SNPs.

(6) The Hairpinanalyzer was programmed by Mathias Bader in our lab. Unfortunately this script is not available online but can be transmitted upon request from the authors.

Bibliography

[1] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[2] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[3] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[4] Stefanie Seisenberger, Simon Andrews, Felix Krueger, Julia Arand, Jörn Walter, Fátima Santos, Christian Popp, Bernard Thienpont, Wendy Dean, and Wolf Reik. The dynamics of genome-wide dna methylation reprogramming in mouse primordial germ cells. *Molecular cell*, 48(6):849–862, 2012.

[5] Pavlo Lutsik, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jörn Walter, and Christoph Bock. Biq analyzer ht: locus-specific analysis of dna methylation by high-throughput bisulfite sequencing. *Nucleic acids research*, 39(suppl_2):W551–W556, 2011.

[6] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

3. THE INFLUENCE OF HYDROXYLATION ON MAINTAINING CPG METHYLATION PATTERNS: A HIDDEN MARKOV MODEL APPROACH

The content of Chapter 3 has been published as:

Giehr, P.[†‡], Kyriakopoulos, C.[*‡], Ficz, G.[§], Wolf, V.[*], & Walter, J.[†] (2016). The influence of hydroxylation on maintaining CpG methylation patterns: a hidden Markov model approach. PLoS computational biology, 12(5), e1004905.

AUTHOR CONTRIBUTIONS

Pascal Giehr: Establishment and further development of hairpin oxidative bisulfite sequencing, as well as conduction of laboratory experiments. Consulting on model design. Data interpretation.

Authoring of abstract and author summary. Authoring of introduction, biological background, creation and design of figure 3.1. Authoring of methods, hairpin oxidative bisulfite sequencing, section 3.2.1, creation and design of figure 3.2. Authoring of results, biological background; presentation and interpretation of data. Joint authoring of discussion together with Charalampos Kyriakopoulos. Authoring of supporting information, section 3.6.5, including figures and tables.

Dr. Charalampos Kyriakopoulos: Design and development of the model. Statistical analysis of the output.

Authoring of introduction i.e. mathematical background. Authoring of methods, hidden Markov model, section 3.2.2, creation and design of figure 3.3 and 3.4. Authoring of results, mathematical aspects and generation of figures 3.5 - 3.7. Joint authoring of discussion together with Pascal Giehr. Authoring of supporting information, section 3.6.1 - 3.6.4, including figures and tables.

[†] Biological Sciences Department, University of Saarland, D-66123 Saarbrücken, Germany

[‡] This Authors contributed equally to tis work

[*] Computer Science Department, Saarland University, D-66123 Saarbrücken, Germany

[§] Barts Cancer Institute, Queen Mary University, London, United Kingdom

Prof. Dr. Verena Wolf: Supervision of stochastic modelling. Study design. Financing. Revision of the manuscript i.e. changes in formulation/wording, structure and appearance of the manuscript - mainly restricted to mathematical background (introduction), as well as data presentation (results) and discussion.

Prof. Dr. Jörn Walter: Supervision of laboratory experiments. Study design. Financing. Revision of the manuscript, i.e. changes in formulation/wording, structure and appearance of the manuscript - mainly restricted to biological background (introduction), as well as data presentation (results) and discussion.

Abstract

DNA methylation and demethylation are opposing processes that when in balance create stable patterns of epigenetic memory. The control of DNA methylation pattern formation by replication dependent and independent demethylation processes has been suggested to be influenced by Tet mediated oxidation of 5mC. Several alternative mechanisms have been proposed suggesting that 5hmC influences either replication dependent maintenance of DNA methylation or replication independent processes of active demethylation. Using high resolution hairpin oxidative bisulfite sequencing data, we precisely determine the amount of 5mC and 5hmC and model the contribution of 5hmC to processes of demethylation in mouse ESCs. We develop an extended hidden Markov model capable of accurately describing the regional contribution of 5hmC to demethylation dynamics. Our analysis shows that 5hmC has a strong impact on replication dependent demethylation, mainly by impairing methylation maintenance.

Author Summary

Oxidation of 5mC by Ten-eleven translocation (Tet) enzymes leads to the formation of 5hmC and other higher oxidized forms in the DNA. Several findings indicate that oxidation induces demethylation processes, but the mechanistic contribution of 5hmC to this process remains unclear. Using an innovative combination of 5hmC detection chemistry and high resolution sequencing, we generate data that can be used for a novel hidden Markov modeling approach. This new model for the first time incorporates 5hmC dynamics and allows to test certain scenarios of demethylation mechanisms. Our findings support the conclusion that 5mC oxidation compromises the copying of DNA methylation patterns across generations in ES-cells.

3.1 Introduction

DNA methylation is an epigenetic modification essential for the regulation of genome stability and genome function [1, 2]. During development the distribution of DNA methylation is under strict control to maintain a temporal and cell type specific persistence of epigenetic information [3]. The methylation of DNA in mammals is restricted to the C-5 position of cytosine and is predominantly found in a CpG sequence context [4, 5].

Our current knowledge suggests that DNA methylation patterns (5mC) are mainly established by DNA methyltransferases Dnmt3a and Dnmt3b [3, 6]. The palindromic nature of a CpG sequence in which 5mC occurs allows a recognition of the 5mC hemimethylated state after semi-conservative replication and a copying of the parental methylation pattern to the newly synthesized DNA strand (see Fig 3.1). A series of experiments revealed that Dnmt1 in conjunction with Uhrf1 are responsible for this copying also referred to as *maintenance* methylation. Dnmt1 and Uhrf1 have a high preference for binding to hemimethylated CpG substrates [7, 8, 9]. Together they assure the maintenance symmetric CpG methylation patterns after each round of replication.

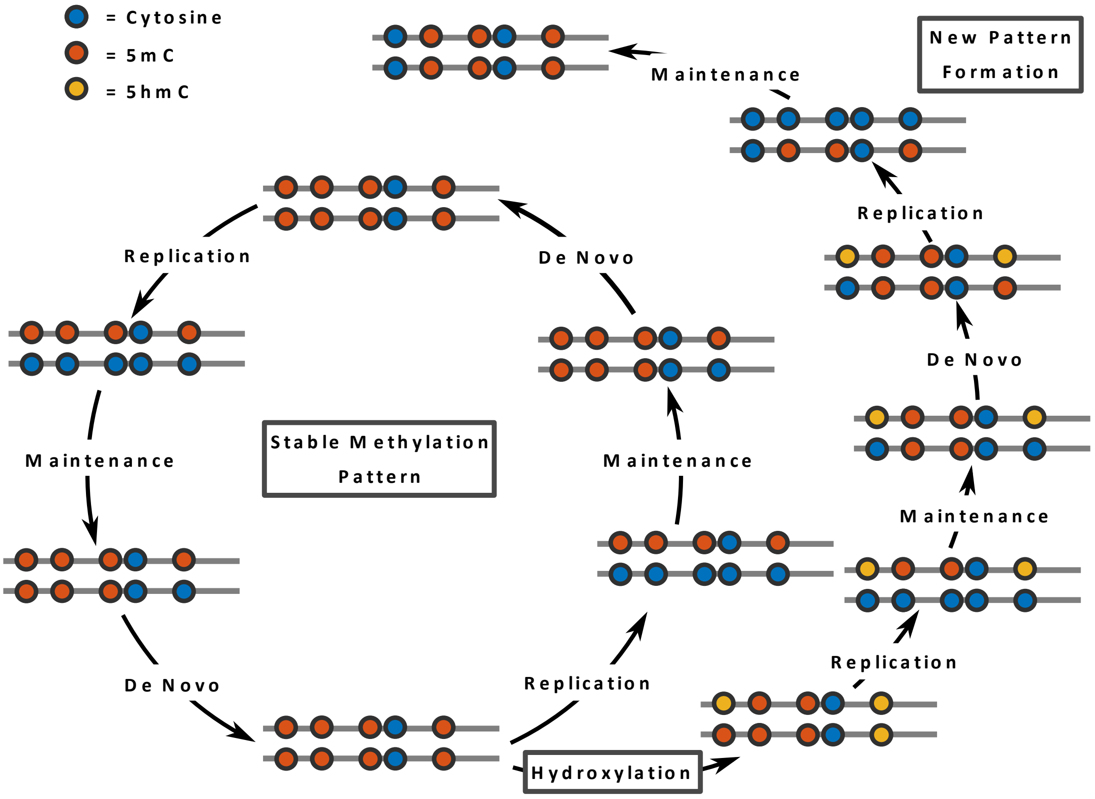


Fig. 3.1: Maintenance and de novo methylation are usually cooperating to maintain a stable methylation pattern (inner circle). The oxidation of 5mC to 5hmC may interfere with the maintenance machinery causing a (partial) loss of CpG methylation after DNA replication. DNA strands are indicated by lines whereas the CpG are shown as colored circles.

In contrast to Dnmt1, Dnmt3a and Dnmt3b act on hemi- as well as unmethylated

CpGs and their activity is not coupled to DNA replication. Both enzymes are highly regulated and regarded as the main enzymes to establish new methylation patterns and are therefore classified as *de novo* DNA methyltransferases. However, recent data shows that Dnmt1 may also de novo methylate unmethylated dyads and that Dnmt3a and Dnmt3b are also involved in reestablishing (thus "maintaining") complete methylation patterns at certain loci [10]. In summary, the persistence of methylation patterns is controlled by a coordinated action of de novo and maintenance functions of all three enzymes.

Besides the establishment and the persistence of methylation its removal is also of great biological importance. Demethylation events can occur on a local scale in case of individual gene activation but also on a global genome wide level like in the early zygote and the germ line, where genomes are reprogrammed for new developmental functions [11, 12]. In both cases demethylation can be achieved either by an active mechanism (direct removal), a passive replication-dependent loss or a combination of both.

Recent findings suggest that the oxidation of 5mC modulates active and passive demethylation processes. 5-hydroxymethyl cytosine (5hmC) is generated by oxidation of 5mC in an enzymatic reaction catalyzed by the oxoglutarate- and Fe(ii)-dependent ten-eleven trans-location dioxygenases (Tet1, Tet2, and Tet3) [13]. Tet enzymes also catalyze further oxidations to 5-formylcytosine (5fC) and to 5-carboxycytosine (5caC), which have been shown to promote processes of active demethylation [14, 15, 16]. Still 5hmC is the most prevalent oxidation type and widely discussed to having an influence on DNA methylation pattern stability in dividing cells. 5hmC not only alters the chemical properties but also the biological recognition of the base. Dnmt1 binds to 5hmC with a much lower efficiency than to 5mC. This may impair the replication dependent copying of 5mC [17].

In mouse ES cells (mESCs), in the early mouse embryo and in the early germ cells DNA demethylation stability is influenced by the conversion of 5mC into 5hmC. Disturbances or depletion of Tet enzymes in these phases result in massive changes of 5hmC and lead to developmental consequences [18, 19, 20]. These findings indicate that the controlled alteration of DNA methylation patterns across DNA replications is dependent on 5hmC. However, the underlying mechanisms are still under debate. Mouse ESCs are a well established system to study these effects as they rapidly lose DNA methylation on a genome wide scale when the cells are transferred from conventional serum medium containing LIF (primed state) to a synthetic 2i medium [21, 22]. This loss of 5mC is coupled to a temporary gain of 5hmC. In our study we follow the dynamic of DNA demethylation in mESCs over time and DNA replications using a novel combination of hairpin sequencing with bisulfite sequencing (BS) and oxidative bisulfite sequencing (oxBS). This method allows us to determine the methylation status of both complementary DNA strands at individual chromosomes and the status of 5hmC levels at given time points [23, 10, 24].

We propose a stochastic model that describes the evolution of both methylation and hydroxylation patterns over time. Our model allows that methylation can be lost due to cell replication and methyl groups can be added due to either maintenance or de novo enzyme activity [10, 25]. In addition, we assume that all methylated sites can be hydroxylated.

Based on these assumptions we define a hidden Markov model (HMM) for each data set and construct likelihood functions on the basis of the two sequencing methods. The combination of the two likelihoods allows us to derive estimations for the levels of (hydroxy-)methylation based on observations at four different time points. Finally, we determine unknown parameters of the model, i.e., methylation and hydroxylation efficiencies as well as the initial distribution of the hidden states. Despite its simplicity, the model accurately predicts the evolution of the (hydroxy-)methylation patterns and allows us to test different assumptions about the activities of the involved enzymes.

3.2 Methods

3.2.1 Hairpin Oxidative Bisulfite Sequencing

Currently no comprehensive data are available allowing to model the fate of 5hmC at a single base resolution level. Therefore, extending the method described in Fitz et al. 2014 and Arand et al. [10, 21] we established a workflow enabling us to produce such data. To obtain base resolution information of the modification status we apply hairpin bisulfite sequencing on DNA samples split into oxidative (oxBS-Seq) and non oxidative standard bisulfite reaction (BS-Seq) data sets. The use of the hairpin linker strategy allows us then to determine the levels of 5hmC and 5mC on both DNA strands [23] and to determine the methylation status (hemimethylated, unmethylated or fully methylated) at each individual CpG dyad within the sequenced loci at single molecule resolution. To obtain a sufficient coverage (>1000x) per CpG we use very deep NGS based sequencing of selected loci. The deep sequencing enables us to determine accurate rates and error rates for each modification. To cover larger parts in the genome we included the analysis of mobile elements which occur in multiple identical copies across the genome and to which we refer as "repetitive elements". In fact our analysis covers about 91% of all annotated IAP(IAPLTR1a_mM) (N = 1635), 20% of L1md_A (N = 3287), 12% for L1md_T (N = 2784) and 30% of MuERVL (N = 725). In this case the >1000x coverage has to be seen as the aggregate of about a 5x coverage of each individual copy of a given repetitive element. Figure 3.2 outlines the main experimental steps of the procedure.

In the first step genomic DNA is digested using restriction enzymes which generate cuts close to the gene/locus selected for methylation analysis. In a following reaction both DNA strands are ligated to a back-folding 'hairpin'-oligonucleotide. Next the DNA is unfolded and subjected to a bisulfite or oxidative bisulfite treatment followed by a locus specific PCR amplification. PCR primers contain Mi-Seq (Illumina) compatible extensions to perform deep (paired end 2x300bp) sequencing (up to 10K/product). Sequencing data are processed using our in house software BiQ-HT and a python script. In the bisulfite only reaction 5mC and 5hmC remain as cytosines, while in the oxidative bisulfite reaction 5hmC is converted to uracil/thymine. Each individual sequence covers the hairpin linker which contains modified and unmodified cytosines at known positions. This allows us to monitor the efficacy of bisulfite and oxidative bisulfite reactions per molecule (note that all unmodified cytosines are converted to thymines) and calculate exact error rates by dividing the number of unconverted bases by the total number of analyzed cytosines. Additional information about the protocol is given in S1 Text together with reference-, primer- and linker-sequences.

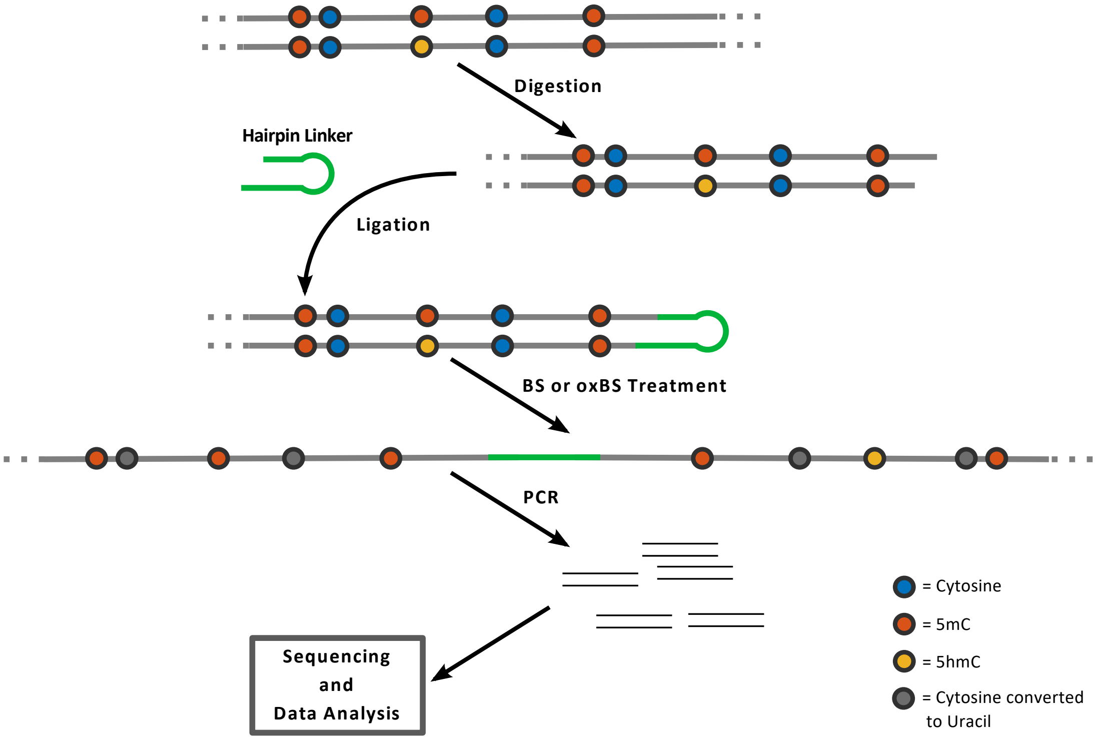


Fig. 3.2: Schematic outline of hairpin bisulfite (BS) and oxidative bisulfite sequencing (oxBS): The method is based on enzymatic digestions of genomic DNA and the covalent connection of upper and lower DNA strands by ligating a hairpin oligonucleotide. PCR enrichment of BS/oxBS treated sample is used for amplicon generation followed by sequencing and data analysis.

3.2.2 Hidden Markov model

Our model considers a CpG site (alternatively dyad) over time and describes its state as a (discrete time) Markov chain $\{\mathcal{X}(t), t \in \mathbb{N}\}$ taking values in $\mathcal{S} = \{u, m, h\}^2$. Each state (s_1, s_2) (for $s_1, s_2 \in \{u, m, h\}$) encodes whether the upper strand (lower strand) is *unmethylated* (u), *methylated* (m) or *hydroxylated* $(h)^*$. For instance, $(s_1, s_2) = (u, h)$ the

*We use u, m, h instead of C, 5mC and 5hmC to shorten the description and avoid confusion with the observable states.

upper strand is unmethylated and the lower strand is hydroxylated. We will often simply write $(s_1 s_2)$ instead of (s_1, s_2).

The time parameter t corresponds to the number of cell divisions and the state transitions are triggered by three consecutive events: cell division, methylation and hydroxylation. The corresponding transition probability matrices are $\mathbf{D}(t)$, $\mathbf{M}(t)$, and $\mathbf{H}(t)$, respectively. Thus, the combined transition probability matrix of $\mathcal{X}$ is defined as

$$\mathbf{P}(t) = \mathbf{D}(t) \cdot \mathbf{M}(t) \cdot \mathbf{H}(t),$$

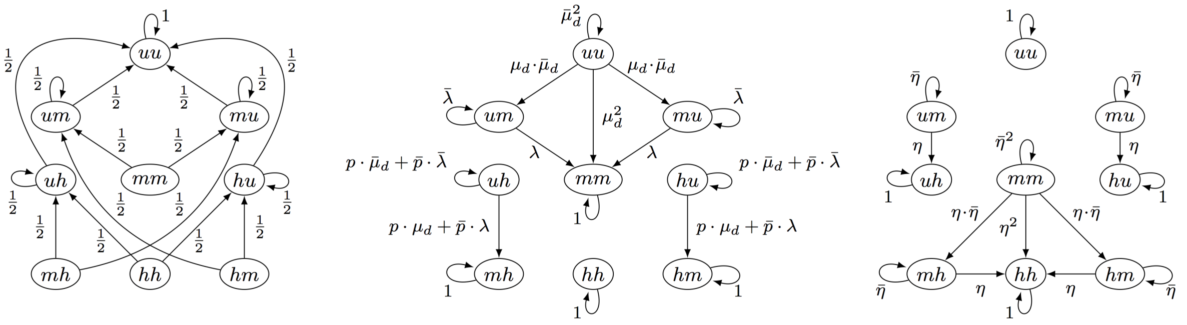
with entries $\mathbf{P}_{ij}(t)$ that equal the probabilities that given $\mathcal{X}(t) = i = (s_1 s_2)$, the next state is $X(t+1) = j = (s'_1 s'_2)$ for all $i, j \in \mathcal{S}$. Note here we assume that hydroxylation occurs after methylation to ensure that between two cell divisions a transition from u to m and then to h is possible. Moreover, note that we allow $\mathbf{P}(t)$ to change over time, so that we capture the case that the (hydroxy-)methylation efficiencies do not remain constant over time. In the sequel we give a detailed description of $\mathbf{D}(t)$, $\mathbf{M}(t)$, and $\mathbf{H}(t)$. For a formal definition of the matrices, we refer to Supplement, Section 3.6.2.

Demethylation through Cell Division.

With each cell division and DNA replication one new DNA strand is synthesized resulting in a temporary situation where only unmodified cytosines are present in the new strand. Since the epigenetic pattern of the parental strand remains unchanged a previously methylated CpG site keeps half of the (hydroxy-)methylated state in the two daughter cells. By averaging over the daughter cells, if the current state is (mm) then after cell division the new state is (um) or (mu) each with probability 0.5 (depending on whether the newly synthesized strand is the upper or the lower strand). Similarly, with probability 0.5 the process enters (uh) or (hu) from (hh). Thus, DNA replication/cell division may result in a direct loss of methyl or hydroxyl groups. The transition probabilities of the remaining states are defined in a similar way and we illustrate the corresponding matrix$\mathbf{D}(t)$ in Fig 3.3a).

Methylation

The loss of methylation by DNA replication is counteracted by a restored methylation due to the combined activity of the three methyltransferases Dnmt1, Dnmt3a and Dnmt3b. We distinguish between maintenance methylation catalyzed by Dnmt1 and de novo methylation catalyzed by Dnmt3a and Dnmt3b. We assume that a cytosine of an unmethylated dyad can only be methylated by a de novo event, while both maintenance and de novo methylation are possible on a hemimethylated dyad. Based on related in vitro experiments [3] and our recently published work [10], we assume that Dnmt3a/b act on hemimethylated sites with the same efficiency as on unmethylated sites.



a) Cell division b) Methylation c) Hydroxylation

Fig. 3.3: Possible transitions of the 9 different states of a CpG site. Methyl groups are a) removed after cell division, b) added due to maintenance (μ_m) or de novo methylation (μ_d) and c) are hydroxylated (η) by Tet enzymes.

We define $\mu_m(t)$ and $\mu_d(t)$ as the probabilities of maintenance and de novo methylation of a cytosine, respectively, where the corresponding methylation event occurs within the t-th cell division cycle ($t \in \{1, 2, \ldots\}$). In addition, we define $\lambda(t)$ to be the total methylation efficiency on a hemimethylated site. It holds that

$$\lambda(t) = \mu_m(t) + \mu_d(t) - \mu_m(t) \cdot \mu_d(t),$$

because maintenance is associated with the replication machinery and happens immediately after replication with efficiency $\mu_m(t)$. In case maintenance methylation by Dnmt1 is not successful the site can still be methylated with de novo methylation efficiency $\mu_d(t)$ which then gives $\lambda(t) = \mu_m(t) + (1 - \mu_m(t)) \cdot \mu_d(t)$. We write $\bar{\mu}_m(t) = 1 - \mu_m(t)$, $\bar{\mu}_d(t) = 1 - \mu_d(t)$ and $\bar{\lambda}(t) = 1 - \lambda(t)$ for the complements of the above probabilities and we omit the time parameter t whenever it is not relevant.

Note that if a CpG site has two unmethylated cytosines then two de novo methylation events are possible. Assuming independence between them, all transition probabilities of the corresponding state (uu) are the product of two event probabilities. We illustrate the corresponding methylation matrix$\mathbf{M}(t)$ in Fig 3.3 b). Here p is the probability that maintenance methylation is not applied to the states (hu) and (uh), i.e., the hydroxyl group prevents the maintenance process, i.e., the methylation of the unmodified cytosine on the opposite strand. As a result, from these states the states (hm) and (mh) can only be entered via de novo methylation. In the opposite case, with probability $\bar{p} = 1 - p$, states (hu) and (uh) are seen as hemimethylated during maintenance and can enter states (hm) and (mh) with probability λ for both maintenance and de novo methylation (see Fig 3.3b). Besides, the states (mh), (hm), and (hh) have only self-loops since the Dnmts do not modify hydroxyl groups.

Hydroxylation

Let $\eta(t)$ be the probability that before the $(t+1)$-th cell division a methylated position becomes hydroxylated, i.e, the probability of a transition from m to h. Similarly as above, we write $\bar{\eta}(t)$ for $1-\eta(t)$ and omit t whenever convenient. Assuming again independence between two hydroxylation events, the corresponding matrix $\mathbf{H}(t)$ is illustrated in Fig 3.3c). Note that without an active hydroxylation mechanism ($\eta > 0$) the level of 5hmC would half after each replication since newly synthesized strands do not inherit the hydroxyl groups of the mother strand.

Hydroxylation is the last modification that we consider before the next cell division. Thus, between two cell divisions an unmethylated position may transition from u to m and then to h.

Observable states and conversion errors.

In order to define the observable states and the corresponding emission probabilities, we first describe the details of hairpin sequencing and (oxidative) bisulfite sequencing. First the DNA is cut by a restriction enzyme. The DNA fragments are then linked covalent to a Hairpin linker resulting in the connection of upper and lower strand. The resulting hairpin fragments are divided into two halves, one is treated with a standard bisulfite reaction and the other is subjected to an oxidation followed by bisulfite treatment. Both 5mC and 5hmC are not affected by the (non-oxidative) bisulfite treatment and appear after sequencing as cytosines. In the oxidative case 5hmC is oxidized to 5fC which is converted during bisulfite treatment to 5fU and represents itself after sequencing as thymine (see Fig 3.4).

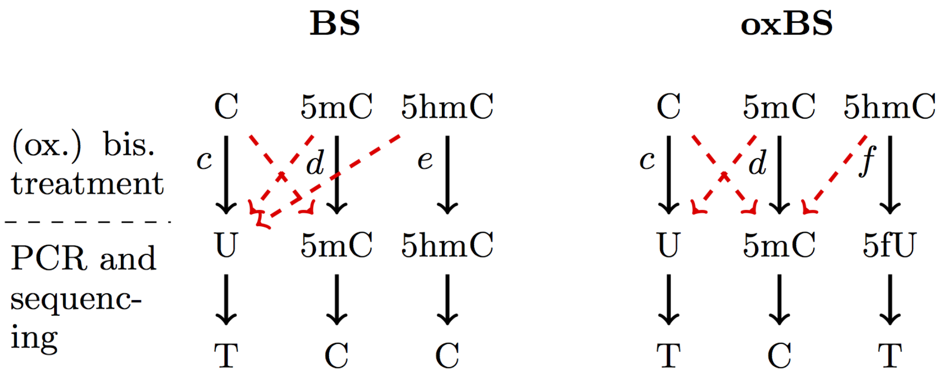


Fig. 3.4: Schematic outline of the conversion of Cytosine, 5mC and 5hmC during BS and oxBS treatment and after sequencing: In the bisulfite reaction a cytosine (C) is converted to uracil (U), whereas 5mC and 5hmC remain untouched. In the oxidative bisulfite sequencing only 5mC remains untouched and cytosine as well as 5hmC is converted to uracil (U). The conversion errors are illustrated as dashed red arrows and c, d, e, f are the conversion probabilities.

We incorporated unmodified cytosine as well as 5mC and 5hmC into the hairpin linker to precisely estimate the conversion errors (see also Supplement, Table 3.14) that influence the transition probabilities between the hidden and the observable states. These controls allow us to correct for technical errors in individual measurements.

In Fig 3.4 the transitions from a site's possible hidden states to the observable ones are shown. Each base will eventually transform into a thymine (T) or a cytosine (C). Thus, the set of the observable states for a CpG site with two cytosines is $\mathcal{S}_{obs} = \{\mathrm{T}, \mathrm{C}\}^2$.. The red dashed arrows correspond to conversion errors and assuming all errors are zero, i.e., the probabilities $c = d = e = f$ of a correct conversion are all one, a C will eventually transform to T and a 5mC will transform to C in both bisulfite and oxidative bisulfite setups. However, a hydroxylated cytosine (5hmC) is ideally mapped to a C during BS and to a T during oxBS. The entries of the corresponding emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ for the transitions from hidden to all observable states can be found in Supplement, Section 3.6.1, Table 3.1 - 3.9 and the values of the conversion errors from all analyzed loci for each of the experimental setups are listed in. Note that the values of c and d can differ between the two treatments and that the conversion probabilities can also differ over time.

Estimation of Model Parameters.

Given the number of times $n_{bs}(j,t)$ and $n_{ox}(j,t)$ that state $j \in \mathcal{S}_{obs} = \{T, C\}^2$ has been observed during independent BS and oxidative BS measurements at time t we use a maximum likelihood approach to estimate the unknown parameters of the HMMs, that is, the initial distribution of the hidden states,$\mathcal{S} = \{u, m, h\}^2$, the unknown functions $\mu_d(t), \mu_m(t)$ and $\eta(t)$, as well as the probability p at which CpG sites with one hydroxyl group are not considered during maintenance.

Formally, let $\pi(t)$ be the row vector of the state probabilities of the hidden states after t cell divisions, i.e., $\pi(0)$ is the initial distribution of the hidden states. For $i \in \mathcal{S}$ let $\pi(i,t) = P(\mathcal{X}(t) = i)$ denote the entry of π(t) that corresponds to state i. The probability of observing state $j \in \mathcal{S}_{obs}$ at time t is given by

$$P(\mathcal{O}(t) = j) = \sum_{i \in \mathcal{S}} P(\mathcal{O}(t) = j \mid \mathcal{X}(t) = i) \cdot \pi(i,t),$$

where $\mathcal{O}(t)$ is the random variable for the state observed at time t and $P(\mathcal{O}(t) = j \mid \mathcal{X}(t) = i)$ is the emission probability. In matrix-vector form this yields

$$\pi_{bs}(t) = \pi(t) \cdot \mathbf{E}_{bs}(t) \quad \text{and} \quad \pi_{ox}(t) = \pi(t) \cdot \mathbf{E}_{ox}(t)$$

for the two sequencing experiments (BS and oxBS, respectively). Here, $\pi_{bs}(t)$ and $\pi_{ox}(t)$ are the vectors with the distribution over the observable states at time t. Note that both HMMs have the same distribution $\pi(t)$ for the hidden states (as for both experiments the same cell population is used) but different emission probabilities and that $\pi(t)$ is given by

$\pi(t) = \pi(0) \cdot \prod_{k=1}^{t} \mathbf{P}(k)$.

First, we estimate the initial distribution $\pi(0)$ based on the initial independent BS and oxidative BS measurements under conventional serum conditions by considering the combined likelihood

$$\mathcal{L}_1(\pi(0)) = \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,0)^{n_{bs}(j,0)} \cdot \pi_{ox}(j,0)^{n_{ox}(j,0)}. \tag{3.1}$$

The above likelihood depends only on the unknown vector $\pi(0)$ and the emission matrices and allows us to determine the initial distribution of the hidden states. We maximize the likelihood subject to the constraint $\sum_i \pi(i,0) = 1$, i.e.

$\pi(0)^* = \arg\max_{\pi(0)} \mathcal{L}_1(\pi(0))$,

where $\pi(0)$ ranges over all vectors that sum to one. Then, given an estimate for $\pi(0)$, we compute for $t \in \{1, 2, \ldots\}$ the state probabilities $\pi(t)$ of the hidden states and consider the common likelihood

$$\mathcal{L}_2(\mathbf{v}) = \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j} \pi_{bs}(j,t)^{n_{bs}(j,t)} \cdot \pi_{ox}(j,t)^{n_{ox}(j,t)} \tag{3.2}$$

for the observations at all remaining observation time points $t \in T_{obs}$. Note that here we assume that the cells divide every 24 hours, hence t ranges over all days at which measurements were made (see also Supplement, Section 3.6.2). In addition, we can assume independence between the observations because during the measurement only a small fraction of cells is taken out of a large pool and thus it is unlikely that we pick two cells with a common descendant.

The likelihood $\mathcal{L}_2(\mathbf{v})$ depends on the matrices $\mathbf{P}(t)$ and thus on the unknown functions $\mu_d(t), \mu_m(t)$, $\eta(t)$ and the probability p. We assume that the enzymes' efficiencies are linear in t, i.e., each function is of the form $\beta_0 + \beta_1 \cdot t$,, which yields a vector $\mathbf{v}$ of seven unknown parameters in total. For estimating $\mathbf{v}$ we use again a maximum likelihood approach, i.e., we determine

$\mathbf{v}^* = \arg\max_{\mathbf{v}} \mathcal{L}_2(\mathbf{v})$,

under the appropriate constraints (see Supplement, Section 3.6.2). The maximization of the likelihoods in Eq. (3.1) and (3.2) is a (global) optimization problem for which it is convenient to minimize the negative logarithm of the likelihood. Deriving expressions for the first and second derivatives of the log-likelihood is straightforward and yields fast convergence of the gradient descent optimization routine with multiple starting values. Due to the large number of samples we expect our maximum likelihood estimators (MLEs) to be approximately unbiased and normally distributed. Moreover, we can compute the observed Fisher information matrix (FIM) and thus derive confidence intervals for all parameters (for details see Supplement, Section 3.6.2).

3.3 Results

Previous genome wide analyses showed a high or moderate decrease of DNA methylation in ESCs transferred from serum into 2i medium [21, 22]. Furthermore, it was shown that the oxidation of 5mC to 5hmC is likely to contribute to this DNA demethylation [21]. The goal of our work was to develop a model which describes the 5hmC dependent molecular mechanisms that cause this loss of DNA methylation upon consecutive rounds of replication. For the modeling we generated an ultra deep DNA methylation data set of selected loci in mouse ES cells (ESCs) collected at defined time points after cultivation in 2i.

For our analysis we chose five multicopy, repetitive elements, IAPs (intracisternal A particle), L1mdA and L1mdT (both Long interspersed nuclear elements), MuERVL (Murine endogenous retrovirus) and mSat (major satellite), as well as four single copy loci in the genes Afp, Snrpn, Ttc25 and Zim3. It was already known that some of these repetitive elements are subject to demethylation. Ttc25 and Zim3 where previously shown to exhibit a less pronounced loss of methylation in the absence of Tet1/Tet2 in 2i medium [21]. Imprinted genes such as Snrpn were shown to be “resistant” to demethylation in 2i.

Deep locus specific DNA methylation profiles were generated from mESCs grown in conventional serum/LIF medium (day0) and after their transfer and cultivation into 2i medium for 24h (day1), 72h (day3) and 144h (day6), respectively. During this period the ESCs undergo a maximum of six cell divisions (as inferred from cell densities). For each time point and locus we performed consecutive bisulfite and oxidative hairpin bisulfite reactions using high coverage Mi-Seq sequencing (see Methods section). Following sequence processing (alignment, trimming, QC filtering) we obtained two data sets for each locus: one describing the combined 5mC+5hmC status (BS-Seq) and one describing the 5mC status alone (oxBs-Seq). The hairpin refolding of sequences then let us determine the accurate double stranded CpG methylation status at a given locus (hemi-, fully- or unmethylated).

With this data we used our HMMs (described in the **Methods** section) to estimate the amount of 5mC and 5hmC in these loci and to predict the efficiencies of maintenance methylation, de novo methylation and hydroxylation over time. In our modeling we analyzed both aggregated and single CpG behavior for each locus. Both average and single CpG modeling gave similar results. The single CpG data, summarized in the supplementary information (see Supplement, Figure 3.10 and 3.11), gave slightly increased confidence intervals compared to averaged data. In our further analysis we use averaged data for model interpretation.

Using the estimated values of the model's unknown parameters we could predict the probabilities of the observable states and compare them to the measured data at various time points. The model accurately describes the dynamics for all loci except for some underestimations of two states CC and TT for oxBs in Ttc25 and Zim3, respectively (Fig 3.5 and Supplement, Figure 3.8).

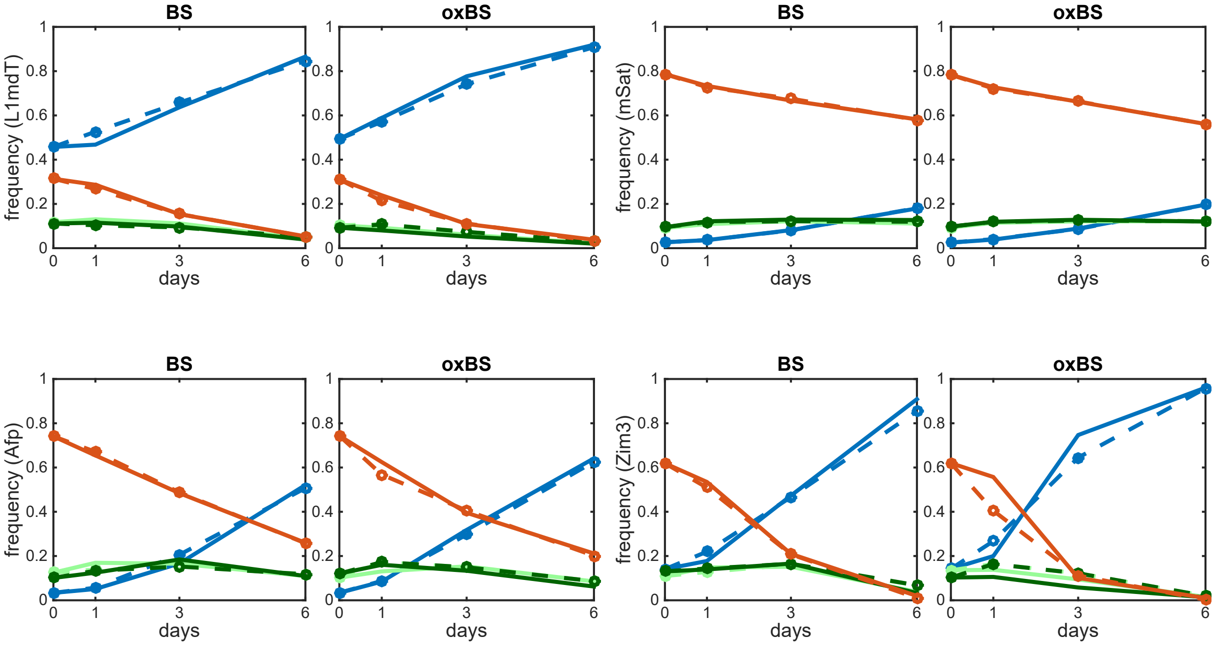


Fig. 3.5: Comparison of predicted modification levels and the obtained sequencing data for BS and oxBS for the loci L1mdT (top-left), mSat (top-right), Afp (bottom-left), Zim3 (bottom-right); TT (blue), TC (light green), CT (dark green), CC (red). The solid lines show the experimentally measured frequencies states and the dashed lines correspond to the values predicted by the two HMMs.

Fig 3.6 shows the probabilities of the hidden states in L1mdT, mSat, Afp, and Zim3, where the parameters are chosen according to the results of the maximum likelihood estimation. The left bar diagram shows the probabilities of all fully methylated (mm), hemimethylated (um and mu) and unmethylated (uu) sites, as well as the total amount of the hydroxylated CpG dyads, i.e., those containing at least one 5hmC. The detailed level of all hydroxylated sites is depicted in the right diagram.

From previous experiments it was known that 5hmC levels initially increase during cultivation in 2i [21, 22]. However, precise levels had not been determined per locus.

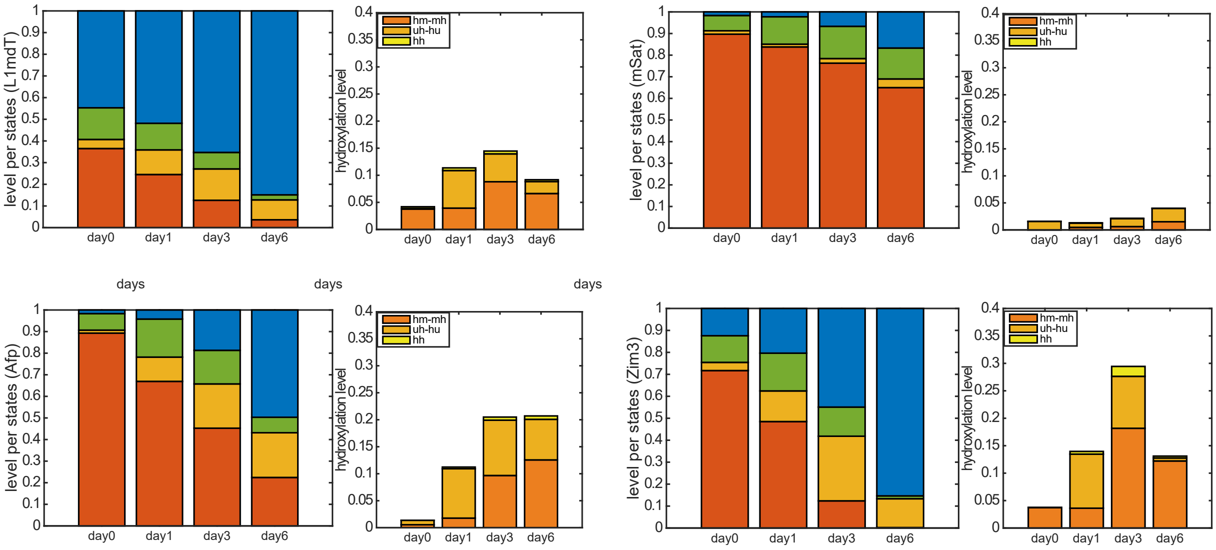


Fig. 3.6: Probabilities of the hidden states for L1mdT (top-left), mSat (top-right), Afp (bottom-left) and Zim3 (bottom-right): The left diagram depicts the amount of fully methylated (mm) hemimethylated (um and mu) 🟩, and unmethylated (uu) 🟦 sites. The orange block 🟧 gives the total amount of CpG sites with at least one 5hmC(hidden states), while the detailed distribution of the hydroxylated states is given by the diagram on the right.

Our analysis provides the first accurate locus specific determination of 5hmC changes. Our estimation of 5hmC confirms an initial increase of hydroxylated cytosines over time for most loci besides L1mdA and Snrpn. L1mdA shows a low level of 5mC and 5hmC, which only slightly decreases in 2i. Snrpn also shows a relatively low level of 5mC and a non significant amount of 5hmC, which do not change in 2i over time (Supplement, Figure 3.9). The highest hydroxylation levels are found in the single copy genes Zim3 and Afp with a maximum level of 0.30 and 0.20. For Afp, mSat, IAP and MuERVL (see Fig 3.6 and Supplement, Figure 3.9), the maximum hydroxylation level is seen at day6, while for L1mdT, Ttc25 and Zim3 at day3. The latter can be explained by the particularly low 5mC levels between day3 and day6 in these loci which naturally reduces the potential substrates for the Tet enzymes. However, the level of 5hmC (orange bar in Fig 3.6 and Supplement, Figure 3.9, left) relative to the total modification level (5hmC + 5mC) (red, orange and green bars), becomes maximal on the sixth day for all loci that show a loss of 5mC. This points towards an increasingly important role of 5hmC in the loss of methylation over time.

Indeed, the probability p (see HMM subsection) that a 5hmC site is not recognized by Dnmt1 (or the Dnmt1/Uhrf1 complex), which corresponds to states (hu) and (uh) in the model, is estimated to be 1 with very small standard deviations for all the loci that show significant 5hmC levels. We estimated smaller values for p only for those loci where hydroxylation is nearly absent (mSat, MuERVL, Snrpn).

In Fig 3.7 we plot the functions $\mu_m(t)$, $\mu_d(t)$, $\eta(t)$ and $\lambda(t)$ over time together with their estimated standard deviations. Note that the estimated standard deviations of all the efficiencies are very small (maximum half width of all confidence intervals is 0.031). For the exact estimates and their standard deviations see Supplement, Table 3.11 and 3.12. From the above efficiencies we can deduce the impact of de novo methylation activity on the hemimethylated dyads as the difference between the total methylation efficiency and maintenance methylation, i.e., $\lambda(t) - \mu_m(t) = \bar{\mu}_m(t) \cdot \mu_d(t)$ (see Fig 3.7). Our data indicates that persistence of DNA methylation at Afp, mSat, IAP and MuERVL elements clearly depends also on de novo enzymes acting on hemimethylated CpGs

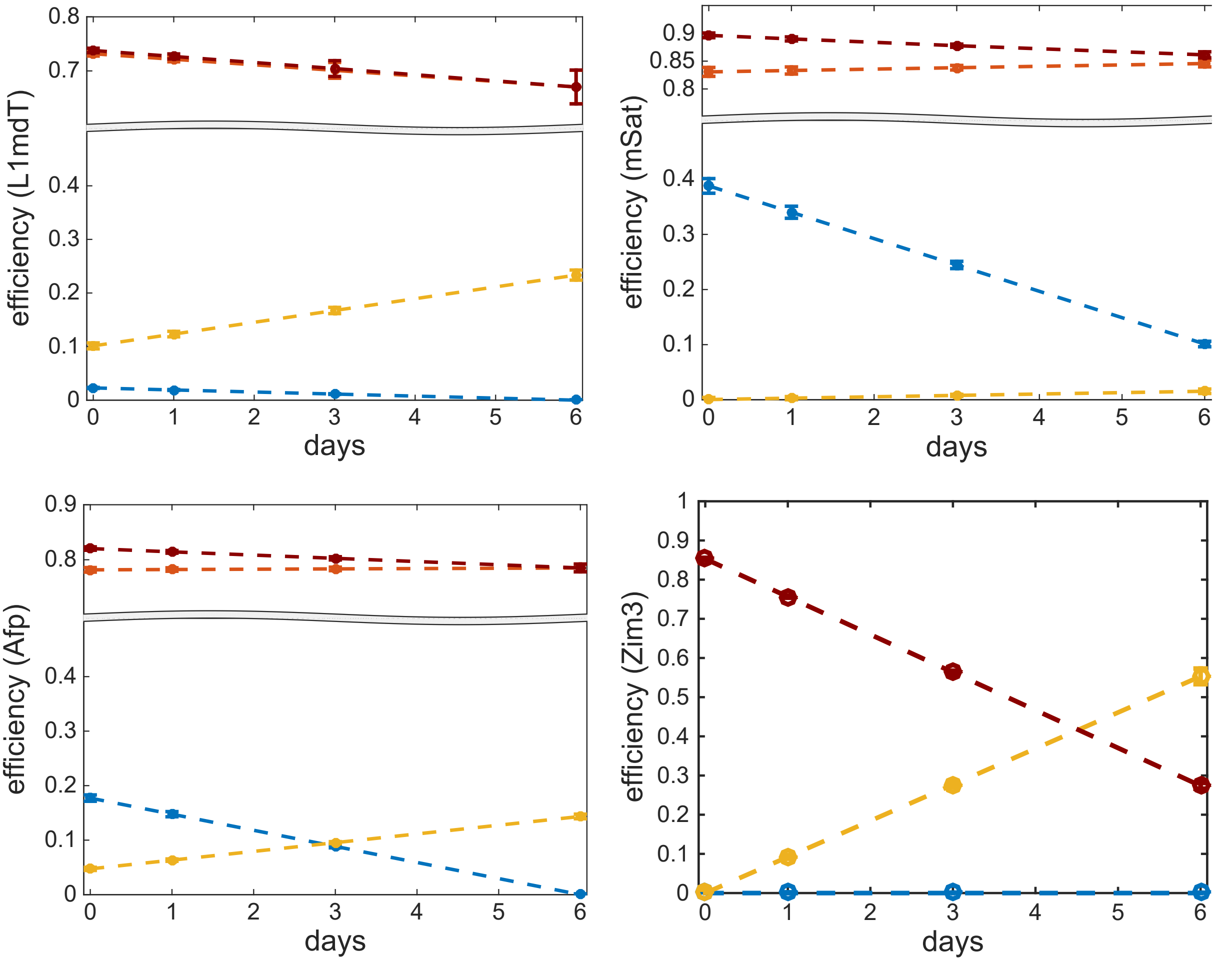


Fig. 3.7: The diagrams show the enzymatic efficiencies and their standard deviations for maintenance (red), de novo (blue), hydroxylation (yellow) and total efficiency on a hemimethylated CpG (dark red). Results are given for L1mdT (top-left), mSat (top-right), Afp (bottom-left) and Zim3 (bottom-right) over time.

For each efficiency, we performed a statistical test with a confidence level of 1% for the null hypothesis that the slope of the corresponding linear function is zero, i.e., that the efficiencies are constant over time (see in addition Supplement, Section 3.6.2). Furthermore, to eliminate the possibility of overfitting due to the linear assumption, we performed leave-one-out cross-validation (LOOCV) to estimate the test error of our model with constant efficiencies against a linear model. Results in Supplement, Table 3.13 show that the linear assumption improves the prediction up to 38.3%. Further tests concerning the sensitivity of the model w.r.t. the parameters showed that the model is also sufficiently robust (see Supplement, Section 3.6.2).

Overall, the estimation of the efficiency functions reveals some common and some locus specific features that accompany the DNA demethylation dynamics over time in 2i. As a common feature we observe that the total methylation on hemimethylated sites, $\lambda(t)$, decreases over time in all examined loci but at different rates. Along with this decrease we observe also a drop of de novo methylation activity at all loci besides Ttc25 and Zim3. In contrast, hydroxylation activity increases for most loci over time (except for Snrpn). Interestingly, the largest increase of $\eta(t)$ occurs in L1mdT and the two DMRs in the genes Ttc25 and Zim3, where we also observe low or even total absence of de novo activity. On the other hand, a weaker hydroxylation activity in mSat, as well as IAP and MuERVL (Supplement, Figure 3.9), is accompanied by a strong decrease of $\mu_d(t)$ in the same loci, while in Afp both de novo methylation and hydroxylation show a moderate decrease and increase, respectively. At last, maintenance methylation seems to differ among loci. For all repetitive multicopy loci and Afp maintenance activity remains nearly constant while for Ttc25 and Zim3 it shows a significant decrease. For the imprinted Snrpn locus, where the methylation level remains constant, our model accurately predicts the apparently constantly high maintenance efficiency of 1.0. Altogether, these findings point towards a major impairment of maintenance methylation by 5hmC. Additionally, for each locus this impairment is modulated by a distinct combination of decreasing (e.g. Dnmt3a,b) or increasing (e.g. Tet) activities in a locus specific manner. Some of the locus specific differences may also have their origin in the particular methylation and (hydroxy-)methylation status present in serum/LIF before the shift into 2i.

3.4 Discussion

The goal of our study was to investigate the role of 5hmC in the process of progressive DNA demethylation at single copy and mulitcopy loci across the genome. As a model system we used the DNA of ES cells grown under conditions where the cells experience a genome wide reduction of DNA methylation [22, 21].

Using time dependent comparative bisulfite and oxidative bisulfite hairpin sequencing data we generated two HMMs: one that represents the dynamics of total modifications (5mC and 5hmC in BS) and the other only representing the 5mC levels (in oxBS). The comparison allowed us to accurately determine the amount and changes of 5hmC at certain genomic loci, to estimate the transient distribution of both 5mC and 5hmC in the DNA and to compute statistically reliable estimates for the efficiencies of maintenance and de novo methylation, as well as for hydroxylation over time.

Our first finding is that 5hmC changes over time and can be modeled along with the overall changes in symmetric DNA methylation at CpGs. Our estimates give us an exact knowledge of 5hmC dynamics, which is congruent with the finding that several Tet enzymes are up-regulated in 2i medium [21, 22]. The calculation of the hidden state probabilities and the efficiencies of the different enzyme-driven processes show that the 5hmC dependent demethylation rates differ considerably However, the dynamics of the (hydroxy-)methylation levels for the CpGs of the same locus show a certain homogeneity (see Supplement, Figure 3.10 and 3.11).

The second major finding is that loci with an enrichment of 5hmC such as Afp, L1mdT and IAP show higher demethylation rates compared to mSat or Snrpn. Hence, 5hmC containing DNA strands are indeed more likely to lose DNA methylation over time. Our modeling strongly supports the hypothesis that 5hmC is less well recognized by the maintenance methylation machinery (Dnmt1/Uhrf1 complex) as indicated by the estimation of the corresponding non-recognition probability p. The accumulation of 5hmC then causes a passive dilution mechanism of CpG methylation with each DNA replication/cell cycle, despite of the fact that the model predicts a constant behaviour of maintenance activity in most of the analyzed loci. In ES cells maintained in 2i medium this mechanism appears to be the main driving force for a rapid and linear DNA demethylation.

Interestingly, in contrast to the previously shown unchanged mRNA expression of Dnmt1 and Uhrf1 in 2i [21, 22] we observe a strong decrease of maintenance function for the single copy genes Ttc25 and Zim3 (see Figure 3.7 and Supplement, Figure 3.9, red line). Since the influence of 5hmC on the maintenance mechanism is reflected by the recognition probability p, the observed decrease is independent of the high 5hmC levels at these loci. This indicates an additional impairment or absence of the maintenance machinery at these loci. However, we cannot exclude the possibility that with the strong decrease in maintenance efficiency our model, at least to some extent, compensates for active demethylation which we cannot capture with our current experimental/model design.

Being able to estimate the de novo methylation impact of Dnmt3a/b on hemimethylated sites, the third observation of our model is that all analyzed elements show a compromised de novo methylation activity as an additional factor contributing to an enhanced local DNA demethylation. The predicted behavior for the involved enzymes' activities appears to follow their relative expression in 2i medium, in which both Dnmt3a and Dnmt3b are clearly down regulated [21, 22]. Our observations, thus, suggest that the down regulation of Dnmt3a and Dnmt3b activities appears to enhance the 5hmC dependent CpG demethylation. This may be either directly due to a decreased methylation efficiency on hemimethylated sites or due to a lower abundance of the enzymes.

In summary, we present a novel HMM method that allows to precisely measure and describe effects related to the influence of 5hmC on the persistence of DNA methylation in the mammalian genome. The modeling allows us to decipher complex DNA methylation patterns and enables us to accurately infer enzymatic activities. In its current form the model already captures a fraction of possible demethylation dynamics and scenarios most likely reflecting many loci in the genome. A genome wide application of our modeling is possible. It comes, though, at the expense of locus specific accuracy since with the existing whole genome hairpin sequencing methods data is difficult to generate and will not reach a sufficient sequencing depth. However, our approach can also be used to accurately model 5hmC dependent methylation dynamics in diseases, e.g. certain cancers and in aging processes of long lived cells. By integrating novel precise sequencing methods, which detect other oxidized modifications the model can be enhanced to additionally capture active demethylation and describe the involved processes.

3.5 Author Contributions

Conceived and designed the experiments: VW JW. Performed the experiments: PG GF. Analyzed the data: PG CK. Contributed reagents/materials/analysis tools: GF. Wrote the paper: PG CK VW JW. Designed/implemented the software used in analysis: CK.

3.6 Supporting Information

3.6.1 BS and oxBS Data

In Tables 3.1-3.9 we show the data for the DNA loci L1mdA, L1mdT, IAP, mSat, MuERVL, Afp, Ttc25, Zim3 and Snrpn taken from bisulfite and oxidative bisulfite sequencing together with the measured conversion errors $\bar{c}$, $\bar{d}$, $\bar{e}$ and $\bar{f}$ for each locus. The conversion errors are calculated using the hairpin linker which is ligated onto the DNA. A more detailed description of the conversion errors' calculation is given in Section 4.1. The measurement times are: 24h after incubation on Serum (day0), and 24h (day1), 72h (day3) and 144h (day6) on 2i. Each table shows the total number of CpGs of the corresponding locus that have been observed in each of the four observable states (TT, TC, CT and CC) for every day of measurement.

Tab. 3.1: IAP

				BS							**oxBS**			
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	39	84	116	890	0.003	0.0709	0.0774	35	70	77	605	0.002	0.0935	0.0701
1	17	89	99	831	0.002	0.0685	0.0411	57	131	115	943	0.002	0.0813	0.0939
3	68	87	111	513	0.001	0.0628	0.0721	77	112	112	449	0.001	0.09	0.0905
6	283	152	178	703	0.003	0.0829	0.0455	210	68	81	365	0.002	0.0737	0.0942

Tab. 3.2: L1mdA

				BS							**oxBS**			
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	41088	3479	4106	8092	0.006	0.0795	0.0734	36286	1968	2203	5094	0.004	0.0853	0.0016
1	30095	2607	2697	5118	0.006	0.078	0.0645	32774	1555	1715	4026	0.004	0.0845	0.0015
3	44382	2819	2953	4769	0.005	0.084	0.0736	35886	1175	1293	2486	0.004	0.0795	0.0913
6	75920	2627	2762	3731	0.005	0.0841	0.0685	54132	965	979	1699	0.004	0.0897	0.083

Tab. 3.3: L1mdT

				BS							**oxBS**			
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	37715	9668	9192	25857	0.007	0.0802	0.0739	30511	6368	5713	19208	0.005	0.0784	0.0729
1	41882	11690	10300	25648	0.008	0.0887	0.0743	43459	6807	5923	17638	0.004	0.0780	0.0738
3	44766	7868	6875	10804	0.007	0.0880	0.0703	31379	2470	2125	4419	0.006	0.0758	0.0683
6	44687	2154	2023	2758	0.006	0.0807	0.0758	56830	1363	1263	2352	0.005	0.0856	0.0714

Tab. 3.4: mSat

				BS							**oxBS**			
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	492	1676	1738	14403	0.004	0.0718	0.0567	315	1170	1221	9804	0.004	0.0663	0.0772
1	448	1337	1495	9029	0.005	0.073	0.0666	568	1678	1748	10654	0.004	0.0727	0.0698
3	1288	1926	2043	10540	0.004	0.0685	0.0642	1171	1602	1697	8746	0.003	0.0685	0.0631
6	3625	2248	2570	11757	0.004	0.0738	0.0605	2618	1619	1604	7471	0.003	0.0725	0.0722

Tab. 3.5: MuERVL

	BS							**oxBS**						
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	111	307	293	1516	0.007	0.0801	0.0695	108	309	295	1381	0.005	0.0646	0.0800
1	149	553	452	1978	0.007	0.0745	0.0632	238	689	597	2607	0.003	0.0492	0.0649
3	448	735	746	2276	0.007	0.1123	0.0740	345	471	420	1262	0.003	0.1321	0.0687
6	798	356	321	702	0.009	0.0606	0.0582	1458	584	470	927	0.003	0.0772	0.0951

Tab. 3.6: Afp

	BS							**oxBS**						
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	1401	5233	4235	31088	0.004	0.0854	0.0852	1208	3652	4307	26568	0.005	0.0982	0.0728
1	2022	6718	4946	25945	0.007	0.0636	0.0646	2821	4367	5366	20886	0.004	0.0836	0.0616
3	4917	4884	5453	14311	0.004	0.0674	0.0765	11285	5443	4739	14034	0.004	0.0636	0.0800
6	29537	6220	6222	14733	0.005	0.0888	0.0523	22516	2989	2182	7421	0.004	0.0638	0.0593

Tab. 3.7: Ttc25

	BS							**oxBS**						
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	16873	5945	6297	22363	0.07	0.0726	0.0751	19490	4338	3926	20641	0.005	0.077	0.1023
1	17013	6342	5340	15431	0.07	0.0625	0.0341	20389	4448	4042	16499	0.006	0.0725	0.0577
3	26107	4950	5705	7472	0.06	0.0813	0.0785	34016	2630	2501	6059	0.004	0.1078	0.058
6	19121	538	627	595	0.06	0.0762	0.059	44122	570	619	1310	0.005	0.0686	0.0933

Tab. 3.8: Zim3

	BS							**oxBS**						
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	14479	11308	13448	63716	0.005	0.065	0.0755	1777	1695	1285	7754	0.007	0.1388	0.1047
1	14295	11947	11222	43046	0.003	0.0717	0.0575	11829	8157	6249	33002	0.007	0.0958	0.0835
3	31291	10020	10965	13864	0.005	0.0666	0.0647	38515	4875	2983	5202	0.008	0.0807	0.0663
6	112883	4761	4100	2434	0.005	0.076	0.0707	1132054	503	457	345	0.006	0.0616	0.0871

Tab. 3.9: Snrpn

	BS							**oxBS**						
day	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{e}$	TT	TC	CT	CC	$\bar{c}$	$\bar{d}$	$\bar{f}$
0	3092	83	109	742	0.0133	0.0757	0.0733	2620	86	125	599	0.0044	0.0785	0.0690
1	3183	100	67	709	0.0135	0.0725	0.0582	3497	48	49	250	0.0038	0.0742	0.0601
3	2571	92	91	557	0.0116	0.0789	0.0717	3357	136	84	503	0.0038	0.0855	0.0731
6	3098	82	98	768	0.0121	0.0779	0.06131	2377	77	127	943	0.0039	0.0759	0.0799

3.6.2 Estimation of Model Parameters

Initial Distribution of the Hidden States

Let $\pi(0)$ be the unknown initial distribution of the hidden states and let $\pi(i,t) = P(\mathcal{X}(t) = i)$ represent the entry of $\pi(t)$ that corresponds to state $i \in S$. In addition, denote by $n_{bs}(j,t)$ and $n_{ox}(j,t)$ the number of times that state $j \in \mathcal{S}_{obs}$ has been observed during independent BS and oxidative BS measurements at time t.

We want to solve the problem: $\pi(0)^* = \arg\max_{\pi(0)} \mathcal{L}_1(\pi(0))$, subject to the constraint $\sum_{i\in\mathcal{S}} \pi(i,0) = 1$, where

$$\mathcal{L}_1(\pi(0)) = \prod_{j\in\mathcal{S}_{obs}} \pi_{bs}(j,0)^{n_{bs}(j,0)} \cdot \pi_{ox}(j,0)^{n_{ox}(j,0)}.$$

We consider the log-likelihood

$$\log \mathcal{L}_1(\pi(0)) = \sum_{j\in\mathcal{S}_{obs}} n_{bs}(j,0) \cdot \log \pi_{bs}(j,0) + n_{ox}(j,0) \cdot \log \pi_{ox}(j,0)).$$

For a gradient descent optimization procedure we need its derivative w.r.t. $\pi(0)$ given by

$$\frac{d}{d\pi(0)} \log \mathcal{L}_1(\pi(0)) = \sum_{j\in\mathcal{S}_{obs}} n_{bs}(j,0) \cdot \frac{\frac{d}{d\pi(0)}\pi_{bs}(j,0)}{\pi_{bs}(j,0)} + n_{ox}(j,0) \cdot \frac{\frac{d}{d\pi(0)}\pi_{ox}(j,0)}{\pi_{ox}(j,0)}.$$

Let $\pi_{bs}(t), \pi_{ox}(t)$ be the vectors with entries $\pi_{bs}(j,t), \pi_{ox}(j,t),\ \forall j \in \mathcal{S}_{obs}, \forall t \in T_{obs}$. Writing the derivatives $\frac{d}{d\pi(0)}\pi_{bs}(j,0)$ and $\frac{d}{d\pi(0)}\pi_{ox}(j,0)$ in vector-matrix notation we get

$$\frac{d}{d\pi(0)}\pi_{bs}(0) = \frac{d}{d\pi(0)}\pi(0) \cdot \mathbf{E}_{bs}(0) = \mathbf{E}_{bs}(0), \quad \frac{d}{d\pi(0)}\pi_{ox}(0) = \frac{d}{d\pi(0)}\pi(0) \cdot \mathbf{E}_{ox}(0) = \mathbf{E}_{ox}(0),$$

which gives us the gradient of the log-likelihood function w.r.t. the initial distribution of the hidden states after insertion into the above equation.

Estimation of the Efficiencies

Let $\mathbf{v} = (\beta_0^{\mu_m}, \beta_1^{\mu_m}, \beta_0^{\mu_d}, \beta_1^{\mu_d}, \beta_0^{\eta}, \beta_1^{\eta}, p)$, be the vector of the seven unknown parameters where μ_m stands for maintenance, μ_d for de novo and η for hydroxylation efficiency, while p is the probability that 5hmC is not considered during maintenance. Recall that we assume that the efficiencies are linear functions of time and so $\mathbf{v}$ contains the coefficients of these functions. E.g. $\mu_m(t) = \beta_0^{\mu_m} + t \cdot \beta_1^{\mu_m}$. The transition matrix of the discrete

Markov chain at time unit t is $\mathbf{P}(t) = \mathbf{D}(t) \cdot \mathbf{M}(t) \cdot \mathbf{H}(t)$, where

$$\mathbf{D}(t) = \begin{array}{cccccccccc} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ uu & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ um & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ mu & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ uh & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ hu & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ hm & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ mh & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ mm & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ hh & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \end{array},$$

$$\mathbf{M}(t) \begin{array}{cccccccccc} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ uu & \bar{\mu}_d^2 & \mu_d \cdot \bar{\mu}_d & \mu_d \cdot \bar{\mu}_d & 0 & 0 & 0 & 0 & \mu_d^2 & 0 \\ um & 0 & \bar{\lambda} & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\ mu & 0 & 0 & \bar{\lambda} & 0 & 0 & 0 & 0 & \lambda & 0 \\ uh & 0 & 0 & 0 & p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda} & 0 & 0 & p \cdot \mu_d + \bar{p} \cdot \lambda & 0 & 0 \\ hu & 0 & 0 & 0 & 0 & p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda} & p \cdot \mu_d + \bar{p} \cdot \lambda & 0 & 0 & 0 \\ hm & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ mh & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ mm & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ hh & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

and

$$\mathbf{H}(t) = \begin{array}{cccccccccc} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ uu & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ um & 0 & \bar{\eta} & 0 & \eta & 0 & 0 & 0 & 0 & 0 \\ mu & 0 & 0 & \bar{\eta} & 0 & \eta & 0 & 0 & 0 & 0 \\ uh & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ hu & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ hm & 0 & 0 & 0 & 0 & 0 & \bar{\eta} & 0 & 0 & \eta \\ mh & 0 & 0 & 0 & 0 & 0 & 0 & \bar{\eta} & 0 & \eta \\ mm & 0 & 0 & 0 & 0 & 0 & \eta \cdot \bar{\eta} & \eta \cdot \bar{\eta} & \bar{\eta}^2 & \eta^2 \\ hh & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}.$$

Note that for $\mathbf{D}(t)$ we can omit the time parameter t since it is time-independent.

Given, now, $\pi(0)$, we want to compute the maximum likelihood estimator (MLE) $\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \log\mathcal{L}_2(\mathbf{v})$, where

$$\mathcal{L}_2(\mathbf{v}) = \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,t)^{n_{bs}(j,t)} \cdot \pi_{ox}(j,t)^{n_{ox}(j,t)}.$$

The only constraint for the above problem is that the efficiencies should be probabilities for all the considered time points, i.e., $0 \leq \beta_0 + \beta_1 \cdot t \leq 1$, $\forall t \in \{0, 6\}$ for all the efficiencies, and the same constraint holds for p, i.e., $0 \leq p \leq 1$.

	bisulfite sequencing				ox. bisulfite sequencing			
	TT	TC	CT	CC	TT	TC	CT	CC
uu	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$
um	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$
mu	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$	$c \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$
uh	$c \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot \bar{e}$	$\bar{c} \cdot e$	$c \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot f$	$\bar{c} \cdot \bar{f}$
hu	$c \cdot \bar{e}$	$\bar{c} \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot e$	$c \cdot f$	$\bar{c} \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot \bar{f}$
hm	$\bar{d} \cdot \bar{e}$	$d \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot e$	$\bar{d} \cdot f$	$d \cdot f$	$\bar{d} \cdot \bar{f}$	$\bar{d} \cdot f$
mh	$\bar{d} \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot \bar{e}$	$d \cdot e$	$\bar{d} \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot f$	$\bar{d} \cdot f$
mm	$\bar{d}^2$	$\bar{d} \cdot d$	$d \cdot \bar{d}$	d^2	$\bar{d}^2$	$\bar{d} \cdot d$	$d \cdot \bar{d}$	d^2
hh	$\bar{e}^2$	$\bar{e} \cdot e$	$e \cdot \bar{e}$	e^2	f^2	$f \cdot \bar{f}$	$f \cdot \bar{f}$	$\bar{f}^2$

Tab. 3.10: Transition probabilities from hidden to the observable states in BS and in oxBS.

It holds

$$\log \mathcal{L}_2(\mathbf{v}) = \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,t) \cdot \log \pi_{bs}(j,t) + n_{ox}(j,t) \cdot \log \pi_{ox}(j,t)$$

and we get the score vector of the log-likelihood function as

$$\frac{d}{d\mathbf{v}} \log \mathcal{L}_2(\mathbf{v}) = \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,t) \cdot \frac{\frac{d}{d\mathbf{v}} \pi_{bs}(j,t)}{\pi_{bs}(j,t)} + n_{ox}(j,t) \cdot \frac{\frac{d}{d\mathbf{v}} \pi_{ox}(j,t)}{\pi_{ox}(j,t)}.$$

Then the matrix-vector form of the derivatives $\frac{d}{d\mathbf{v}} \pi_{bs}(j,t)$ and $\frac{d}{d\mathbf{v}} \pi_{ox}(j,t)$ can be written as

$$\frac{d}{d\mathbf{v}} \pi_{bs}(t) = \frac{d}{d\mathbf{v}} \pi(t) \cdot \mathbf{E}_{bs}(t) \text{ and } \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}} \pi_{\mathrm{ox}}(\mathrm{t}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}} \pi(\mathrm{t}) \cdot \mathbf{E}_{\mathrm{ox}}(\mathrm{t}), \ \forall \mathrm{t} \in \mathrm{T}_{\mathrm{obs}},$$

where the entries of the emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ are given in Table 5.3.

Considering, now, the forward Kolmogorov equation for the discrete Markov chain and its derivative w.r.t. the parameters it suffices to simultaneously solve the following two equation systems.

$$\begin{aligned} \pi(t) &= \pi(t-1) \cdot \mathbf{P}(t) \\ \frac{d}{d\mathbf{v}} \pi(t) &= \frac{d}{d\mathbf{v}} \pi(t-1) \cdot \mathbf{P}(t) + \pi(t-1) \frac{d}{d\mathbf{v}} \mathbf{P}(t), \ \forall t \geq 1 \end{aligned} \tag{3.3}$$

with $\frac{d}{d\mathbf{v}} \pi(0) = 0$ and $\pi(0) = \pi(0)^*$. The derivative of the transition matrix is

$$\frac{d}{d\mathbf{v}} \mathbf{P}(t) = \frac{d}{d\mathbf{v}} (\mathbf{D} \cdot \mathbf{M}(t) \cdot \mathbf{H}(t)) = \mathbf{D} \cdot \left(\frac{d}{d\mathbf{v}} \mathbf{M}(t) \cdot \mathbf{H}(t) + \mathbf{M}(t) \cdot \frac{d}{d\mathbf{v}} \mathbf{H}(t) \right)$$

E.g. applying the chain rule and writing μ_m instead of $\mu_m(\beta_0^{\mu_m}, \beta_1^{\mu_m}, t)$ we get

$$\frac{d}{d\beta_0^{\mu_m}}\mathbf{M}(\mu_m) = \frac{d}{d\mu_m}\mathbf{M}(\mu_m) \cdot \frac{d}{d\beta_0^{\mu_m}}\mu_m = \frac{d}{d\mu_m}\mathbf{M}(\mu_m)$$

and

$$\frac{d}{d\beta_1^{\mu_m}}\mathbf{M}(\mu_m) = \frac{d}{d\mu_m}\mathbf{M}(\mu_m) \cdot \frac{d}{d\beta_1^{\mu_m}}\mu_m = \frac{d}{d\mu_m}\mathbf{M}(\mu_m) \cdot t.$$

In the same way we get the first derivatives w.r.t. all the other components of parameter vector $\mathbf{v}$. Applying once more the product rule in Eq. (5.3), and using similar arguments as above we can also compute the second partial derivatives $\frac{d}{d\mathbf{v}_i d\mathbf{v}_j} \log \mathcal{L}_2(\mathbf{v})$ which will give us the (i, j)-th entry of the Hessian matrix $\mathcal{H} = \nabla\nabla^{\mathrm{T}} \log \mathcal{L}_2(\mathbf{v})$.

Standard Deviations and Confidence Intervals

The observed Fisher information is defined as $\mathcal{J}(\mathbf{v}^*) = -\mathcal{H}(\mathbf{v}^*)$, where $\mathbf{v}^*$ is the maximum likelihood estimator. The expected Fisher information is $\mathcal{I}(\mathbf{v}) = \mathbb{E}[\mathcal{J}(\mathbf{v})]$ and its inverse is a lower bound for the covariance matrix of the MLE. Thus, here we approximate the standard deviations of the estimates as $\sigma(\mathbf{v}^*) = \sqrt{\mathrm{Var}(\mathbf{v}^*)} = \sqrt{\mathrm{diag}(-\mathcal{H}^{-1}(\mathbf{v}^*))}$. In order to approximate the standard deviations of the efficiencies over time, i.e. $\sigma(\mu_m(t)), \sigma(\mu_d(t))$ and $\sigma(\eta(t))$, we exploit the fact that if $f(t) = \beta_0 + \beta_1 \cdot t$ then $\mathrm{Var(f(t))} = \mathrm{Var}(\beta_0 + \beta_1 \cdot \mathrm{t}) = \mathrm{Var}(\beta_0) + \mathrm{t}^2\mathrm{Var}(\beta_1) + 2\mathrm{tCov}(\beta_0, \beta_1)$.

Given, now, the variances of the estimated efficiencies we can compute the variance $\lambda(t)$, for any t as

$$\mathrm{Var}(\lambda) = \mathrm{Var}(\mu_\mathrm{m}) + \mathrm{Var}(\mu_\mathrm{d}) + 2\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{d}) + \mathrm{Var}(\mu_\mathrm{m}\mu_\mathrm{d}) - 2\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{m}\mu_\mathrm{d}) - 2\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{m}\mu_\mathrm{d}),$$

where the last four terms are computed as follows:

$$\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{d}) = \mathrm{Cov}(\beta_0^{\mu_\mathrm{m}}, \beta_0^{\mu_\mathrm{d}}) + \mathrm{tCov}(\beta_0^{\mu_\mathrm{m}}, \beta_1^{\mu_\mathrm{d}}) + \mathrm{tCov}(\beta_1^{\mu_\mathrm{m}}, \beta_0^{\mu_\mathrm{d}}) + \mathrm{t}^2\mathrm{Cov}(\beta_1^{\mu_\mathrm{m}}, \beta_1^{\mu_\mathrm{d}}),$$

and

$$\mathrm{Var}(\mu_\mathrm{m}\mu_\mathrm{d}) = \mathbb{E}[\mu_\mathrm{m}^2\mu_\mathrm{d}^2] - \mathbb{E}[\mu_\mathrm{m}\mu_\mathrm{d}]^2 \tag{3.4}$$

$$\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{m}\mu_\mathrm{d}) = \mathbb{E}[\mu_\mathrm{m}^2\mu_\mathrm{d}] - \mathbb{E}[\mu_\mathrm{m}]\mathbb{E}[\mu_\mathrm{m}\mu_\mathrm{d}], \tag{3.5}$$

$$\mathrm{Cov}(\mu_\mathrm{d}, \mu_\mathrm{m}\mu_\mathrm{d}) = \mathbb{E}[\mu_\mathrm{d}^2\mu_\mathrm{m}] - \mathbb{E}[\mu_\mathrm{d}]\mathbb{E}[\mu_\mathrm{m}\mu_\mathrm{d}] \tag{3.6}$$

Since the MLEs are approximately normally distributed and for any two random variables X, Y, $\mathbb{E}[XY] = \mathrm{Cov(X, Y)} + \mathbb{E}[\mathrm{X}]\mathbb{E}[\mathrm{Y}]$, we get

$$\mathbb{E}[\mu_m^2\mu_d] = E[\mu_m]^2E[\mu_d] + \mathrm{Var}(\mu_\mathrm{m})\mathrm{E}[\mu_\mathrm{d}] + 2\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{d})\mathrm{E}[\mu_\mathrm{m}]$$

$$\mathbb{E}[\mu_d^2\mu_m] = E[\mu_d]^2E[\mu_m] + \mathrm{Var}(\mu_\mathrm{d})\mathrm{E}[\mu_\mathrm{m}] + 2\mathrm{Cov}(\mu_\mathrm{m}, \mu_\mathrm{d})\mathrm{E}[\mu_\mathrm{d}]$$

$$\begin{aligned}\mathbb{E}[\mu_m^2\mu_d^2] &= E[\mu_m]^2E[\mu_d]^2+\mathrm{Var}(\mu_\mathrm{m})\mathrm{Var}(\mu_\mathrm{d})+\mathrm{Var}(\mu_\mathrm{d})\mu_\mathrm{m}^2+\mathrm{Var}(\mu_\mathrm{m})\mathrm{E}[\mu_\mathrm{d}]^2\\ &\quad +2\mathrm{Cov}(\mu_\mathrm{m},\mu_\mathrm{d})^2+4\mathrm{Cov}(\mu_\mathrm{m},\mu_\mathrm{d})\mathrm{E}[\mu_\mathrm{m}]\mathrm{E}[\mu_\mathrm{d}],\end{aligned}$$

where the expectations and thus all terms in Eq. (3.4) - (3.6) are now known. Obtaining this way the standard deviations of all the efficiencies over time one can create the corresponding confidence intervals for a fixed confidence level, here $\beta = 95\%$ was chosen. For instance the confidence interval for the total methylation on hemimethylated sites will be

$$\lambda \pm z\cdot\sigma(\lambda) = \lambda \pm z\cdot\sqrt{\mathrm{Var}(\lambda)},$$

where $z = F^{-1}\left(\frac{\beta+1}{2}\right)$ and F is the cummulative distribution function (cdf) of the standard normal distribution. Similarly we get the confidence intervals for all remaining parameters.

3.6.3 Hypothesis Test

We carried out a number of hypotheses tests related to the estimated parameters (see Section S.3). Here we briefly describe the details of the Wald statistic that we used to conduct these tests.

Given a maximum likelihood estimate $\mathbf{v}^*$ of an unknown parameter vector $\mathbf{v}_0 \in V \subseteq \mathbb{R}^p$ we want to test the null hypothesis H_0 that $g(\mathbf{v}_0) = 0$, where $g : \mathbb{R}^p \to \mathbb{R}^r$ is a vector valued function with $r \leq p$. We define the Wald statistic for this estimate as

$$w = g(\mathbf{v}^*)^\intercal\left[J_g(\mathbf{v}^*)\cdot\widehat{\Sigma}\cdot J_g(\mathbf{v}^*)^\intercal\right]^{-1}g(\mathbf{v}^*),$$

where $J_g(\mathbf{v}^*)$ is the Jacobian of g, i.e., the $r \times p$ matrix of the partial derivatives of the entries of g with respect to the entries of $\mathbf{v}$, and $\widehat{\Sigma}$ is a consistent estimate of the assymptotic covariance matrix, here equal to the negative Hessian, of $\mathbf{v}^*$. Note that w here is a realization of a random variable $W_{\mathbf{v}^*}$ as it is a function of $\mathbf{v}^*$ which is a random variable itself depending on the observed data.

Under the regularity assumptions that for all $\mathbf{v} \in V$, the entries of g are continuously differentiable w.r.t. all entries of $\mathbf{v}$ and that $J_g(\mathbf{v})$ has rank r, the following holds. If the null hypothesis is true, i.e. $g(\mathbf{v}_0) = 0$, then the Wald statistic $W_{\mathbf{v}^*}$ converges to a Chi-square distribution with r degrees of freedom [26].

Thus, conducting the Wald test consists of comparing the Wald statistic with a critical threshold $z = F^{-1}(1-a)$, where F is the cdf of a Chi-square random variable with r degrees of freedom and a is a predefined confidence level, e.g. $a = 1\%$. If $w > z$ then we can safely reject the null hypothesis. The p-value of the test is the probability p $= P(W_{\mathbf{v}^*} > w) = 1 - P(W_{\mathbf{v}^*} \leq w) \approx 1 - F(w)$ and so equivalently one also rejects the null hypothesis if p $\leq a$.

For estimates taken from maximum likelihood alternative tests, such as likelihood ratio or score test, are also possible. The Wald statistic, however, is convenient in case of testing multiple hypotheses in parallel. In addition, the use of all tests mentioned before for our estimates returned similar p-values and did not lead to a different result regarding the cases that one rejects H_0.

3.6.4 Results

In Table 3.11 we present the MLEs returned by our global optimization routine for the parameter vector $\mathbf{v}$ and the corresponding vector of standard deviations $\sigma(\mathbf{v})$, given the data of section S.1 for each of the five genome regions. The p-value of the efficiencies μ_m, μ_d and η corresponds to the null hypothesis $H_0 : \beta_1 = 0$, where β_1 is the gradient of the corresponding efficiency, and for the total methylation λ it takes the form $H_0 : \beta_1^\lambda = 0 \wedge \beta_2^\lambda = 0$. On the other hand, the null hypothesis for the probability p, that 5hmC is not considered during maintenance, is $H_0 : p = 1$. The confidence level α has been set to 1% for deciding H_0 for each of the above parameters.

In Table 3.12 we show the computed coefficients of the quadratic total methylation $\lambda(t)$, which can be implicitly taken from the maintenance and de novo estimated coefficients.

In Figure 3.8 we see the predicted probabilities of the observable states that have been taken using the estimated values of Table 3.11 for each region. We compare them to the measured data (frequency) at the various days. Figure 3.9 shows the predicted probabilities of the hidden states and the detailed hydroxylation levels, as well as the estimated (hydroxy-)methylation efficiencies over time for the regions IAP, L1mdA, MuERVL, Zim3 and Snrpn that do not appear in the manuscript.

In order to measure the test error of the model we performed leave one out cross validation (LOOCV) and tested two competing assumptions: “1) The enzyme efficiencies are constant” and “2) The enzyme efficiencies can also be linear”. For each region we tested the prediction of the model for each single CpG, having trained it on the data of the other CpGs and we averaged at the end the test error. For comparing the prediction ability of the model for each of the two cases 1) and 2) we used two different distribution distance measures (Kullback-Leibler divergence and Bhattacharyya distance) between the output distribution and the data. Our results in Table 3.13 show that for all regions the test error (i.e. the above distance) becomes evidently smaller for the case where we allow efficiencies to be linear over time. In the two columns where we report the improvement (”gain”) $\frac{\text{KL-const - KL-linear}}{\text{KL-const}}$ of the test error, we see that the decrease of the test error using the linear model over the constant varies from 0.6% (in mSat) to 38.3% (in Zim3) for the Kullback-Leibler distance. The predictive potential of the model, and consequently the above gain ratio, depends on the available number of CpGs for the training data and on how much the efficiencies deviate from constant behavior over time.

In Figures 3.10, 3.11 we show the (hydroxy-)methylation efficiencies and the (hydroxy-

)methylation levels for all CpGs of all the examined loci, in case the data of each locus is not aggregated and separate estimations are taken for each of the single CpG dyads. Although, the absolute (hydroxy-)methylation levels at distinct CpGs can be slightly different, one observes that the tendency of the demethylation process has clearly homogeneous characteristics between CpGs of the same locus. Particularly, the increase of the hydroxylation level in relation to the methylated substrates is always present. Also, the day with the highest absolute 5hmC level is, in the majority of the cases, the same for the CpGs of a locus. Similarly, the predicted behavior of the enzymes' efficiencies within a locus is in principle homogeneous with some differences in the absolute estimated values that come with larger confidence intervals due to the smaller number of samples.

Finally, to validate the robustness of the model sensitivity analysis of the parameters has been examined. Perturbing one parameter at a time (OAT) by ±1% we get a maximum (over all regions, time points and parameters) absolute change of 0.0053 for the total hydroxylation level and 0.0198 for the total methylation level.

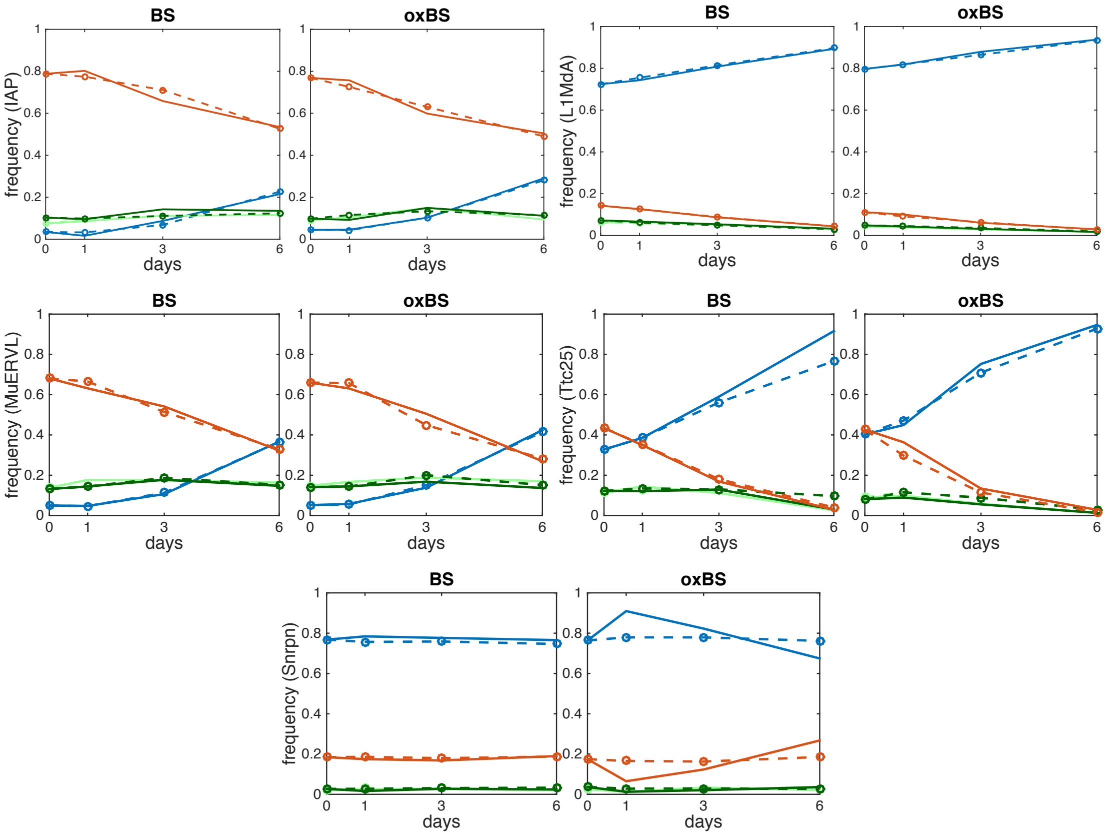


Fig. 3.8: Comparison of prediction and data for IAP, L1mdA, MuERVL, Ttc25 and Snrpn: probabilities of the observable states TT (blue), TC (light green), CT (dark green), CC (red) in BS and oxidative BS. The solid lines show the experimentally measured frequencies states and the dashed lines correspond to the values predicted by the two HMMs.

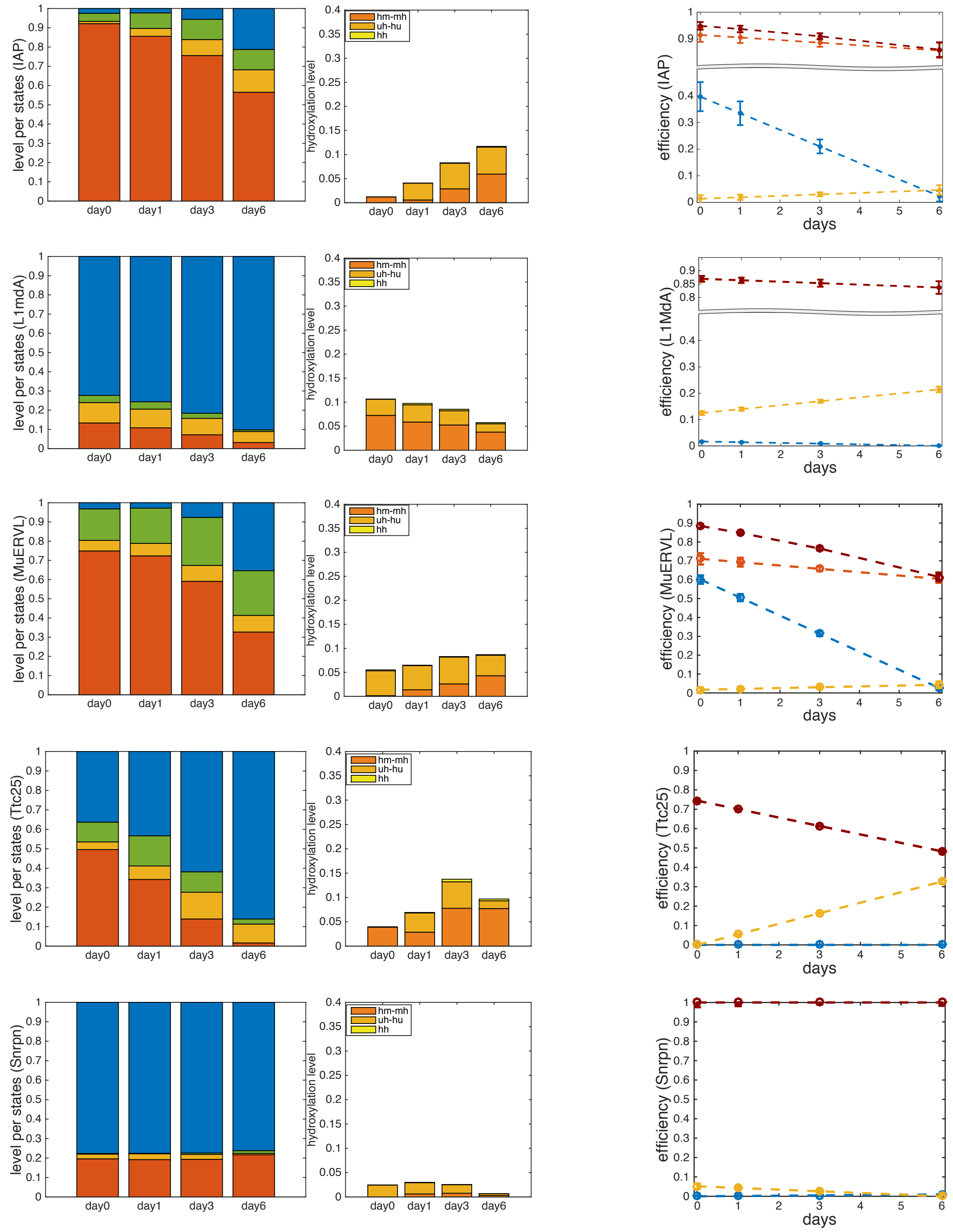


Fig. 3.9: Results for regions IAP, L1mdA, MuERVL, Ttc25 and Snrpn: Left: Probabilities of the hidden states. The amount of fully methylated (mm) 🟥, hemimethylated (um and mu) 🟩, and unmethylated (uu) 🟦 sites. The orange block 🟧 gives the total amount of CpG sites with at least one 5hmC, while the detailed distribution of the hydroxylated states is given by the diagram on the right. Right: Estimated efficiencies and standard deviations over time. Maintenance (red), de novo (blue), hydroxylation (yellow) and total efficiency on a hemimethylated CpG (dark red).

Tab. 3.11: Estimated coefficients of the functions $\mu_d(t), \mu_m(t)$ and $\eta(t)$ and their approximate standard deviations. The p-values have been taken conducting a hypothesis test $H_0 : \beta_1 = 0$ using the Wald statistic.

IAP: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.9155	0.0256	-0.0097	0.0072	0.180
μ_d	0.3977	0.0545	-0.0624	0.0106	$< 10^{-5}$
η	0.0134	0.0132	0.0055	0.0045	0.226
p	1.0000	0.2577	-	-	1

L1mdA: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.8682	0.0104	-0.0052	0.0040	0.190
μ_d	0.0168	0.0007	-0.0027	0.0002	$< 10^{-5}$
η	0.1249	0.0074	0.0149	0.0023	$< 10^{-5}$
p	1	0.0238	-	-	1

L1mdT: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.7317	0.0040	-0.0102	0.0044	0.020
μ_d	0.0229	0.0010	-0.0038	0.0002	$< 10^{-5}$
η	0.1013	0.0046	0.0220	0.0015	$< 10^{-5}$
p	1	0.0468	-	-	1

mSat: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.8304	0.0080	0.0026	0.0019	0.186
μ_d	0.3879	0.0133	-0.0478	0.0025	$< 10^{-5}$
η	0.0002	0.0038	0.0026	0.0011	0.024
p	0.8025	0.1966	-	-	0.315

MuERVL: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.7106	0.0300	-0.0177	0.0076	0.019
μ_d	0.6006	0.0221	-0.0955	0.0039	$< 10^{-5}$
η	0.0172	0.0119	0.0044	0.0045	0.336
p	0.5428	0.2858	-	-	0.11

Afp: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.7817	0.0041	0.0006	0.0015	0.717
μ_d	0.1772	0.0058	-0.0295	0.0011	$< 10^{-5}$
η	0.0473	0.0028	0.0160	0.0010	$< 10^{-5}$
p	1.000	0.0208	-	-	1

Ttc25: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.7440	0.0064	-0.0435	0.0003	$< 10^{-5}$
μ_d	0.0000	0.0018	-0.0000	0.0003	1
η	0.0000	0.0072	0.0544	0.0023	$< 10^{-5}$
p	1.000	0.0670	-	-	1

Zim3: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	0.8530	0.0027	-0.0965	0.0014	$< 10^{-5}$
μ_d	0.0000	0.0022	-0.0000	0.0005	1
η	0.0000	0.0087	0.0922	0.0047	$< 10^{-5}$
p	1.000	0.0255	-	-	1

Snrpn: (hydroxy)					
methylation prob.	β_0	$\sigma(\beta_0)$	β_1	$\sigma(\beta_1)$	p-value
μ_m	1.0000	0.0253	0.0000	0.0076	1
μ_d	0.0000	0.0029	0.0016	0.0008	0.047
η	0.0517	0.0170	-0.0086	0.0038	0.030
p	0.5	0.0807	-	-	1

Tab. 3.12: Estimated coefficients of the function $\lambda(t)$ and their approximate standard deviations. The p-values have been taken conducting a hypothesis test $H_0 : \beta_1^\lambda = 0 \wedge \beta_2^\lambda = 0$ using the Wald statistic.

DNA region	β_0^λ	β_1^λ	β_2^λ	p-value
IAP	0.9491	-0.0111	$6.05 \cdot 10^{-4}$	$< 10^{-5}$
L1mdA	0.8705	-0.0055	$1.40 \cdot 10^{-5}$	0.187
L1mdT	0.7378	-0.0011	$3.89 \cdot 10^{-5}$	0.005
mSat	0.8962	-0.0065	$1.21 \cdot 10^{-4}$	$< 10^{-5}$
MuERVL	0.8440	-0.0347	$1.69 \cdot 10^{-3}$	$< 10^{-5}$
Afp	0.8203	0.0059	$1.68 \cdot 10^{-5}$	$< 10^{-5}$
Ttc25	0.7440	-0.0435	$-2.95 \cdot 10^{-14}$	$< 10^{-5}$
Zim3	0.8530	-0.0965	$-1.16 \cdot 10^{-14}$	$< 10^{-5}$
Snrpn	1.0000	$-2.89 \cdot 10^{-11}$	$-4.44 \cdot 10^{-14}$	1.000

Tab. 3.13: Computed Kullback-Leibler divergence and Bhattacharya distance values given by LOOCV data to compare the test error for assuming linear vs constant efficiencies.

DNA region	KL-const	KL-linear	KL gain	BC-const	BC-linear	BC gain
IAP	0.164	0.131	20.1 %	5.33e-03	4.38e-03	17.8 %
L1mdA	0.026	0.023	11.5 %	8.10e-04	7.18e-04	11.4 %
L1mdT	0.101	0.099	1.9 %	3.18e-03	3.17e-03	0.3 %
mSat	0.163	0.162	0.6 %	5.09e-03	5.00e-03	1.8 %
MuERVL	0.497	0.321	35.4 %	1.62e-02	1.02e-02	37.0 %
Afp	0.149	0.114	23.5 %	4.79e-03	3.66e-03	23.6 %
Ttc25	0.209	0.171	18.2 %	7.03e-3	6.07e-3	13.7 %
Zim3	0.342	0.211	38.3 %	1.13e-2	7.00e-3	38.1 %
Snrpn	0.194	0.192	1 %	1.13e-2	7.00e-3	1 %

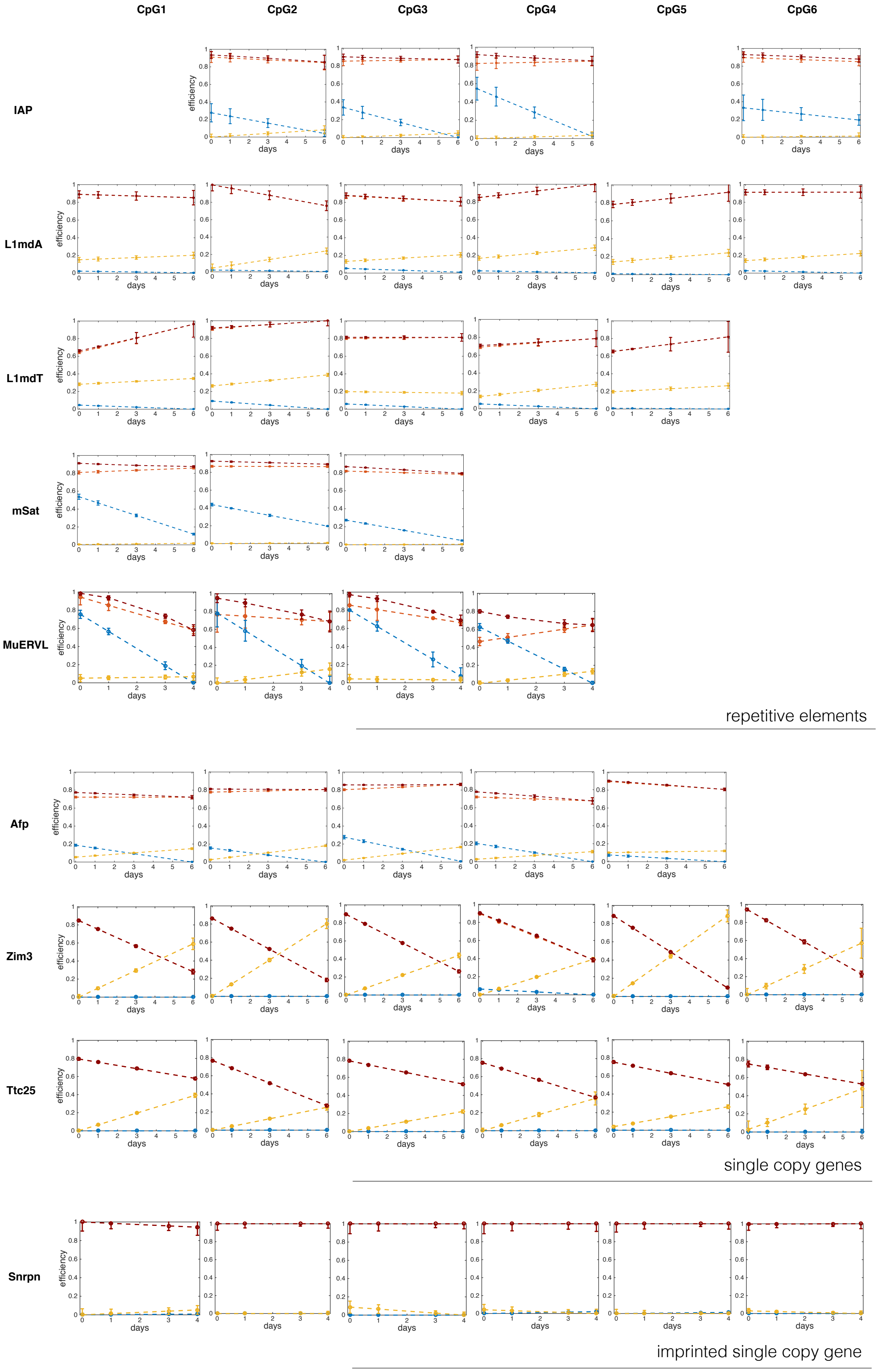


Fig. 3.10: Estimated efficiencies and standard deviations for each single CpG dyad of regions IAP, L1mdA, L1mdT, mSat, MuERVL and the single copy genes Afp, Ttc25, Zim3, Snrpn over time. In the case of IAP we cover six CpG positions. However, during evolution CpG one and five underwent a transition resulting in a loss of the CpG positions in this particular IAP class. Furthermore, due to the lack of space we only show the first 6 CpGs out of 13 CpGs analyzed in L1mdA and out of 8 CpGs analyzed in Zim3. The colors are the same as in Fig. 3.9 (right)

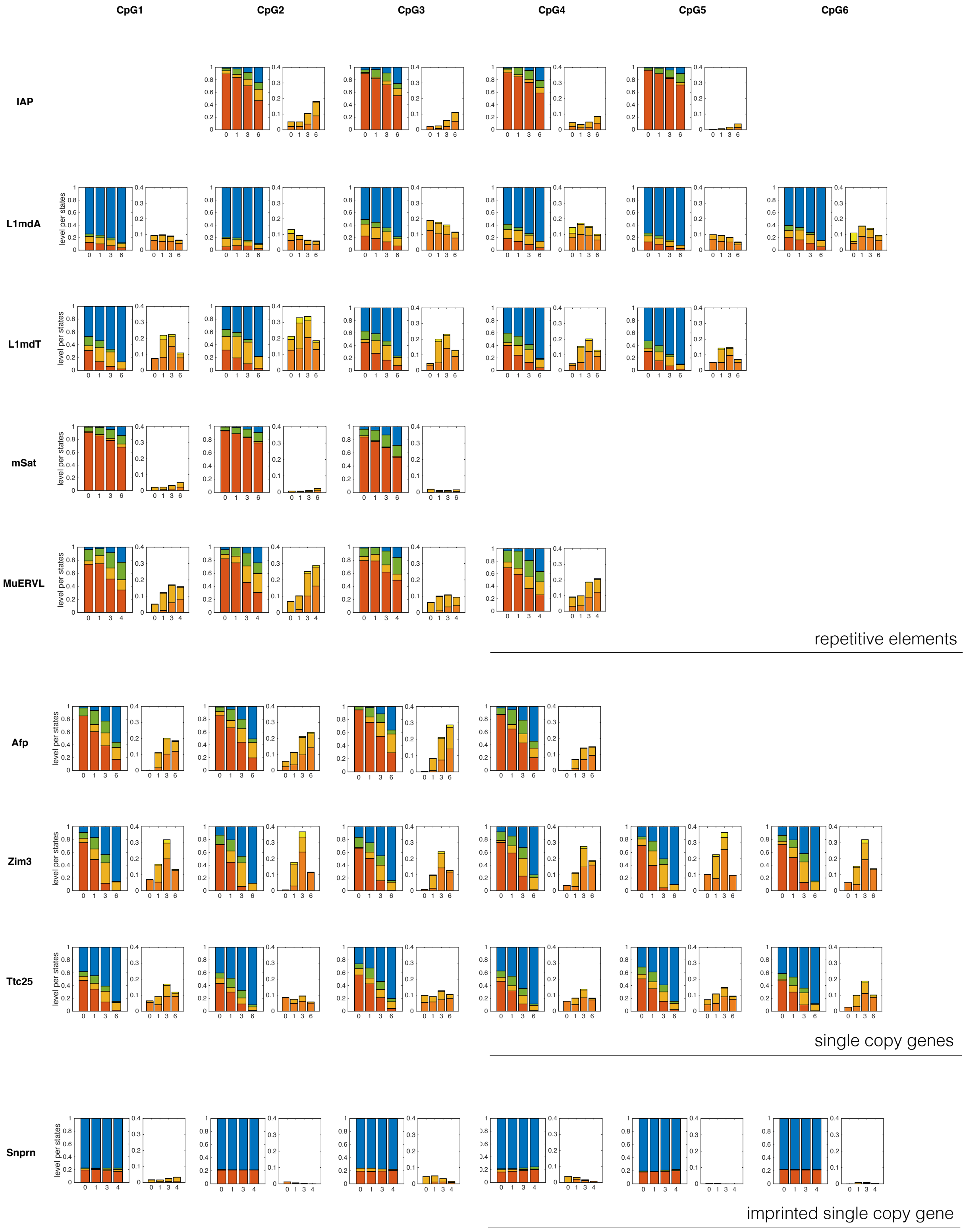


Fig. 3.11: Prediction of the (hydroxy-)methylation levels for each single CpG dyad of regions IAP, L1mdA, L1mdT, mSat, MuERVL and the single copy genes Afp, Ttc25, Zim3, Snrpn over time. The colors are the same as in Fig. 3.9 (left)

3.6.5 Hairpin Oxidative Bisulfite Sequencing

500 ng of mESC DNA was cleaved with 10 units of restriction enzymes for 5h in a 30µl reaction. For IAP L1mdA the DNA was cut with DdeI (New England Biolabs; NEB), for mSat and MuERVL with Eco47I (Thermo Fisher Scientific), Afp, Ttc25, Zim3 with TaqI (Thermo Fisher Scientific) and in case of Snrpn with NlaIII (NEB). The restriction was stopped by a 20 min heat inactivation at 80°C. The restricted DNA was then subjected to a 16 h or overnight ligation with T4-DNA Ligase (New England Biolabs). 200 units of T4-DNA Ligase, 4 µl 10mM ATP and 1µl 100 µM hairpin linker was added directly into the restriction reaction and the volume was adjusted to 40 µl using ddH2O. During ligation the hairpin linker becomes covalent attached to the restriction site of the DNA. Purification and oxidative BS treatment was carried out using the chemicals and protocols provided by Cambridge Epigenetix. Amplicons were generated by PCR using Hotfire Taq polymerase from Solis Biodyne. Sequencing was carried out using the MiSeq Illumina system (paired end sequencing 2x250bp reads). After Sequecning in a first informatics step the adapter sequence is removed from the reads (Trimming). The resulting read information is then analyzed analyzed using the BiQAnalyzerHT and a python script. For the repeats the sequences were filtered by sequence identity score, meaning that only reads which matched the reference sequence to at least 80% were used for the analysis. For single copy genes this score was set to 90% and in addition only reads with maximum 10% missing CpG sites were analyzed.

Primer- and Reference sequences

Table 3.14 shows the sequence of the nine different hairpin linkers used to covalent link both DNA strands. We included unmodified cytosine, 5mC(X) and 5hmC(y) into the hairpin linker to follow the conversion of these modifications during BS and oxBS treatment. Mapping the sequencing information to this reference sequences we determine the states of each cytosine which allows us to calculate all possible measurement errors for each time point and each genomic region. For example: 5hmC should be converted after oxBS treatment to 5fU and will after sequencing seen as T. We check for each sequenced hairpin molecule the state of the 5hmC position which can be either C or T. We divide then the number of T by the total number of T and C at this position to get the conversion error of 5hmC during oxBS treatment. The conversion error for cytosine and 5mC is calculated in the same way. For Snrpn we had to use a hairpin linker without 5mC or 5hmC and could therefore not calculate the conversion errors for this sample probably. However, to correct for more general errors we used the mean conversion error of all other loci. In addition table C and table D give the primer sequences and the corresponding reference sequence for each regions respectively.

Tab. 3.14: Sequence of the hairpin linker for Afp, L1mdT, L1mdA, mSat, IAP; *M* indicates the localization of 5mC, *H* the position of 5hmC in the sequence. All hairpin linker carry a 5'-phosphorylation.

Hairpin	Linker Sequence
IAP-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC
L1mdA-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC
L1mdT-HP	*Pho*-CCGGAGGG*M*CCATDDDDDDDDATGGG*H*CCT
mSat-HP	*Pho*-GNCGGG*M*CCATDDDDDDDDATGGG*H*CC
MuERVL-HP	*Pho*-GNCGGG*M*CCATDDDDDDDDATGGG*H*CC
Afp-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
Ttc25-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
Zim3-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
Snrpn-HP	*Pho*-GGGCCTADDDDDDDDTAGGCCCCATG

Tab. 3.15: Primer for amplification of the analyzed regions after bisulfite and oxidative bisulfite treatment.

Primer	Sequence
IAP-HP-Forward	TTTTTTTTTTAGGAGAGTTATATTT
IAP-HP-Revers	ATCACTCCCTAATTAACTACAAC
L1mdA-HP-Forward	GTGAGTGGATTATAGTGTTTGTTTTAA
L1mdA-HP-Revers	AAATAAATCACAATACCTACCCCAAT
L1mdT-HP-Forward	TGGTAGTTTTTAGGTGGTATAGAT
L1mdT-HP-Revers	TCAAACACTATATTACTTTAACAATTCCCA
mSat-HP-Forward	GGAAAATTTAGAAATGTTTAATGTAG
mSat-HP-Revers	AACAAAAAAACTAAAAATCATAAAAA
MuERVL-HP-Forward	TAAGGGTTAGGTGGTAGTATTGAAT
MuERVL-HP-Revers	CAAAAACCAAATAACAACATTAAAT
Afp-HP-Forward	TTTTGTTATAGGAAAATAGTTTTTAAGTTA
Afp-HP-Revers	AAATCACAAAACATCTTACCTATCC
Ttc25-HP-Forward	TGAAAGAGAATTGATAGTTTTTAGG
Ttc25-HP-Revers	AAAACAAAAATCTATTCCATCACTC
Zim3-HP-Forward	TTTATTTATTTGTGTGTGGTTTTTG
Zim3-HP-Revers	CACATATCAAAATCCACTCACCTAT
Snrpn-HP-Forward	AGAATTTATAAGTTTAGTTGATTTTTT
Snrpn-HP-Revers	TAATCAAATAAAATACACTTTCACTACT

TGTCACTCCCTGATTGGCTGCAGCCCATCGGCCGAGTTGACGTCACGGGGAAGGCAGAGCACATGGAGTAGAGAACCACCCTC
GGCATATGCGCAGATTATTTGTTTACCAC**TNAGGGMCCATDDDDDDDDATGGGHCC**TAAGTGGTAAACAAATAATCTGCGCAT
ATGCCGAGGGTGGTTCTCTACTCCATGTGCTCTGCCTTCCCCGTGACGTCAACTCGGCCGATGGGCTGCAGCCAATCAGGGAG
TGACA

Fig. 3.12: Reference Sequence used for 5mC and 5hmC analysis of IAP; M = 5mC, H = 5hmC

TCCAATCGCGCGGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTC
GACTCGAGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGCCCACACGGGACTCCCCACGGG
ACCC**TNAGGGMCCATDDDDDDDDATGGGHCC**TNAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTTGGGCAAGAC
TCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTGCCCCAGATCAGGCCCGGGTAGCCTGCT
TCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGATTGGGGCAGGCACTGTGATCCACTC

Fig. 3.13: Reference Sequence used for 5mC and 5hmC analysis of L1mdA; M = 5mC, H = 5hmC

CCCGGGACCAAGATGGCGACCGCTGCTGCTGTGGCTTAGGCCGCCTCCCCAGCCGGGTGGGCACCTGT
CCT**CCGGAGGGMCCATDDDDDDDDATGGGHCCT**CCGGAGGACAGGTGCCCACCCGGCTGGGGAGGCGG
CCTAAGCCACAGCAGCAGCGGTCGCCATCTTGGTCCCGGG

Fig. 3.14: Reference Sequence used for 5mC and 5hmC analysis of L1mdT; M = 5mC, H = 5hmC

GGAAAATTTAGAAATGTTTAATGTAGGACGTGGAATATGGCAAGAAAACTGAAAATCATGGGAAATGA
GAAACATCCACTTGTCGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGAGAAATGCACA
CTGAAG**GWCGGGMCCATDDDDDDDDATGGGHCC**GWCCTTCAGTGTGCATTTCTCATTTTTCACGTTTT
TTAGTGATTTCGTCATTTTTCAAGTCGACAAGTGGATGTTTCTCATTTTTTATGATTTTTAGTTTTTT
TGTT

Fig. 3.15: Reference Sequence used for 5mC and 5hmC analysis of mSat; M = 5mC, H = 5hmC

CGCCCGAGACAAGGTGATTCTAGTTATTATAATGGACAGCGTAGACAAAAGAATGTTTATAATAACAT
ACCCAGTAATGGTCAGCACAGGAGAGGTGAAATTTATAATGGCATGACTCGGTTG**GWCGGGMCCATDD**
DDDDDDATGGGHCCGWTTCAACCGAGTCATGCCATTATAAATTTCACCTCTCCTGTGCTGACCATTAC
TGGGTATGTTATTATAAACATTCTTTTGTCTACGCTGTCCATTATAATAACTAGAATCACCTTGTCTC
GGGCG

Fig. 3.16: Reference Sequence used for 5mC and 5hmC analysis of MuERVL; M = 5mC, H = 5hmC

TTTTGTTATAGGAAAATAGTTTTTAAGTTACAAAGCATCTTACCTATCCCAAACTCATTTTCGTGCAA
TGCTTTGGACGCAGCGAAATGTAGCAGGAGGATGAGGGAAGCGGGTGTGATCCACTTCATGGCTGCTG
GTTCCTTCACCGCAGGCAGTGCTGGAAGTGGGATGTTT**CGGGGMCCATDDDDDDDDATGGGHCC**CGAA
ACATCCCACTTCCAGCACTGCCTGCGGTGAAGGAACCAGCAGCCATGAAGTGGATCACACCCGCTTCC
CTCATCCTCCTGCTACATTTCGCTGCGTCCAAAGCATTGCACGAAAATGAGTTTGGGATAGGTAAGAT
GtTTTGTGATTT

Fig. 3.17: Reference Sequence used for 5mC and 5hmC analysis of Afp; M = 5mC, H = 5hmC

CCAGTAGATCCTCAGCTGGGGGCAGGGATCTATTCCATCACTCCCCTTCCGTGTCGGGATTTCGTGCA
GCTCAGACGGGTCCAAGTCTTACACAAGCTGTCCTAACTGCTGTGCGTTTATATAACAACTACCCGGT
TGTGTTTAGAAAACACTGTTTT**CGGGGMCCATDDDDDDDDATGGGHCC**CGAAAACAGTGTTTTCTAAA
CACAACCGGGTAGTTGTTATATAAACGCACAGCAGTTAGGACAGCTTGTGTAAGACTTGGACCCGTCT
GAGCTGCACGAAATCCCGACACGGAAGGGGAGTGATGGAATAGATCCCTGCCCC

Fig. 3.18: Reference Sequence used for 5mC and 5hmC analysis of Ttc25; M = 5mC, H = 5hmC

CCCGGCCACCATAGTCGGATTATCCGTGGGCGGGGTGAGATGGACGGAGCGCCTTGCAGACCTCAGGA
AAACCTCCCCACGCCTGTCCGGCCTTGGCTTGGTGACAGGGAAACTGGCTGGACT**CGGGGMCCATDDD**
DDDDDATGGGHCCCGAGTCCAGCCAGTTTCCCTGTCACCAAGCCAAGGCCGGACAGGCGTGGGGAGGT
TTTCCTGAGGTCTGCAAGGCGCTCCGTCCATCTCACCCCGCCCACGGATAATCCGACTATGGTGGCCG
GGCAAGGACCACAC

Fig. 3.19: Reference Sequence used for 5mC and 5hmC analysis of Zim3; M = 5mC, H = 5hmC

AGAATTTACAAGTTTAGTTGATTTTTTTCGCTCCATTGCGTTGCAAATCACTCCTCAGAACCAAGCGT
CTGGCATCTCCGGCTCCCTCTCCTCTCTGCGCTAGTCTTGCCGCAATGGCTCAGGTTTGTCGCGCGGC
TCCCTACGCATG**GGGCCTADDDDDDDDTAGGCCCCATG**CGTAGGGAGCCGCGCGACAAACCTGAGCCA
TTGCGGCAAGACTAGCGCAGAGAGGAGAGGGAGCCGGAGATGCCAGACGCTTGGTTCTGAGGAGTGAT
TTGCAACGCAATGGAGCGAGGAAGGTCAGCTGGGCTTGTGGATTCTAGTAGTGAAAGTGTATTTTATT
TGATTA

Fig. 3.20: Reference Sequence used for 5mC and 5hmC analysis of Snrpn; M = 5mC, H = 5hmC

Bibliography

[1] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96, 2004.

[2] En Li, Timothy H Bestor, and Rudolf Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

[3] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[4] Adrian Bird, Mary Taggart, Marianne Frommer, Orlando J Miller, and Donald Macleod. A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, 40(1):91–99, 1985.

[5] Melanie Ehrlich, Miguel A Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C Kuo, Roy A McCune, and Charles Gehrke. Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research*, 10(8):2709–2721, 1982.

[6] Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature genetics*, 19(3):219, 1998.

[7] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 2004.

[8] Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.

[9] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908, 2007.

[10] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[11] Junjie U Guo, Yijing Su, Chun Zhong, Guo-li Ming, and Hongjun Song. Hydroxylation of 5-methylcytosine by tet1 promotes active dna demethylation in the adult brain. *Cell*, 145(3):423–434, 2011.

[12] Konstantin Lepikhov, Mark Wossidlo, Julia Arand, and Joern Walter. Dna methylation reprogramming and dna repair in the mouse zygote. *International Journal of Developmental Biology*, 54(11-12):1565–1574, 2011.

[13] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[14] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[15] Liang Zhang, Xingyu Lu, Junyan Lu, Haihua Liang, Qing Dai, Guo-Liang Xu, Cheng Luo, Hualiang Jiang, and Chuan He. Thymine dna glycosylase specifically recognizes 5-carboxylcytosine-modified dna. *Nature chemical biology*, 8(4):328, 2012.

[16] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[17] Hideharu Hashimoto, Yiwei Liu, Anup K Upadhyay, Yanqi Chang, Shelley B Howerton, Paula M Vertino, Xing Zhang, and Xiaodong Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849, 2012.

[18] Mark Wossidlo, Toshinobu Nakamura, Konstantin Lepikhov, C Joana Marques, Valeri Zakhartchenko, Michele Boiani, Julia Arand, Toru Nakano, Wolf Reik, and Jörn Walter. 5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications*, 2:241, 2011.

[19] Tian-Peng Gu, Fan Guo, Hui Yang, Hai-Ping Wu, Gui-Fang Xu, Wei Liu, Zhi-Guo Xie, Linyu Shi, Xinyi He, Seung-gi Jin, et al. The role of tet3 dna dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366):606, 2011.

[20] Meelad M Dawlaty, Achim Breiling, Thuc Le, Günter Raddatz, M Inmaculada Barrasa, Albert W Cheng, Qing Gao, Benjamin E Powell, Zhe Li, Mingjiang Xu, et al. Combined deficiency of tet1 and tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Developmental cell*, 24(3):310–323, 2013.

[21] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[22] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[23] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[24] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, 2012.

[25] Laura B Sontag, Matthew C Lorincz, and E Georg Luebeck. Dynamics, stability and inheritance of somatic dna methylation imprints. *Journal of theoretical biology*, 242(4):890–899, 2006.

[26] Marco Taboga. *Lectures on probability theory and mathematical statistics.* CreateSpace Independent Publishing Platform, 2012.

4. TWO ARE BETTER THAN ONE: HPOXBS - HAIRPIN OXIDATIVE BISULFITE SEQUENCING

The content of chapter 4 has been published as:

Giehr, P.[†], Kyriakopoulos, C.[*], Lepikhov, K.[†], Wallner, S.[§], Wolf, V.[*], & Walter, J.[†] (2018). Two are better than one: HPoxBS-hairpin oxidative bisulfite sequencing. Nucleic acids research.

AUTHOR CONTRIBUTIONS

Pascal Giehr: Conduction of hairpin oxBS experiments. Authoring of the manuscript including the generation of all figures and tables, as well as the entire supplement of the manuscript.

Dr. Charalampos Kyriakopoulos: Application of hidden Markov models. Revision of the mathematical aspects of the manuscript concerning the background and the results of the hidden Markov model *i.e.* changes in formulation/wording and structure of the text.

Dr. Konstantin Lepikhov: Isolation of primordial germ cells from mouse embryos.

Dr. Stefan Wallner: Isolation, cultivation and differentiation of human monocytes. Isolation of genomic DNA.

Prof. Dr. Verena Wolf: Supervision of hidden Markov modelling. Revision of the manuscript *i.e.* changes in formulation/wording and structure of the text.

Prof. Dr. Jörn Walter: Supervision of wet-lab experiments. Financing. Revision of the manuscript *i.e.* changes in formulation/wording and structure of the text.

[†] Department of Biological Sciences, Saarland University, Campus A2.4, 66123 Saarbrücken, Saarland, Germany

[*] Computer Science Department, Saarland University, Campus E1.3, 66123 Saarbrücken, Saarland, Germany

[§] Institute for Clinical Chemistry and Laboratory Medicine, University Hospital, Franz-Josef-Strauß-Allee 11, 93053 Regensburg, Bayern, Germany

Abstract

The controlled and stepwise oxidation of 5mC to 5hmC, 5fC and 5caC by Tet enzymes is influencing the chemical and biological properties of cyto- sine. Besides direct effects on gene regulation, oxidised forms influence the dynamics of demethylation and re-methylation processes. So far, no combined methods exist which allow to precisely determine the strand specific localisation of cytosine modifications along with their CpG symmetric distribution. Here we describe a comprehensive protocol combining conventional hairpin bisulfite with oxidative bisulfite sequencing (HPoxBS) to determine the strand specific distribution of 5mC and 5hmC at base resolution. We apply this method to analyse the contribution of local oxidative effects on DNA demethylation in mouse ES cells. Our method includes the HPoxBS workflow and subsequent data analysis using our developed software tools. Besides a precise estimation and display of strand specific 5mC and 5hmC levels at base resolution we apply the data to predict region specific activities of Dnmt and Tet enzymes. Our experimental and computational workflow provides a precise double strand display of 5mC and 5hmC modifications at single base resolution. Based on our data we predict region specific Tet and Dnmt enzyme efficiencies shaping the distinct locus levels and patterns of 5hmC and 5mC.

4.1 Introduction

In mammals, DNA methylation is restricted to the C5 position of cytosine and is predominantly found in a CpGcontext [1, 2, 3]. The precise control of its establishment and maintenance is tightly controlled by the DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. All three enzymes catalyse the transfer of a methyl group from s-adenosyl methionine to cytosine.

Dnmt1 is associated with the replication machinery by directly interacting with Uhrf1 and PCNA [4, 5, 6, 7]. The protein complex modulates the preferred recognition of Dnmt1 for hemimethylated CpGs, such that Dnmt1 acts as a copying enzyme for existing methylation patterns from the old to the newly synthesised DNA strand, maintaining original methylation patterns across cell divisions [8, 9]. This process is one of the key mechanisms of epigenetic inheritance.

On the other hand, Dnmt3a and Dnmt3b are the key enzymes to methylate CpG dinucleotides in the first place. They are called "*de novo*" methyltransferases and mainly act on unmethylated DNA during epigenetic programming phases of development and differentiation [10, 11]. However, there are numerous indications that the strict separation of *de novo* and maintenance methylation functions between Dnmt1 and Dnmt3a/Dnmt3b is not definite. Instead, under certain conditions, these enzymes exhibit overlapping functions [12, 13, 14].

Moreover, 5-methyl cytosine (5mC) can be oxidised by a group of oxigenases called ten-eleven translocation enzymes (Tets) [15, 16]. Under consumption of oxygen and 2-oxoglutarate, these Fe(II) dependent dioxigenases oxidise 5mC in a first reaction to 5-hydroxymethyl cytosine (5hmC), followed by 5-formyl cytosine (5fC) and finally 5-carboxy cytosine (5caC) [17, 18]. The most abundant form of these oxidised cytosine variants is 5hmC. Recent publications show that 5hmC can be found in numerous cell types such as embryonic stem cells (ESC), neurons or liver cells [19, 20, 21, 22]. The current knowledge suggests that 5hmC, like 5mC, imposes an epigenetic regulatory function through the recognition of specific reader proteins.

In zygotes 5mC is extensively converted into 5hmC mainly on the paternal (sperm derived) chromosomes [23, 24]. Furthermore, in subsequent cell divisions, DNA methylation decreases, suggesting that 5hmC under certain conditions promotes genome wide DNA methylation reprogramming [25, 26]. Based on this, and other observations, several mechanisms have been proposed how 5hmC contributes to a passive (replication dependent) and active (non-replicative) loss of DNA methylation [26, 27, 28, 29, 30, 31].

In order to better understand and comprehensively follow such processes over cell divisions, accurate base resolution detection methods discriminating 5mC and 5hmC are essential.

One such method is the oxidative bisulfite conversion (oxBS). In addition to a standard bisulfite treatment, where both 5mC and 5hmC remain unconverted and indistinguishable as cytosine after sequencing, a pre-bisulfite oxidation reaction converts 5hmC to 5fC, which will be converted by bisulfite to 5f-uracil and to thymine in the subsequent PCR [32, 33]. By comparing the readout of standard bisulfite sequencing (BS) and oxBS, one can determine the amount of 5mC and 5hmC for each modified cytosine within the DNA.

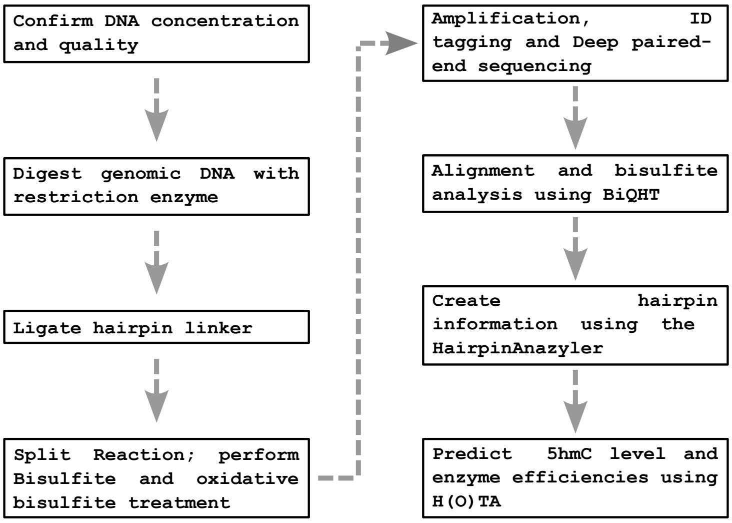

Confirm DNA concentration and quality
Digest genomic DNA with restriction enzyme
Ligate hairpin linker
Split Reaction; perform Bisulfite and oxidative bisulfite treatment
Amplification, ID tagging and Deep paired-end sequencing
Alignment and bisulfite analysis using BiQHT
Create hairpin information using the HairpinAnazyler
Predict 5hmC level and enzyme efficiencies using H(O)TA

Fig. 4.1: HPoxBS pipeline. Individual steps of HPoxBS starting from DNA quality assessment to 5hmC prediction and enzyme efficiency estimation.

Since bisulfite modification based methods only work efficiently on single stranded

DNA, the subsequent sequencing information only covers the methylation status of one DNA strand. It is therefore impossible to deduce the symmetry of modifications at CpG dyads in double stranded (ds) DNA. To overcome this limitation, Laird et al. developed a method of bisulfite sequencing which physically links DNA strands by the addition of a hairpin linker, in other words a short hairpin oligo nucleotide is attached onto the DNA to prevent a physical separation of the upper (Wat- son) and lower (Crick) strand during bisulfite treatment [34, 35, 36].

In order to monitor the distribution of 5mC and 5hmC in ds-DNA, we here describe a protocol which combines conventional hairpin bisulfite sequencing (HPBS) with oxBS [30].

4.2 Materials and Methods

Hairpin oxidative bisulfite sequencing (HPoxBS) comprises a series of biochemical reaction and purification steps. First, fragmented genomic DNA is ligated to a synthetic hairpin linker [37]. The ligated DNA is then used for BS and oxBS treatment, sequence specific PCR amplification and finally next generation based sequencing (NGS). Figure 4.1 provides a general outline of the individual steps of the method.

4.2.1 Digestion of Genomic DNA

The first step of the experimental procedure is the digestion of genomic DNA (Figure 4.2). The optimal restriction enzyme (RE) used for HPoxBS should fulfil the following conditions: The restriction site should be in close proximity to the CpGs of interest, as it provides the anchor for the hairpin linker ligation. The distance between restriction site and region of interest should be ≥250bp when using a 2x300bp paired sequencing mode on an Illumina MiSeq platform.

We recommend the usage of type II REs generating 3' or 5' overhangs to increase the ligation efficiency of the hairpin linker. Alternatively, blunt end DNA can be A-tailed using Klenow Fragment ($3' \rightarrow 5'$ exo-). The RE should not be sensitive against both 5mC and 5hmC, thereby avoiding a bias of the analysis by blocked restriction. Ideally, no CpG should be present within the restriction site. We have successfully used the enzymes *R.BsaWI*, *R.DdeI* and *R.TaqI*.

Following DNA preparation and hairpin linker ligation, the oxBS treatment includes two harsh chemical modifications, which strongly increase the risk of damaging the input DNA. It is recommended to use sufficient amounts of high quality (desalted, pure) DNA to compensate for the loss of amplifiable DNA, inevitably caused by chemical fragmentation and depurination.

We recommend to start with 300 - 500 ng genomic DNA digested in a buffered 20μl reaction using a 5-10x excess of RE (units/μg). To ensure complete digestion of the DNA,

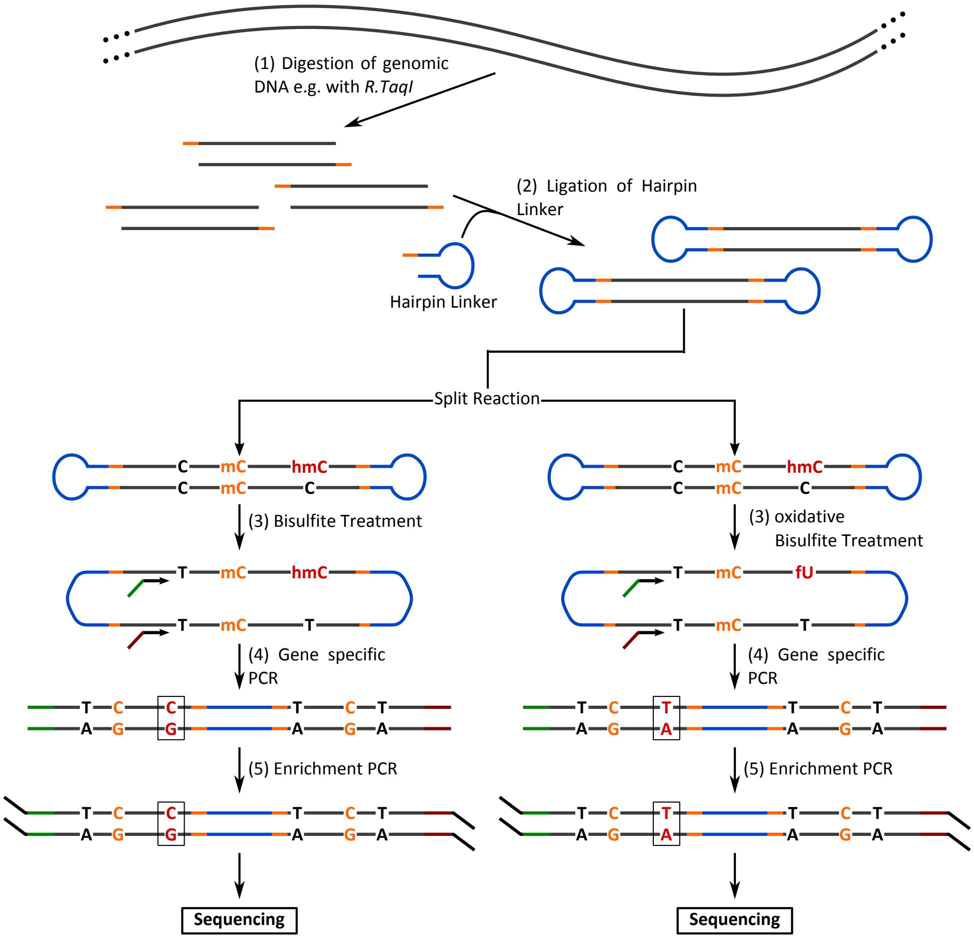


Fig. 4.2: Experimental workflow of HPoxBS. (1) Genomic DNA is enzymatically digested; (2) DNA strands are linked covalently by ligation of a hairpin linker; (3) after ligation the reaction is split and treated with BS or oxBS; (4) region of interest is amplified and sequencing adapters are introduced; (5) multiplexed enrichment PCR including ID tagging

the incubation should be performed overnight (12 h). After digestion is complete, the enzyme must be inactivated. We only use temperature and no chemical inactivation, as in our experience this negatively affected the ligation of the hairpin linker. The amount of DNA might be reduced in case the DNA has sufficient quality and integrity. Alternatively, when operating with very low cell numbers, HPoxBS can be applied using a 'one tube' reaction without prior DNA isolation. We demonstrate this application for the analysis of primordial germ cells (see Supplement, Sections S4.6.3). However, if possible, we advice the use of sufficient amount of isolated high quality DNA to obtain optimal results.

4.2.2 Hairpin Linker Design and Ligation

The hairpin linker contains a single stranded overhang complementary to the genomic DNA overhangs generated by the RE. Figure 4.3 shows an example with a two base 5'-CpG overhang generated by the RE *R.TaqI*.

The hairpin linker comprises the following features: (i) A unique sequence (molecu-

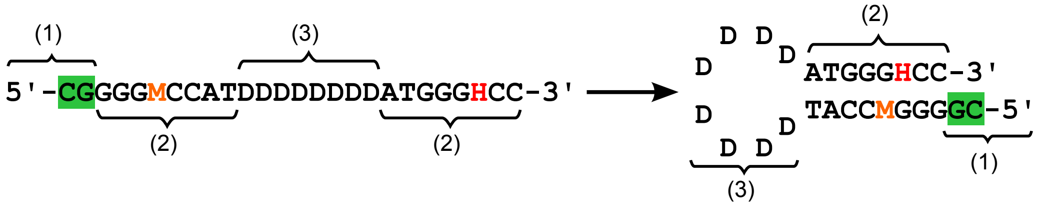


Fig. 4.3: Hairpin linker structure. Example of a hairpin linker in unfolded (left) and annealed (right) conformation matching a 5'-overhang created by the restriction enzyme *R.TaqI*. (1) green = restriction enzyme complementary 5'-CG overhang; (2) = stem structure facilitating the folding; (3) = loop structure with unique molecular identifier sequence; M = 5mC, H = 5hmC, D = A, T or G.

lar) identifier (UMI) which allows to identify individual original ligation events and to bioinformatically remove clonal PCR amplificates from the pool of sequences [35]. (ii) Unmodified cytosines at defined positions allowing to determine the overall C to T (G to A) bisulfite conversion rates. (iii) 5mC and 5hmC to deduce the rates of unwanted conversion of both modified bases due to either BS or oxBS treatment. Note that this could be expanded by including 5fC and 5caC modified bases for e.g. fCAB or MAB-Seq analysis [38, 39, 40].

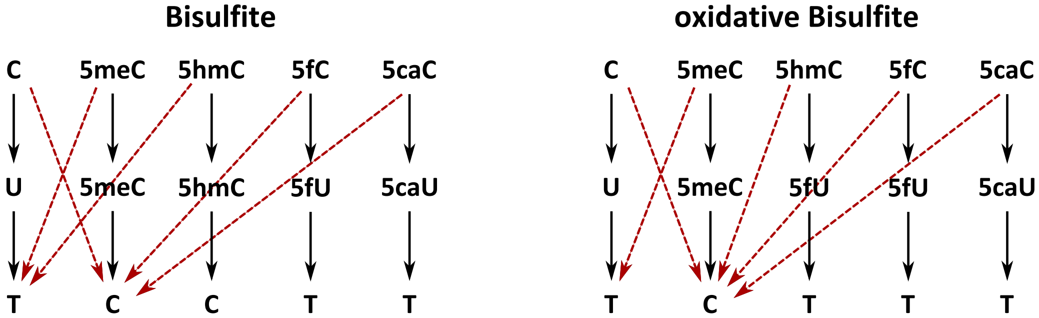


Fig. 4.4: Conversion during BS and oxBS. Conversions of cytosine and its modified derivatives (upper row) during BS and oxBS (middle row) as well as their appearance after sequencing (lower row). Black straight arrows indicate the intended conversion reaction; red dashed arrows indicate possible conversion errors.

Ligation of the hairpin linker, will generate closed DNA fragments (Figure 4.2). To minimise self-ligation of DNA fragments, hairpin linker is given in excess.

The ligation reaction occurs for >4h (or overnight) at 16°C.

4.2.3 Bisulfite and Oxidative Bisulfite Treatment

The oxBS conversion includes an oxidation step prior to the bisulfite treatment. For this oxidation (which we perform according to the manufacturer's manual (Cambridge Epigenetix (CEGX)) the purity of the sample is of great importance as traces of salt or ethanol will cause the reaction to fail. For this reason, it is essential to purify (and desalt) the sample after the ligation reaction. We then continue with a bisulfite reaction using

Tab. 4.1: Typical ligation reaction for HPoxBS.

20μl	digested chromosomal DNA (15-25ng/μl)
2.5μl	10mM ATP
1.0μl	100pmol/μl hairpin linker
0.5μl	400U/μl T4 DNA ligase (New England Biolabs)
1.0μl	ddH_2O

the TrueMethyl Kit provided by Cambridge Epigenetix (CEGX). We usually perform the following protocol:

(1) After ligation, transfer the solution into a 1.5 ml reaction tube and adjust the volume to 50µl using ddH_2O.

(2) Add 100µl(2x) AMPure XP beads and incubate for 15min at room temperature (RT).

(3) Place the tube onto a magnetic stand and incubate for 10min at RT.

(4) Carefully discard the supernatant without disturbing the beads.

(5) Keep on the magnetic stand and add 1 ml freshly prepared 80% acetonitrile and wait for 30 s. Then carefully remove and discard the supernatant.

(6) Repeat the wash step from (5) three more times for a total of four wash steps.

(7) Let the beads dry for 5min on the magnetic stand.

(8) Without removing the tube from the magnet, add 20µl 0.05M NaOH.

(9) Remove the reaction from the magnetic stand and resuspend the beads completely by pipetting. Incubate for 10min at RT to elute the DNA.

(10) lace the tube back onto the magnetic stand and incu bate for 5min until the suspension becomes clear.

(11) Without disturbing the beads, remove 9µl of the supernatant for BS and 9µl for oxBS and put each into a new reaction tube. Proceed with the oxBS workflow according to Booth *et al.*

Note, for the preparation of 80% Acetonitrile and 0.05M NaOH ensure high purity of the used ddH_2O.

For each purified DNA we then perform two separate conversion reactions: (i) a conventional bisulfite conversion reaction and (ii) a combined oxidation and bisulfite reaction

(Figure 4.2 and Figure 4.4). The single treatment with sodium bisulfite allows to simultaneously detect 5mC and 5hmC. All unmodified cytosines (as well as 5fC and 5caC, see below) are converted into uracils, while 5mC and 5hmC are not converted. In the subsequent PCR amplification and sequencing, converted cytosines will be read as thymine instead of cytosine. In the case of oxBS, 5fC will be oxidised to 5fU and converted to 5fU during bisulfite treatment. Following subsequent PCR, 5hmC will appear as thymine after sequencing. We recommend to use the TrueMethyl Kit (CEGX) to perform the bisulfite treatment. Note, when using other bisulfite protocols, ensure that the method achieves a complete conversion of 5fC and 5caC.

4.2.4 Amplification of Target Genes

After BS and oxBS treatment, the targeted regions are amplified by PCR in which also the first part of the sequencing adapters are introduced (Table 4.2 and Figure 4.2 The PCR uses gene specific primers to amplify a specific target region ligated to the hairpin linker. For PCR we use the HOT FIREPol® DNA Polymerase from Solis BioDyne, which performs well on uracil containing bisulfite templates.

After incubation, the amplified product needs to be purified to remove PCR residues, such as nucleotides, salt and primers which would interfere with downstream processes. Routinely, we perform purification using AMPure XP beads in a ratio of 1:1 (µl PCR:µl beads) or agarose gel purification using Geneaid "Gel/PCR DNA Fragments Kit" following manufacturer's instructions.

Tab. 4.2: Typical PCR protocol for HPoxBS using HOT FIREPol®

PCR Protocol	**PCR Conditions**	
2.0μl BS/oxBS hairpin sample		
3.0μl 10x Buffer BD	95°C - 15min	
3.0μl 25mM $MgCl_2$	95°C - 1min	40x
2.4μl 10mM dNTPs	X°C - Xmin	
0.5μl Forward Primer	72°C - 1min	
0.5μl Reverse Primer	72°C - 7min	
0.7μl HOT FIREPol®		
19.1μl ddH_2O		

4.2.5 Amplicon Preparation and Sequencing

Amplicon preparation for sequencing is finalised by subjecting the purified product to a second PCR Table 4.3 Figure 4.2. In this amplification, primers are not gene specific, but bind to the adapter part introduced during the first PCR. The second primer pair provides the adapter sequence which facilitates the binding to the sequencing platform and in addition carries a sample specific ID. The Reaction can be performed as a multiplex PCR, where several distinct amplicons can be flagged with the same ID.

Tab. 4.3: Multiplex-PCR protocol for sequencing preparation

PCR Protocol	**PCR Conditions**	
25.0μl BS/oxBS hairpin sample		
5.0μl 10x Buffer HotStarTaq	95°C - 15min	
2.0μl 25mM $MgCl_2$	95°C - 30sec	
4.0μl 10mM dNTPs	60°C - 30sec	5x
2.5μl Forward Primer	72°C - 30sec	
2.5μl Reverse Primer	72°C - 5min	
0.6μl HotStarTaq®		
8.4μl ddH_2O		

In our case, subsequent sequencing is performed on an Illumina MiSeq platform using a multiplexed 2x300 bp paired-end sequencing. For this, the products of the second PCR are again purified using AMPure XP beads with a ratio of 1:1.1 (µlPCR:µl beads). All amplicon pools are then adjusted to a concentration of 5 nM and joined for multiplexed sequencing. Following manufacture's instructions, the pooled library is further diluted to a final concentration of 18 pM.

4.2.6 Sequence Alignment and Methylation Calling

Following demultiplexing and quality control, sequence alignment and extraction of methylation information is performed using BiQAnalyzer HT (BiQHT) (`http://biq-analyzer-ht.bioinf.mpi-inf.mpg.de/`). BiQHT is a Java based tool with graphical user interface which has been developed for locus specific DNA methylation analysis [41]. The program aligns the sequencing reads against a given reference sequence and determines the methylation state for each cytosine.

To exploit all the information contained in the hairpin amplicon, four individual analysis steps have to be performed:

CpG Methylation Analysis. CpG methylation is analysed by providing a genomic reference sequence without for each locus, consisting of the unconverted DNA sequence from top and bottom strand with the converted (C replaced by T) hairpin linker sequence in between (Supplement, Section S4.6.6). Cytosines of the hairpin linker will be analysed independently and therefore have to be replaced by Ts. The analysed methylation context has to be set to 'CpG'. BiQHT provides several filter options to dispose unwanted sequencing reads. Routinely, we use a sequence identity of $\geq$0.9 for single copy genes and

$\geq$0.8 for repetitive elements due to their sequence variability. However, filtering might be optimised for individual amplicons by including additional parameters such as conversion rate, alignment score or fraction of unrecognized sites.

nonCpG Methylation Analysis. For an unbiased detection of nonCpG methylation, all CpGs in the reference file and also in the sequencing read file must be replaced by NpNs. Furthermore, to allow nonCpG methylation detection, the analysed methylation context has to be changed to ‘C’. Usually, the filter conditions from the CpG methylation analysis can be applied.

Linker Conversion Rate. The unmodified cytosines in the hairpin linker allow the determination of cytosine conversion, unbiased by nonCpG methylation. To extract the information, sequencing reads are aligned to the genomic sequence of the hairpin linker (Supplement, Section S4.6.6). The variable loop sequence creates UMIs which cannot be described by one reference sequence alone. Therefore, the sequencing identity filter must be reduced to $\geq$ 0.6 in order to prevent the loss of sequencing reads. Analysed methylation context has to be set to ‘C’.

SNP Detection. BiQHT annotates single nucleotide polymorphisms (SNPs), if specified in the reference sequence. This function can also be used to determine the state of 5mC and 5hmC within the hairpin linker. Both cytosines have to be replaced by ‘N’ in the reference sequence for the analysis. BiQHT then annotates the occurring base (C or T) for both cyotsines in each read. The output can be used to calculate the conversion rate of 5mC and 5hmC during BS and oxBS Figure 4.4. Applied are the same settings as in the CpG methylation analysis but in addition the option “output results of the SNP analysis” must be selected.

4.2.7 Restoration of Double Strand Information

Subsequently to BiQHT, we use *Hairpinanalyzer* to restore the ds information. The *Hairpinanalyzer* is a python based script which accepts the output of BiQHT, restores the ds information and generates the following output:

(1) A map of methylation pattern in form of a portable network graphic (png).

(2) A text file for each sample containing the CpG methylation information of each read and in addition, position specific nonCpG methylation, conversion rates and SNPs

(3) A text summary file for all samples related to the same reference sequence.

Note that the results of BS and oxBS are stored as individual files and the level of 5hmC must be calculated by comparing both outputs. The most simple calculation is the subtraction of the mean methylation level of oxBS from BS results. Additionally, to also gain the distribution of 5hmC, this calculation has to be done for fully methylated CpGs, hemimethylated CpG on the top- and hemimethylated CpGs on the bottom strand, respectively. The script of the Hairpinanalyzer is available on GitHub `https://github.com`.

4.2.8 Estimation of 5hmC and Enzyme Efficiency

The conversion scheme in Figure 4.4 suggests that 5hmC levels are simply determined by subtracting mean methylation levels of BS reactions from those of oxBS reactions. We observe that this may lead to inaccurate estimates due to random sampling of different cells and the omission of conversion errors. Therefore, we propose a more accurate approach by combining two Hidden Markov Models(HMMs), one for BS and one for oxBS, that take into account all possible conversions as outlined in Figure 4.4. We then link the two HMMs and calibrate the model's parameters, such that they simultaneously fit the results of BS and oxBS. By this, we can accurately estimate 5hmC levels. For this purpose we developed H(O)TA, a MAT-LAB based tool which uses ds information for such calculations `https://mosi.uni-saarland.de/HOTA` [42]. H(O)TA works with classical HPBS data, but also with data from HPoxBS experiments. Based on the information provided, H(O)TA considers the ds information and conversion rates to estimate accurate 5mC and 5hmC level as well as their ds distribution. In addition, it predicts the efficiencies of Dnmts and Tets. Furthermore, based on the ds information, H(O)TA provides a more accurate discrimination of maintenance and *de novo* methylation compared to single strand based models.

All tools come with detailed instruction for easy use. In addition, we included a test data set to the supplement information, which includes raw data from MiSeq sequencing, BiQHT and Hairpinanalyzer output as well as the input files and the results of the H(O)TA analysis.

4.3 Results

In this section, we outline the complete HPoxBS workflow (including a H(O)TA analysis) for the analysis of demethylation dynamics in mouse embryonic stem cells (ESCs). This section is followed by a brief summary of use cases on mouse primordial germ cells and human monocytes, respectively. A full description for the additional data can be found in supplement sections S4.6.3 and S4.6.4

Mouse ESCs have a high genome wide methylation status when cultivated on serum/LIF, while loosing DNA-methylation in a replication dependent manner under 2i conditional medium [43, 44, 45, 46]. We analysed ESCs under Serum/LIF (day0) conditions as well as after their transition into 2i after 24h (day1), 72h (day3) and 144h (day6). Our goal was to monitor the progressive changes in DNA-methylation and DNA-hydroxymethylation levels at three single- (Afp, Ttc25 and Zim3) and five multi copy loci (IAP, L1mdA, L1mdT, mSat and MuERVL) using HPoxBS. Following the method outlined in Figure 4.2 we sequenced PCR products on an Illumina MiSeq platform obtaining a mean read coverage of 5188 per locus. Figure 4.5 shows the CpG methylation maps for ESCs, generated after BiQHT alignment and Hairpinanalyzer refolding for BS and oxBS samples separately. Each column represents one CpG position and each row one unique sequence read, which corresponds to the region specific pattern of one chromosome. CpG positions modified on both DNA strands are shown in red, hemimethylated CpGs in green and unmodified CpG positions in blue. The Hairpinanalyzer script also generates a text file for each sample, containing the read ID, CpG methylation pattern, nonCpG Methylation and, if provided in the BiQHT reference sequence, information on SNPs.

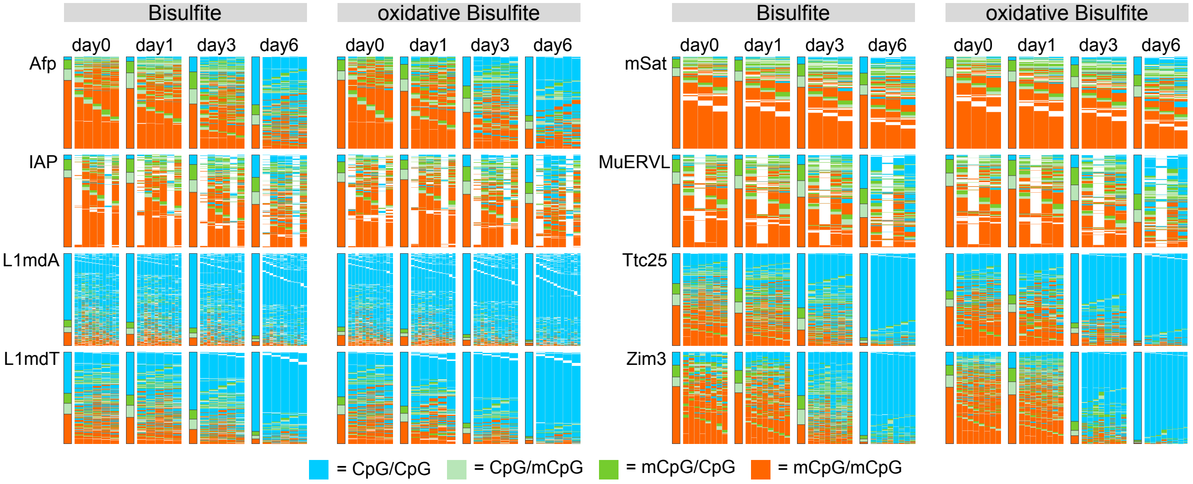


Fig. 4.5: Hairpin Methylation Pattern Maps. Methylation patterns for the single copy genes Afp, Ttc25 and Zim3, as well as the retrotransposable elements IAP, L1mdT, L1mdA, mSat and MuERVL for BS and oxBS of ECS cultivated under Serum/LIF (d0) and 2i medium (d1 = 24h 2i, d3 =72h 2i, d6 = 144h 2i). Each column represents one CpG dyad, each row one sequenced chromosome. The very left column gives the mean methylation pattern over all analysed CpGs. Red = CpG dyad is modified on both DNA strands (BS = 5mC or 5hmC; oxBS = 5mC only); Dark green = CpG dyad is only modified on the plus strand (BS = 5mC or 5hmC; oxBS = 5mC only); Light green = CpG dyad is only modified on the lower strand (BS = 5mC or 5hmC; oxBS = 5mC only); Blue = CpG dyad is unmodified on both strands (BS = C only; oxBS = C or 5hmC); White = CpG dyad was not analysable.

In line with our previous findings, we observe that the overall level of 5mC/5hmC decreases with region specific dynamics upon prolonged culturing of ESCs in 2i medium. This decrease occurs at retrotransposable (repetitive) elements (IAP, L1MdA, LmdT, mSat, MuERVL), as well as at single copy genes (Afp, Tct25, Zim3) (Figure 4.5). Using H(O)TA, we find considerable levels of 5hmC at CpG positions in most of these regions (Figure 4.6).

Concerning conversion quality of the oxBS reactions, we determined the conversion

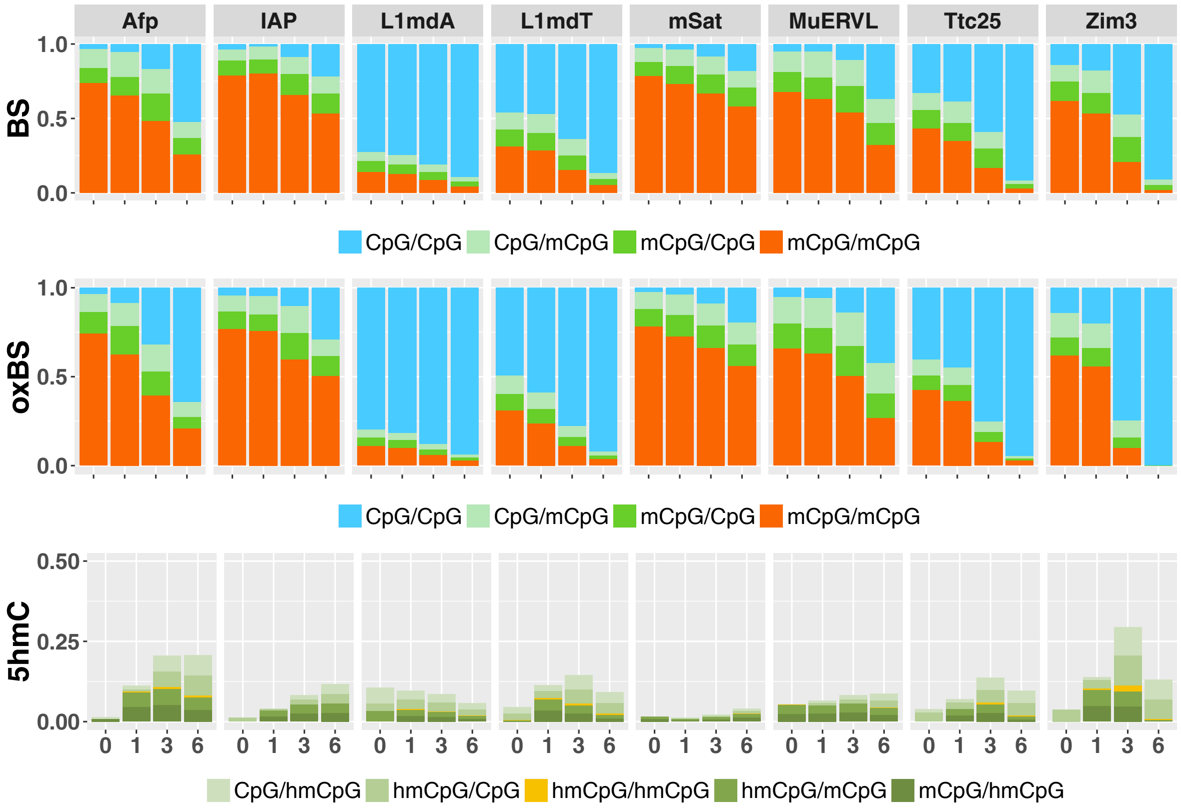


Fig. 4.6: Average modification level. Mean methylation level of BS (upper panel) and oxBS (middle panel) samples as well as the predicted 5hmC amount and distribution (lower panel). x-axis = days; y-axis = 5mC/5hmC level; red = CpG dyad is modified on both DNA strands (BS = 5mC or 5hmC; oxBS = 5mC only); dark green = CpG dyad is only modified on the plus strand (BS = 5mC or 5hmC; oxBS = 5mC only); light green = CpG dyad is only modified on the lower strand(BS = 5mC or 5hmC; oxBS = 5mC only); blue = CpG dyad is unmodified on both strands (BS = C only; oxBS = C or 5hmC).

rate of the known C, 5mC and 5hmC positions within the hairpin linker (Figure 4.3). The conversion rates were calculated by dividing the number of sequenced thymines at given cytosine positions by the total number of obtained reads (Table 4.4 and Table 4.5).

$$Conversion\ Rate = \frac{Number\ of\ T\ at\ C\ Positions}{Number\ of\ Reads\ at\ C\ Positions}$$

Conversion of C during BS and oxBS was found to be highly efficient with a conversion rate of $\geq$ 99% (Table 4.4 Table 4.5). However, this almost complete conversion comes at the expense of an unwanted conversion of 5mC/5hmC in the range of 5-10% due to the harsh bisulfite reaction conditions. Conversion of 5hmC after oxBS was 93%.

Based on the rates, individual conversion erros for BS and oxBS were calculated. A scheme of all possible conversions and conversion errors are given in Figure 4.4. The

Tab. 4.4: Conversion rates of C, 5mC and 5hmC of BS samples

	Afp	IAP	L1mdA	L1mdT	mSat	MuERVL	Ttc25	Zim3
C	0.996	0.999	0.995	0.993	0.996	0.993	0.994	0.995
5mC	0.0674	0.0628	0.084	0.088	0.0685	0.0819	0.0813	0.0763
5hmC	0.0765	0.0721	0.0736	0.0703	0.0642	0.0662	0.0785	0.0696

Tab. 4.5: Conversion rates of C, 5mC and 5hmC of oxBS samples

	Afp	IAP	L1mdA	L1mdT	mSat	MuERVL	Ttc25	Zim3
C	0.996	0.999	0.996	0.994	0.997	0.997	0.996	0.996
5mC	0.0636	0.0900	0.0795	0.0758	0.0685	0.0808	0.1078	0.0773
5hmC	0.920	0.9095	0.909	0.9323	0.93693	0.922	0.942	0.9315

precise BS/oxBS values including conversion errors were then used for HMM as described in our H(O)TA tool to predict the level and distribution of 5hmC, as well as the enzyme efficiencies. H(O)TA allows to perform these predictions for individual CpGs. However, for simplicity, we here predicted the mean levels over all CpGs across one amplicon.

Figure 4.6 shows the mean methylation level of BS and oxBS together with the predicted 5hmC levels. The ds information demonstrates that 5hmC in most cases occurs in an asymmetric pattern paired either with C (Figure 4.6, lower diagram, light green) or 5mC (Figure 4.6, lower diagram, dark green). Only the minority of CpGs contain 5hmC in a symmetrical state (Figure 4.6, lower diagram, yellow).

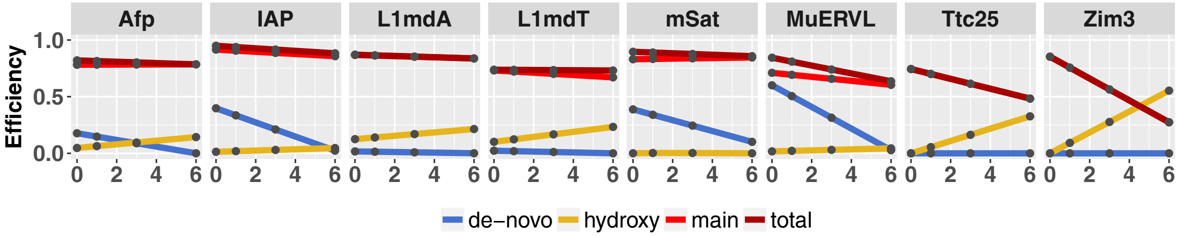


Fig. 4.7: Enzyme efficiencies. Predicted enzyme efficiencies for Dnmts and Tets. Dark red = total methylation activity of Dnmts at hemimethylated CpG dyads (maintenance plus *de novo*); red = maintenance methylation of Dnmts at hemimethylated CpG dyads; blue = *de novo* activity of Dnmts at CpG dyads; yellow = hydroxylation efficiency of Tet enzymes at methylated CpG dyads. X-axis = days; Y-axis = efficiency.

In addition to the 5hmC distribution, H(O)TA calculates the enzyme efficiencies for Dnmts (maintenance and *de novo* methylation) and Tets (hydroxylation) for each time point. Our analysis shows that the efficiencies differ clearly between the distinct regions. In general, we can observe a loss in maintenance and *de novo* methylation activity together with an increase in hydroxylation activity. For some regions *de novo* methylation/hydroxylation efficiency is almost zero.

As a second use case we present our HPoxBS analysis of rare PGCs and nonPGC control cells, isolated from embryos at E10.5 and E11.5 of development. At this time point, PGCs are known to undergo a rapid replication dependent demethylation, probably supported by Tet mediated oxidation [47]. We performed HPoxBS on repeat regions and indeed find indications for the presence of 5hmC in PGCs albeit at low levels. Our analysis demonstrates that it is possible to downscale the amount of sample material (in our case 50-80 cells/sample).

Finally, as a third application we demonstrate the anal- ysis of human monocytes / macrophages following the dynamics of 5hmC during an "active" demethylation process. In previous work we identified several deferentially methylated regions (DMRs) derived from such active demethylation and showed that the loss of 5mC is likely to be caused by Tet mediated oxidation [48]. Here, we show HPoxBS results for two DMRs along a time course of 24 h (0, 12 and after 24 h) following an established differentiation protocol [49] (Supplement, Section S4.6.4). We indeed detect a region specific presence and dynamic change of 5hmC during this time course (Supplement, Section S4.6.4).

4.4 Discussion

The understanding of dynamic changes of DNA methylation during development and disease is a major research area in the field of epigenetics. Such a task can only be realised if DNA modifications can be measured accurately. This is especially challenging for oxidative derivatives of 5mC considering their low abundance and unequal distribution in the genome. Furthermore there is a clear lack of reproducible and easy-to-handle assays for determination of their distribution at single base resolution. The precise knowledge however would allow to model their presumed influence on epigenetic inheritance and temporal stability. In addition, most chemical assays only allow measuring DNA methylation on one DNA strand, making it impossible to determine the precise rates of symmetric methylation and its consequences. HPoxBS is the first method to combine HPBS and oxBS for the simultaneous detection of 5mC and 5hmC levels and their distribution at both complementary DNA strands.

Our workflow not only describes the generation of sequencing data, but also bioinformatic tools applicable for data analysis and modelling. A key element in our method is the ligation of a hairpin linker to "fix Watson and Crick" strands to be able to simultaneously monitor modifications in CpG dyads. In addition, we combine this approach with a double bisulfite chemistry, allowing the discrimination between 5mC and 5hmC in the DNA [32, 33]. We also introduce the novel concept to incorporate 5mC and 5hmC nucleotides into the hairpin linker. This allows us to directly measure their conversion rates following

BS and oxBS treatment, respectively. Typically, conversion rates are determined using spike ins, i.e. small ds oligos. Such oligos are difficult to titrate and frequently perform with a different conversion efficiency. As an integrated part of the analysed DNA region, the ligated hairpin linker improves the sample specific conversion rate detection and at the same time serves as a UMI.

Demonstrating the strength of HPoxBS, we analysed the DNA methylation of several multi- and single copy sequences in embryonic stem cells (ESCs) under growth conditions in which the ESCs strongly demethylate their genome [43, 45, 46].

We show that HPoxBS represents a unique novel method to determine the distribution of 5mC and 5hmC as fully or hemimethylated CpG dyads. Such ds data provide a new resource for mathematic modelling of proposed DNA methylation maintenance and *de novo* methylation activities as well as active processes of DNA demethylation. More importantly, ds information allows a more accurate discrimination of maintenance and *de novo* methylation compared to singe strand data.

Using our recently developed HMM based H(O)TA tool we predict the enzyme efficiencies in discrete regions of the genome. We observe that indeed individual loci display individual combinations of enzyme efficiencies and DNA demethylation dynamics. During the ESC culturing, i.e. the transition from serum to 2i medium, we detect a general reduction of *de novo* methylation, accompanied by an increase in hydroxylation activity (Figure 4.7). This observation is in concordance with the loss of Dnmt3a/3b protein and the simultaneous increase in the expression of Tet enzymes in the presence of 2i [43].

In addition, the analysis of PGCs evidences that HPoxBS can be used in experiments, where only limited amounts of cells or DNA is available, e.g. when analysing reprogramming events during early embryonic or germ cell development (Supplement, Section S4.6.3). Here, both active and passive demethylation processes are known to take place but the exact involvement of oxidation processes is still debated [47, 49].

Rapid locus specific demethylation can also be found in somatic cells and are likewise thought to be Tet mediated. One such example is the generation of region specific demethylation during monocyte-to-macrophage maturation. Our analysis shows that indeed the active loss of 5mC clearly relies on a strong increase of 5hmC level (Supplement, Section S4.6.4).

All three examples show the broad application possibilities for HPoxBS. Moreover, these three examples demonstrate possible variations in design (DNA vs cells), molecular performance (high or low amount of material) and data analysis (and modeling).

4.5 Conclusion

Taken together, we present a step by step protocol of HPoxBS which allows the detection and distribution of both 5mC and 5hmC. Overall, the outlined procedures can be modified and implemented for a number of biological questions, e.g. to understand and model the dynamic loss and gain of DNA methylation in non dividing aging cells, to characterise the heterogeneity of epihaplotypes (epigenetic chromosomal patterns) and most importantly to understand changes occurring during development and differentiation with and without DNA replication. Ultimately, in combination with new detection methods, our pipeline could easily be adjusted to likewise, describe the distribution and the behavior 5fC or 5caC [38, 39, 40, 40, 50, 51].

4.6 Supplementary Data

4.6.1 S.1 Hairpin Oxidative Bisulfite Sequencing

500 ng of mESC DNA was cleaved with 5-10 units of restriction enzyme for 5h or overnight in a 30µl reaction at the enzyme specific incubation temperature. For Afp, Ttc2, Zim3 the enzyme TaqI (ThermoScientific) was used. For L1MdT, DNA was digested with BsaWI (New England Biolabs), for mSat with Eco47I (Thermo Scientific). For L1MdA and IAP, DdeI (New England Biolabs) was used. The restriction was stopped by a 20 min heat inactivation at 80°C. The restricted DNA was then subjected to ligation with T4-DNA Ligase (New England Biolabs) for 16h or overnight. We used 200 units of T4-DNA Ligase, 4µl 10mM ATP and 1µl 100µM hairpin linker. All chemicals were added directly into the restriction reaction. Finally, the volume was adjusted to 40µl using ddH2O. BS and oxidative BS treatment was carried out using the TrueMethyl Kit from Cambridge Epigenetix (now provided by NuGEN) following manufacture's instructions. The target genes were amplified using HOTFIREPol® polymerase from Solis BioDyne. Sequencing was performed on a MiSeq Illumina system (using 2x300bp paired end sequencing). The computational analysis was done as described in the main manuscript using BiQAnalyzerHT, python script (HairpinAnalyzer) and H(O)TA.

4.6.2 S.2 Hairpin Oxidative Bisulfite Sequencing for Low Cell Numbers

Primordial germ cells (PGCs) were collected in 2µl M2 medium. The amount of PGCs or nonPGCs for each experiment is provided in Table 4.6. Cells were solubilised by adding 1µl lysis buffer (10mM TrisHCl, 5mM EDTA), 1µl salmon sperm DNA (100 ng) and 1µl Proteinase K (1mg/ml). The reaction was incubated at 55°C overnight. Proteinase K was inactivated by adding 0.7µl 8.14mM Pefabloc®SC and incubation for 1h at room temperature. Subsequently, DNA was digested in a 8µl reaction using 5U Eco47I (ThermoFisher), 0.8µl digestion buffer and 1µl 5mM MgCl2. Digestion was performed overnight at 37°C and inactivated at 65°C for 20min. The hairpin linker was ligated by adding 250U T4 DNA ligase (New England Biolabs), 1µl 10mM ATP and 0.5 °l 100µM Eco47I specific hairpin linker in an overnight reaction at 16°C. Purification, oxidation and bisulfite treatment were performed using the TrueMethyl Kit from Cambride Epigenetix. Again, amplification was performed with HOTFIREPol® polymerase from Solis BioDyne.

4.6.3 S.3 HPoxBS on Primordial Germ Cells

PGCs show a global loss of methylation during their development. The demethylation seems mostly driven by passive, replication dependent mechanisms, but with a possible influence of 5hmC. To test wherever HPoxBS could also be used for low cell numbers, we

Tab. 4.6: Cell numbers of nonPGCs and PGCs

Sample	Number of Cells
nonPGCs	50
PGCs E10.5	60
PGCs E11.5 -1	70
PGCs E11.5 -2	80

collected PGCs at different developmental stages (E10.5 and E11.5). We successfully generated hairpin constructs for major satellites (Figure 4.8). Applying our analysis pipeline, we find considerable levels of 5hmC at both time points (Figure 4.8). We used about 60 cells in case of nonPGCs, approximate 70 cells from E10.5 and between 80 and 90 cells for PGCs from E11.5.

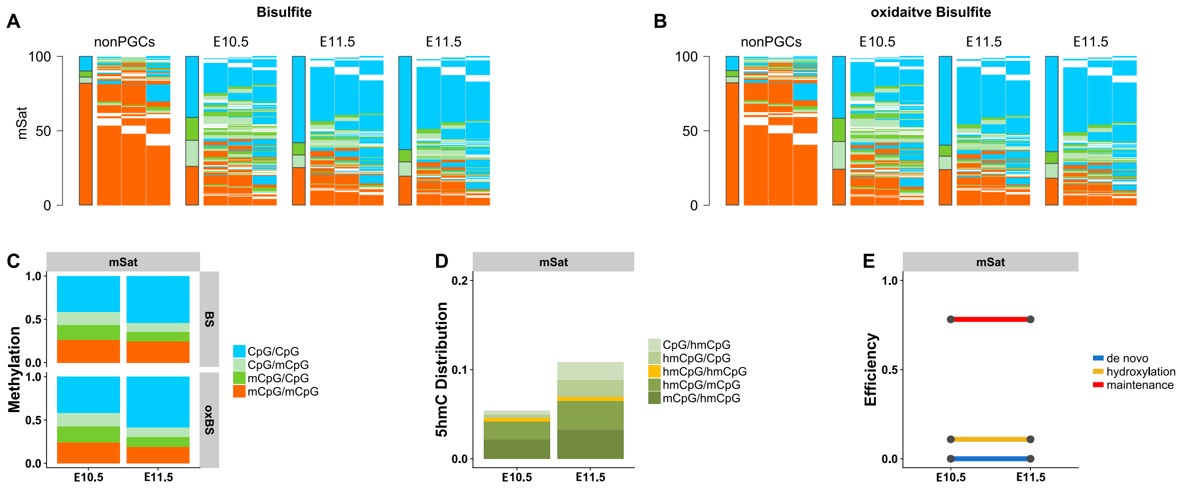


Fig. 4.8: HPoxBS results for PGCs. Methylation pattern maps for BS samples, generated by the Hairpinanalyzer (A); methylation pattern maps for oxBS samples, generatd by the Hairpinanalyzer (B); average 5mC level and distribution estimated by H(O)TA (C); average 5hmC level and distribution calculated by H(O)TA (D); enzyme efficiencies for maintenance methylation, de novo methylation and hydroxylation predicted by H(O)TA.

4.6.4 S.4 Active Demethylation in Monocytes

The differentiation of monocytes to macrophages requires the activation of several cell type specific genes. Demethylation of the corresponding gene promoter accompany the stable activation of transcription. Interestingly, the differentiation occurs without cell division or replication which means, that the observed loss of DNA methylation corresponds to active removal of 5mC. We analysed two previously identified DMRs at the beginning (0h), in the middle (12h) and the end (24h) of monocyte-to-macrophage differentiation. Figure 4.9 displays methylation pattern maps created by the Hairpinanalyser as well as the average methylation level of BS and oxBS. So far, the process of active demethylation cannot be described by H(O)TA, but will be soon implemented.

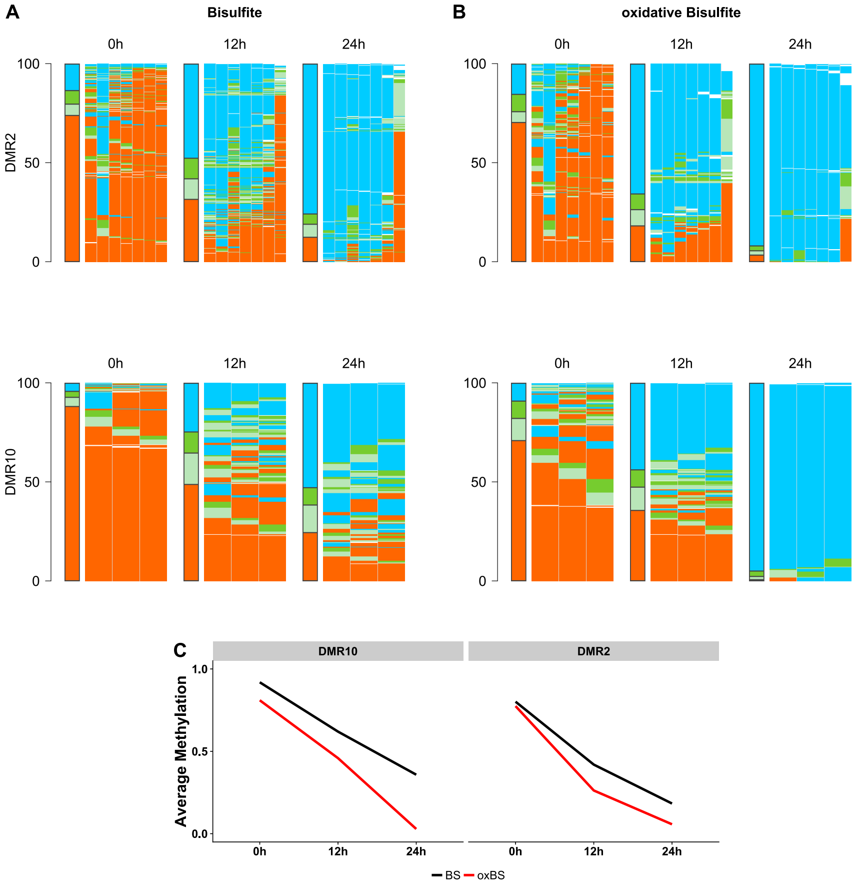


Fig. 4.9: Demethylation of Monocyte DMR2 and DMR10. A: methylation pattern of hairpin BS; B: methylation pattern of hairpin oxBS; each column represents one CpG dyad, each row one sequence read i.e. one analysed DNA strand (chromosome); C: average methylation levels of DMR2 and DMR10 for BS (5mC+5hmC) and oxBS (5mC only).

4.6.5 S.5 BS and oxBS Data

S.5.1 BS and oxBS Data for mESCs

Tab. 4.7: Obtained reads and methylation calls for Afp

sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	8459	41898	31049	5225	4230	1394
Serum (d0)-oxBS	7200	35683	26537	3646	4300	1200
24h 2i (d1)-BS	8004	39541	25901	6693	4928	2019
24h 2i (d1)-oxBS	6732	33393	20851	4362	5363	2817
72h 2i (d3)-BS	5961	29525	14294	4879	5448	4904
72h 2i (d3)-oxBS	7163	35420	13988	5438	4736	11258
144h 2i (d6)-BS	11406	56386	14537	6203	6210	29436
144h 2i (d6)-oxBS	7024	34696	7339	2986	2177	22194

Tab. 4.8: Obtained reads and methylation calls for IAP

sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	205	874	553	148	130	43
Serum (d0)-oxBS	130	536	338	82	79	37
24h 2i (d1)-BS	254	1091	665	185	180	61
24h 2i (d1)-oxBS	144	586	368	69	107	42
72h 2i (d3)-BS	233	994	553	174	156	111
72h 2i (d3)-oxBS	159	673	367	95	109	102
144h 2i (d6)-BS	356	1561	604	259	282	416
144h 2i (d6)-oxBS	209	876	391	153	128	204

Tab. 4.9: Obtained reads and methylation calls for L1MdA

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	4673	55024	7833	3376	3963	39852
Serum (d0)-oxBS	3774	44457	4955	1916	2135	35451
24h 2i (d1)-BS	3334	39390	4968	2561	2571	29290
24h 2i (d1)-oxBS	3314	39237	3935	1515	1671	32116
72h 2i (d3)-BS	4528	53522	4568	2749	2869	43336
72h 2i (d3)-oxBS	3385	39909	2420	1150	1276	35063
144h 2i (d6)-BS	7033	82765	3646	2555	2697	73867
144h 2i (d6)-oxBS	4824	56647	1643	954	961	53089

Tab. 4.10: Obtained reads and methylation calls for L1MdT

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	6214	29632	9570	3497	3274	13291
Serum (d0)-BS	5708	27314	8720	2823	2426	13345
24h 2i (d1)-BS	8969	42648	12289	5589	4837	19933
24h 2i (d1)-oxBS	13697	64778	15498	5917	5237	38126
72h 2i (d3)-BS	7203	34411	5404	3889	3399	21719
72h 2i (d3)-oxBS	3105	14873	1678	906	774	11515
144h 2i (d6)-BS	4560	21775	1138	880	898	18859
144h 2i (d6)-oxBS	5406	25748	1000	590	515	23643

Tab. 4.11: Obtained reads and methylation calls for mSat

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	6845	18268	14372	1673	1732	491
Serum (d0)-oxBS	4662	12477	9776	1168	1220	313
24h 2i (d1)-BS	4618	12267	8998	1332	1492	445
24h 2i (d1)-oxBS	5489	14598	10625	1666	1741	566
72h 2i (d3)-BS	5878	15757	10513	1924	2036	1284
72h 2i (d3)-oxBS	4931	13195	8734	1597	1694	1170
144h 2i (d6)-BS	7563	20150	11724	2240	2562	3624
144h 2i (d6)-oxBS	4985	13265	7449	1613	1593	2610

Tab. 4.12: Obtained reads and methylation calls for MuERVL

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	767	2227	1516	293	307	111
Serum (d0)-oxBS	732	2093	1381	295	309	108
24h 2i (d1)-BS	1068	3132	1978	452	553	149
24h 2i (d1)-oxBS	1391	4131	2607	597	689	238
72h 2i (d3)-BS	847	2498	1262	420	471	345
72h 2i (d3)-oxBS	1411	4205	2276	746	735	448
144h 2i (d6)-BS	743	2186	702	321	365	798
144h 2i (d6)-oxBS	1182	3439	927	470	584	1458

Tab. 4.13: Obtained reads and methylation calls for Ttc25

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	8702	51478	22363	6297	5945	16873
Serum (d0)-oxBS	8180	48395	20641	3926	4338	19490
24h 2i (d1)-BS	7473	44126	15431	5340	6342	17013
24h 2i (d1)-oxBS	7679	45378	16499	4042	4448	20389
72h 2i (d3)-BS	7481	44234	7472	5705	4950	26107
72h 2i (d3)-oxBS	7669	45206	6059	2630	2501	34016
144h 2i (d6)-BS	3541	20881	595	627	538	19121
144h 2i (d6)-oxBS	7925	46621	1310	619	570	44122

Tab. 4.14: Obtained reads and methylation calls for Zim3

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
Serum (d0)-BS	1574	12484	7754	1258	1695	1777
Serum (d0)-oxBS	12986	102951	63716	13448	11308	14479
24h 2i (d1)-BS	7467	59235	33002	6249	8157	11827
24h 2i (d1)-oxBS	10156	80510	43046	11222	11947	14295
72h 2i (d3)-BS	6507	51557	5202	2983	4875	38497
72h 2i (d3)-oxBS	8345	66140	13864	10965	10020	31291
144h 2i (d6)-BS	4203	33359	345	457	503	32054
144h 2i (d6)-oxBS	15675	124178	2434	4100	4761	112883

S.5.2 BS and oxBS Data for Monocytes

Tab. 4.15: Obtained reads and methylation calls for DMR2

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
0h-BS	6467	44577	32911	2589	3035	6042
0h-oxBS	5423	37282	26198	2109	3249	5726
12h-BS	5710	39094	12315	4091	4070	18618
12h-oxBS	163	1115	204	90	87	734
24h-BS	4171	28480	3523	1893	1492	21572
24h-oxBS	2459	16734	556	359	441	15378

Tab. 4.16: Obtained reads and methylation calls for DMR10

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
0h-BS	12612	37464	32993	1805	1077	1589
0h-oxBS	19022	56570	40167	6293	5007	5103
12h-BS	16604	49431	24057	7936	5210	12228
12h-oxBS	15623	46510	16607	5427	4019	20457
24h-BS	16330	48658	11901	6786	4264	25707
24h-oxBS	17878	53324	397	795	1470	50662

S.5.3 BS and oxBS Data for PGCs

Tab. 4.17: Obtained reads and methylation calls for mSat

Sample	# of obtained reads	# of analysed CpG	5mC/5mC	5mC/C	C/5mC	C/C
nonPGCs-BS	12401	32721	26745	1348	1397	3231
nonPGCs-oxBS	8113	21430	17603	880	884	2063
PGCs-E10.5-BS	11964	31610	8253	5556	4796	13005
PGCs-E10.5-oxBS	9824	26035	6324	4781	4092	10838
PGCs-E11.5-BS	11658	30722	7767	2635	2441	17879
PGCs-E11.5-oxBS	7828	20661	4946	1842	1527	12346
PGCs-E11.5-BS	11300	29901	5824	2916	2396	18765
PGCs-E11.5-oxBS	11362	30000	5449	2924	2426	19201

4.6.6 S.6 Linker Sequences

Tab. 4.18: Sequences of the used hairpin linker for Afp, IAP, L1mdA, L1mdT, mSat, MuERVL, Ttc25 and Zim3; *M* indicates 5mC, *H* 5hmC. All hairpin linker carry a 5'-phosphorylation.

Hairpin	Linker Sequnce
Afp-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
IAP-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC
L1mdA-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC
L1mdT-HP	*Pho*-CCGGAGGG*M*CCATDDDDDDDDATGGG*H*CCT
mSat-HP	*Pho*-GNCGGG*M*CCATDDDDDDDDATGGG*H*CC
MuERVL-HP	*Pho*-GNCGGG*M*CCATDDDDDDDDATGGG*H*CC
Ttc25-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
Zim3-HP	*Pho*-CGGGG*M*CCATDDDDDDDDATGGG*H*CC
DMR2-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC
DMR10-HP	*Pho*-TNAGGG*M*CCATDDDDDDDDATGGG*H*CC

4.6.7 S.7 Primer Sequences

Tab. 4.19: Primer for amplification of the analysed regions after BS or oxBS treatment.

Primer	Sequnce
Afp-HP-Forward	TTTTGTTATAGGAAAATAGTTTTTAAGTTA
Afp-HP-Reverse	AAATCACAAAACATCTTACCTATCC
IAP-HP-Forward	TTTTTTTTTTAGGAGAGTTATATTT
IAP-HP-Reverse	ATCACTCCCTAATTAACTACAAC
L1mdT-HP-Forward	TGGTAGTTTTTAGGTGGTATAGAT
L1mdT-HP-Reverse	TCAAACACTATATTACTTTAACAATTCCCA
L1mdA-HP-Forward	GTGAGTGGATTATAGTGTTTGTTTTAA
L1mdA-HP-Reverse	AAATAAATCACAATACCTACCCCAAT
mSat-HP-Forward	GGAAAATTTAGAAATGTTTAATGTAG
mSat-HP-Reverse	AACAAAAAAACTAAAAATCATAAAAA
MuERVL-HP-Forward	TAAGGGTTAGGTGGTAGTATTGAAT
MuERVL-HP-Reverse	CAAAAACCAAATAACAACATTAAAT
Ttc25-HP-Forward	TGAAAGAGAATTGATAGTTTTTAGG
Ttc25-HP-Reverse	AAAACAAAAATCTATTCCATCACTC
Zim3-HP-Forward	TTTATTTATTTGTGTGTGGTTTTTG
Zim3-HP-Reverse	CACATATCAAAATCCACTCACCTAT
DMR2-HP-Forward	TGAGTAATTGGGTTATAGGGAATAAAAAATTTT
DMR2-HP-Reverse	CTTCCTATATAAACAACTAAATCACAAAAAACA
DMR10-HP-Forward	GTTAGTATTGGTTTTGGGGTGGATTTT
DMR10-HP-Reverse	ATCTAAACTAACCTAAACCCTTACCCT

4.6.8 S.8 Hairpin Reference Sequence

TTTTGTTATAGGAAAATAGtTTTTAAGTTACAAAGCATCTTACCTATCCCAAACTCATTTTCG
TGCAATGCTTTGGACGCAGCGAAATGTAGCAGGAGGATGAGGGAAGCGGGTGTGATCCACTTC
ATGGCTGCTGGTTCCTTCACCGCAGGCAGTGCTGGAAGTGGGATGTTT**CGGGGMCCTADDDDD**
DDDTAGGGHCCCGAAACATCCCACTTCCAGCACTGCCTGCGGTGAAGGAACCAGCAGCCATGA
AGTGGATCACACCCGCTTCCCTCATCCTCCTGCTACATTTCGCTGCGTCCAAAGCATTGCACG
AAAATGAGTTTGGGATAGGTAAGATGtTTTGTGATTT

Fig. 4.10: Reference Sequence for Afp

TGTCACTCCCTGATTGGCTGCAGCCCATCGGCCGAGTTGACGTCACGGGGAAGGCAGAGCACA
TGGAGTAGAGAACCACCCTCGGCATATGCGCAGATTATTTGTTTACCAC**CTAGGGMCCATNNN**
NNNNNATGGGHCCTAAGTGGTAAACAAATAATCTGCGCATATGCCGAGGGTGGTTCTCTACTC
CATGTGCTCTGCCTTCCCCGTGACGTCAACTCGGCCGATGGGCTGCAGCCAATCAGGGAGTGA
CA

Fig. 4.11: Reference Sequence for IAP

TCCAATCGCGCGGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGG
GGCACAAGTCCCTTCCGCTCGACTCGAGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTT
GCCCAACACCCGCAAGGGCCCACACGGGACTCCCCACGGACC**CTNAGGGMTTATDDDDDDDD**
ATGGGHCCTNAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTTGGGCAAGACTCT
GCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTGCCCCAGATCA
GGCCCGGGTAGCCTGCTTCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGATTGGGGCA
GGCACTGTGATCCACTC

Fig. 4.12: Reference Sequence for L1mdA

CCCGGGACCAAGATGGCGACCGCTGCTGCTGTGGCTTAGGCCGCCTCCCCAGCCGGGTGGGCA
CCTGTCCT**CCGGAGGGMCCATDDDDDDDDATGGGHCC**CCGGAGGACAGGTGCCCACCCGGCTG
GGGAGGCGGCCTAAGCCACAGCAGCAGCGGTCGCCATCTTGGTCCCGGG

Fig. 4.13: Reference Sequence for L1mdT

GGAAAATTTAGAAATGTTTAATGTAGGACGTGGAATATGGCAAGAAAACTGAAAATCATGGGA
AATGAGAAACATCCACTTGTCGACTTGAAAAATGACGAAATCACTAAAAAACGTGAAAAATGA
GAAATGCACACTGAAG**GNCGGGMCCATDDDDDDDDATGGGHCC**GNCCTTCAGTGTGCATTTCT
CATTTTTCACGTTTTTTAGTGATTTCGTCATTTTTCAAGTCGACAAGTGGATGTTTCTCATTT
TTTATGATTTTTAGTTTTTTTGTT

Fig. 4.14: Reference Sequence for mSat

CGCCCGAGACAAGGTGATTCTAGTTATTATAATGGACAGCGTAGACAAAAGAATGTTTATAAT
AACATACCCAGTAATGGTCAGCACAGGAGAGGTGAAATTTATAATGGCATGACTCGGTTG**GNC**
GGGMCCATDDDDDDDATGGGHCCGNTTCAACCGAGTCATGCCATTATAAATTTCACCTCTCCT
GTGCTGACCATTACTGGGTATGTTATTATAAACATTCTTTTGTCTACGCTGTCCATTATAATA
ACTAGAATCACCTTGTCTCGGGCG

Fig. 4.15: Reference Sequence for MuERVL

CCAGTAGATCCTCAGCTGGGGGCAGGGATCTATTCCATCACTCCCCTTCCGTGTCGGGATTTC
GTGCAGCTCAGACGGGTCCAAGTCTTACACAAGCTGTCCTAACTGCTGTGCGTTTATATAACA
ACTACCCGGTTGTGTTTAGAAAACACTGTTTT**CGGGGMCCTADDDDDDDDATGGGHCCC**GAAA
ACAGTGTTTTCTAAACACAACCGGGTAGTTGTTATATAAACGCACAGCAGTTAGGACAGCTTG
TGTAAGACTTGGACCCGTCTGAGCTGCACGAAATCCCGACACGGAAGGGGAGTGATGGAATAG
ATCCCTGCCCC

Fig. 4.16: Reference Sequence for Ttc25

CCCGGCCACCATAGTCGGATTATCCGTGGGCGGGGTGAGATGGACGGAGCGCCTTGCAGACCT
CAGGAAAACCTCCCCACGCCTGTCCGGCCTTGGCTTGGTGACAGGGAAACTGGCTGGACT**CGG**
GGMCCATDDDDDDDDATGGGHCCCGAGTCCAGCCAGTTTCCCTGTCACCAAGCCAAGGCCGGA
CAGGCGTGGGGAGGTTTTCCTGAGGTCTGCAAGGCGCTCCGTCCATCTCACCCCGCCCACGGA
TAATCCGACTATGGTGGCCGGGCAAGGACCACAC

Fig. 4.17: Reference Sequence for Zim3

AGTATACACAGCGGACGTCAGCAGAGGTGGGCGGCAGGCGAGCCTCCTGCAGGAGCAGGCGGTCCCCTGAAGAAACTCCTTTC
GGAGTTGGCTCCTCCCCGACTTTTCAGGGAGGGATGTGGAGCAGACTCTGTGCCACCTGCCC**TNAGGGMTTATDDDDDDDDAT**
GGGHTTCTNAGGGCAGGTGGCACAGAGTCTGCTCCACATCCCTCCCTGAAAAGTCGGGGAGGAGCCAACTCCGAAAGGAGTTT
CTTCAGGGGACCGCCTGCTCCTGCAGGAGGCTCGCCTGCCGCCCACCTCTGCTGACGTCCGCTGTGTATACTGA

Fig. 4.18: Reference Sequence for DMR2

TGTGAGGCTGTGTGGTTGCCAGGGAAGCCAGAAGAAATGACTTACTCCTGCCCCTGCCTCTAATGTCATGCGGTCACAAGTCC
CCAGAAGGTCTGGGCTGGCCTGGGCCCTTGCCCTCCCCACGGTGGGGGCTCACCCAGCCTGGGCGCGCTGGTCACAC**TNAGGG**
MCCATDDDDDDDDATGGGHCCNTGAGTGTGACCAGCGCGCCCAGGCTGGGTGAGCCCCCACCGTGGGG

Fig. 4.19: Reference Sequence for DMR10

4.6.9 S.9 R-script Hairpinizer V2

The Hairpinizer V2 is a small R script which builds hairpin reference sequences of a region of interest. For this, the user has to provide the genomic sequence of the region of interest, possible restriction enzymes, including their cutting sequence, as well as the sequence of the corresponding hairpin linker. Hairpinizer V2 screens the upper strand for restriction sites, cuts the DNA, attaches the hairpin linker to the restriction site and adds the lower DNA strand sequence to the other side of the hairpin linker. Subsequently, this hairpin construct can then be used for primer design and building of reference sequences for BiQHT. The complete code of the Hairpinizer V2 is provided on the following pages:

The underlying code of the Hairpinizer V2 can be found in the supplement of the original publication [52]

Bibliography

[1] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, 2000.

[2] MJ Ziller, F Muller, J Liao, Y Zhang, and H Gu. Bock c., boyle p., epstein cb, bernstein be, lengauer t., et al. *PLoS Genet*, 7(12):e1002389, 2011.

[3] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.

[4] Heinrich Leonhardt, Andrea W Page, Heinz-Ulrich Weier, and Timothy H Bestor. A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei. *Cell*, 71(5):865–873, 1992.

[5] Linda S-H Chuang, Hang-In Ian, Tong-Wey Koh, Huck-Hui Ng, Guoliang Xu, and Benjamin FL Li. Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1. *Science*, 277(5334):1996–2000, 1997.

[6] Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.

[7] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908, 2007.

[8] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 2004.

[9] Kyohei Arita, Mariko Ariyoshi, Hidehito Tochio, Yusuke Nakamura, and Masahiro Shirakawa. Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. *Nature*, 455(7214):818, 2008.

[10] Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature genetics*, 19(3):219, 1998.

[11] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[12] Daniela Meilinger, Karin Fellinger, Sebastian Bultmann, Ulrich Rothbauer, Ian Marc Bonapace, Wolfgang EF Klinkert, Fabio Spada, and Heinrich Leonhardt. Np95 interacts with de novo dna methyltransferases, dnmt3a and dnmt3b, and mediates epigenetic silencing of the viral cmv promoter in embryonic stem cells. *EMBO reports*, 10(11):1259–1264, 2009.

[13] Gangning Liang, Matilda F Chan, Yoshitaka Tomigahara, Yvonne C Tsai, Felicidad A Gonzales, En Li, Peter W Laird, and Peter A Jones. Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology*, 22(2):480–491, 2002.

[14] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[15] Ryoichi Ono, Tomohiko Taki, Takeshi Taketani, Masafumi Taniwaki, Hajime Kobayashi, and Yasuhide Hayashi. Lcx, leukemia-associated protein with a cxxc domain, is fused to mll in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). *Cancer research*, 62(14):4075–4080, 2002.

[16] RB Lorsbach, J Moore, S Mathew, SC Raimondi, ST Mukatira, and JR Downing. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia*, 17(3):637, 2003.

[17] Lakshminarayan M Iyer, Mamta Tahiliani, Anjana Rao, and L Aravind. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell cycle*, 8(11):1698–1710, 2009.

[18] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[19] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[20] Daniel Globisch, Martin Münzel, Markus Müller, Stylianos Michalakis, Mirko Wagner, Susanne Koch, Tobias Brückl, Martin Biel, and Thomas Carell. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one*, 5(12):e15367, 2010.

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

[22] Aleksandra Szwagierczak, Sebastian Bultmann, Christine S Schmidt, Fabio Spada, and Heinrich Leonhardt. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic dna. *Nucleic acids research*, 38(19):e181–e181, 2010.

[23] Mark Wossidlo, Toshinobu Nakamura, Konstantin Lepikhov, C Joana Marques, Valeri Zakhartchenko, Michele Boiani, Julia Arand, Toru Nakano, Wolf Reik, and Jörn Walter. 5-hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nature communications*, 2:241, 2011.

[24] Tian-Peng Gu, Fan Guo, Hui Yang, Hai-Ping Wu, Gui-Fang Xu, Wei Liu, Zhi-Guo Xie, Linyu Shi, Xinyi He, Seung-gi Jin, et al. The role of tet3 dna dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477(7366):606, 2011.

[25] Hideharu Hashimoto, Yiwei Liu, Anup K Upadhyay, Yanqi Chang, Shelley B Howerton, Paula M Vertino, Xing Zhang, and Xiaodong Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849, 2012.

[26] Victoria Valinluck and Lawrence C Sowers. Endogenous cytosine damage products alter the site selectivity of human dna maintenance methyltransferase dnmt1. *Cancer research*, 67(3):946–950, 2007.

[27] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[28] Atanu Maiti and Alexander C Drohat. Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011.

[29] Liang Zhang, Xingyu Lu, Junyan Lu, Haihua Liang, Qing Dai, Guo-Liang Xu, Cheng Luo, Hualiang Jiang, and Chuan He. Thymine dna glycosylase specifically recognizes 5-carboxylcytosine-modified dna. *Nature chemical biology*, 8(4):328, 2012.

[30] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

[31] Debin Ji, Krystal Lin, Jikui Song, and Yinsheng Wang. Effects of tet-induced oxidation products of 5-methylcytosine on dnmt1-and dnmt3a-mediated cytosine methylation. *Molecular bioSystems*, 10(7):1749–1752, 2014.

[32] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, 2012.

[33] Michael J Booth, Tobias WB Ost, Dario Beraldi, Neil M Bell, Miguel R Branco, Wolf Reik, and Shankar Balasubramanian. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature protocols*, 8(10):1841, 2013.

[34] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[35] Brooks E Miner, Reinhard J Stöger, Alice F Burden, Charles D Laird, and R Scott Hansen. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite pcr. *Nucleic acids research*, 32(17):e135–e135, 2004.

[36] Alice F Burden, Nathan C Manley, Aaron D Clark, Stanley M Gartler, Charles D Laird, and R Scott Hansen. Hemimethylation and non-cpg methylation levels in a promoter region of human line-1 (l1) repeated elements. *Journal of Biological Chemistry*, 280(15):14413–14419, 2005.

[37] Pascal Giehr and Jörn Walter. Hairpin bisulfite sequencing: Synchronous methylation analysis on complementary dna strands of individual chromosomes. In *DNA Methylation Protocols*, pages 573–586. Springer, 2018.

[38] Xingyu Lu, Chun-Xiao Song, Keith Szulwach, Zhipeng Wang, Payton Weidenbacher, Peng Jin, and Chuan He. Chemical modification-assisted bisulfite sequencing (cab-seq) for 5-carboxylcytosine detection in dna. *Journal of the American Chemical Society*, 135(25):9315–9317, 2013.

[39] Chun-Xiao Song, Keith E Szulwach, Qing Dai, Ye Fu, Shi-Qing Mao, Li Lin, Craig Street, Yujing Li, Mickael Poidevin, Hao Wu, et al. Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell*, 153(3):678–691, 2013.

[40] Francesco Neri, Danny Incarnato, Anna Krepelova, Stefania Rapelli, Francesca Anselmi, Caterina Parlato, Claudio Medana, Federica Dal Bello, and Salvatore Oliviero. Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter dna methylation dynamics. *Cell reports*, 10(5):674–683, 2015.

[41] Pavlo Lutsik, Lars Feuerbach, Julia Arand, Thomas Lengauer, Jörn Walter, and Christoph Bock. Biq analyzer ht: locus-specific analysis of dna methylation by high-throughput bisulfite sequencing. *Nucleic acids research*, 39(suppl_2):W551–W556, 2011.

[42] Charalampos Kyriakopoulos, Pascal Giehr, and Verena Wolf. H (o) ta: estimation of dna methylation and hydroxylation levels and efficiencies from time course data. *Bioinformatics*, 33(11):1733–1734, 2017.

[43] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[44] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[45] Marius Walter, Aurelie Teissandier, Raquel Perez-Palacios, and Deborah Bourc'his. An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. *Elife*, 5:e11418, 2016.

[46] Ferdinand von Meyenn, Mario Iurlaro, Ehsan Habibi, Ning Qing Liu, Ali Salehzadeh-Yazdi, Fátima Santos, Edoardo Petrini, Inês Milagre, Miao Yu, Zhenqing Xie, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861, 2016.

[47] Peter WS Hill, Harry G Leitch, Cristina E Requena, Zhiyi Sun, Rachel Amouroux, Monica Roman-Trufero, Malgorzata Borkowska, Jolyon Terragni, Romualdas Vaisvila, Sarah Linnett, et al. Epigenetic reprogramming enables the transition from primordial germ cell to gonocyte. *Nature*, 555(7696):392, 2018.

[48] Stefan Wallner, Christopher Schröder, Elsa Leitão, Tea Berulava, Claudia Haak, Daniela Beißer, Sven Rahmann, Andreas S Richter, Thomas Manke, Ulrike Bönisch, et al. Epigenetic dynamics of monocyte-to-macrophage differentiation. *Epigenetics & chromatin*, 9(1):33, 2016.

[49] Julia Arand, Mark Wossidlo, Konstantin Lepikhov, Julian R Peat, Wolf Reik, and Jörn Walter. Selective impairment of methylation maintenance is the major cause of dna methylation reprogramming in the early embryo. *Epigenetics & chromatin*, 8(1):1, 2015.

[50] Michael J Booth, Giovanni Marsico, Martin Bachman, Dario Beraldi, and Shankar Balasubramanian. Quantitative sequencing of 5-formylcytosine in dna at single-base resolution. *Nature chemistry*, 6(5):435, 2014.

[51] Chenxu Zhu, Yun Gao, Hongshan Guo, Bo Xia, Jinghui Song, Xinglong Wu, Hu Zeng, Kehkooi Kee, Fuchou Tang, and Chengqi Yi. Single-cell 5-formylcytosine landscapes of mammalian early embryos and escs at single-base resolution. *Cell Stem Cell*, 20(5):720–731, 2017.

[52] Pascal Giehr, Charalampos Kyriakopoulos, Konstantin Lepikhov, Stefan Wallner, Verena Wolf, and Jörn Walter. Two are better than one: Hpoxbs-hairpin oxidative bisulfite sequencing. *Nucleic acids research*, 46(15):e88–e88, 2018.

5. GENOME WIDE EFFICIENCY PROFILING OF SINGLE CPGS REVEALS MODULATION OF MAINTENANCE AND *DE NOVO* METHYLATION BY TET DI-OXYGENASES

The content of chapter 5 is taken from a provisional manuscript under the working title: Giehr, P.[†], Kyriakopoulos, C.[*], Nordström, K.[†], Salhab, A.[†], Müller, F. [§], von Meyenn, F. [¶], Ficz, G. [‖], Reik, W.[**], Wolf, V.[*], & Walter, J.[†] (2019). Genome Wide Efficiency Profiling of Single CpGs Reveals Modulation of Maintenance and *De Novo* Methylation by Tet Di-Oxygenases.

AUTHOR CONTRIBUTIONS

Pascal Giehr: Development and conduction of reduced representative hairpin oxidative bisulfite sequencing. Data comparison to replication time points, segmentation profiles and genomic context. Profiling of nonCpG methylation. Model's output analysis including DEEP-Tool application, as well as data comparison using R.

Authoring of abstract, background, methods - section 5.2.1 “Reduced Representation Hairpin Oxidative Bisulfite Sequencing” and section 5.2.6 “LOLA Analysis”, Results including the generation of figures 5.1 - 5.8, 5.10, 5.11, as well as Discussion and Conclusion. In addition, authoring of supplement section 5.6.1 “RRHPoxBS”, as well as parts of section 5.6.5 concerning additional model's output comparisons and biological interpretation of the data including the generation of figures 5.23 - 5.25, as well as 5.34 - 5.35.

Dr. Charalampos Kyriakopoulos: Design and development of the model. Design and development of the statistical and computational tools for genome wide enzymatic activities' prediction and (hydroxy-)methylation levels profiling. Model's output analysis including spatial correlations and enzymatic activities clustering.

[†] Department of Biological Sciences, Saarland University, D-66123 Saarbrücken, Germany

[*] Computer Science Department, Saarland University, Campus E1.3, D-66123 Saarbrücken, Saarland, Germany

[§] Department of Genetics, Stanford University, 5120 Stanford, USA

[¶] Department of Health Sciences and Technology, ETH Zürich, 8603 Schwerzenbach, Switzerland

[‖] Centre of Haemato-Oncology, Barts Cancer Institute, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ

[**] Epigenetics Program, The Babraham Institute, Cambridge, CB22 AT, UK

Authoring of Methods section 5.2.3 “Modeling” and 5.2.4 “Clustering of Single CpG Efficiencies”, results concerning spatial clustering, including generation of figure 5.9. Authoring of supplement, section 5.6.2 “Hidden Markov Model” including figures 5.13 - 5.22, section 5.6.3 “Maximum Likelihood Estimation (MLE)”, as well as part of section 5.6.5, “Additional Results” concerning “ES Cell Chromosomes Results” (Figure 5.26 - 5.33).

Dr. Karl Nordström: Processing of sequencing data and generation of primary hairpin sequencing output, meta data analysis in form of genetic annotation of CpGs. Contributing in authoring of Methods concerning the processing of sequencing data.

Abdulrahman Salhab: Reprocessing of chromatin immuno precipitation sequencing data, preparation of methylation segmentation data. Contributing in authoring of Methods concerning the application of MethylSeekR.

Dr. Fabian Müller: Conduction of LOLA enrichment analysis. Revision of the manuscript, commenting and advising on data analysis and formulations within the manuscript.

Prof. Dr. Ferdinand von Mayenn: Cultivation of embryonic stem cells and isolation of genomic DNA. Revision of the manuscript, commenting and advising on data analysis and formulations within the manuscript.

Dr. Gabriella Ficz: Providing relevant DNA samples to analyse. Revision of the manuscript, commenting and advising on data analysis and formulations within the manuscript.

Prof. Dr. Wolf Reik: Supervision. Consulting. Revision of the manuscript, commenting and advising on data analysis and formulations within the manuscript.

Prof. Dr. Verena Wolf: Supervision of Hidden Markov Modelling, clustering and spatial correlation. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Prof. Dr. Jörn Walter: Supervision of experimental work. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Abstract

Background DNA methylation is an essential epigenetic modification, which is set and maintained by DNA methyl transferases (Dnmts) and removed via active and passive mechanisms involving Tet mediated oxidation. Yet, activity profiles of Dnmts and Tets at base resolution have not been achieved so far. Here we present a novel combination of precise genome wide mapping of 5mC and 5hmC and a HMM based analysis to determine the enzymatic contributions during developmental changes at CpG resolution.

Results We describe a novel reduced representation hairpin oxidative bisulfite sequencing (RRHPoxBS) approach, which allows to measure and compare 5mC and 5hmC at single CpGs in DNA double strands. We apply this method to follow the genome wide 5mC/5hmC changes during a Serum-to-2i shift in ES cells comparing the progressive demethylation of WT and Tet Triple KO ES cells over time. On these data we apply an extended genome-wide version of our previously presented Hidden Markov Model to calculate the efficiencies of Dnmts and Tets along the genome at CpG resolution. We find that Dnmts and Tets exhibit antagonistic effects on methylation stability in a very contextual manner with strong links to gene structure and other genetic and epigenetic control factors.

Conclusions Our data show a very strong dynamic and contextual control of complementary and antagonistic Dnmt and Tet activities. *De novo* and maintenance methylation activities are most pronounced across gene bodies and promoters of inactive genes, while Tets exhibit their highest activity around unmethylated regulatory elements, i.e. at active promoters and enhancers. The absence of Tets leads to a misregulation of Dnmts resulting in a more persistent *de novo* methylation activity and an ectopic maintenance efficiency.

5.1 Background

Transcriptional access to genetic information encoded in the DNA is regulated by epigenetic mechanisms, such as DNA methylation [1, 2, 3, 4]. In mammals, the methylation of DNA is restricted to cytosine and is almost exclusively found in a palindromic CpG di-nucleotide context [5, 6, 7]. Generation of 5-methylcytosine (5mC) is controlled by the DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. These enzymes catalyse the transfer of a methyl group from s-adenosyl methionine to the fifth carbon atom of cytosine.

Dnmt1 is responsible for maintaining an existing methylation pattern after replication. Via interaction with Uhrf1 and PCNA, Dnmt1 is tightly associated with the replication machinery [8, 9]. Furthermore, the cooperation with Uhrf1 modulates Dnmt1 to be receptive for hemimethylated DNA generated after replication [10, 11]. Thus, the protein complex post-replicatively copies the methylation pattern from the inherited to the newly synthesised DNA strand [12, 13]. Dnmt3a and Dnmt3b methylate DNA independently of its methylation status (hemimethylated or unmethylated) and are mainly responsible for the establishment of new methylation patterns during development [14, 15].

However, several studies indicate that the strict separation of Dnmt1 and Dnmt3a/b activity is not coherent and that under certain conditions, these enzymes exhibit overlapping functions [16, 17, 18].

Once established, 5mC can be further processed by a family of di-oxigenases, the ten-eleven translocation enzymes Tet1, Tet2 and Tet3 [19, 20, 21]. These Fe(II) and oxo-glutarate-dependent enzymes consecutively oxidise 5mC to 5-hydroxymethyl cytosine (5hmC), 5-formyl cytosine (5fC) and ultimately to 5-carboxy cytosine (5caC) [22, 23]. 5hmC is the most abundant oxidative variant and can be found in numerous cell types [24, 25, 26]. Each oxidation step changes the chemical properties of the base and with it its biological function [27, 28, 29]. Several mechanisms have been proposed in which oxidative cytosine derivatives (oxC) serve as an intermediate during the course of active or passive demethylation [30, 31, 32, 33, 34]. Such removal of 5mC occurs locally during cell differentiation, but also on a genome wide scale in the zygote, as well as during the maturation of primordial germ cells (PGCs) [35, 36, 37].

Global loss of 5mC can also be observed in cultivated mouse embryonic stem cells (ES cells) during their transition from Serum to 2i medium. Under classical Serum/LIF conditions, ES cells exhibit DNA hypermethylation, whereas upon transition to GSK3β and Erk1/2 inhibitors (2i) containing medium, the cells experience a gradual genome wide loss of 5mC [38, 39, 40]. Even though the enzymatic mechanisms of oxCs generation are well characterised, the question how oxCs affect the maintenance of DNA methylation or to which extent they contribute in changing methylation patterns remains still elusive.

In order to address this question we developed **R**educed **R**epresentation **H**air**p**in **ox**idative **B**isulfite **S**equencing (RRHPoxBS). Our method combines three essential features: (i) a genome wide detection of a defined number of CpGs using restriction enzymes (REs), (ii) a strand-specific detection of 5mC based on covalent bonding of top and bottom strand by ligating a short hairpin oligo (HP) and lastly, (iii) the simultaneous detection of 5mC and 5hmC using oxidative bisulfite sequencing (oxBS) [41, 42, 43, 44, 45]. We applied RRHPoxBS to the above mentioned Serum-to-2i-shift of wild type (WT) and Tet triple knock-out (Tet TKO) mouse ES cells to investigate the role of Tets and 5hmC in the progress of demethylation.

Using an extended version of a Hidden Markov Model (HMM), first presented in [46, 47], we predict the levels and the strand specific distribution of 5mC and 5hmC. In addition, we estimate efficiencies for maintenance methylation, *de novo* methylation, and hydroxylation of Dnmts and Tets over time in 2i. We perform a combined HMM analysis of the observations of oxidative and non-oxidative bisulfite sequencing to obtain a very accurate estimation of methylation and hydroxylation activities. We then develop a sophisticated clustering approach for the corresponding methylation and hydroxylation efficiencies in order to identify the main profiles of enzymatic activity.

We find that Tet activity is highest at unmethylated regions such as promoters and transcription factor binding sites(TFBS). By computing spatial cross-correlations, we show that the methylation and hydroxylation efficiencies are negatively correlated throughout the genome. Finally, we observe that the absence of Tet enzymes in TKO cells leads to a change in the activity profile of maintenance methylation, which is further pronounced at regions originally protected by Tet enzymes in WT ES cells.

5.2 Methods

5.2.1 Reduced Representation Hairpin Oxidative Bisulfite Sequencing (RRHPoxBS)

1.2μg DNA is divided equally into three 0.5ml reaction tubes and digested in a 20μl reaction using 20U of either HaeIII (NEB), AluI (NEB) or HpyCH4V (NEB), respectively. The reactions are incubated over night for a minimum of 12h at 37℃. Restriction enzymes are inactivated at 80℃ for 30min. The reactions are pooled and subjected to a ligation step. During this process, hairpin linker and sequencing adapter are introduced to the opposing ends of each restriction fragment. For this, 200mM biotin labeled hairpin linker and 100mM sequencing adapter are added to the digested DNA, incubated with 1mM ATP and 4000U T4 DNA Ligase (NEB) for 16h at 16℃. The reaction is purified using AMPureXP beads followed by enrichment for hairpin containing fragments with streptavidin beads. The library is then subjected to BS/oxBS work-flow of the TrueMethyl kit from CEGX according to manufacturer's instructions. Amplification of the library was done with HotStarTaq® polymerase (QIAGEN) and sequencing was performed on an Illumina HiSeq2500 platform in a 150bp paired-end sequencing mode. (see Figure 5.12 for schematic overview)

5.2.2 Read Mapping and Methylation Calling

The sequences are aligned as suggested by Porter *et al.* [48]. In detail; reads are trimmed for adapter, hairpin-linker and 3' quality (Q$\geq$20) with TrimGalore! [49] and cutadapt [50]. Trimmed read pairs are aligned with the Smith-Waterman algorithm allowing for bisulfite induced mismatches. The two bisulfite converted strands are used to deduce the original genomic sequence. Mismatches other than G-to-A and T-to-C are replaced with N. The resulting sequences are aligned to the mouse genome (mm10) with GEM-mapper (beta build 1.376) [51], after which the methylation information is reintroduced with a custom pileup function based on HTSJDK [52] and ratios for the four methylation states are calculated for each cytosine. The pipeline was implemented with the Ruffus pipeline framework [53].

5.2.3 Modeling

Estimation of (Hydroxy-)Methylation Levels

For CpGs with observations at up to two time points, we combined information from BS and oxBS experiments to arrive at maximum likelihood estimates (MLEs) for strand specific (hydroxy-)methylation levels for each observation time point. The derived MLEs take into account the conversion errors of each experiment and we estimate their accuracy by approximating the corresponding standard deviations. For details see Section S5.6.2

Estimation of Enzyme Efficiencies

For those CpGs, for which the maximal number of three observation time points is available, we defined an underlying discrete time Markov process that shapes the demethylation dynamics. The state space of the process is the set of possible CpG site's state $s \in \mathcal{S} = \{u, m, h\}^2$, where state s encodes whether the upper and the lower strand of the site is *unmethylated* (u), *methylated* (m) or *hydroxylated* (h). *E.g.* in state (h, u)* the upper strand is hydroxylated and the lower strand is unmethylated. The model's time parameter corresponds to the number of cell divisions and the transitions of a state are being triggered by consecutive division or (hydroxy-) methylation events. Getting measurements along with the conversion errors from two different experiments (BS and oxBS) allows us to define one HMM for each experiment and get estimates for the model's parameters. The last are linear functions that represent the enzymes' efficiencies over time. In addition, a parameter related to the recognition of 5hmC by Dnmts (passive demethylation) is being estimated for each CpG. For a detailed presentation of the above model we refer the reader to Giehr *et al.* 2016, and Kyriakopoulos *et al.*, 2017 [47, 46], as well as to Section S5.6.2.

In case of an adequate number of samples per time point, when a very deep sequencing is possible, the MLE provides accurate estimates with narrow confidence intervals [47]. On the other hand, MLE is known to give imprecise results for a smaller number of samples [54, 55] and in particular in cases where the true values are close to the boundary constraints [56]. Since a consistently deep sequencing ($\geq$100x) is under the current methods impossible on a genome wide level, we develop here a combination of MLE and Bayesian Inference (BI) methods in order to get accurate estimates even in genome regions with less deeper sequencing. In particular, we use a MLE step as initial information to be given to a Metropolis-Hastings MCMC sampler, from which we get the posterior distribution of the parameters. We do verify that a BI method is indeed necessary here to get meaningful results. The approach is being described in detail in Section S5.6.4

*For simplicity we will often write *hu* etc. instead.

5.2.4 Clustering of Single CpG Efficiencies

We cluster the genome wide output of our model, meaning the efficiencies of the enzymes responsible for maintenance, *de novo* and hydroxylation over time for $1.5 \cdot 10^6$ CpGs, uniformly located over the entire genome. Since we aim to cluster parameter estimates, we consider a sophisticated clustering approach that takes into account the uncertainty, *i.e*, posterior's covariance matrix, around the BI estimators, *i.e.*, posterior's mean. The clustering approach we apply here gives a different and an evidently more natural 'optimal' number of clusters than a typical k-means clustering algorithm would return.

5.2.5 Segmentation

The whole genome bisulfite data of primed mouse ES cells (Ficz *et al.* 2013) was segmented into low methylation regions (LMRs), unmethylated regions (UMRs) and partially methylated domains (PMDs) [38], using MethylSeekR [57]. The rest of the genome, after filtering gaps annotated by UCSC, was called highly methylated regions (HMRs) [58]. The aggregated strand information per CpG was used as an input for MethylSeekR. The used parameters were the following; coverage of $\geq$5x, $\leq$50% methylation and FDR <0.05 for calling hypomethylation regions and consequently a cutoff of ≥ 4 CpGs per LMR.

5.2.6 LOLA Analysis

We performed a standard LOLA analysis against the regular LOLA universe, extended by ChIP-Seq profiles from von Meyenn *et al.*, 2016 (GSE70724, GSE77420) and Walter *et al.*, 2016 (GSE71593) [40, 59, 60].

5.3 Results

Previous studies showed a dynamic loss of DNA methylation in mouse ES cells during their transition from Serum/LIF to 2i containing medium and only partially address the contribution of Dnmt and Tet efficiencies [38, 39, 60]. In this context, the estimation of enzyme efficiencies relies on aggregated methylation data and almost exclusively on classical single strand information [60]. The aim of this study was to model the methylation changes of single CpGs during Serum-to-2i transition using precise strand specific information, obtained by genome wide hairpin sequencing under conventional BS and oxidative bisulfite (oxBS) conditions. To achieve this, we here describe the first application of a combined hairpin BS and oxBS sequencing pipeline at a reduced representation level. Following our approach we reach in total around 3 million CpGs in both WT and Tet TKO cells across the mouse genome with a sequencing depth sufficient for comparative modeling. After modelling, we obtain enzymatic efficiencies for individual CpGs, which permits a more detailed investigation of Dnmt and Tet activities.

We sequenced six hairpin libraries of WT ES cells at three different time points: Serum/LIF (d0), 72h 2i (d3) and 144h 2i (d6), and four Tet triple KO cells (Serum/LIF, 24h 2i, 48h 2i and 96h 2i). For WT we sequenced one BS and one oxBS library for each sample, respectively. In the case of Tet TKO ES cells, four HPBS libraries were sequenced: Serum/LIF (d0), 48h 2i (d1), 96h 2i (d4) and 168h (d7). Using an extended version of our previously described HMM [47], we calculate the (hydroxy-) methylation levels and the detailed distribution of 5hmC, and in addition, we estimate the efficiencies of Dnmt and Tet enzymes for each individual CpG. Taking advantage of the strand specific information, we distinguish in the case of Dnmts between maintenance and *de novo* methylation events. At last, the comparison of WT and TKO cells allows us to determine any changes in maintenance and *de novo* methylation efficiency in the absence of Tet enzymes and oxidised cytosine derivatives.

5.3.1 Tet TKO Cells Display Reduced Demethylation Rates

First, we determined the level and distribution of 5mC within the obtained RRHPoxBS data. In line with previous reports [38, 39, 60] we observe an overall level of 65% CpG methylation in primed ES cells (Serum/LIF) and a consecutive loss of methylation upon cultivation in 2i to 20% (Figure 5.1A). The majority of methylated CpGs is present in a symmetric methylation state under both cultivation conditions. Furthermore, hemimethylation of CpGs seems to be equally distributed among both DNA strands. However, the level of hemimethylated CpGs is always lower in oxBS samples, indicating that a considerable amount of 5hmC is present in a hemi(hydroxy)methylated (5hmC/C or C/5hmC) state.

Tet TKO cells present a higher methylation level under both, primed (Serum/LIF) (75%) and naive (2i) (40%) conditions. Hemimethylation in Tet TKO cells at d0 is strongly reduced compared to WT cells. In addition, we observe that Tet TKO cells exhibit a reduced demethylation rate in comparison to WT ES cells. In order to determine the difference in the demethylation kinetics, we calculated the increase of unmethylated CpGs per day (Supplement Section S5.6.5, Figure S5.35).WT ES cells show an increase of unmethylated CpGs of more than 8%, where as Tet TKO cells exhibit demethylation rates of 4.2%. Hence, demethylation in the absence of Tet enzymes is reduced by around 50%, indicating that Tet oxidation notably contributes to DNA demethylation in the present system.

We observe a similar behavior for nonCpG methylation. WT ES cells show approximately 1% nonCpG methylation at d0, which quickly declines in 2i (Figure 5.1B). In contrast, Tet TKO cells exhibit twice as much nonCpG methylation under Serum/LIF conditions and furthermore, nonCpG methylation seems to be more stable during 2i cultivation under naive conditions (5.1 C). Even at day7, we still detect considerable levels of methylated CpGs in a nonCpG context (Figure 5.1D). In accordance with previous ob-

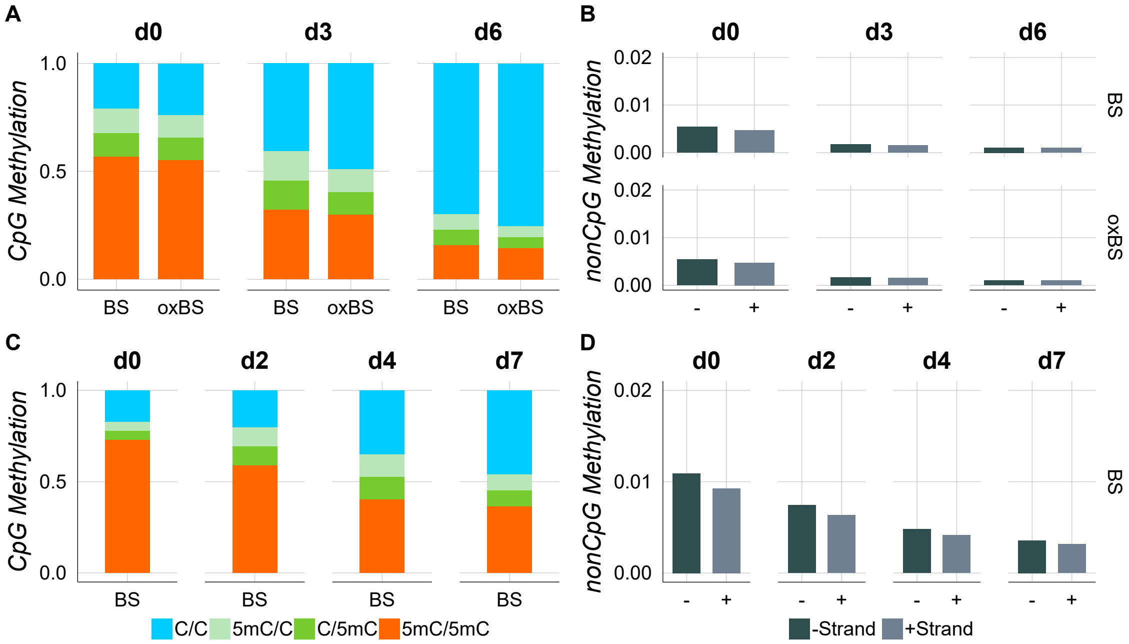


Fig. 5.1: Average CpG and nonCpG Methylation. **(A)** Genome wide average CpG methylation level of wild type ES cells cultivated under Serum/LIF conditions (d0), and their shift to 2i after 72h (d3), 144h (d6). **(B)** Average nonCpG methylation level of WT cells. **(C)** Average CpG methylation pattern of Tet triple knockout ES cells. **(D)** Average nonCpG methylation level of Tet triple knockout ES cells.

servations, the majority of nonCpG methylation is present in a CpA context and mostly located in regions with high CpG methylation levels (Supplement, Section S5.6.5, Figure S5.43-S5.45).

5.3.2 Tets are More Active at Accessible Chromatin

Using Bayesian Inference (BI) on the HMM parameters (Section S5.6.4), we predicted the efficiencies of maintenance methylation, *de novo* methylation and hydroxylation activity based on BS and oxBS data for WT ES cells. The efficiencies describe the enzymatic activities of Dnmts and Tets in 2i, which facilitate the continuous methylation loss under these conditions. First, we investigated the efficiency profiles of Dnmts and Tets across genes. Both, maintenance and *de novo* methylation activity show initially high efficiencies at the gene body ($\geq$ 0.6 for maintenance, and $\geq$ 0.2 for *de novo*), whereas at the transcription start site, efficiencies are strongly reduced (Figure 5.2A). In the case of *de novo* methylation, efficiency drops almost to zero. The maintenance efficiency of 0.6 is relatively low compared to previous estimations under Serum/LIF conditions (minimum 0.9), indicating a notable reduction of maintenance methylation in 2i medium [60]. Hydroxylation on the other hand shows an inverse behaviour: Reduced activity at the gene body and high efficiency at the transcription start site (TSS). Over time we observe an increase of Tet activity at the gene body, whereas *de novo* activity shows a strong reduc-

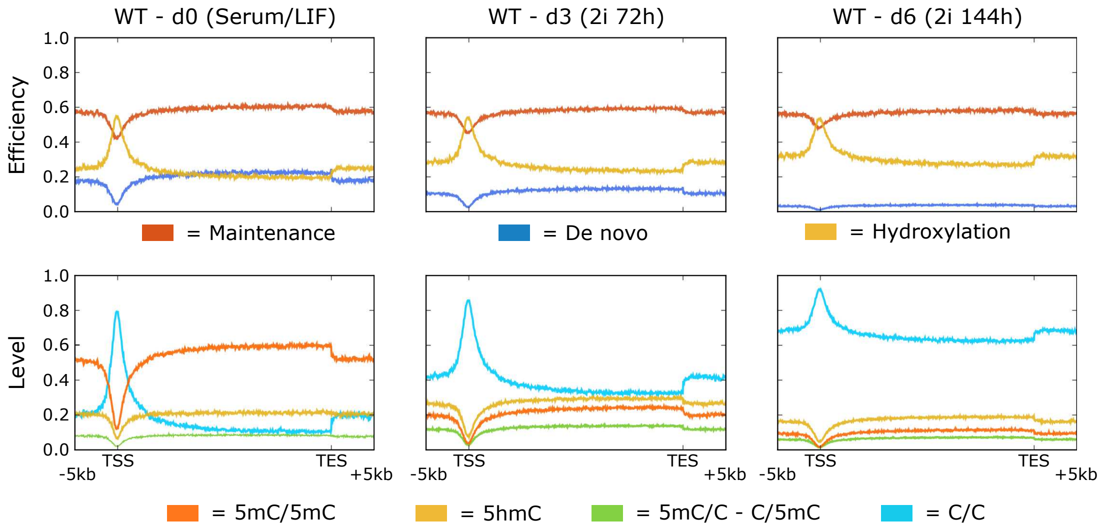
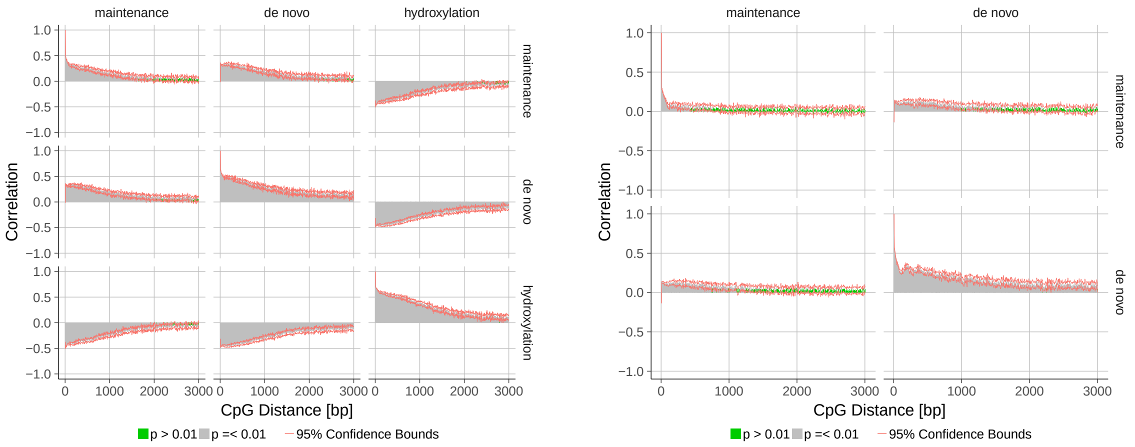


Fig. 5.2: Average enzyme efficiencies and CpG methylation level of WT ES cells. Red = Average enzyme maintenance, blue = *de novo* and yellow = hydroxylation activity across genes during Serum-to-2i transition **(upper panel)**. Average CpG methylation level across genes during Serum-to-2i shift **(lower panel)**. Red = Symmetric methylated CpG dyads (mCpG/mCpG); Dark green asymmetric CpG methylation (mCpG/CpG), Light green (CpG/mCpG), Blue = unmethylated CpG dyads (CpG/CpG). TSS = transcription start site, TES = transcription end site



(*a*) WT - day 0

(*b*) Tet TKO - day 0

Fig. 5.3: Spatial correlation of enzyme efficiencies in WT and Tet TKO cells. **(A)** Auto- and cross correlation of maintenance, *de novo* and hydroxylation efficiency at day0. Y-axis displays correlation, x-axis distance of CpGs in base pairs. Red lines indicate confident intervals. grey bars = correlation with p-value ≤ 0.01, green = correlation with p-value > 0.01. Correlation in WT cell, **(B)** correlation in Tet TKO cells.

tion. The temporal profile of maintenance activity suggests no further changes at the gene body, which suggests that the physiological changes affecting maintenance methylation represent early events and are completed within the first 24h.

Concerning their profiles, 5mC and 5hmC level resemble those of *de novo* and maintenance methylation efficiencies. Both modifications are enriched at the gene body and reduced at the TSS. The reduced levels of 5hmC at already d0 across the TSS, despite the strong Tet activity is on the one hand the consequence of missing substrate, but may also suggest a further processing and constant turnover of 5mC and 5hmC (Figure 5.2B).

To further investigate the observed enzymatic antagonism, we calculated the spatial auto and cross-correlations of methylation and hydroxylation efficiencies. In line with the efficiency profiles (Figure 5.2), we consistently see a positive correlation between maintenance and *de novo* efficiency and, in addition, a negative correlation of hydroxylation with both methylation efficiencies. With increasing distance of CpGs, correlations get closer to zero (Figure 5.3a, and Supplement Section S5.6.4, Figure 5.23). Maintenance autocorrelation drops rather quickly and becomes almost zero at around 1500bp. After this point, correlation is also no longer significant (p-value > 0.01). In contrast, autocorrelation of *de novo* and hydroxylation efficiency show initially higher values but also seem to smoothen out after around 2000bp on average. The Pearson correlation between the (hydroxy-)methylation levels and the enzyme activities hints towards a causal relationship of hydroxylation and unmethylated CpGs (C/C), indicating that Tet activity might maintain the unmethylated state (Section S5.6.4, Figure 5.24).

Interestingly, in Tet TKO cells, autocorrelation of maintenance methylation efficiency is initially strongly reduced (≈ 0.25), shows a strong spatial decrease and a loss of statistical significance ($p > 0.01$) much earlier, at around 500bp, whereas *de novo* autocorrelation seems to remain unaffected, suggesting a misregulation of maintenance methylation in the absence of Tets (Figure 5.3b).

The inverse behaviour of methylation and hydroxylation activity can also be observed at protein binding sites and selected histone marks (Figure 5.4). In case of H3K9me2, known to recruit Dnmt1, we see higher maintenance and *de novo* methylation efficiency together with a strongly reduced hydroxylation activity [61, 60]. Similar profiles were also observed for H3K9me3 and H3K36me3. In case of the open chromatin mark H3K4me3, the model reveals a high hydroxylation and reduced methylation activity, which is in agreement with the observation, that Tet1 preferentially binds to H3K4me3 rich regions [62, 63]. In addition, we observe a high hydroxylation efficiency for binding sites of Tet1, Sox2, Nanog and Oct4, accompanied by a reduction of maintenance and *de novo* methylation efficiency. Again, *de novo* methylation is almost zero at the centre of these binding sites. Taken together, the efficiency profiles indicate a higher activity for Tet enzymes at open and accessible chromatin.

In Tet TKO cells, the methylation efficiencies show similar tendencies across genes, histone modifications and protein binding sites. However, compared to WT data, there are distinct differences in maintenance and *de novo* methylation activity in the absence of Tet enzymes. For instance, maintenance methylation activity in Tet TKO is elevated at the TSS compared to WT ES cells (Figure 5.5). This increase of efficiency becomes even

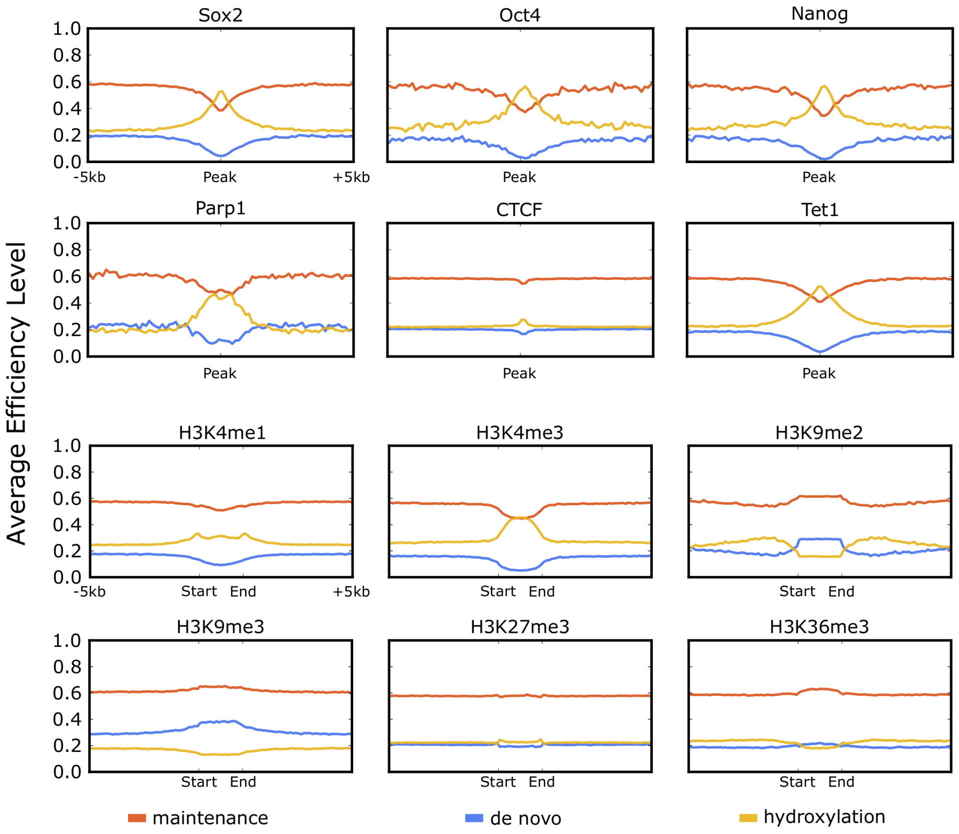


Fig. 5.4: Enzyme efficiencies at TFBS and across histone marks. Average efficiency profiles for Dnmts and Tets at TFBS and across histone modification in WT ES cells. red = maintenance methylation, blue = *de novo* methylation, yellow = hydroxylation.

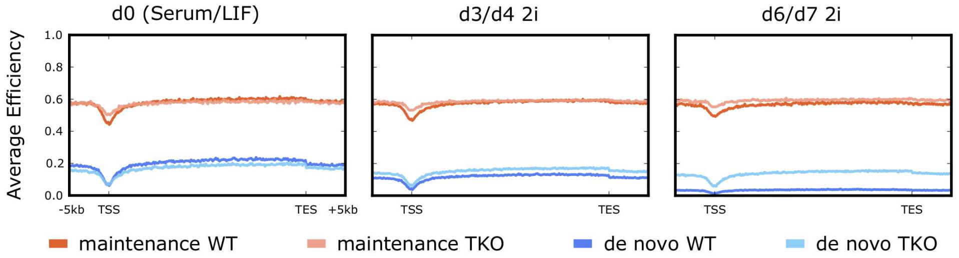


Fig. 5.5: Efficiencies in WT and TKO ES cells. Comparison of maintenance and *de novo* methylation efficiencies in WT and TKO ES cells. red = maintenance WT, light red = maintenance TKO, blue = *de novo* WT, light blue = *de novo* TKO.

more pronounced over time. Initially, *de novo* methylation in Tet TKO cells shows no visible increase either at the TSS or the gene body. It is particularly compelling, that it seems to decrease much slower over time than it does in WT ES cells. At d6 of the WT, *de novo* efficiency is almost zero, whereas in the Tet TKO cells, there is a considerable amount of *de novo* activity at d7.

The seemingly very accurate prediction of remaining *de novo* activity of our model can be independently verified by the observed elevated nonCpG methylation in Tet TKO ES cells (see Figure 5.1D). Consequently, the observed increase in maintenance and *de novo* methylation efficiencies in Tet TKO ES cells indicate an inhibiting effect of Tet enzymes towards maintenance and *de novo* methylation events.

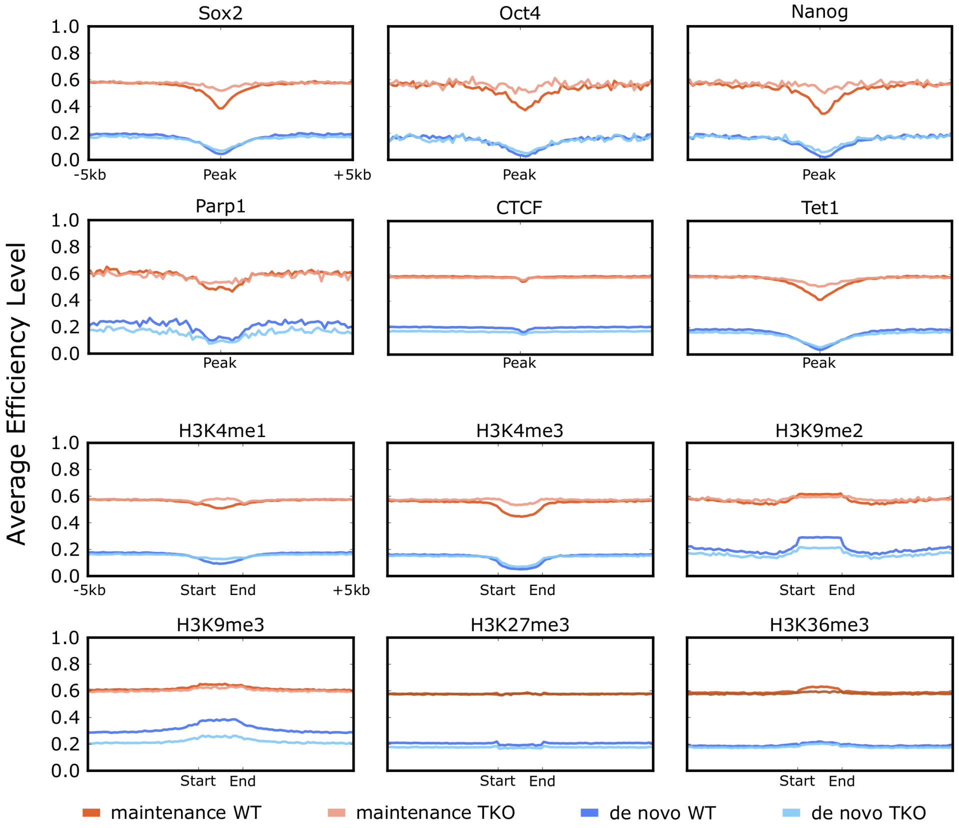


Fig. 5.6: Efficiencies in WT and TKO ES cells. Comparison of maintenance and *de novo* methylation efficiencies in WT and TKO ES cells. Red = maintenance WT, light red = maintenance TKO, blue = *de novo* WT, light blue = *de novo* TKO.

The most pronounced increase in maintenance methylation efficiency can be observed at the binding sites of Tet1 and the pluripotency factors Nanog, Sox2 and Oct4. In addition, we see a strong increase in regions which display H3K4me1 and H3K4me3 in WT ES cells (Figure 5.6).

5.3.3 Distinct Profiles at Highly and Lowly Expressed Genes

Methylation at promoters and TSS is known to correlate with gene expression [64, 65, 66]. We investigated whether the enzyme efficiencies show similar relations. For our analysis, we used a previous published transcriptome of mouse ES cells under Serum/LIF conditions [38]. Calculating the median of transcripts per million reads (TPM), we considered genes with a TPM ≥ 0.065 as highly expressed and genes with a TPM < 0.065 as none/lowly expressed.

Profiles of highly expressed genes (Figure 5.7) match nicely those of the averaged efficiency profiles across all genes (Figure 5.2). However, none/lowly transcribed genes show a diverse pattern, particularly at the TSS. Compared to the expressed genes, we

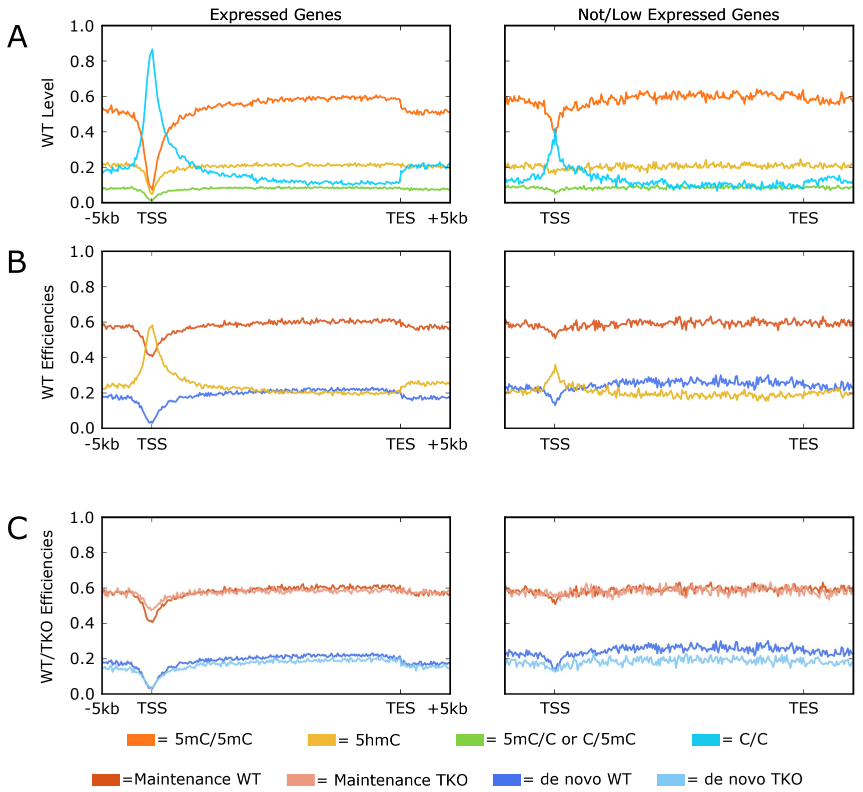


Fig. 5.7: Methylation and enzyme efficiency of expressed and non-expressed genes. Methylation level across expressed and none/lowly expressed genes under Serum/LIF conditions; red = 5mC/5mC, yellow = 5hmC, green = 5mC/C or C/5hmC, blue = C/C **(A)**. Average efficiency profile across expressed and non/lowly expressed genes **(B)**. Comparison of WT and TKO cell efficiency across expressed and none/lowly expressed genes **(C)**. Red = maintenance WT, light red = maintenance TKO, blue = *de novo* WT, light blue = *de novo* TKO, yellow = hydroxylation.

observe higher maintenance and *de novo* activity, but reduced efficiency of Tet enzymes, which nicely agrees with the higher methylation state of ‘inactive’ gene promoters (Figure 5.7). In addition, we capture an increase in *de novo* efficiency across the gene body.

In Tet TKO cells, we observe again compared to WT an increase of maintenance efficiency at the TSS for expressed, but only a mild change for none/lowly expressed genes. However, none/lowly expressed genes show a reduced *de novo* activity across the gene body.

5.3.4 Enzyme Efficiencies Shape the Large-Scale Methylome

Based on the methylation frequency of CpGs, the genome can be segmented into large scale methylated domains and small regulatory regions with low methylation levels [7, 67]. Thus, we used MethylSeekR, a computational method published by Burger *et al.* 2013, to subset the genome into four distinct segment classes: Highly methylated regions (HMRs), partially methylated domains (PMDs), low methylated regions (LMRs) as well as unmethylated regions (UMRs)[57]. For optimal segmentation, we used a whole genome bisulfite sequencing data set of primed mES cells, published by Ficz *et al.* in 2013 and subsequently compared the segmentation to RRHPoxBS data set.

Considering the number and the size of the individual segments, we find that under primed conditions the majority of the genome is assigned to HMRs. This is expected, since ES cells kept under Serum/LIF exhibit a hypermethylated phenotype (HMRs: 85.5%, PMDs: 12.6%, LMRs: 0.4%, UMRs: 1.5%) (Figure 5.8A, 5.8B). Next, we assigned the methylation level derived from double strand information, as well as the distribution of 5hmC and the enzyme efficiencies to each segment type. Consistent with the genome wide methylation data, we see high levels of 5mC at HMRs and PMDs, whereas LMRs and UMRs exhibit low methylation levels (Fig. 5.8E). Despite their low methylation levels, LMRs exhibit relatively high levels of 5hmC, which also occurs more frequently as a fully hydroxylated CpG dyad (5hmC/5hmC state) (Figure 5.8E, 5.8G).

In terms of enzymatic activity, we observe high maintenance and *de novo* methylation efficiencies together with moderate hydroxylation activity in HMRs and PMDs, while LMRs and UMRs display high hydroxylation activity and strongly reduced methylation efficiencies.

5.3.5 Late Replication Accompanies High Methylation

The truthful inheritance of DNA methylation patterns can only be ensured by correct maintenance activity. Potentially, the timing of DNA replication might influence the efficiency of maintenance methylation. Therefore, we compared the replication timing of the distinct segments using the replication information of three ES cell lines published by Hiratani *et al.* [68]. In this context, we observe that HMRs tend to replicate late, whereas PMDs, LMRs and UMRs are replicating earlier (Figure 5.8D).

5.3.6 Individual Efficiency Profiles of CpGs

The gene wide profiles suggest that CpGs display distinct combinations of efficiencies depending on their genomic location. Therefore, we clustered individual CpGs based on their efficiencies and temporal changes in order to identify distinct enzyme kinetics during the Serum-to-2i transition. We consider a clustering method [69] that takes into account the uncertainty, the covariance matrix of the posterior distribution, around the estimated parameters in our case. While a typical k-means clustering algorithm would return four as the optimal number of clusters, our approach clearly decides for two clusters suggesting that the remaining clusters are probably the result of noise on the estimation due to the insufficient depth of the sequencing (see Figure 5.9). For details on the clustering method, taking into account the estimation uncertainty, we refer the reader to Supplement, Section S5.6.4.

Cluster 1 contains 855201 CpGs and it is characteried by high maintenance (about

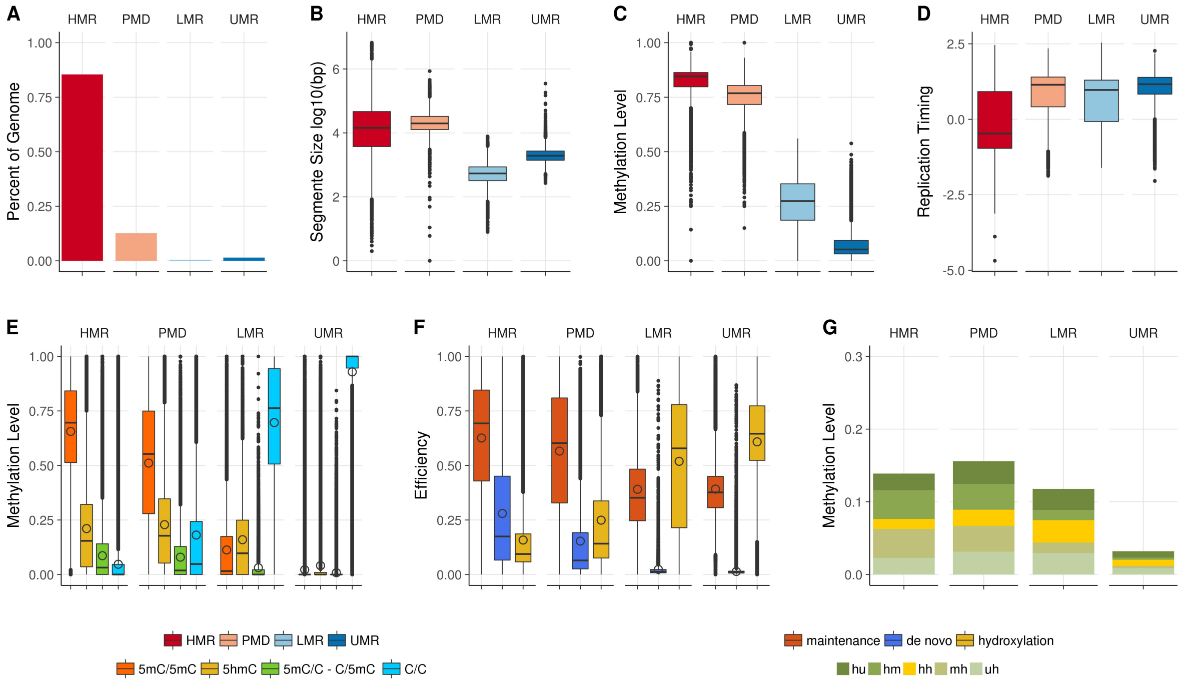


Fig. 5.8: Methylome Segmentation. **(A)** Outcome of the segmentation using MethylSeekR of mouse ES cells under Serum/LIF conditions. **(B)** Number of HMRs, PMDs, LMRs and UMRs after segmentation. Size distribution of the individual segments. **(C)** Methylation level of segments according to Ficz *et al.* 2013. **(D)** Replication timing based on the data from Hiratani *et al.* 2008. **(E)** Methylation distribution based on RRHPoxBS. **(F)** maintenance (red), *de novo* (blue) and hydroxylation (yellow) efficiency. **(G)** 5hmC distribution in HMRs, PMDs, LMRs and UMRs.

0.6%) and *de novo* activity at d0, whereas the activity of Tet enzymes is rather low (Figure 5.10B). At the same time, we observe high methylation levels at day0 (Figure 5.10A). Over time, we observe a strong decrease in *de novo* methylation together with a nearly stable maintenance and an increasing hydroxylation efficiency. This observation again indicates that the change in maintenance methylation appears to be an early event, while the reduction in *de novo* methylation is a gradual process. In terms of methylation, these changes in efficiency are accompanied by transient increase of 5hmC and hemimethylated CpGs and result in a hypomethylated phenotype at day6 (Figure 5.10A).

Cluster 2, containing 702901 CpGs, is mainly characterised by a high hydroxylation activity, which further increases over time (Figure 5.10B). Maintenance efficiency (about 0.5%) is considerably lower compared to cluster 1 and appears to slightly decrease during the transition to 2i. Additionally, we here observe a very low *de novo* activity of Dnmts, which decreases over time as well. The initial methylation level of cluster 2 is lower compared to cluster 1, but also displays a transient increase in hemimethylated CpGs and 5hmC and a general loss of methylation over time. Interestingly and despite the difference in the absolute hydroxylation efficiency in the two clusters, their demethylation rates appear to be equal (Figure 5.10A).

In addition to the shared temporal increase of 5hmC from d0 to d3, we observe com-

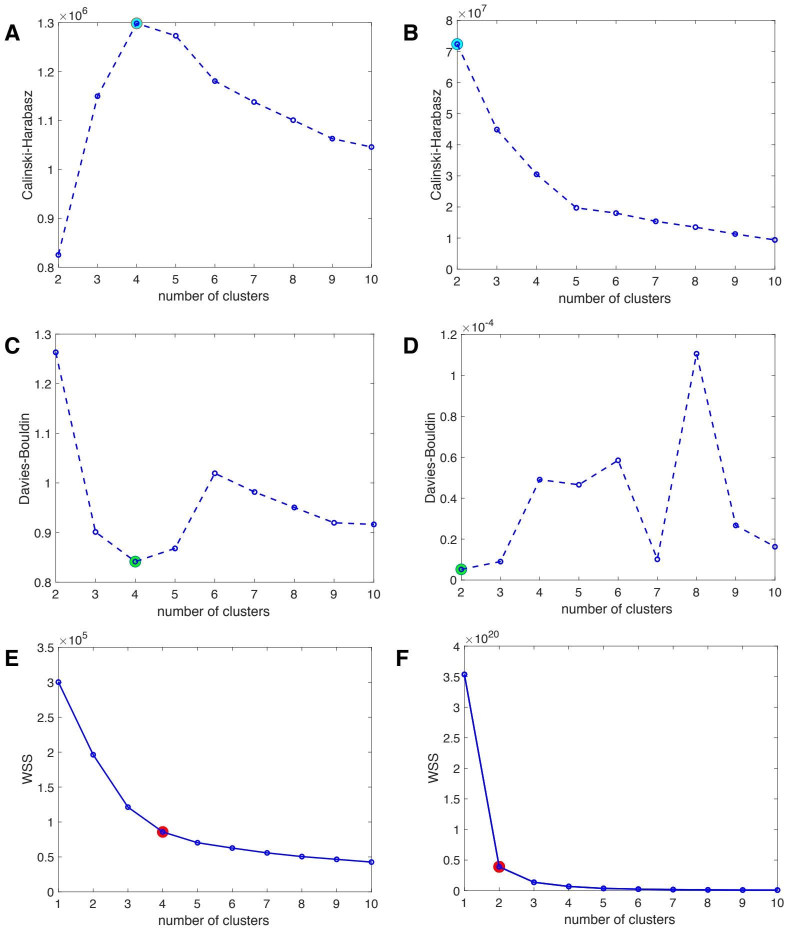


Fig. 5.9: Optimal number of clusters for k-means and k-error algorithms according to three clustering validity metrics. Calinski-Harabasz criterion for **(A)** k-means and **(B)** k-error . Davies-Bouldin criterion for **(C)** k-means and **(D)** k-error . Elbow method (WSS) for **(F)** k-means and **(E)** k-error .

parable average 5hmC levels in both clusters. In both clusters, 5hmC is symmetrically distributed between both DNA strands, meaning that the individual 5hmC states appear with the same frequency at Watson and Crick strands. Nevertheless, the distribution of 5hmC is distinct for each cluster. Whereas most CpGs in cluster1 exhibit a 5hmC/5mC or 5mC/5hmC state, the majority of 5hmC in cluster 2 is paired with unmethylated cytosine on the opposite strand (5hmC/C or C/5hmC).

LOLA enrichment analysis of both clusters revealed for cluster 2 an enrichment of TFBS, euchromatic histone modifications and CpG islands [59]. The list of TFBS includes typical stem cell markers such as Oct4, Nanog and Sox2, as well as transcription activators involved in stem cell self-renewal, e.g. Stat3, Stat5a/b and Taf3 [70, 71]. Yet, the strongest enrichment can be observed for Dpy30. Being a subunit of the Set1/MLL complex, Dpy30 is involved in the methylation of lysine 4 of histone 3, particularly trimethylation (H3K4me3) at bivalent enhancers [72]. Overall, these observations suggest a function for Tets in maintaining a stem cell phenotype and, furthermore, in predefining promoters of important developmental genes.

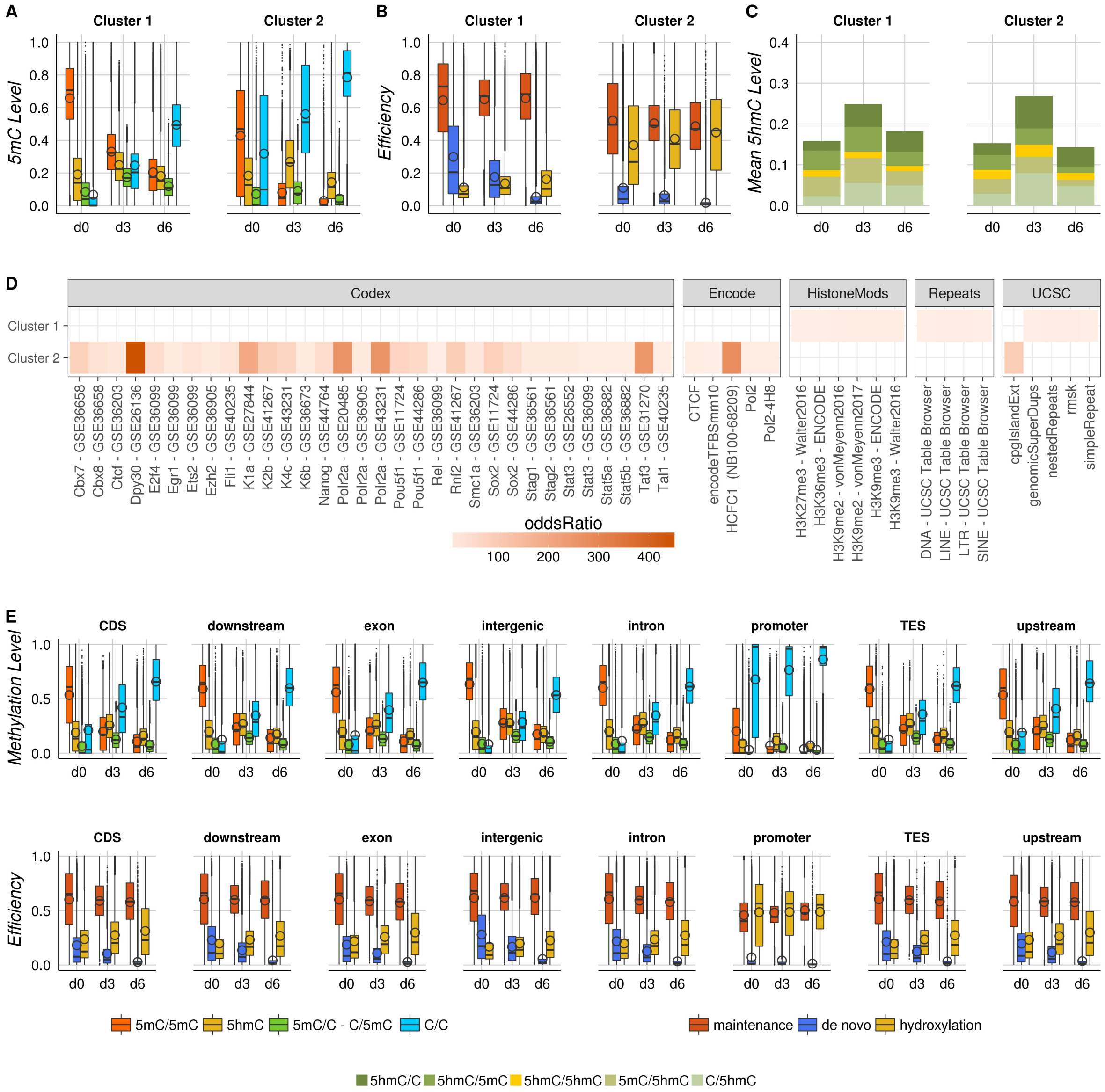


Fig. 5.10: Single CpG clustering premised on enzyme efficiency. **(A)** Methylation profile of identified efficiency clusters. **(B)** Efficiency profiles of identified clusters. **(C)** Mean 5hmC level and distribution. **(D)** LOLA enrichment analysis of clustered CpGs. **(E)** Methylation and efficiency profiles of annotated genomic features.

In Tet TKO cells both clusters show higher methylation levels, and retain a notably amount of 5mC even at d7 in 2i containing medium. This is particularly evident for cluster 2 (Figure 5.11A). Additionally, we observe a reduced *de novo* efficiency for d0 compared to WT cells, which however stays rather stable over time. Maintenance methylation efficiency of cluster 1 seems to be unchanged in Tet TKOs, which indicates that misregulation of maintenance in the absence of Tets remains restricted to un/low methylated CpGs.

Cluster 2, on the other hand, exhibits a notable increase in both maintenance and *de novo* methylation activity for all time points. Again, *de novo* methylation efficiency persists even for late time points.

Grouping the CpGs based on their genomic context reveals conserved methylation and efficiency patterns. In particular, maintenance methylation, which appears to be stable over time, exhibits the same behavior for almost all genetic features (Figure 5.10E). Small differences between the individual features can be observed in *de novo* and hydroxylation efficiency. However, tendencies and temporal profiles are shared between all attributes. The only exception are promoter regions. Here, we observe high hydroxylation efficiency, moderate maintenance and only marginal *de novo* methylation efficiencies. This observation is in agreement with the profile plots across genes, which unveiled similar dynamics at the TSS (Figure 5.2).

In TKO cells, *de novo* efficiency is more equally distributed between the distinct genomic features and CpGs in promoters exhibit a small increase in *de novo* methylation activity (Fig. 5.11D). In addition promoters display a clear increase in maintenance methylation. Again, this demonstrates a considerable misregulation of methylation efficiencies in the absence of Tets.

5.4 Discussion

In our study, we investigated how Dnmts and Tets contribute to a stable methylome constructed by alternating unmethylated and methylated domains and, furthermore, examine how changes in the enzyme activity shape new methylation patterns.

To address these questions, we developed RRHPoxBS, a method that comprises three features: (i) genome wide analysis of a subset of about 3 million CpGs with an adequate coverage, (ii) simultaneous analysis of 5mC and 5hmC, as well as (iii) the combined detection of both strands of one individual DNA molecule. This unique data set is best suited to investigate the relationship between Dnmts and Tets. In combination with our HMM analysis, we are able to calculate the detailed distribution of 5hmC states and the enzymatic activity for each individual CpG.

5.4.1 RRHPoxBS - A Robust Method for 5mC/5hmC Detection

The first genome wide hairpin approach developed by Zhao *et al.* in 2014 presents a powerful technique for the detection of double strand methylation information [44]. However, this method comes with high sequencing costs and demands large amounts of DNA. In case of RRHPoxBS we restrict our analysis to 3 million CpGs equally distributed across the genome, which reduces the sequencing costs and provides high coverage of informative CpGs. In addition, our pipeline only uses about one tenth of the DNA amounts formerly needed and can probably be scaled down further.

Our first observation is that the methylation levels of RRHPoxBS for Serum/LIF (60%) and 2i (20%) conditions are in line with previous described methylation levels from RRBS and WGBS [38, 60]. Furthermore, we observe a reduction of nonCpG methylation after

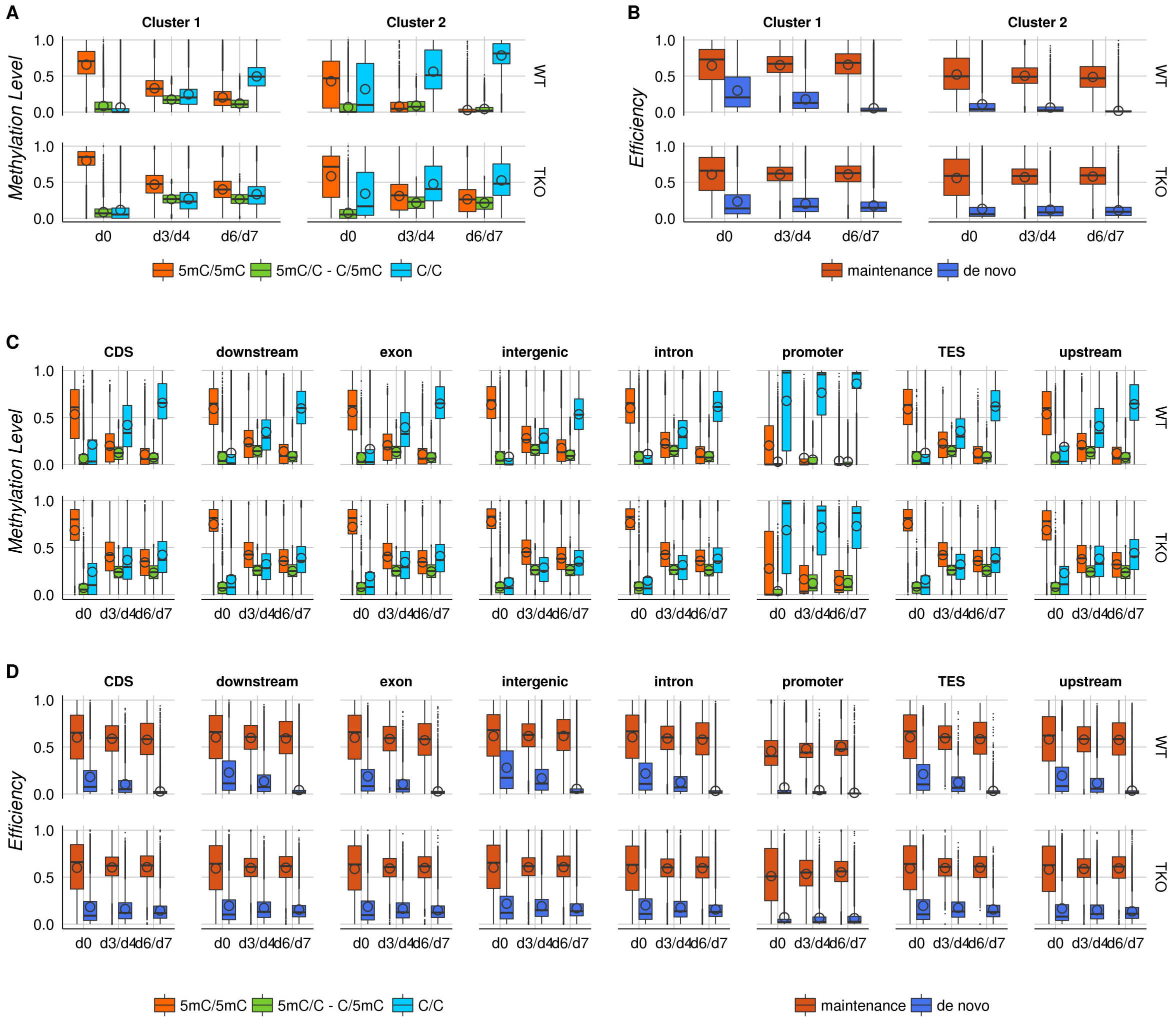


Fig. 5.11: Comparison of clustered CpGs in WT and TKO ES cells. **(A)** Methylation profile of clustered CpGs. **(B)** Efficiency profile of clustered CpGs. **C** Methylation profile of annotated genomic features. **(D)** Efficiency profile of annotated genmoic features.

the incubation in 2i, which is in agreement with the previous reported loss of Dnmt3a and 3b under naive conditions [38]. In addition, the readout of the used spike-in oligos shows a good conversion of C, 5fC in BS and C, 5hmC, 5fC in oxBS libraries, demonstrating that RRHPoxBS presents a reliable method for the detection of 5mC.

5.4.2 Asymmetric CpG Methylation - Intention or Accident

Potentially, hemimethylated CpGs can present a selective, strand specific epigenetic information. For example, the orientation of hemimethylated CpGs could mark the coding strand of RNA and enforce the transcription of either Watson or Crick strand. However, evaluation of the double strand information obtained from RRHPoxBS does not reveal any strand specific distribution of hemimethylated CpGs in relation to transcription (Supplement, Section 5.6.5, Figure S5.34). Instead, hemimethylation is equally distributed between both strands and follows the behaviour of symmetric CpG methylation, suggesting that hemimethylation is more likely the result of *de novo* methylation or active and passive demethylation events. In this context, we observe a temporary increase of hemimethylated CpGs during demethylation in 2i. Interestingly, we observe more hemimethylation in WT compared to Tet TKO cells, which indicates that Tet enzymes enhance the passive loss of 5mC. Indeed, our model predicts that 5hmC is very likely not recognised by Dnmt1 after replication (on average with a probability of 65%) and by that enhances passive demethylation.

5.4.3 Reduction of Maintenance Efficiency is an Early 2i Event

The combination of Dnmt and Tet enzyme activity defines the methylation status of each CpG. Thus, we calculated the enzyme efficiencies of Dnmts and Tets for individual CpGs. Our first observation is that our model in general agrees with previous findings from us and others, which on average suggest a reduced but stable maintenance activity ($\approx$0.6) in 2i, a continuous decline of *de novo*- and a slightly increasing hydroxylation efficiency [47, 60]. The reduction of maintenance efficiency was recently related to the reduction of H3K9me2 under 2i conditions [60]. Since we do not observe major changes in maintenance methylation over time, we reasoned that the reduction of H3K9me2 and, consequently, maintenance methylation efficiency, are early events and completed within the first 24h upon the transition to 2i medium. In contrast, *de novo* methylation activity progressively decreases, which fits to a gradual degradation and transcriptional halt of Dnmt3a and 3b [38].

5.4.4 Dnmts and Tet Act Opposed, but not Mutually Exclusive

The model clearly suggests that methylation and hydroxylation efficiencies are not exclusive for a given CpG, but show an antagonistic behaviour. The spatial cross-correlation verifies that a low methylation efficiency is usually accompanied by a high hydroxylation efficiency and *vice versa*, defining alternating domains of low and high methylation levels, respectively. In general, we observe high maintenance and *de novo* efficiency at the majority of the genome. The activity of Tet enzymes, on the other hand, is highest at UMRs and LMRs, i.e. promoters, TFBS (Sox2, Pou5f1) and particularly at the TSS. Very recent studies based on chromatin immunoprecipitation support our findings revealing that binding of Dnmt3s is higher at the gene body and HMRs, whereas Tet1 binding was predominantly observed across methylation valleys (LMRs and UMRs) [73, 74].

5.4.5 Local Control of Tets - Creation of Stable 5hmC

In general, we observe that 5hmC always represents a fraction of 5mC. However, in LMRs, which represent mostly enhancers [67], the level of 5hmC exceeds those of 5mC. Our findings are in accordance with previous observations, which link enhancer functions to the presence of 5hmC [75, 76, 77]. Our model reveals, that a specific combination of maintenance and hydroxylation efficiency is sufficient to maintain constantly a high amount of 5hmC at these locations. Overall, the distinct hydroxylation efficiencies observed at HMRs, PMDs, LMRs and UMRs suggest a tight regulation of Tet enzymes. In this context, several mechanisms are possible, histone modifications, which attract (H3K4me3) or repel Tets, the expression of Tet isoforms, but also post-translational modifications or the interaction with cofactors.

5.4.6 Active Tet Enzymes Promote ES cell Self-Renewal and Differentiation

Comparison of CpGs with high hydroxylation efficiency and ChIP profiles using LOLA, identifies two roles for active Tet enzymes in mouse ES cells (Figure 5.10 D). Firstly, CpGs with high Tet efficiency are located at TFBS, known to be involved in stem cell self-renewal, such as Oct4 (Pou5fI) and Sox2. Secondly, a strong overlap of CpGs with high hydroxylation efficiency and with binding sites of Dpy30 can be observed. Dpy30 is not involved in ES cell self-renewal but appears to be essential for differentiation of ES cells. As part of the SET1/MLL complex, Dpy30 is involved in the generation of H3K4me3 particularly at bivalent promoters.

These observations indicate, that active Tets might not be essential, but clearly contribute to both self renewal and the generation of bivalent promoters, probably by preventing the creation of DNA methylation. Previous studies already proposed a dual function of Tet enzymes in mouse ES cells. KD experiments show that the absence of Tet enzymes can lead to partial loss of the ES cell phenotype, while depletion of Tets from outlived KO ES cells prevent proper differentiation.

5.4.7 Tets - Guardians Against Methylation Spreading

In the absence of Tet we observe a clear misregulation in both, maintenance and *de novo* methylation efficiency. In particular, with the exception of day0, we see a strong increase of *de novo* activity for the entire genome and an increase of maintenance activity limited to regions exhibiting a high hydroxylation efficiency in WT ES cells. A misregulation of Dnmt1 is further supported by the spatial autocorrelation of maintenance efficiency in Tet TKO cells (Figure 5.3b and Section S5.6.4, Figure 5.23).

The almost stable estimated *de novo* efficiency under 2i conditions in Tet TKO is surprising, considering the downregulation of Dnmt3a/3b in WT ES cells. However, the apparent presence of Dnmt3a/3b under 2i condition in Tet TKO cells is strongly supported by the persistent nonCpG methylation in these cells. Moreover, we observe a strongly reduced demethylation rate in Tet TKO cells compared to WT ES cells, showing the importance of Tet enzymes in the demethylation kinetics.

Taken together, we hypothesise that Tet enzymes work against methylation in three ways. A high hydroxylation efficiency (i) guarantees an instant conversion of 5mC and acts against its establishment during a cell replication either via passive or active demethylation, (ii) inhibits the effectiveness of the maintenance machinery over regions that should remain unmethylated. At last, it seems that Tet enzymes (iii) ensure an efficient down-regulation of the *de novo* enzymes, which can not be observed in their absence.

5.5 Conclusion

We developed RRHPoxBS, a method which allows simultaneous detection of 5mC and 5hmC, as well as their strand specific distribution. In combination with an extended version of our hidden Markov model we present a robust and powerful method for the investigation of enzyme efficiencies across the genome. We find, that Dnmts and Tets act cooperatively on CpGs and generate distinct methylation domains with clear boundaries across the genome. In this context, Tet enzymes shield unmethylated CpGs from accidental methylation and, in addition, prevent the inheritance of ectopic 5mC by antagonising maintenance methylation. Furthermore, modulation of Tet activity leads to the generation of stable 5hmC rather than unmethylated CpGs. As future prospects, one could consider the analysis of single KO systems to decrypt distinct roles of Tet1 and Tet2 in these processes. Additionally, integration of further oxidised cytosine variants would allow a clear separation of passive and active demethylation [78, 79].

5.6 Supplementary Data

5.6.1 Reduced Representation Hairpin Oxidative Biuslfite Sequencing (RRHPoxBS)

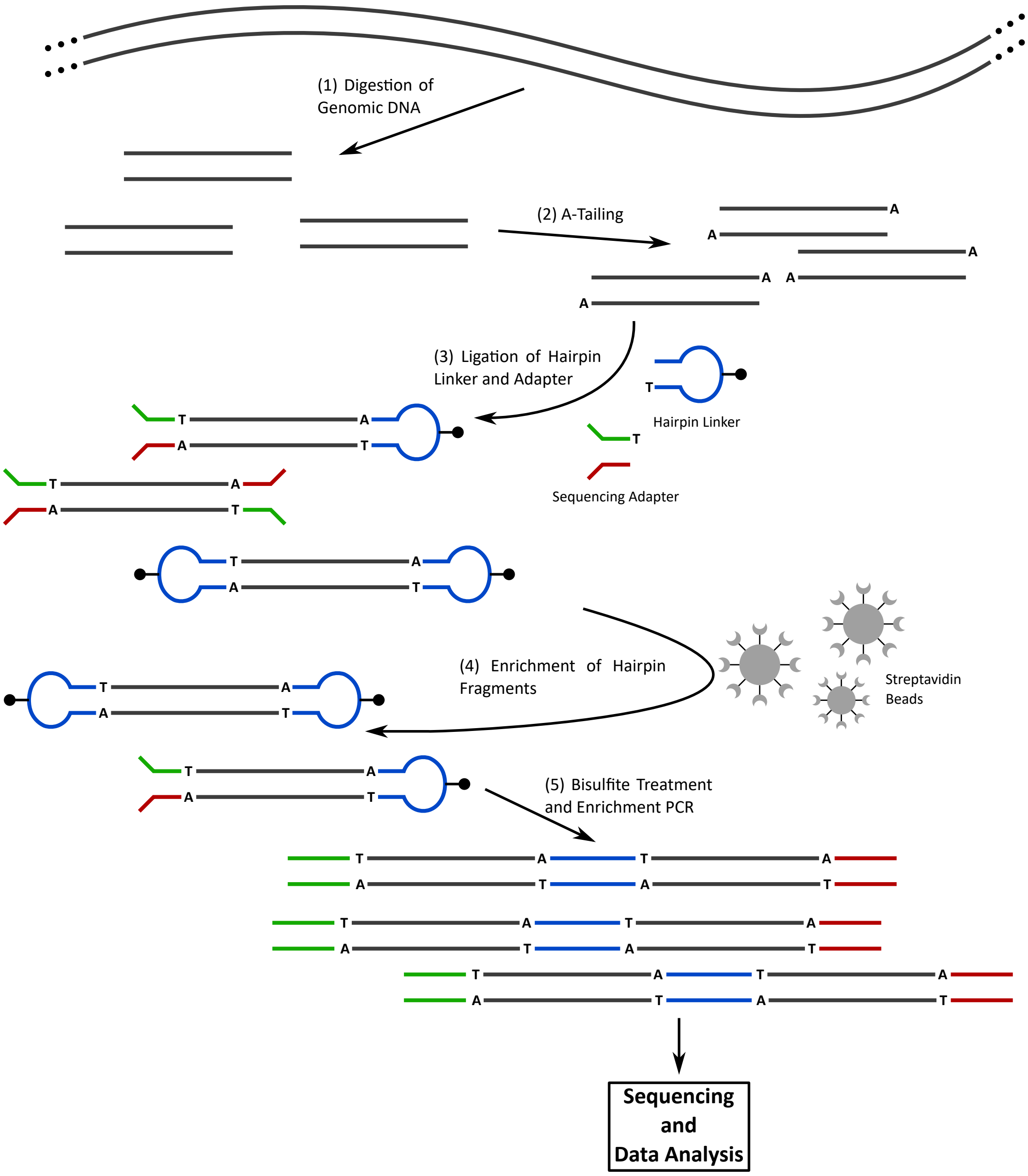

(1) Digestion of Genomic DNA
(2) A-Tailing
A
T
(3) Ligation of Hairpin Linker and Adapter
Hairpin Linker
Sequencing Adapter
(4) Enrichment of Hairpin Fragments
Streptavidin Beads
(5) Bisulfite Treatment and Enrichment PCR
Sequencing and Data Analysis

Fig. 5.12: Schematic representation of RRHPoxBS. (1) Digestion of genomic DNA using endonucleases generates blunt-end DNA fragments, (2) generation of single 3' Adenine overhangs(A-Tailing), (3) ligation of hairpin linker and Illumina® sequencing adapter, (4) enrichment of HP fragments by biotin-streptavidin-purification, (5) BS and oxBS treatment of HP library followed by PCR amplification, sequencing and data analysis.

First, DNA (1.2 µg) is split equally into three reaction tubes (400ng each). Each DNA sample is subjected to enzymatic digestion using one of the three restriction enzymes, HaeIII(NEB), AluI(NEB) or HpyCH4V (NEB). Each reaction is performed in 20 µl with 20U restriction enzyme and 2 µl buffer CutSmart®(NEB) and incubated at 37°C over night. Restriction enzymes are inactivated by subsequent incubation at 80°C for 30min. Following inactivation, the reactions are pooled and subjected to ligation step. 200mM biotin labelled hairpin linkers (Biomers), 100mM sequencing adapters (Biomers), 1mM ATP (NEB) and 4000U T4 DNA ligase (NEB) are added to the pooled sample. Ligation is then performed overnight at 16°C. The ligation of hairpin linkers and sequencing adapters are undirected processes and will generate three distinct types of fragments: (i) non-hairpin fragments, with sequencing adapter on both ends, (ii) hairpin fragments, with hairpin linker on both ends, as well as (iii) hairpin fragments with hairpin linker on one side and sequencing adapter on the other side. In order to deplete unwanted non-hairpin fragments from the library, the ligation is first subjected to a purification using AmpureXP® beads with a ratio of 1:2 (library:beads), followed by a purification using streptavidin coated magnetic beads (Dynabeads™ M-280 Streptavidin; ThermoFisher Scientific). Only hairpin fragments carrying the biotinylated hairpin linker will bind to the M280 beads, while non-hairpin fragments remain in the supernatant. After removal of the supernatant and three subsequent wash steps (1xBW buffer according to manufacturer's specifications), the beads are incubated in 20µl 1xTE buffer containing 1% SDS and incubated at 100Â°°C for 15min, to release hairpin fragments from M280 beads. Next, the hairpin library is subjected to BS and oxBS treatment using the TruMethyl Kit from CEGX accroding to manufacturer's instructions, followed by PCR amplification (Table 5.1) and paired end sequencing on an Illumina HiSeq2500 system.

Tab. 5.1: Typical PCR protocol for RRHPoxBS using HOTStarTaq® from QIAGEN

PCR Protocol	**PCR Conditions**	
10.0μl RRHPoxBS sample		
5.0μl 10x Buffer HotSarTaq®	95°C - 15min	
2.0μl 25mM $MgCl_2$	95°C - 1min	12-15x
4μl 10mM dNTPs	60°C - 1min	
0.8μl Forward Primer	72°C - 1min	
0.8μl Reverse Primer	72°C - 7min	
0.8μl HOT FIREPol®		
26.6μl ddH_2O		

Tab. 5.2: HP-Linker, Adapter and Enrichment-Primer Sequence. All oligonucleotides were purchased from Biomers.

Primer	Sequnce
HP-Linker	phoGGGCCTADDDBDDDTAGGCCCT B = biotinylated Thymine
Upper-Adpter	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Lower-Adpter	GTTCGTCTTCTGCCGTATGCTCTAGCACTACACTGACCTCAAGTCTGCACACGAGAAGGCTAG
Forward-Primer	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Reverse-Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

5.6.2 Hidden Markov Model (HMM)

We use the hidden Markov model (HMM) presented in [47] to describe the temporal evolution of a single CpG dyad over time for each experiment (bisulfite or ox. bisulfite). The hidden states of the model correspond to the different modifications, e.g. the cytosines (C) on both strands are unmethylated or the C on the upper strand is methylated while the C on the lower strand is unmethylated, etc. The observable states are those that we measure after bisulfite (BS) or oxidative bisulfite (oxBS) hairpin sequencing. Hence, we include the conversion errors of the measurement process and we link two HMMs that describe oxidative and non-oxidative hairpin bisulfite sequencing to accurately determine hydroxymethylation levels and the efficiencies of the involved enzymes over time. Formally, we define the sets of hidden states $\mathcal{S} = \{u, m, h\}^2$ and the set of observable states $\mathcal{S}_{obs} = \{\mathrm{T}, \mathrm{C}\}^2$. A state $s \in \mathcal{S}$ describes whether the upper and the lower strand of the site is *unmethylated* (u), *methylated* (m) or *hydroxylated* (h). E.g. in state (h, u) the upper strand is hydroxylated and the lower strand is unmethylated. Similarly, a state $j \in \mathcal{S}_{obs}$ encodes whether the upper strand and the lower strand of the site has been transformed after the BS or the oxBS treatment to a thymine (T) or a cytosine (C). We use the abbreviation hu for state (h, u) and similar for all other states.

Distribution of Hidden and Observable States

Let the vector $\pi(t)$ be the hidden states distribution at time t and let $\pi(i, t) = P(\mathcal{X}(t) = i)$ represent the entry of $\pi(t)$ that corresponds to state $i \in S$.

The transition matrix of the hidden states is defined as $\mathbf{P}(t) = \mathbf{D}(t) \cdot \mathbf{M}(t) \cdot \mathbf{H}(t)$, where $\mathbf{D}(t)$ describes the modifications due to cell division, $\mathbf{M}(t)$ the modifications due to methylation, and $\mathbf{H}(t)$ the modifications due to hydroxymethylation. As in [47], we define

$$\mathbf{D}(t) = \begin{array}{c|ccccccccc|} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ \hline uu & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ um & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ mu & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ uh & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ hu & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ hm & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ mh & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ mm & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ hh & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \end{array},$$

$$\mathbf{M}(t) = \begin{array}{c|ccccccccc|} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ \hline uu & \bar{\mu}_d^2 & \mu_d \cdot \bar{\mu}_d & \mu_d \cdot \bar{\mu}_d & 0 & 0 & 0 & 0 & \mu_d^2 & 0 \\ um & 0 & \bar{\lambda} & 0 & 0 & 0 & 0 & 0 & \lambda & 0 \\ mu & 0 & 0 & \bar{\lambda} & 0 & 0 & 0 & 0 & \lambda & 0 \\ uh & 0 & 0 & 0 & p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda} & 0 & 0 & p \cdot \mu_d + \bar{p} \cdot \lambda & 0 & 0 \\ hu & 0 & 0 & 0 & 0 & p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda} & p \cdot \mu_d + \bar{p} \cdot \lambda & 0 & 0 & 0 \\ hm & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ mh & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ mm & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ hh & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array},$$

and

$$\mathbf{H}(t) = \begin{array}{c|ccccccccc|} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ \hline uu & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ um & 0 & \bar{\eta} & 0 & \eta & 0 & 0 & 0 & 0 & 0 \\ mu & 0 & 0 & \bar{\eta} & 0 & \eta & 0 & 0 & 0 & 0 \\ uh & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ hu & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ hm & 0 & 0 & 0 & 0 & 0 & \bar{\eta} & 0 & 0 & \eta \\ mh & 0 & 0 & 0 & 0 & 0 & 0 & \bar{\eta} & 0 & \eta \\ mm & 0 & 0 & 0 & 0 & 0 & \eta \cdot \bar{\eta} & \eta \cdot \bar{\eta} & \bar{\eta}^2 & \eta^2 \\ hh & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}.$$

Here, μ_m stands for the maintenance efficiency, μ_d for *de novo* and η for the hydroxylation efficiency, while p is the probability that 5hmC is not considered during maintenance (see [47] for details).

Note, that for $\mathbf{D}(t)$ we can omit the time parameter t since it is time-independent, while the other two matrices depend on t as explained later. Note also, that the HMMs of BS and oxBS experiments have both the same distribution $\pi(t)$ for the hidden states (as for both experiments the same cell population is used) but different emission probabilities and that $\pi(t)$ is given by

$$\pi(t) = \pi(0) \cdot \prod_{k=1}^{t} \mathbf{P}(k).$$

	BS				oxBS			
	TT	TC	CT	CC	TT	TC	CT	CC
uu	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$
um	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot d$	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$
mu	$c \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$	$c \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$
uh	$c \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot \bar{e}$	$\bar{c} \cdot e$	$c \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot f$	$\bar{c} \cdot \bar{f}$
hu	$c \cdot \bar{e}$	$\bar{c} \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot e$	$c \cdot f$	$\bar{c} \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot \bar{f}$
hm	$\bar{d} \cdot \bar{e}$	$d \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot e$	$\bar{d} \cdot f$	$d \cdot f$	$\bar{d} \cdot \bar{f}$	$\bar{d} \cdot f$
mh	$\bar{d} \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot \bar{e}$	$d \cdot e$	$\bar{d} \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot f$	$\bar{d} \cdot f$
mm	$\bar{d}^2$	$\bar{d} \cdot d$	$d \cdot \bar{d}$	d^2	$\bar{d}^2$	$\bar{d} \cdot d$	$d \cdot \bar{d}$	d^2
hh	$\bar{e}^2$	$\bar{e} \cdot e$	$e \cdot \bar{e}$	e^2	f^2	$f \cdot \bar{f}$	$f \cdot \bar{f}$	$\bar{f}^2$

Tab. 5.3: Transition probabilities from hidden to the observable states in bisulfite sequencing (BS) and in ox. bisulfite sequencing (oxBS).

Let the vectors $\pi_{bs}(t), \pi_{ox}(t)$ be the observable states distribution at time t, with entries $\pi_{bs}(j,t)$ and $\pi_{ox}(j,t)$, $j \in \mathcal{S}_{obs}$, for the BS and oxBS experiments, respectively. We then get:

$$\pi_{bs}(t) = \pi(t) \cdot \mathbf{E}_{bs}(t) \quad \text{and} \quad \pi_{ox}(t) = \pi(t) \cdot \mathbf{E}_{ox}(t),$$

where the entries of the emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ are given in Table 5.3.

5.6.3 Maximum Likelihood Estimation (MLE)

Initial Distribution of the Hidden States

Let $n_{bs}(j,t)$ and $n_{ox}(j,t)$ be the number of times that state $j \in \mathcal{S}_{obs}$ has been observed during independent hairpin bisulfite (BS) and oxidative hairpin bisulfite (oxBS) measurements out of a certain number of reads (mean coverage of all samples $\approx$ 20x) at time t.

Since we assume that $t = 0$ is the time of the first measurement, we have observations at $t = 0$ and can estimate the unknown initial distribution over the hidden states using maximum likelihood estimation (MLE). For this, we have to solve the optimization problem: $\pi(0)^* = \arg\max_{\pi(0)} \mathcal{L}_1(\pi(0))$, subject to the constraint $\sum_{i \in \mathcal{S}} \pi(i,0) = 1$, where

$$\mathcal{L}_1(\pi(0)) = \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,0)^{n_{bs}(j,0)} \cdot \pi_{ox}(j,0)^{n_{ox}(j,0)}.$$

During the optimization procedure, we use the log-likelihood

$$\ln \mathcal{L}_1(\pi(0)) = \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,0) \cdot \ln \pi_{bs}(j,0) + n_{ox}(j,0) \cdot \ln \pi_{ox}(j,0)).$$

Moreover, to allow gradient descent optimization we also compute the derivative w.r.t.

$\pi(0)$ given by

$$\frac{d}{d\pi(0)} \ln \mathcal{L}_1(\pi(0)) = \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,0) \cdot \frac{\frac{d}{d\pi(0)} \pi_{bs}(j,0)}{\pi_{bs}(j,0)} + n_{ox}(j,0) \cdot \frac{\frac{d}{d\pi(0)} \pi_{ox}(j,0)}{\pi_{ox}(j,0)}. \qquad (5.1)$$

Writing the vectors of partial derivatives $\frac{d}{d\pi(0)}\pi_{bs}(j,0)$ and $\frac{d}{d\pi(0)}\pi_{ox}(j,0)$ in a vector-matrix notation including all $j \in \mathcal{S}_{obs}$ we get

$$\frac{d}{d\pi(0)}\pi_{bs}(0) = \frac{d}{d\pi(0)}\pi(0) \cdot \mathbf{E}_{bs}(0) = \mathbf{E}_{bs}(0), \quad \frac{d}{d\pi(0)}\pi_{ox}(0) = \frac{d}{d\pi(0)}\pi(0) \cdot \mathbf{E}_{ox}(0) = \mathbf{E}_{ox}(0),$$

which after insertion into Eq. 5.1 gives us the gradient of the log-likelihood function w.r.t. the initial distribution of the hidden states.

Estimation of the Efficiencies

Let $\mathbf{v} = (\beta_0^{\mu_m}, \beta_1^{\mu_m}, \beta_0^{\mu_d}, \beta_1^{\mu_d}, \beta_0^{\eta}, \beta_1^{\eta}, p) \in \mathbb{R}^v$, be the vector of seven, i.e., $v = 7$, unknown parameters. We assume here that the efficiencies are linear functions of time (except for p) and so $\mathbf{v}$ contains the coefficients of these functions, e.g., $\mu_m(t) = \beta_0^{\mu_m} + t \cdot \beta_1^{\mu_m}$.

Now, after determining $\pi(0)$, (see section 5.6.3) we want to compute the MLE $\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \log\mathcal{L}_2(\mathbf{v})$, where

$$\mathcal{L}_2(\mathbf{v}) = \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,t)^{n_{bs}(j,t)} \cdot \pi_{ox}(j,t)^{n_{ox}(j,t)}. \qquad (5.2)$$

Note here we assume that the cells divide every 24 hours, hence t ranges over all days at which measurements were made after day0. In addition to derive the likelihood of Eq. 5.2 we assume that all observations made at time points $t \in T_{obs} \setminus \{0\}$ are independent. The independence assumption is well justified since during the measurement only a very small fraction of cells is taken out of a large pool and hence it is unlikely that we pick two cells with a common descendant.

Since the efficiencies are probabilities we have the constraint that for all time points in T_{obs} and all efficiencies we have $0 \leq \beta_0 + \beta_1 \cdot t \leq 1$. In addition, $0 \leq p \leq 1$.

It holds

$$\ln \mathcal{L}_2(\mathbf{v}) = \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,t) \cdot \ln \pi_{bs}(j,t) + n_{ox}(j,t) \cdot \ln \pi_{ox}(j,t)$$

and we get the score vector of the log-likelihood function as

$$\frac{d}{d\mathbf{v}} \ln \mathcal{L}_2(\mathbf{v}) = \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,t) \cdot \frac{\frac{d}{d\mathbf{v}}\pi_{bs}(j,t)}{\pi_{bs}(j,t)} + n_{ox}(j,t) \cdot \frac{\frac{d}{d\mathbf{v}}\pi_{ox}(j,t)}{\pi_{ox}(j,t)}.$$

Then the matrix-vector form of the derivatives $\frac{d}{d\mathbf{v}}\pi_{bs}(j,t)$ and $\frac{d}{d\mathbf{v}}\pi_{ox}(j,t)$ can be written as

$$\frac{d}{d\mathbf{v}}\pi_{bs}(t) = \frac{d}{d\mathbf{v}}\pi(t)\cdot\mathbf{E}_{bs}(t) \text{ and } \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}}\pi_{\mathrm{ox}}(\mathrm{t}) = \frac{\mathrm{d}}{\mathrm{d}\mathbf{v}}\pi(\mathrm{t})\cdot\mathbf{E}_{\mathrm{ox}}(\mathrm{t}),\ \ \forall\mathrm{t}\in\mathrm{T}_{\mathrm{obs}}.$$

Considering now, the forward Kolmogorov equation for the HMM and its derivative w.r.t. the parameters it suffices to simultaneously solve the following two equation systems.

$$\begin{aligned}\pi(t) &= \pi(t-1)\cdot\mathbf{P}(t)\\ \frac{d}{d\mathbf{v}}\pi(t) &= \frac{d}{d\mathbf{v}}\pi(t-1)\cdot\mathbf{P}(t)+\pi(t-1)\frac{d}{d\mathbf{v}}\mathbf{P}(t),\ \ \forall t\geq 1\end{aligned}\tag{5.3}$$

with $\frac{d}{d\mathbf{v}}\pi(0)=0$ and $\pi(0)=\pi(0)^*$. The derivative of the transition matrix is

$$\frac{d}{d\mathbf{v}}\mathbf{P}(t) = \frac{d}{d\mathbf{v}}(\mathbf{D}\cdot\mathbf{M}(t)\cdot\mathbf{H}(t)) = \mathbf{D}\cdot\Big(\frac{d}{d\mathbf{v}}\mathbf{M}(t)\cdot\mathbf{H}(t)+\mathbf{M}(t)\cdot\frac{d}{d\mathbf{v}}\mathbf{H}(t)\Big).$$

Now, applying the chain rule and taking into account that $\mu_m = \beta_0^{\mu_m}+\beta_1^{\mu_m}t$ we get for the entry that corresponds to $\beta_0^{\mu_m}$

$$\frac{d}{d\beta_0^{\mu_m}}\mathbf{M}(\mu_m) = \frac{d}{d\mu_m}\mathbf{M}(\mu_m)\cdot\frac{d}{d\beta_0^{\mu_m}}\mu_m = \frac{d}{d\mu_m}\mathbf{M}(\mu_m)$$

and

$$\frac{d}{d\beta_1^{\mu_m}}\mathbf{M}(\mu_m) = \frac{d}{d\mu_m}\mathbf{M}(\mu_m)\cdot\frac{d}{d\beta_1^{\mu_m}}\mu_m = \frac{d}{d\mu_m}\mathbf{M}(\mu_m)\cdot t.$$

In a similar fashion we get the first derivatives w.r.t. all the other components of parameter vector $\mathbf{v}$. Applying once more the product rule in Eq. (5.3), and using similar arguments as above we can additionally compute the second partial derivatives $\frac{d}{d\mathbf{v}_i d\mathbf{v}_j}\ln\mathcal{L}_2(\mathbf{v})$, which will give us the (i,j)-th entry of the Hessian matrix $\mathcal{H}=\nabla\nabla^{\mathrm{T}}\ln\mathcal{L}_2(\mathbf{v})$.

Standard Deviations and Confidence Intervals

The observed Fisher information is defined as $\mathcal{J}(\mathbf{v}^*) = -\mathcal{H}(\mathbf{v}^*)$, where $\mathbf{v}^*$ is the maximum likelihood estimator. We use the inverse of the expected Fisher information $\mathcal{I}(\mathbf{v}) = \mathbb{E}[\mathcal{J}(\mathbf{v})]$ to estimate the covariance matrix of the MLE. Hence, we approximate the covariance of our ML estimator as $\Sigma_{\mathbf{v}^*} = -\mathcal{H}^{-1}(\mathbf{v}^*))$. Then, in order to approximate the standard deviations of the efficiencies' functions over time, i.e., $\sigma(\mu_m(t)),\sigma(\mu_d(t))$ and $\sigma(\eta(t))$, we exploit the identity that if $f(t)=\beta_0+\beta_1\cdot t$ then

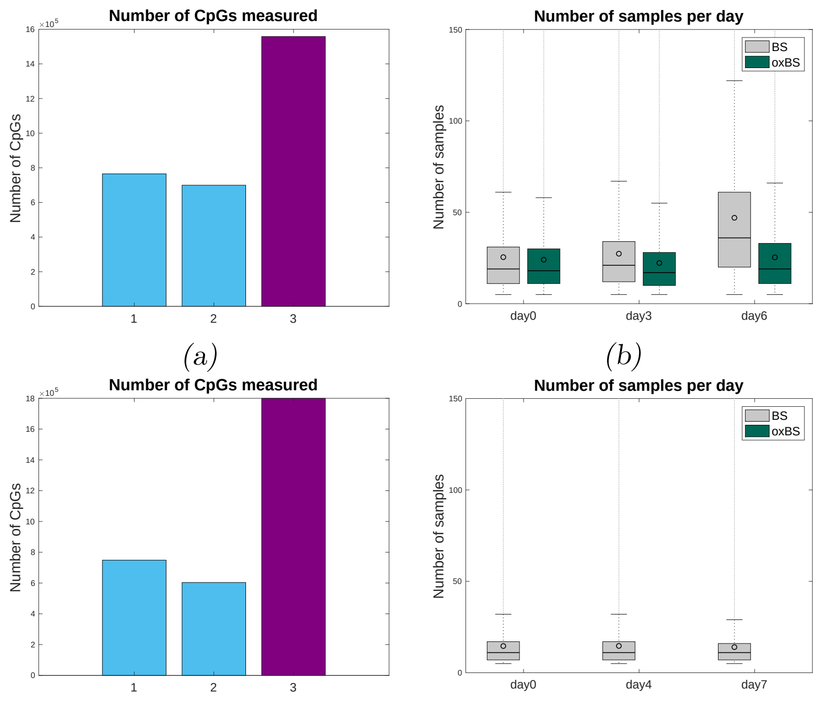
$$\sigma(f(t)) = \sqrt{\mathrm{Var}(\beta_0+\beta_1\cdot\mathrm{t})} = \sqrt{\mathrm{Var}(\beta_0)+\mathrm{t}^2\mathrm{Var}(\beta_1)+2\mathrm{t}\mathrm{Cov}(\beta_0,\beta_1)}.$$

We then determine the confidence intervals for a fixed confidence level $\beta = 95\%$. For instance the confidence interval for the maintenance methylation function will be

$$\mu_m(t) \pm z \cdot \sigma(\mu_m(t))$$

where $z = F^{-1}\left(\frac{\beta+1}{2}\right)$ and F is the cumulative distribution function (cdf) of the standard normal distribution. Similarly, we get the confidence intervals for all remaining parameters.

5.6.4 Running the Whole Genome Using Bayesian Inference



(a) (b)

(c) (d)

Fig. 5.13: Number of CpGs with observations at one, two, or three days in WT (a) and Tet TKO (c). Average number of independent single CpG samples (sequencing depth) per day for BS and oxBS of WT (b) and for BS of Tet TKO (d) data.

We have double stranded single base pair resolution data from bisulfite (BS) and oxidative bisulfite sequencing (oxBS) for 3,022,903 CpGs in wild type (WT) cells and for 3,151,985 from BS data in Tet triple TKO (Tet TKO) cells. In case of each of 1,464,801 CpGs in WT and of 1,352,297 in Tet TKO with only one or two observation time points available we predict for every measurement time point only the levels of the hidden states by performing a MLE for the (hydroxy-)methylation levels as described in Section 5.6.3 for estimating the initial distribution. In case of a CpG with three observation time points (1,558,102 in WT and 1,799,688 in Tet TKO, see purple column in Figure 5.13a, 5.13c) we assume a linear behavior of the efficiencies over time and we analyse the HMM as described in Section 5.6.2 for estimating both the values of (hydroxy-)methylation efficiencies and levels over time. Using a computer cluster consisting of 32 machines with 16 physical kernels each, we are able to efficiently parallelize the computations for large bunches of all available CpGs.

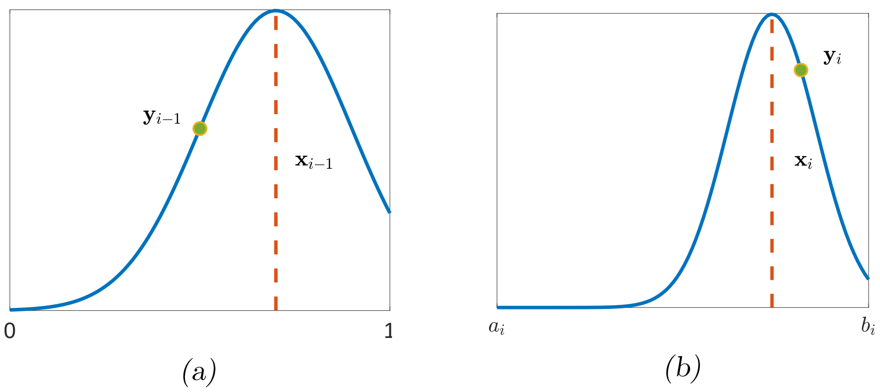
Due to the low depth sequencing per time point and experiment (40 for BS, 29 for oxBS in WT, and 14 for BS in Tet TKO on average, see Figure 5.13b) we assume that the asymptotic properties of the MLE around the true parameter value do not hold [54, 55], especially in cases where the true parameter values are close to boundary constraints [56].

For that reason, we additionally use a Bayesian Inference (BI) approach to get the posterior distribution of the model parameters, i.e, the efficiencies over time. For all CpGs we choose as prior distribution the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where the mean $\boldsymbol{\mu}$ is the average of the estimated efficiencies in [47]. Similarly, Σ is the average of the corresponding covariance matrices. Note that in [47] MLE was sufficient due to the better coverage. Finally, we make a comparison between the MLE and the BI methods and we confirm that a BI method that incorporates an informative prior distribution should be preferable for epigenome-wide analysis especially for the regions where the coverage is low [80, 81].

Metropolis-Hastings

We apply BI by sampling from the multi-dimensional posterior $P(\mathbf{v}|\text{data}) = \frac{\mathcal{L}_2(\text{data}|\mathbf{v})P(\mathbf{v})}{\int_{\mathbf{v}} P(\text{data},\mathbf{v})}$ and avoid to approximate the normalizing factor $\int_{\mathbf{v}} P(\text{data}, \mathbf{v})$. Hence, we apply a Metropolis-Hastings MCMC approach using an asymmetric and truncated proposal distribution. The bounds of the truncation are determined s.t. the constraints for the efficiencies constantly hold for the time span of the observations, i.e., efficiencies are in $[0, 1]$ for all $t \in [0, t_{\max}]$. Hence, in every state $\mathbf{x} \in \mathbb{R}^v$ (with $v = 7$) of the MCMC we generate the next sample from a product of truncated univariate normals $\mathcal{N}(\mathbf{y}) = \prod_i f(\mathbf{y}_i|\mathbf{x}_i, \sigma_i^2/c, a_i, b_i)$, around the current MCMC point $\mathbf{x}$, where $\mathbf{x}_i$ refers to the $i-$th entry of the parameter vector for $i = 1, \dots, 7$, σ_i^2/c is the univariate normal variance and a_i, b_i are the truncation bounds for parameter $\mathbf{x}_i$. Consider position i where $\mathbf{y}_i$ refers to the gradient of an efficiency and $\mathbf{y}_{i-1}$ to the corresponding intercept. We sample the next value for each efficiency by sampling first the intercept $\mathbf{y}_{i-1}$ value from the truncated normal distribution within the interval $[a_{i-1}, b_{i-1}] = [0, 1]$ and based on this realization we sample the gradient $\mathbf{y}_i$ value from the truncated normal in $[a_i, b_i]$, where $a_i = -\mathbf{y}_{i-1}/t_{\max}, b_i = (1 - \mathbf{y}_{i-1})/t_{\max}$ as it is being illustrated in Figure 5.14. The bounds of probability p are set as those of an intercept, i.e., $[a_i, b_i] = [0, 1]$.

Note that the variance of parameter $\mathbf{x}_i$ we used for the proposal distribution is the same as the variance of the prior distribution $\sigma_i^2 = \Sigma_{i,i}$ normalized by a scale factor c. Since it is well known that the efficiency of Metropolis-Hastings algorithm crucially depends on the scaling of the proposal density, we empirically choose a $c = 50$ to normalize the standard



(a) (b)

Fig. 5.14: Metropolis Hasting's update step: We sample a new efficiency vector using two truncated normal distributions in two steps: (a) Step 1: We sample the intercept $\mathbf{y}_{i-1}$ from the truncated normal with mean $\mathbf{x}_{i-1}$ and bounds [0, 1]. (b) Step 2: We sample the gradient $\mathbf{y}_i$ from the truncated normal distribution with mean $\mathbf{x}_i$ and bounds $[a_i, b_i]$, which depend on the sampled intercept $\mathbf{y}_{i-1}$ of Step 1.

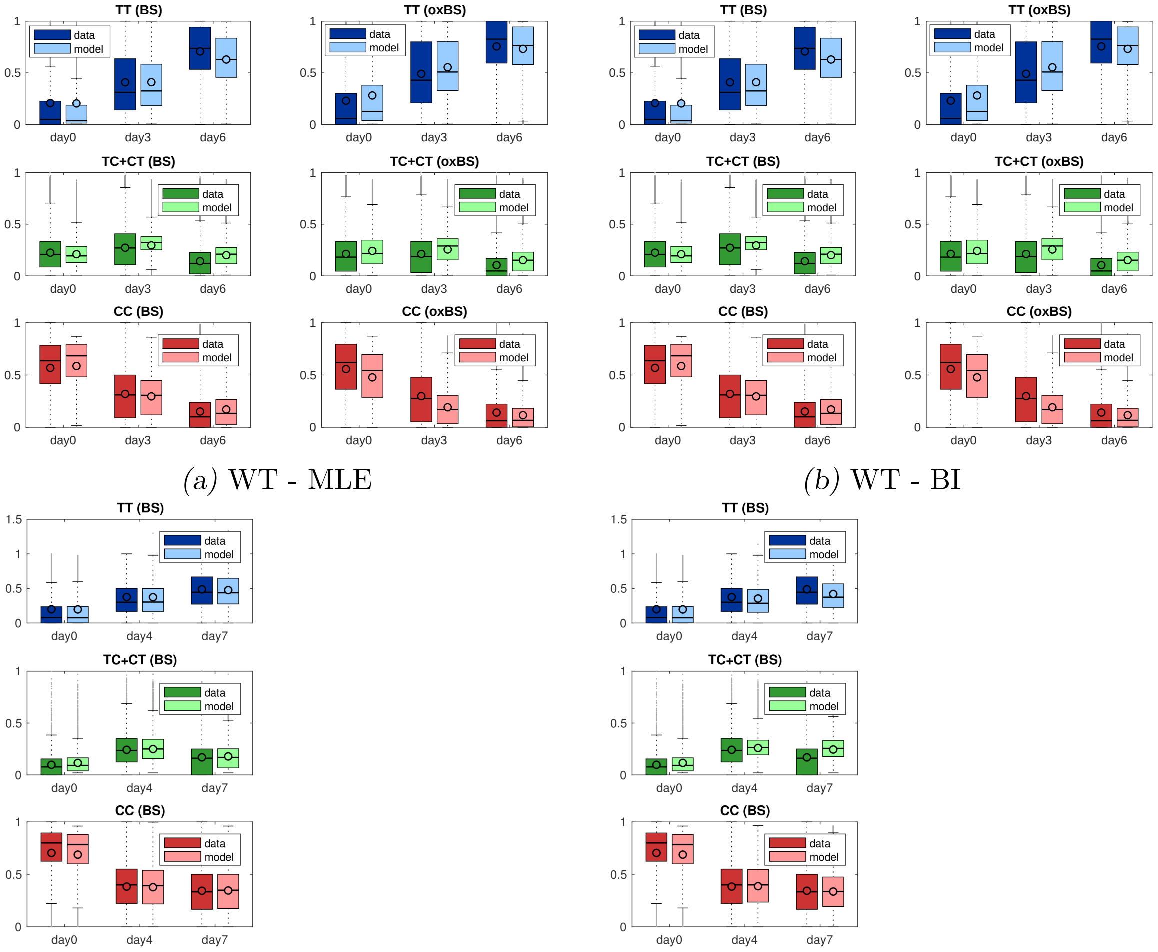
deviation of the proposal distribution[†] s.t. the average MCMC acceptance ratio is around 25% of the total number of generated samples [82]. As final estimators of the BI method we get the sample mean of the posterior distribution and we build credible intervals using the corresponding sample covariance.

Fit of Whole Genome Data vs the Model

Using box plots, we compare the levels of CC, TT and CT-TC CpG dyads for the whole genome present in the data of BS and oxBS in WT (Figure 5.15a, 5.15b) and of BS in Tet TKO (Figure 5.15c, 5.15d) and the probabilities of the observable states predicted by the two HMMs using MLE or BI for estimating the model's parameters. The circles inside the plots correspond the the mean value of each box plot and the horizontal lines to the medians. The bottom and the top of the boxes are the first and the third quartiles. The values for the whiskers correspond to the $\pm 2.7 \cdot s_{\text{data}}$ interval from the sample mean, where s_{data} is the sample standard deviation of the data. To quantify the goodness of the fit for each estimation method we report in Table 5.4 the average Kullback-Leibler divergence $D_{KL}(P||Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$ between the data distribution P and the distribution Q predicted by the model. Note that the model fit to the data reported by the average Kullback-Leibler divergence metric is better for the MLE than for BI for both WT and Tet TKO data. This is to be expected since MLE always tries to maximize the likelihood of the data no matter how well the data samples represent the true underlying distribution.

In Figure 5.16 we plot the average efficiencies computed by the two estimation methods (MLE vs BI) at days 0,3,6 for WT and days 0,4,7 for Tet TKO. We average over all CpGs along the DNA for which we sampled at all three measurement time points. We observe

[†] A low acceptance ratio indicates a wide proposal, while a high acceptance ratio indicates a narrow proposal and in both extreme cases the convergence is slow.



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

Fig. 5.15: Comparison between data and prediction of observable states after fitting the HMMs based on MLE (a), (c) and BI (b), (d). Dark box plots show the experimentally measured frequencies states and light box plots correspond to the values predicted by the two HMMs.

that there are some major differences between the MLE and the BI estimates. First in WT the ML estimates show an evident decrease of maintenance over time while BI estimates show maintenance to be almost constant. In addition, the hydroxylation activity seems to slightly drop using MLE while BI estimates that it increases. In the Tet TKO experiment, the ML estimates give a completely unexpected increase of maintenance activity, while *de novo* seems to be not affected compared with its WT behavior. On the contrary, BI estimators for maintenance in Tet TKO remain almost unchanged comparing with their WT - BI behavior, while interestingly *de novo* seems to drop in a much slighter rate in the absence of Tet enzymes. Looking carefully at the prediction of the enzymatic activity we have several reasons to trust more the results of the BI method than those of MLE. In the WT data we observe that the BI estimates are in line with the genome wide behavior being described in the literature for the vast majority of the examined regions [60, 47]. Furthermore, the prediction of the remaining *de novo* activity being present mainly in the BI and not in the ML estimates for the Tet TKO data is in line with the detection of remaining nonCpG methylation in our RRHPoxBS data set which is not part of the model and therefore presents an independent readout of Dnmt3a and 3b activity. In addition, looking at the box plots we note that the dispersion of the efficiencies values is evidently smaller for the BI estimates. This shows the higher precision of the BI estimates for the efficiencies comparing with the MLE estimates.

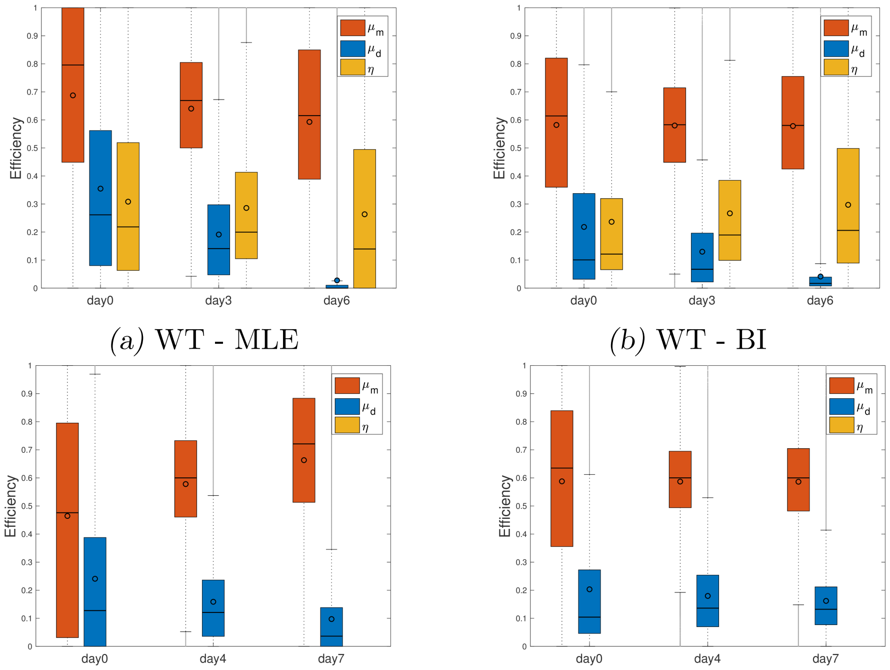
To quantify the improvement of BI compared to MLE regarding the decrease in the uncertainty of the parameter estimators we computed the average hypervolume corresponding to the covariance matrices of the estimators in each case. The volume of the hyper-ellipse of a multivariate-normal distribution is proportional to the square root of the generalized variance, i.e., the square root of the determinant of the covariance matrix, and it is given by the function

$$V = \frac{2\pi^{v/2}}{v\Gamma(v/2)}(\chi^2_{crit})^{v/2}|\Sigma_{\mathbf{v}^*}|^{1/2},$$

where v is the number of parameters, $|\Sigma_{\mathbf{v}^*}|$ is the determinant of the estimators' covariance matrix, χ^2_{crit} is the critical value for $\chi^2(v)$ and $\Gamma(x)$ is the gamma function (see Figure 5.17 for details). In WT the average volume of the hyper-ellipse in case of MLE is 0.0024 while the average hyper-ellipse volume in BI is $3.5162 \cdot 10^{-5}$. In Tet TKO the average volume of the hyper-ellipse for ML estimates is 0.0480 while in case of BI only $9.6 \cdot 10^{-4}$. In Figure 5.18 we plot the levels of the hidden states of the HMM for each combination of statistical estimation method (MLE vs BI) and cell type (WT vs Tet TKO). Overall we see small differences on the prediction of the hidden states even though there is some evident difference in the enzyme's efficiency estimators in particular for the Tet TKO case. This indicates again how critical an ML estimation bias can be for an accurate estimation of the efficiencies. For all the aforementioned reasons we use the BI estimates as the output of our model for all the analysis we present in the main manuscript as well as for the clustering that we describe in the sequel.

Tab. 5.4: Computed Kullback-Leibler divergence between the data and the model distribution for MLE and BI, where P_{bs} and P_{ox} is the data distribution for BS and oxBS experiment respectively.

experiment - method	$\hat{D}_{KL}(P_{bs}\|\|\pi_{bs})$	$\hat{D}_{KL}(P_{ox}\|\|\pi_{ox})$
WT - MLE	0.1802	0.2369
WT - BI	0.2904	0.3941
Tet TKO - MLE	0.154	-
Tet TKO - BI	0.277	-



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

Fig. 5.16: Bar plots for maintenance, *de novo* and hydroxylation efficiencies over time taken by MLE (a), (c) and BI (b), (d) methods. Red = maintenance methylation efficiency (μ_m), blue = *de novo* methylation efficiency (μ_d), yellow = hydroxylation efficiency (η).

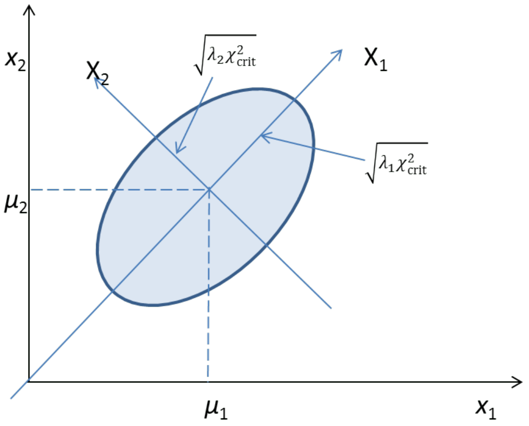
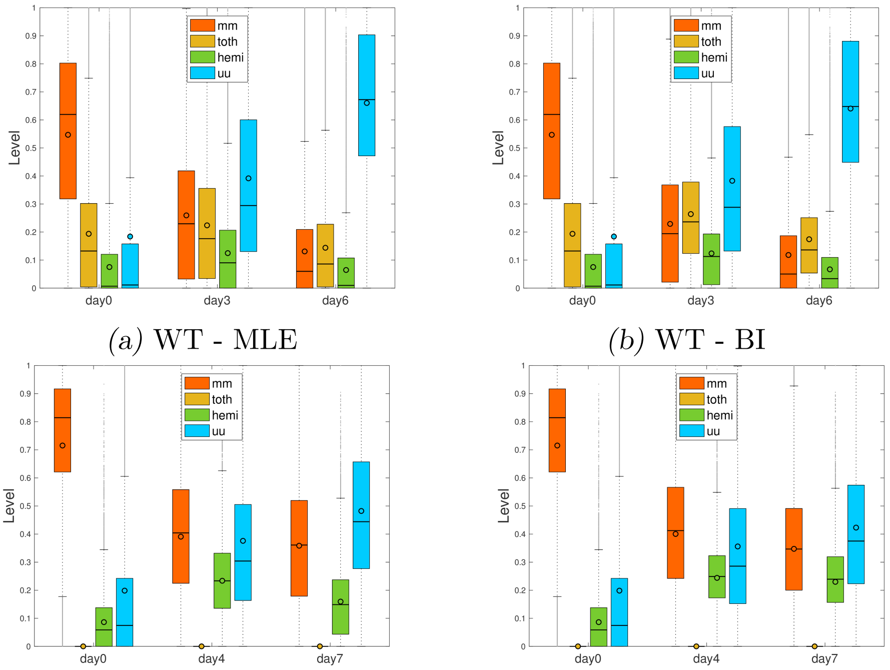

x_2
X_2
$\sqrt{\lambda_2\chi^2_{\text{crit}}}$
X_1
$\sqrt{\lambda_1\chi^2_{\text{crit}}}$
μ_2
μ_1
x_1

Fig. 5.17: The ellipse has axes pointing in the directions of the eigenvectors $X_1, X_2, ..., X_p$ of the covariance matrix Σ. Here, for the bivariate normal, the longest axis of the ellipse points in the direction of the first eigenvector X_1 and the shorter axis is perpendicular to the first, pointing in the direction of the second eigenvector X_2. The half length of the axis corresponding to eigenvector X_i is given by the formula $l_i = \sqrt{\lambda_i \chi^2_{crit}}$.

Clustering the Enzymatic Activity of Different CpGs

k-Means Clustering

We use a modification of the k-means algorithm called k-error clustering [69] that takes into account the uncertainties of each data point, i.e, the covariance matrix $\Sigma_{\mathbf{v}}$ of the



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

Fig. 5.18: Bar plots for the hidden states levels for all CpGs in the genome estimating the parameters with MLE (a), (c) and BI (b), (d). Red = symmetric methylated CpG (mm - 5mC/5mC), yellow = 5hmC in all possible combinations (toth - 5hmC/C, C/5hmC, 5hmC/5mC, 5mC/5hmC, 5hmC/5hmC), green = hemi methylated CpGs (hemi - 5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

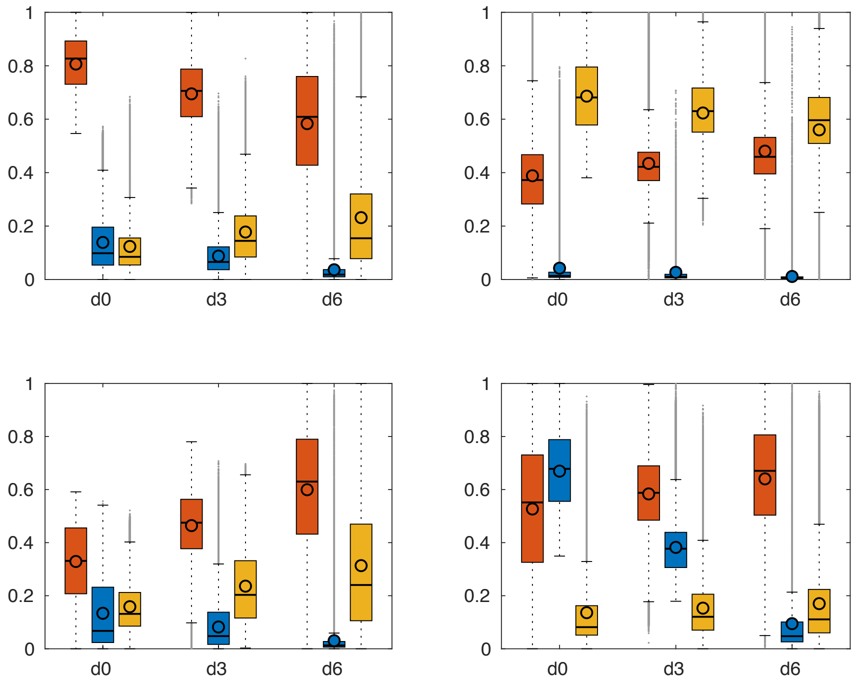


Fig. 5.19: The optimal clustering of enzymatic efficiencies over time based on the k-means algorithm and the squared euclidean distance. Red = maintenance methylation efficiency (μ_m), blue = *de novo* methylation efficiency (μ_d), yellow = hydroxylation efficiency (η).

parameter vector of the efficiencies $\mathbf{v}$.

If $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^v$ are the estimated parameter vectors and $\Sigma_1, \ldots, \Sigma_N \in \mathbb{R}^{v\times v}$ the associated covariance matrices for all input CpGs. If we assume that the estimated parameter vectors are independent and each arises from a $v-$variate normal distribution with one of k possible means $\theta_1, \ldots, \theta_k$, that is $\mathbf{v}_i \sim N_p(\mu_i, \Sigma_i)$, where $\mu_i \in \{\theta_1, \ldots, \theta_k\}$ for $i = 1, \ldots, N$. We seek to find the clusters $C_1, \ldots, C_k$ such that the parameter vectors that have the same mean $\mu_i = \theta_j$ all belong to the same cluster C_j, for $j = 1 \ldots, k$.

Let $S_j = \{i \mid \mathbf{v}_i \in C_j\}$, hence $\mu_i = \theta_j$ for $j = 1 \ldots k$ and $\forall i \in S_j$. Given N parameter vectors $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ and their error matrices $\Sigma_1, \ldots, \Sigma_N$ we search for a partition $S = (S_1, \ldots, S_k)$ and $\theta = (\theta_1, \ldots, \theta_N)$ that maximizes the following likelihood:

$$\mathcal{L}_c(\mathbf{v}) = \prod_{j=1}^{k} \prod_{i \in S_j} \frac{1}{2\pi}^{p/2} |\Sigma_i|^{-1/2} e^{-1/2(\mathbf{v}_i - \theta_j)\Sigma_i^{-1}(\mathbf{v}_i-\theta_j)^\intercal}, \tag{5.4}$$

where $|\Sigma_i|$ is the determinant of matrix Σ_i for $i = 1, \ldots, N$. Maximizing the likelihood of Eq. 5.4 is equivalent as minimizing the total squared Mahalanobis distance of the points that belong to a cluster from the cluster centroid [69], i.e.,

$$\min_S \sum_{j=1}^{k} \sum_{i \in S_j} (\mathbf{v}_i - \hat{\theta}_j)\Sigma_i^{-1}(\mathbf{v}_i - \hat{\theta}_j),$$

where $\hat{\theta}_j$ is the ML estimate of θ_j given by

$$\hat{\theta}_j = \Big(\sum_{i \in S_j} \Sigma_i^{-1}\Big)^{-1} \Big(\sum_{i \in S_j} \mathbf{v}_i \Sigma_i^{-1}\Big) \tag{5.5}$$

for $j = 1, \ldots, k$. Notice that the estimated centroid $\hat{\theta}_j$ is a weighted mean of the point in cluster C_j. We refer to it as the Mahalanobis mean of C_j.

In addition, by using simple matrix algebra we can compute that the covariance matrix Ψ_j associated with the centroid $\hat{\theta}_j$ equals

$$\Psi_j = \mathrm{Cov}(\hat{\theta}_\mathrm{j}) = \mathrm{Cov}\left(\left(\sum_{\mathrm{i} \in \mathrm{S_j}} \Sigma_\mathrm{i}^{-1}\right)^{-1}\left(\sum_{\mathrm{i} \in \mathrm{S_j}} \mathbf{v}_\mathrm{i}\Sigma_\mathrm{i}^{-1}\right)\right) = \left(\sum_{\mathrm{i} \in \mathrm{S_j}} \Sigma_\mathrm{i}^{-1}\right)^{-1}.$$

Hence, after randomly choosing an initial set of k centroids (Forgy method) the k-error method follows as an iteration over the next two steps until no change happens to the assignment of the points.

a. Assign each data point x_i to the cluster whose centroid is the closest using the squared Mahalanobis distance, i.e,

$$\arg\min_{j} \mathrm{d}_{i,j} = \arg\min_{j} (x_i - \hat{\theta}_j)\Sigma_i^{-1}(x_i - \theta_j)^{\intercal}. \tag{5.6}$$

b. For clusters $C_1, \ldots, C_k$ compute the new cluster centroids $\hat{\theta}_1, \ldots, \hat{\theta}_k$ as the Mahalanobis means of the clusters (Eq. 5.5).

The choice of the distance function used in Eq. 5.6 guarantees the decrease in the objective function in each iteration of k-error as it is shown in [69].

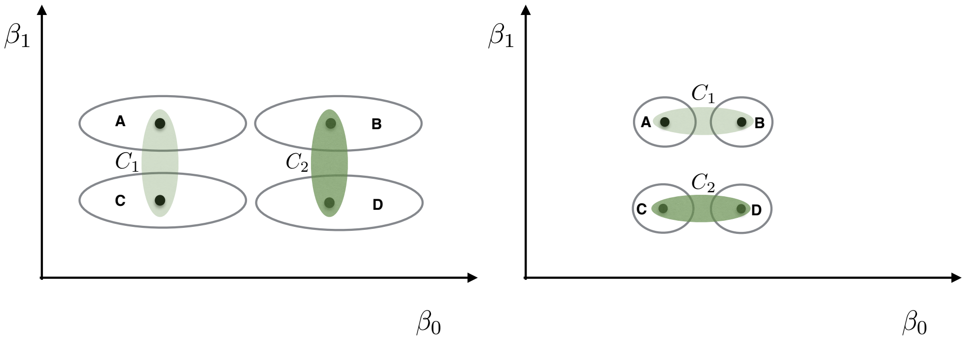


Fig. 5.20: Illustration of the clustering of an estimated enzymatic efficiency (with intercept β_0 and gradient β_1) for CpGs A, B, C, D using k-means clustering (Left) vs k-error clustering (Right).

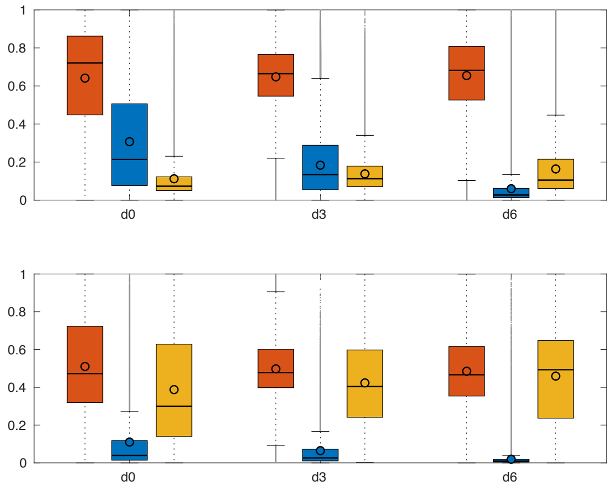


Fig. 5.21: The optimal clustering of the enzymatic efficiencies over time based on the k-error algorithm and the squared Mahalanobis distance. Cluster 1 contains 855201 CpGs while Cluster 2 contains 702901 CpGs.

Metrics for Deciding the Number of Clusters

In order to identify the "optimal" number of clusters we use Davies-Bouldin and Calinski-Harabasz criteria. These metrics evaluate the overall within-to-between cluster variability, each in a slightly different fashion. In addition, we use the simple but commonly used elbow method that considers the sum of squared errors (SSE) of a certain clustering. The goal of this method is to identify the number of clusters after which adding more clusters results only to a minor decrease of the SSE.

Davies-Bouldin Criterion

Let $R_{i,j}$ be the within-to-between cluster distance ratio for clusters i and j defined as

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}},$$

where S_i is a measure of within cluster i variance, i.e.,

$$S_i = \frac{1}{|C_i|} \sum_{x \in C_i} \mathrm{d}(x, m_i)$$

and $M_{i,j} = \mathrm{d}(m_i, m_j)$ is a measure of separation between clusters i and j defined as the distance between the clusters' centroids m_i, m_j. We define $D_i = \max_{j \neq i} R_{i,j}$, i.e., the $R_{i,j}$ of the most similar cluster to cluster i, and we get Davies-Bouldin index as the average over all D_i indices,

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i.$$

Since the value of DB represents the (worst-case) average within-to-between cluster distance ratio we decide the optimal number of clusters to be the one that provides the smallest DB.

Calinski-Harabasz Criterion

The Calinski-Harabasz criterion, alternatively called Variance Ratio Criterion (VRC), is defined as

$$CH_k = \frac{SS_B}{SS_W} \frac{(N-k)}{k-1},$$

where SS_B is the overall between-cluster variance, SS_W is the overall within-cluster variance, k is the number of clusters and N is the total number of observations. The overall between-cluster variance is defined as

$$SS_B = \sum_{i=1}^{k} |C_i| \, \mathrm{d}(m_i, m)$$

where m_i is the centroid of cluster i and m is the overall sample mean. The overall within-cluster variance is defined as

$$SS_W = \sum_{i=1}^{k} \sum_{x \in C_i} \mathrm{d}(x, m_i),$$

where the second sum goes over all points x that belong to cluster C_i. Intuitively, clusterings with well defined clusters have a large SS_B and a small SS_W. Hence, the larger the CH_k for varying k, the better the clustering. Consequently, to determine the optimal number of clusters we target to maximize CH_k w.r.t. k.

Elbow Method

We compute the sum of squared errors (SSE) for a range of number of clusters k. We choose the optimal k to be the point where the graph starts to flatten significantly. In Figure 5.22 the optimal number of clusters is clearly two.

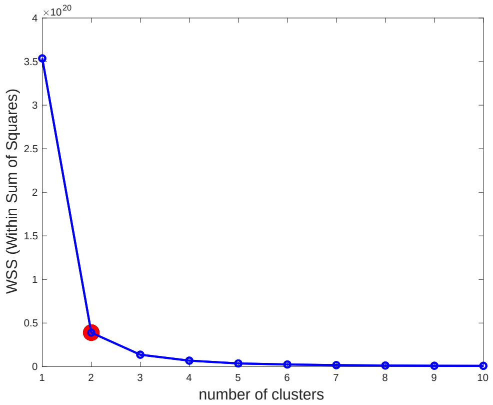


Fig. 5.22: Elbow method: The "optimal" number of clusters is the point where the graph starts to smooth out, i.e., the "elbow" of the graph.

Choice of Distance Function of the Metrics

For the evaluation of the clusterings we plug in as the distance function of the above criteria the same distance function that we used for performing the clustering. Hence, in case of k-means we use the square euclidean distance $\mathrm{d}(x, y) = \|x - y\|^2$ while for k-error we use the squared Mahalanobis distance $\mathrm{d}(x, y) = (x - y)^\intercal \Sigma_x^{-1} (x - y)$, where Σ_x is the covariance matrix of point x.

Spatial Correlations

Let X_s be the discrete space random process describing the dispersion of an enzymatic activity over the whole genome at a certain time point. For a space interval τ its spatial autocorrelation is defined as

$$R(\tau) = \frac{\mathbb{E}[(X_s - \mu_{X_s})(X_{s+\tau} - \mu_{X_{s+\tau}})]}{\sigma_{X_s}\sigma_{X_{s+\tau}}}.$$

Similarly the spatial cross-correlation between two random processes X, Y that describe the dispersion of two different enzymatic activities over the genome is defined as

$$\rho_{X,Y} = \frac{\mathbb{E}[(X_s - \mu_{X_s})(Y_{s+\tau} - \mu_{Y_{s+\tau}})]}{\sigma_{X_s}\sigma_{Y_{s+\tau}}}.$$

We compute the sample spatial autocorrelation $\hat{R}$ and the cross-correlations $\hat{\rho}$ for all enzymatic processes in both WT and Tet TKO experiments as follows. Let genome position $s \in S(\tau)$ when both CpGs of positions s and $s + \tau$ are included in our data. Then

$$\hat{R}(\tau) = \frac{1}{|S(\tau) - 1|\hat{\sigma}_{X_s}\hat{\sigma}_{X_{s+\tau}}} \sum_{s \in S(\tau)} (X_s - \bar{X}_s)(X_{s+\tau} - \bar{X}_{s+\tau})].$$

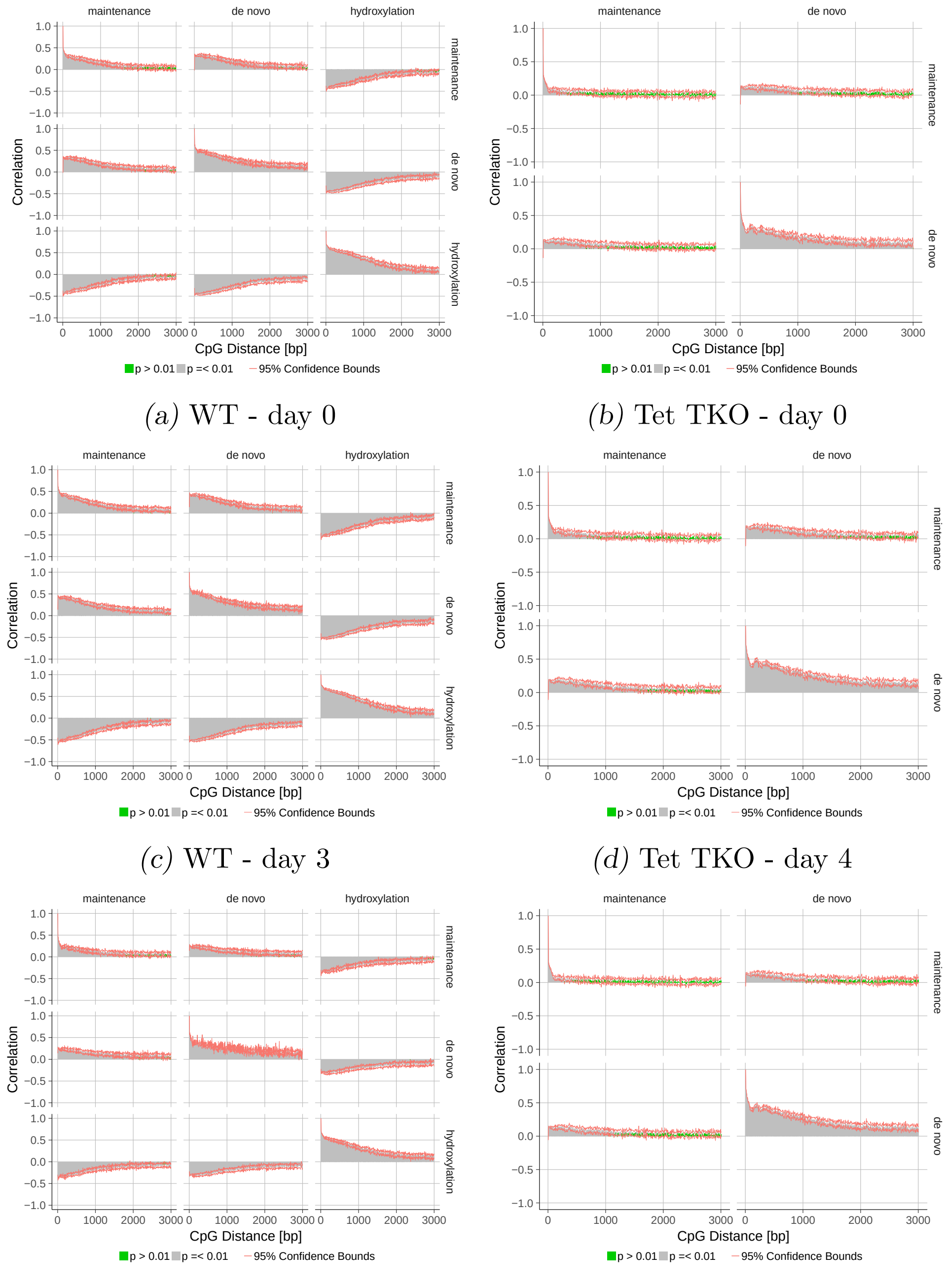
In the above sample estimator $\bar{X}_s$ and $\hat{\sigma_{X_s}}$ are the sample mean and the sample standard deviation respectively of all measurements X_s for which $s \in S(\tau)$. The same way we compute

$$\hat{\rho}(\tau) = \frac{1}{|S(\tau) - 1|\hat{\sigma}_{X_s}\hat{\sigma}_{Y_{s+\tau}}} \sum_{s \in S(\tau)} (X_s - \bar{X}_s)(Y_{s+\tau} - \bar{Y}_{s+\tau})].$$

Fixing $\tau = 5$ we plot in Figure 5.23 the sample autocorrelations and sample cross-correlations between all efficiencies at all time points in WT (Figure 5.23a, 5.23c, 5.23e) and Tet TKO (Figure 5.23b, 5.23d, 5.23f) experiments. Together with the sample correlations we report 95% confidence intervals following the approach of [83] and p-values for the null hypothesis that the auto or the cross-correlation is zero.

We observe a negative correlation of methylation and hydroxylation efficiencies across the entire genome and a positive correlation between *de novo* and maintenance efficiencies. Furthermore, we observe for all efficiencies a positive autocorrelation which constantly declines as the distance between CpGs increases. In WT maintenance autocorrelation flattens out at about 1500bp distance between CpGs and at the same time it starts losing its significance (p>0.01). Both *de novo* and hydroxylation activity show a high autocorrelation in the beginning (more than 0.5) and seem to influence windows of larger sizes, around 2000 bp.

In Tet TKO cells maintenance methylation activity seems to flatten out earlier than in WT, around 300 bp, which is in agreement with the observation that maintenance activity appears misregulated in Tet TKOs, in particular showing an increase at the TSS. On the contrary, the activity area of *de novo* enzymes seems to be stable compared to WT cells (around 2000 bp). Overall the spatial autocorrelations do not indicate any change of the activity window size of the enzymes over time.



(a) WT - day 0

(b) Tet TKO - day 0

(c) WT - day 3

(d) Tet TKO - day 4

(e) WT - day 6

(f) Tet TKO - day 7

Fig. 5.23: Spatial auto- and cross correlation of maintenance, *de novo*- and hydroxylation efficiency across the genome. Grey bars indicate correlations with a p value < 0.01, green bars correlations with p values > 0.01, red line shows the confidence bounds. Y-axis displays correlation, x-axis gives the distance of CpG in base pairs.

Pearson Correlation

In addition to the spatial correlation, we calculated a simple pearson correlation for average efficiency and estimated modification levels.

In WT ES cells, we observe a positive correlation of fully methylated CpGs with *de novo* and maintenance efficiency at day 0. For later time points, this correlation increases. In addition, there is a positive correlation between the two methylation efficiencies and hemimethylated CpGs for day 3 and day 6, which is not present at day 0. This is likely due to an increase in hemimethylated CpGs which temporally overlap with the remaining Dnmt activity in 2i.

Interestingly, we observe no correlation between hydroxylated CpGs and hydroxylation efficiency. Instead, hydroxylation activity correlates with unmethylated CpGs. Thus, we conclude that high hydroxylation activity is not sufficient to generate stable 5hmC but will rather result in methylation free CpGs.

In Tet TKO, the correlation between maintenance methylation efficiency and fully methylated CpG dyads are reduced and in case of day 0 only comes to 0.13. However, we observe a stronger correlation of fully methylated CpGs with *de novo* methylation efficiency which points towards a misregulated methylation activity in the absence of Tet enzymes.

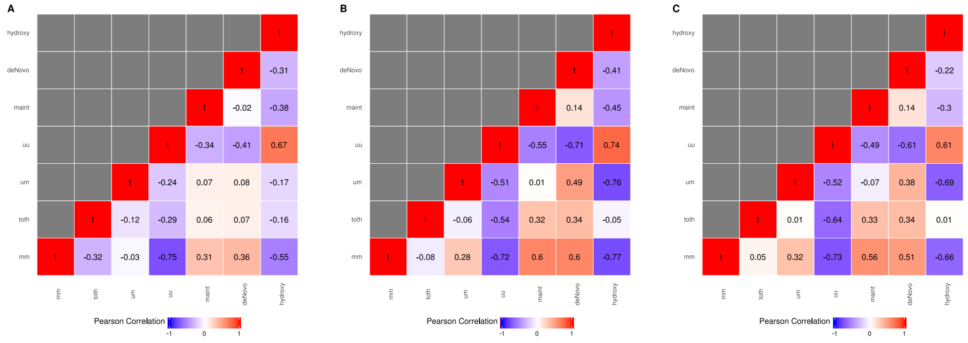


Fig. 5.24: Pearson correlation of enzyme efficiencies and methylation level in WT ES cells for d0, d3 and d6. mm = fully methylated (5mC/5mC), toth = hydroxylated CpG of all possible states, um = hemimethylated (5mC/C or C/5mC), uu = unmethylated (C/C), maint = maintenance methylation efficiency, deNovo = *de novo* methylation efficiency, hydroxy = hydroxylation efficiency

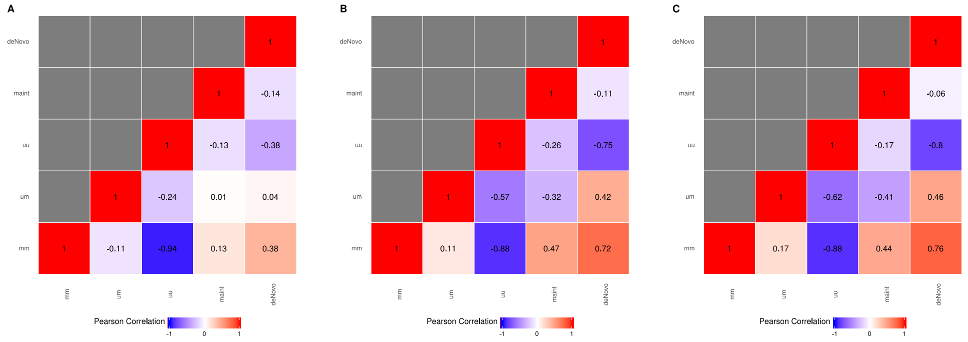


Fig. 5.25: Pearson correlation of enzyme efficiencies and methylation level in Tet TKO ES cell for d0, d4 and d7. mm = fully methylated (5mC/5mC), um = hemimethylated (5mC/C or C/5mC), uu = unmethylated (C/C), maint = maintenance methylation efficiency, deNovo = *de novo* methylation efficiency.

5.6.5 Additional Results

ES Cell Chromosomes Results

In this section we provide the input-data information plots as well as the output of our model for each of the 21 main chromosomes of the ESCs. In Figure 5.26, 5.27 we plot the number of CpGs for each chromosome with one, two or three observation days in WT and Tet TKO cells, respectively. We plot the average number of samples (depth sequencing) for each chromosome in WT (Figure 5.28) and Tet TKO (Figure 5.29). In Figure 5.30, 5.31 we show the efficiencies over time computed by BI and in Figure 5.32, 5.33 we report the prediction of the model for the hidden states probabilities in each chromosome in WT and Tet TKO cells.

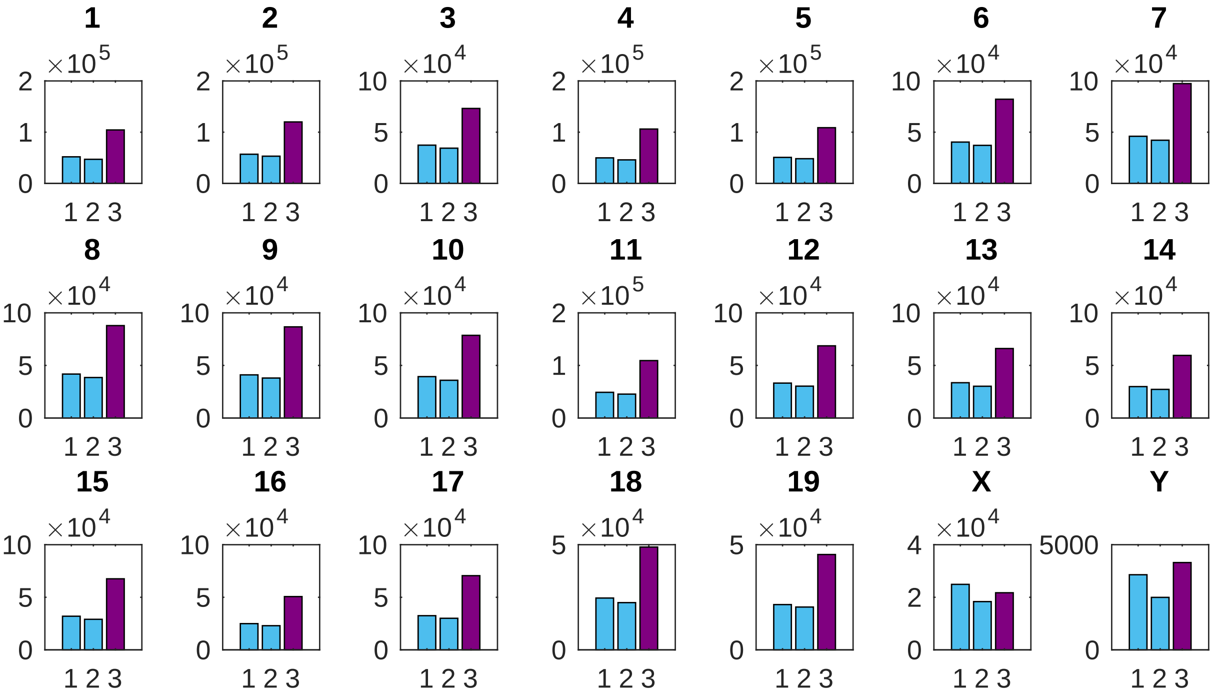


Fig. 5.26: Number of CpGs (y-axis) with one, two or three observation days (x-axis) for each chromosome in WT data.

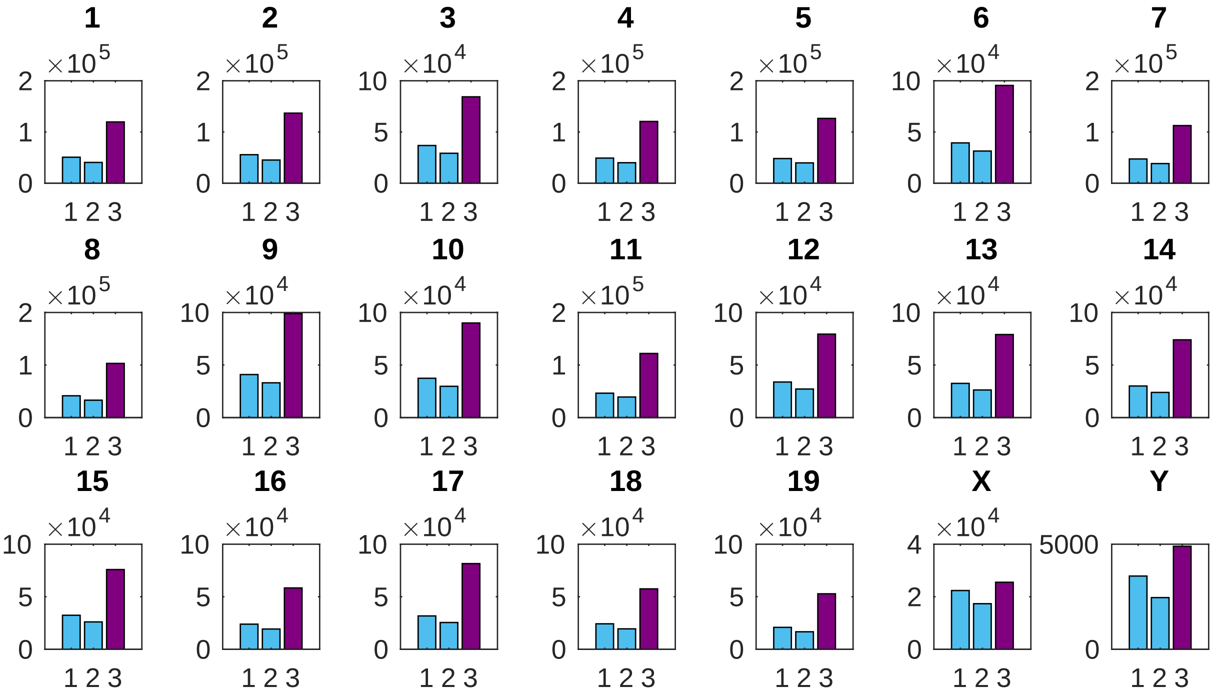


Fig. 5.27: Number of CpGs (y-axis) with one, two or three observation days (x-axis) for each chromosome in Tet TKO data.

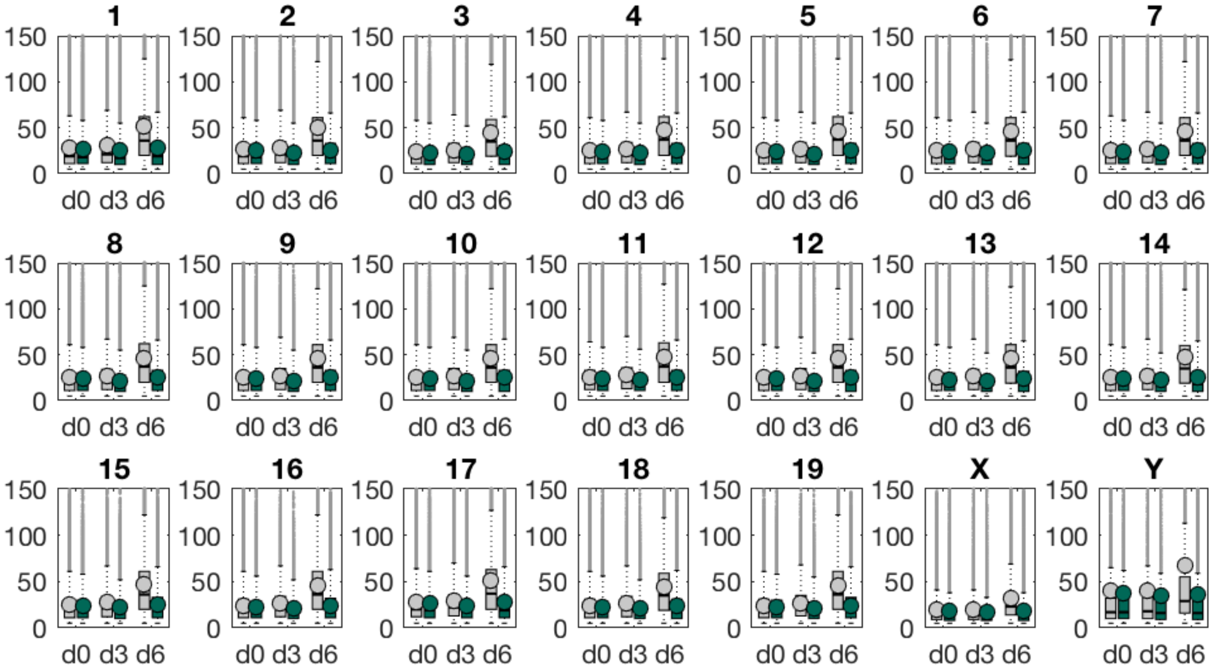


Fig. 5.28: Average number of single CpG independent samples, i.e, depth sequencing, (y-axis) per day (x-axis) for each chromosome in WT data.

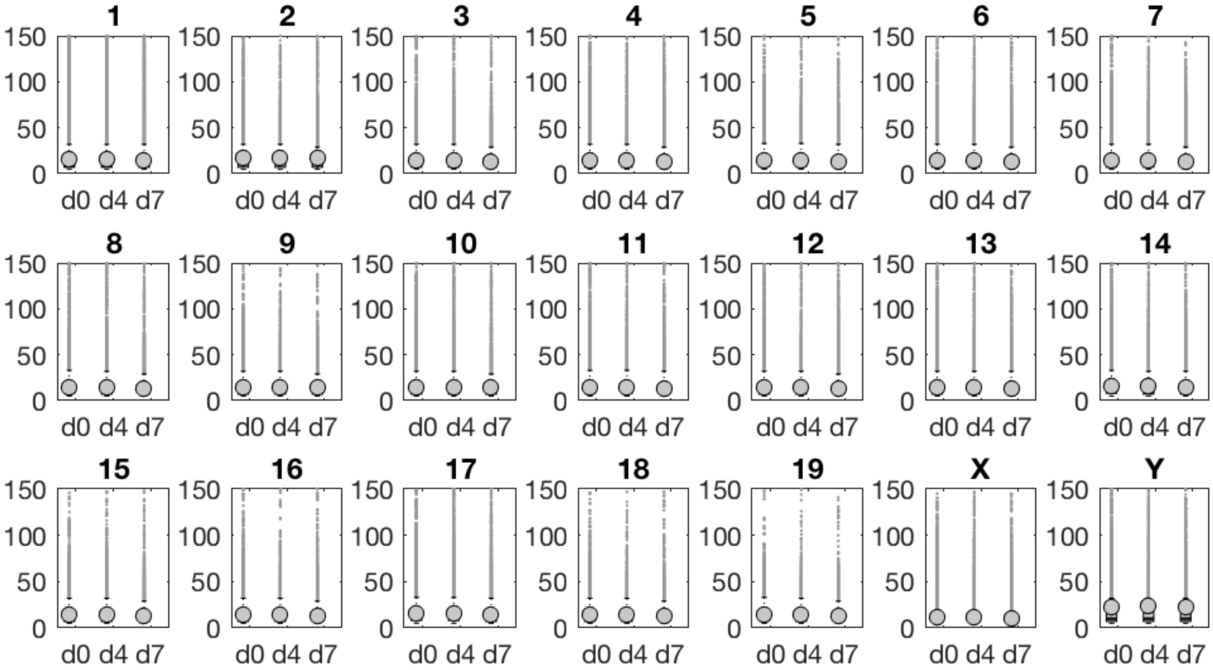


Fig. 5.29: Average number of single CpG independent samples, i.e, depth sequencing, (y-axis) per day (x-axis) for each chromosome in Tet TKO data.

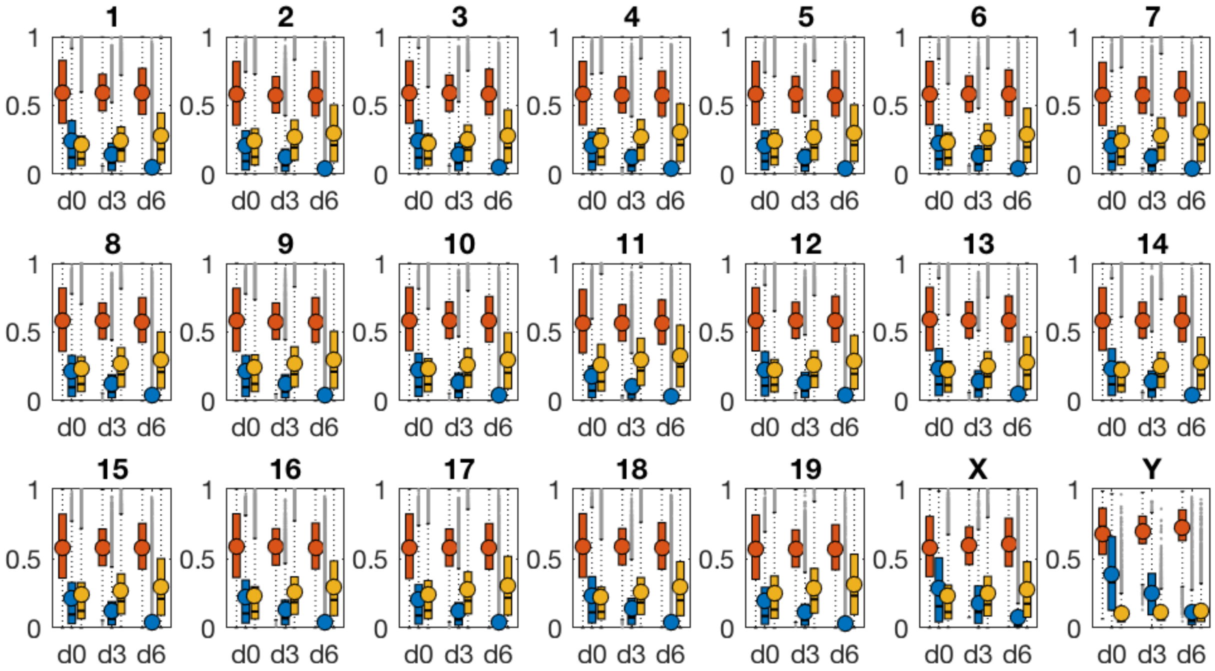


Fig. 5.30: Bar plots for the maintenance (red), *de novo* (blue) and hydroxylation (yellow) efficiencies over time taken by BI method for each individual chromosome in WT cells.

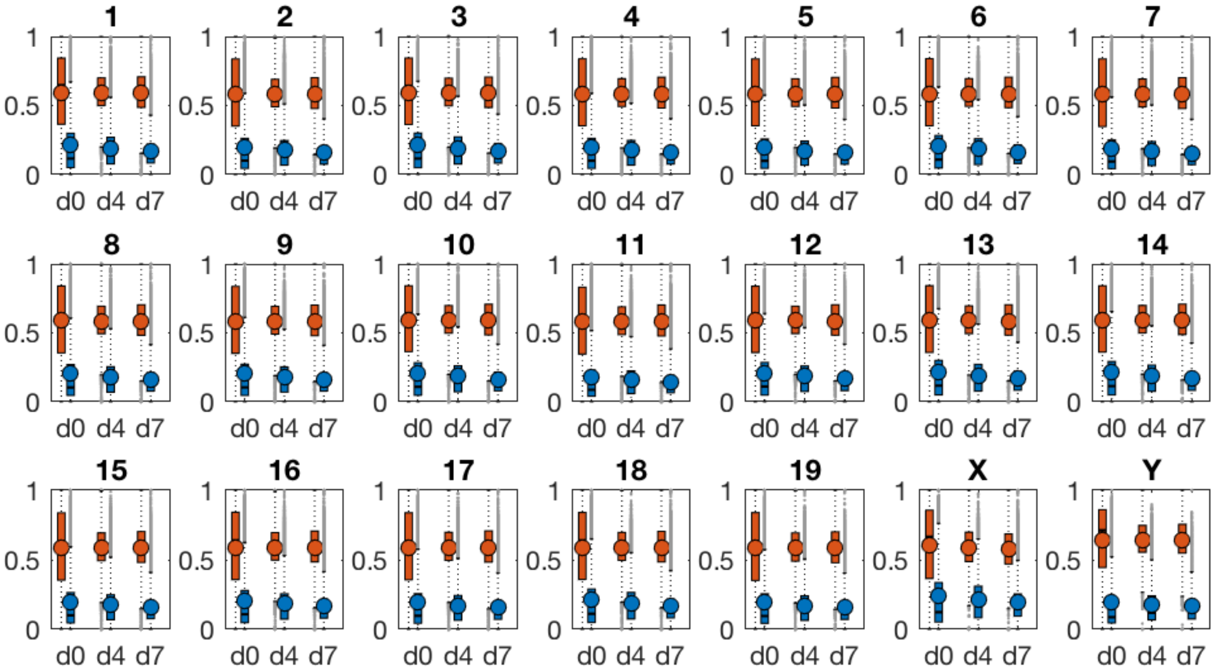


Fig. 5.31: Bar plots for the maintenance (red) and *de novo* (blue) efficiencies over time taken by BI method for each individual chromosome in Tet TKO cells.

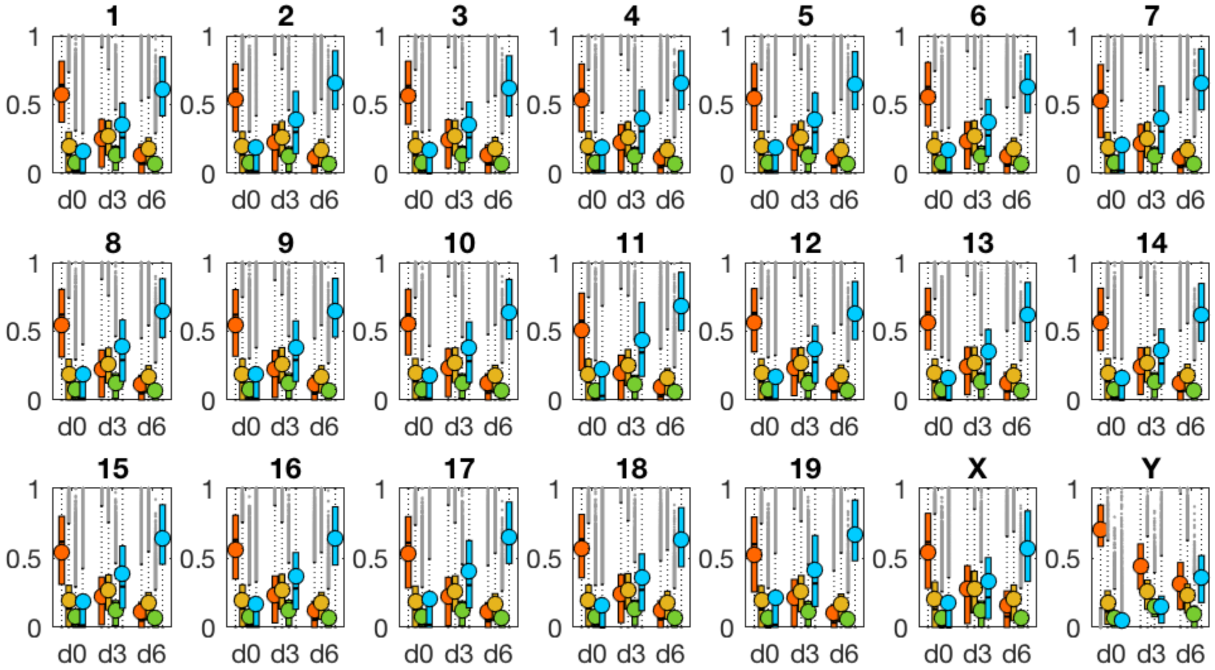


Fig. 5.32: Bar plots for the hidden states levels over time of each individual chromosome in WT. Red = symmetric methylated CpG (mm - 5mC/5mC), yellow = 5hmC in all possible combinations (toth - 5hmC/C, C/5hmC, 5hmC/5mC, 5mC/5hmC, 5hmC/5hmC), green = hemi methylated CpGs (hemi - 5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

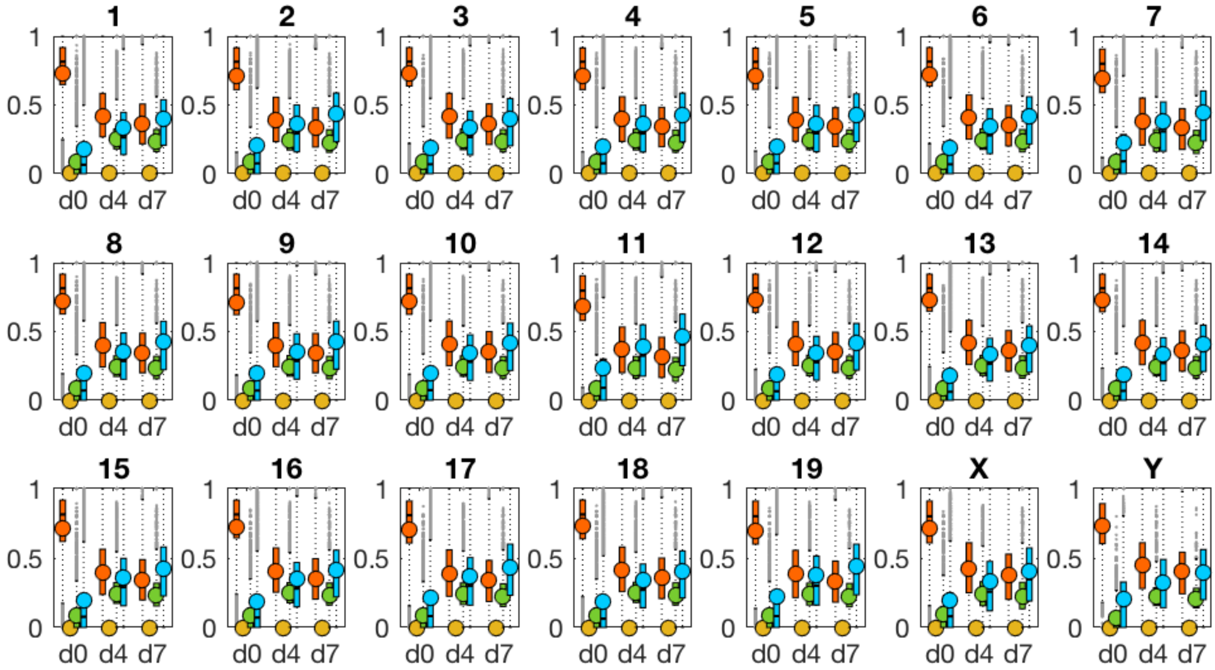


Fig. 5.33: Bar plots for the hidden states levels over time of each individual chromosome in TET TKO. Red = symmetric methylated CpG (mm - 5mC/5mC), yellow = 5hmC in all possible combinations (toth - 5hmC/C, C/5hmC, 5hmC/5mC, 5mC/5hmC, 5hmC/5hmC), green = hemi methylated CpGs (hemi - 5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

Spike-In Analysis

To determine the conversion rate of BS and oxBS we included short oligonucleotides into our RRHPoxBS libraries. The oligo mix is part of the TrueMethyl kit from Cambridge Epigenetix and includes C, 5mC, 5hmC and 5fC at known positions. After sequencing, we calculated the conversion rates for each cytosine variant, which were than included into our model to compensate for conversion errors.

Tab. 5.5: Conversion rate of cytosine variants included in the TruMethyl Spike in after BS treatment

	C	5mC	5hmC	5fC
Serum	0.996332	0.0699681	0.0673588	0.75626
72h	0.996165	0.0725858	0.0715434	0.762992
144h	0.995809	0.0696952	0.0682802	0.739254

Tab. 5.6: Conversion rate of cytosine variants included in the TruMethyl Spike in after oxBS treatment

	C	5mC	5hmC	5fC
Serum	0.99687	0.0662679	0.964215	0.968836
72h	0.99656	0.0670022	0.967298	0.9663
144h	0.996901	0.0534113	0.949588	0.932044

Hemimethylated CpGs

Hemimethylated CpGs are the result of *de novo* methylation events and/or active and passive demethylation. Theoretically, selective methylation of a DNA strand could provide a strand specific gene regulation mechanism. Thus, we analysed the strand specific methylation of genes transcribed from plus- and minus strand. Results are displayed in Figure 5.34.

We cannot observe any methylation differences between genes expressed from upper or lower DNA strands. In both cases we observe the same amount of hemimethylation at both strands. The same holds true for low/not expressed genes.

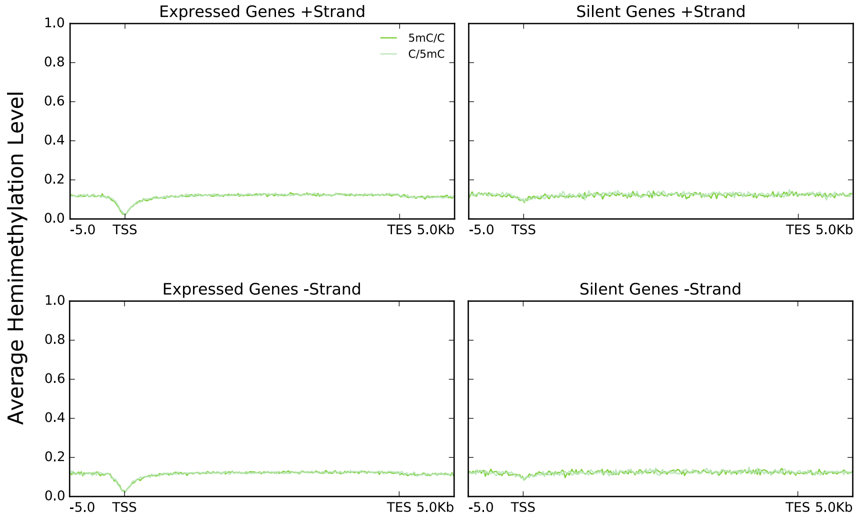


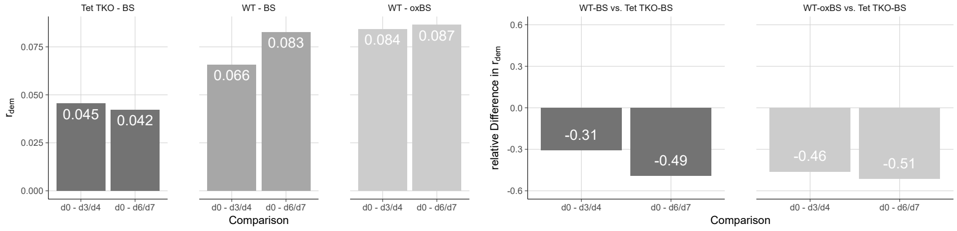
Fig. 5.34: Average hemimethylated CpGs detected by RRHPoxBS across expressed and not/low expressed genes. Dark green = 5mC/C, light green = C/5mC

Demethylation Kinetics

Previous studies indicated that Tet TKO cells exhibit the same demethylation kinetics as WT ES cells during their transition from Serum to 2i [60]. However, our RRHPoxBS data shows a noticeable difference in the methylation levels of WT and Tet TKO cells. Thus, we calculated the demethylation rate r_{dem} for each cell type to further investigate the distinct demethylation kinetics. For this, we calculated the increase of unmethylated cytosines for time points and for WT $t = \{3, 6\}$ and Tet TKO cells for time points $t = \{4, 7\}$ using the equation:

$$r_{\text{dem}}(t) = \frac{\text{TT}(t) - \text{TT}(0)}{t}.$$

Results are displayed in Figure 5.35.



(a) Demethylation Rates WT

(b) Demethylation Rates Tet TKO

Fig. 5.35: (a) Demethylation rate in WT and Tet TKO cells (b) Relative difference in demethylation rate between WT and Tet TKO cells.

In contrast to the previous observations, we observe distinct demethylation rates for WT and Tet TKO cells. WT cells exhibit a demethylation rate between 6 and 8%, whereas Tet TKO ES cells show a reduced demethylation rate of around 4% (Fig. 5.35 (a)). Consequently, the demethylation rate in Tet TKO cells w.r.t. to WT cells is reduced by 30 to

50%, demonstrating the considerable contribution of Tet enzymes to DNA demethylation (Fig. 5.35 (b)).

Repetitive Elements

The majority of the mammalian genome is composed of repetitive elements (REs). Thus, we examined whether a subset of REs would reflect the average behavior of the genome. For this, we assign CpGs to individual REs. Figures 5.36 to 5.40 show methylation level and efficiencies for the 25 most frequent repetitive elements in our data set for WT and Tet triple TKO ES cells. Indeed, we observe that the majority of REs resemble closely the level and efficiency profile of individual chromosomes as well as the average genome profiles. However, we also observe some exceptions. Intracisternal A particle and major satellites exhibit considerable higher methylation level and methylation efficiency compared to the mean genome profile. In addition, GC rich elements show almost no 5mC/5hmC, low methylation efficiency but high hydroxylation activity of Tets. Thus, they resemble more the behavior of promoters and TSS.

In case of the Tet TKO cells, we observe, that the maintenance efficiency in the distinct repetitive elements converge. In addition, *de novo* methylation appears again reduced for the first time point, but remains present even after continuous incubation in 2i media (Fig.5.40).

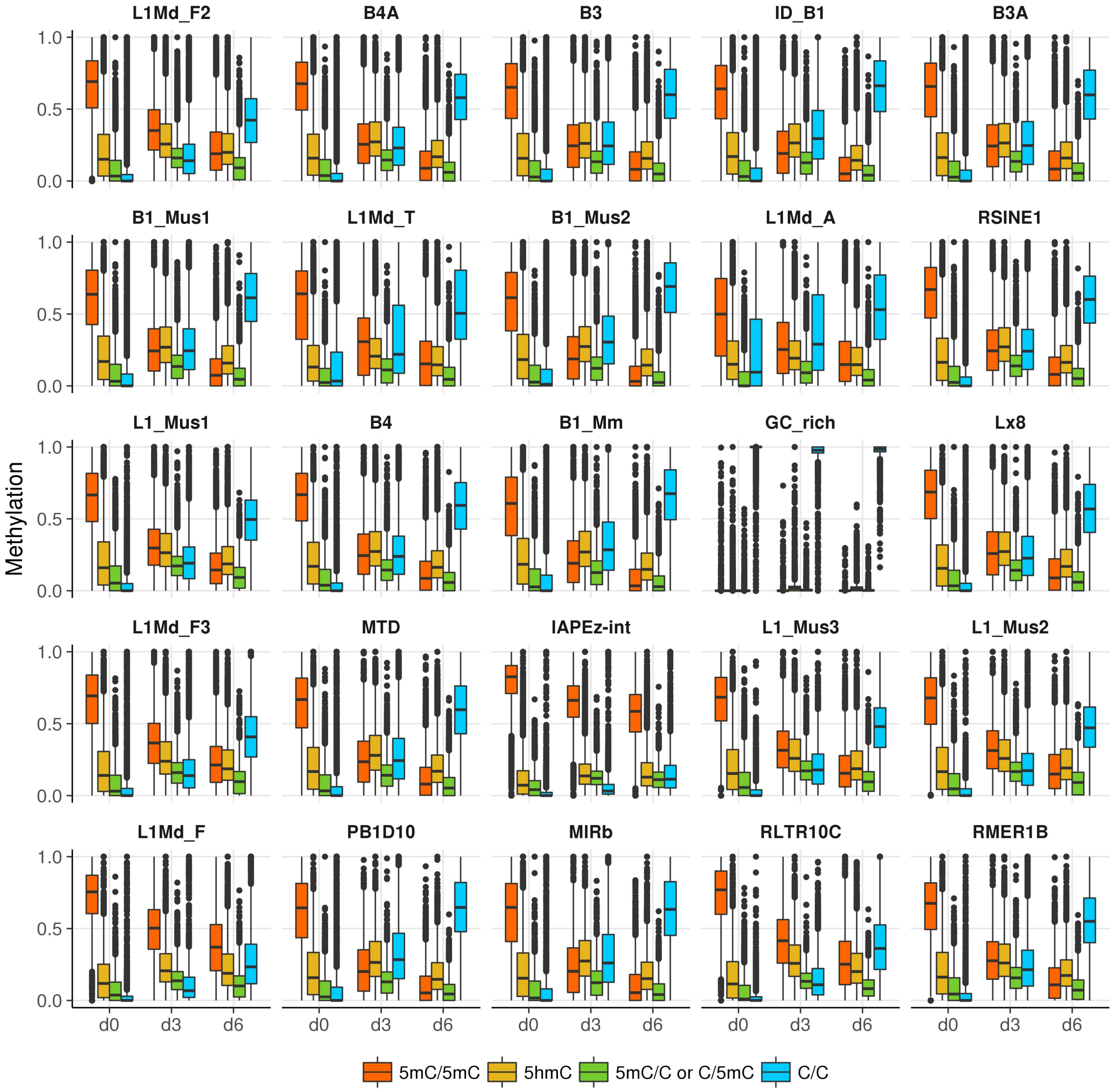


Fig. 5.36: Methylation level at the 25 most frequent repetitive elements in our analysis for WT ES cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. y-axis = methylation frequency, x-axis = time in days (d0, d3, d6). Red = fully methylated CpGs (5mC/5mC), green = hemimethylated CpGs (5mC/C or C/5mC), yellow = 5hmC, blue = unmethylated CpGs (C/C).

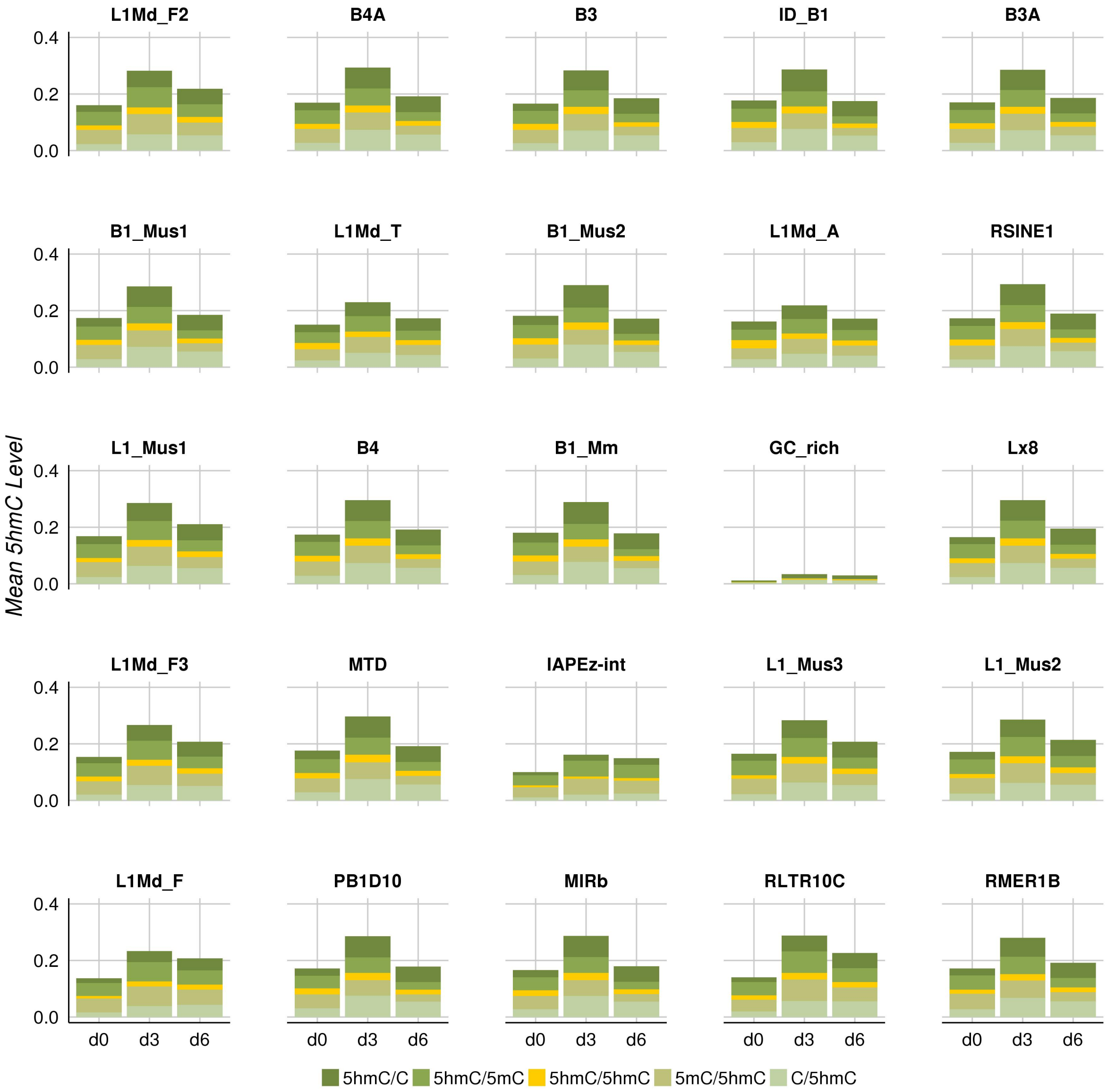


Fig. 5.37: Level and distribution of 5hmC within the 25 most frequent repetitive elements in our data set for WT ES cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. y-axis = mean 5hmC level, x-axis = time in days (d0, d3, d6).

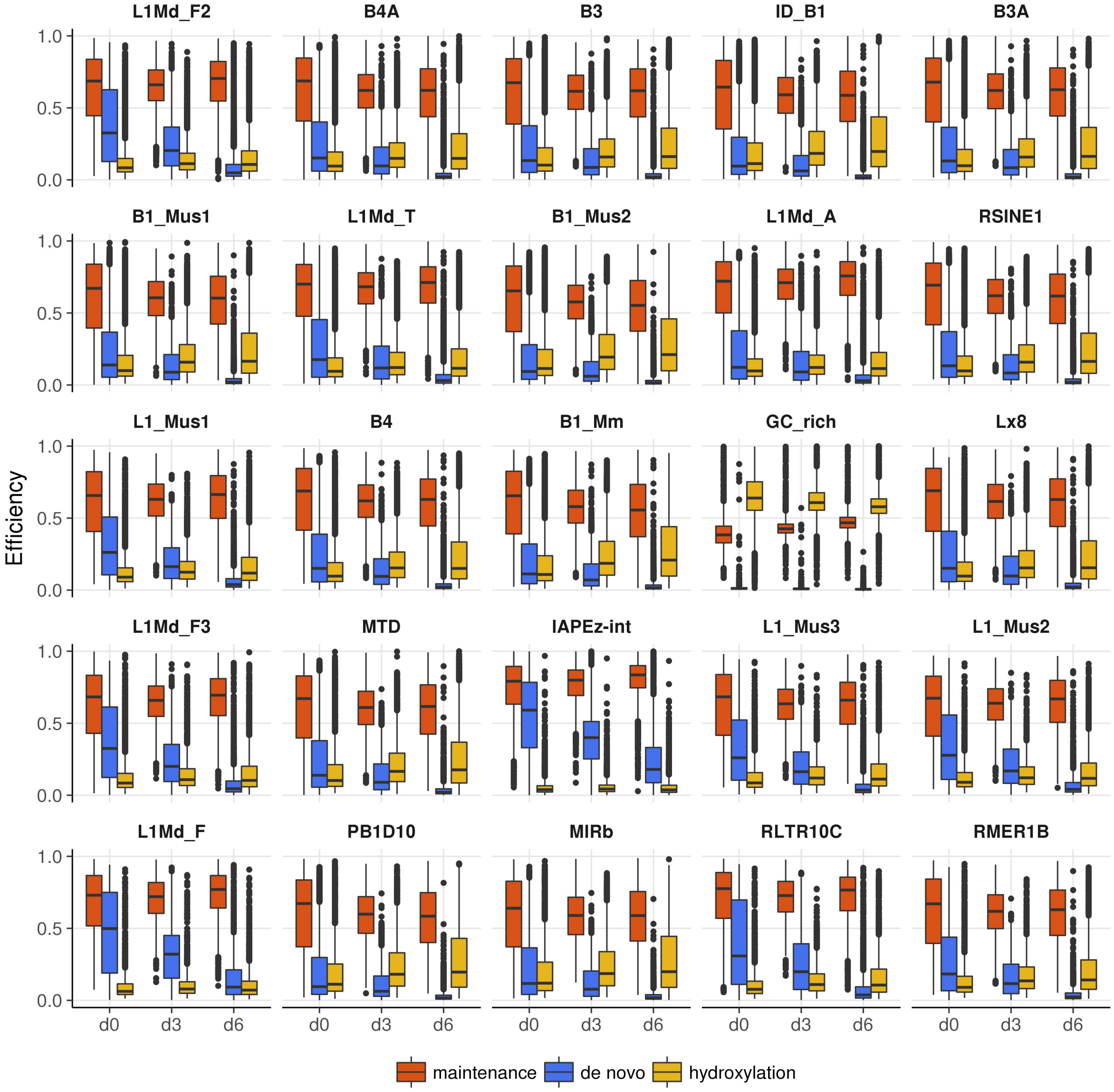


Fig. 5.38: Efficiency profiles of the 25 most frequent repetitive elements in our analysis for WT ES cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. y-axis = efficiency; x-axis = time in days (d0, d3, d6), red = maintenance efficiency, blue = *de novo* efficiency, yellow = hydroxylation efficiency.

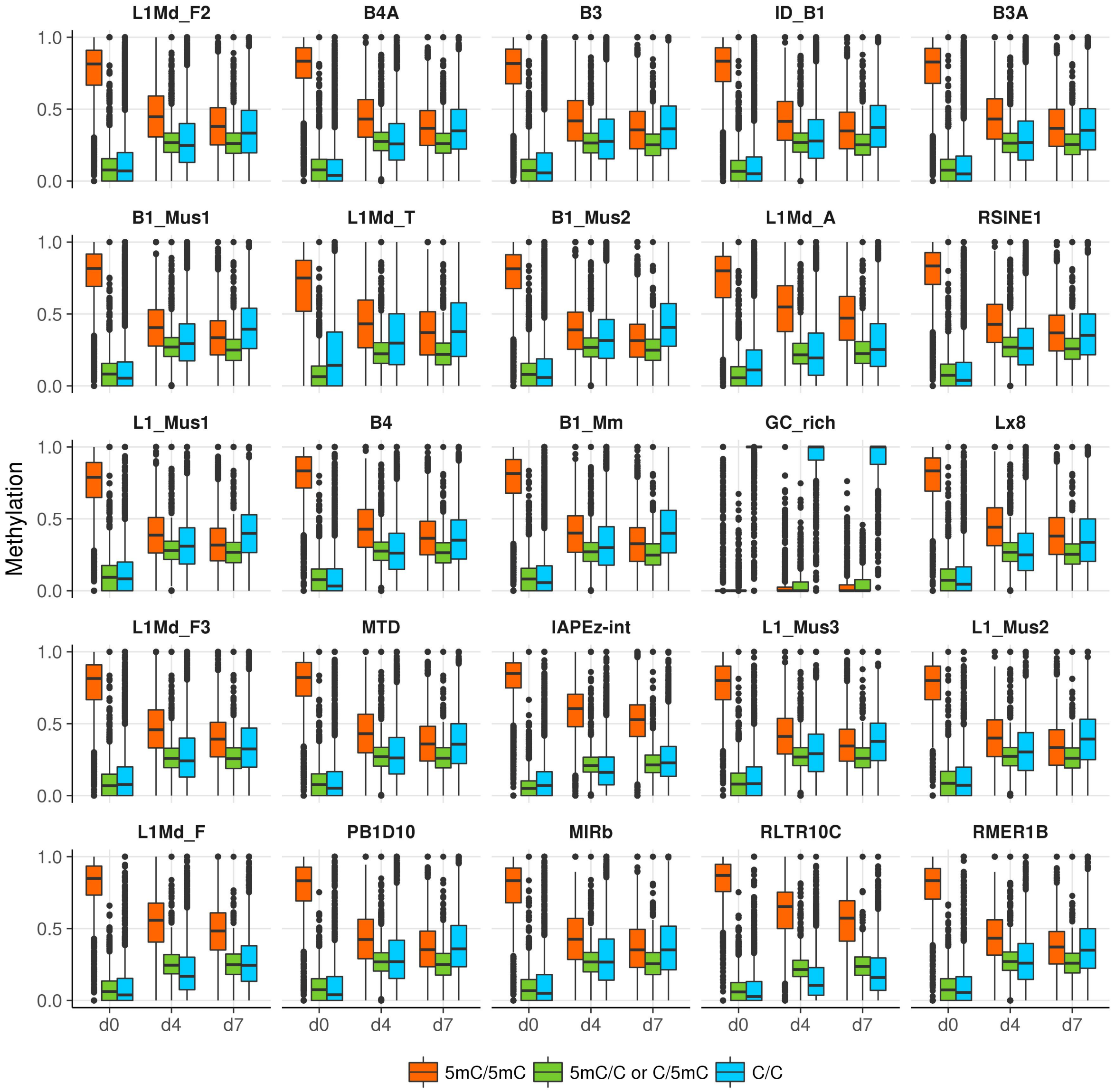


Fig. 5.39: Methylation level at the 25 most frequent repetitive elements in our analysis for Tet TKO cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. y-axis = methylation frequency, x-axis = time in days (d0, d4, d7). Red = fully methylated CpGs (5mC/5mC), green = hemimethylated CpGs (5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

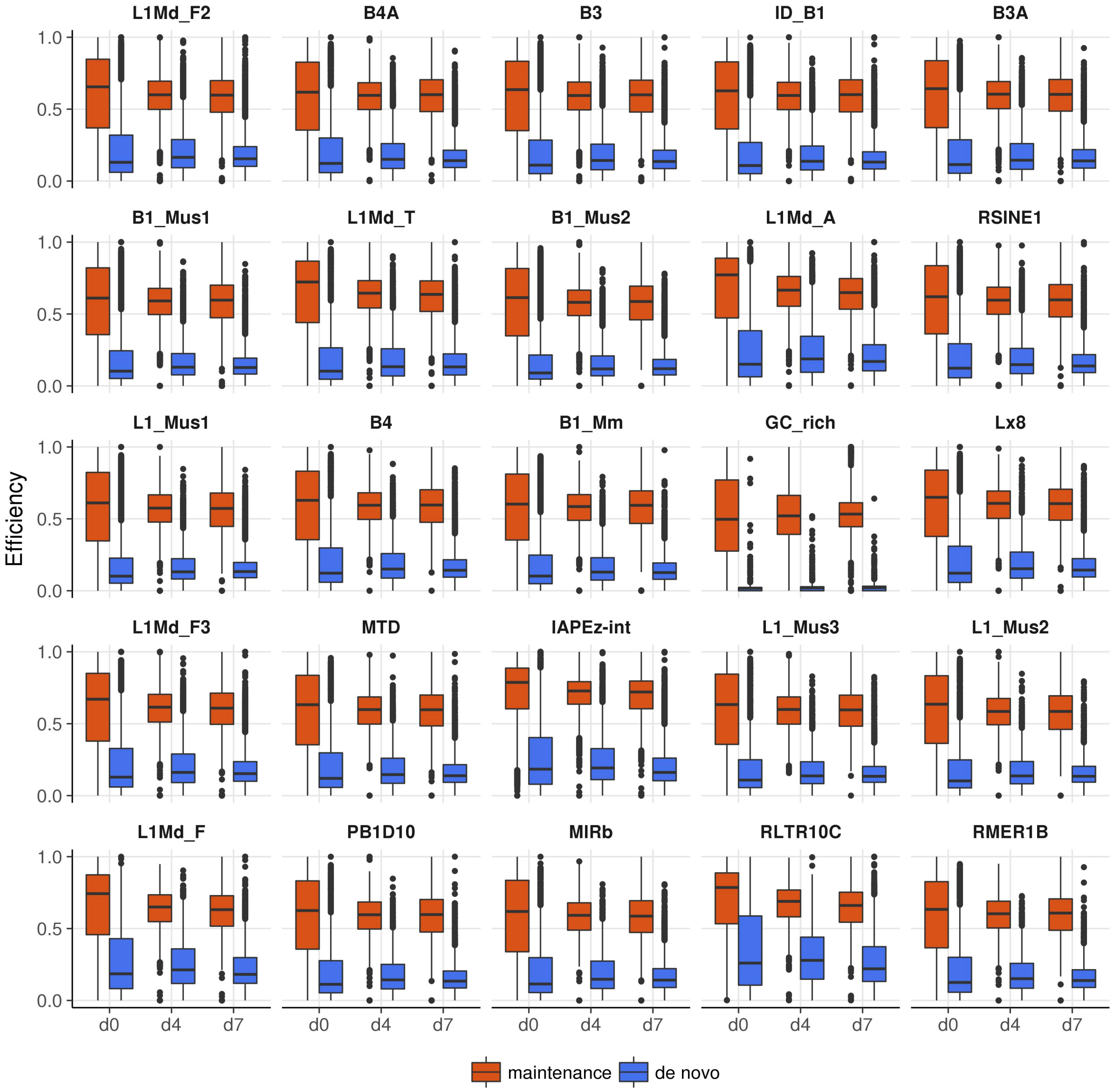


Fig. 5.40: Efficiency profiles of the 25 most frequent repetitive elements in our analysis for Tet TKO cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. y-axis = efficiency; x-axis = time in days (d0, d4, d7), red = maintenance efficiency, blue = *de novo* efficiency.

Efficiency and Binding Profiles of DNA Modifiers

In Figure 5.41 and 5.42 we compared the binding profile (ChIP-Seq) of DNA modifiers from previous publications to maintenance, *de novo* and hydroxylation efficiencies estimated by our model(GSM659799, GSE57413, GSE100957) [62, 73, 74]. ChIP profiles for Dnmt3 iso-forms reveal a reduced binding around the TSS, which is in concordance with our models prediction in reduced *de novo* efficiency. Interestingly, when comparing the ChIP profiles of Dnmt3s to transcriptome sequencing from Ficz et al. we observe distict profiles for expressed and non-expressed genes 5.41. In case of expressed genes, we observe a strong enrichment across the gene body and reduced binding around the TSS, in particular for Dnmt3a1, while non-expressed genes display the strongest binding precisely at the TSS.

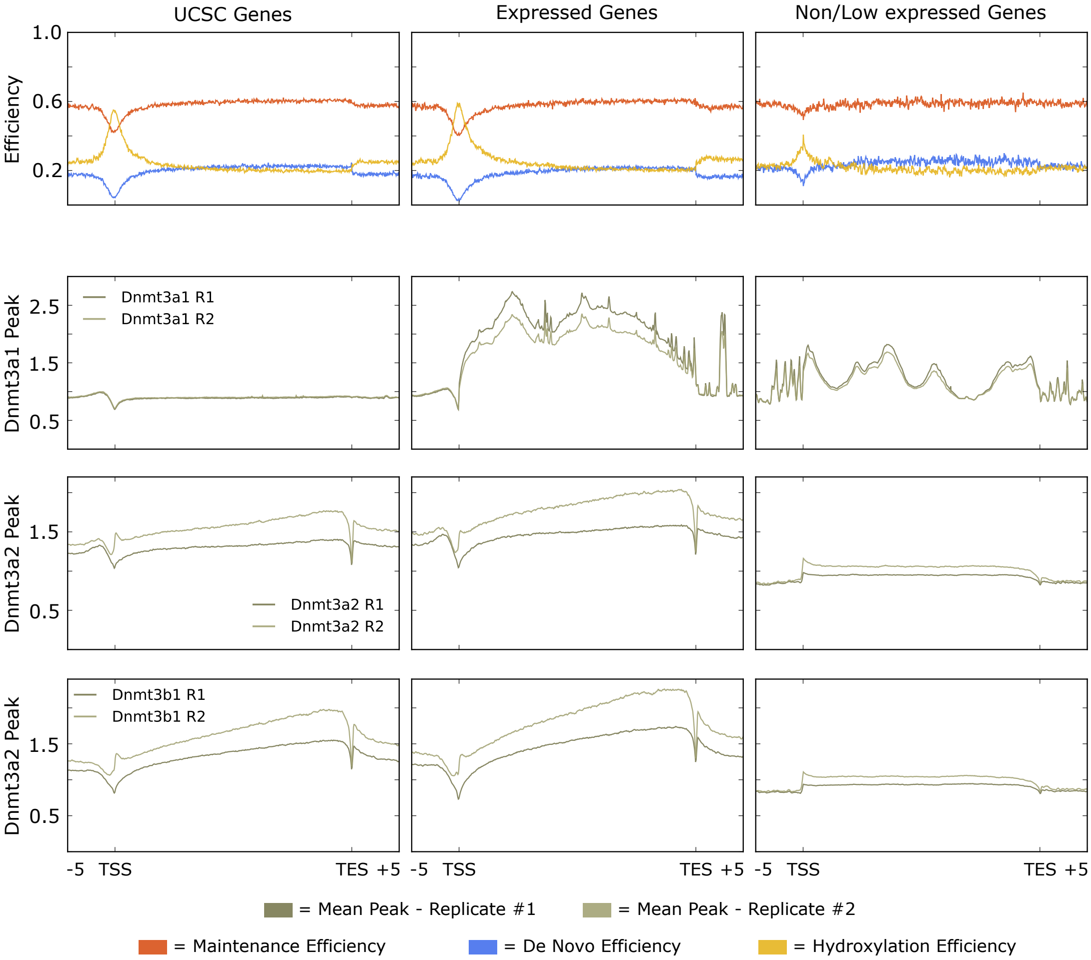


Fig. 5.41: Estimated Efficiencies of Dnmts and Tets in form of maintenance methylation(red), *de novo* methylation and hydroxylation as well as binding of Dnmt3a and 3b isoforms.

Additionally, we compared the efficiencies to ChIP binging profiles of Tet1 and Uhrf1, as one essential subunit of the maintenance machinery. Again, the ChIP profiles correspond nicely to our efficiencies of Dnmts and Tets 5.42. While Uhrf1 is less frequent at TSS similar to maintenance and *de novo* methylation, Tet1 in contrast displays a high

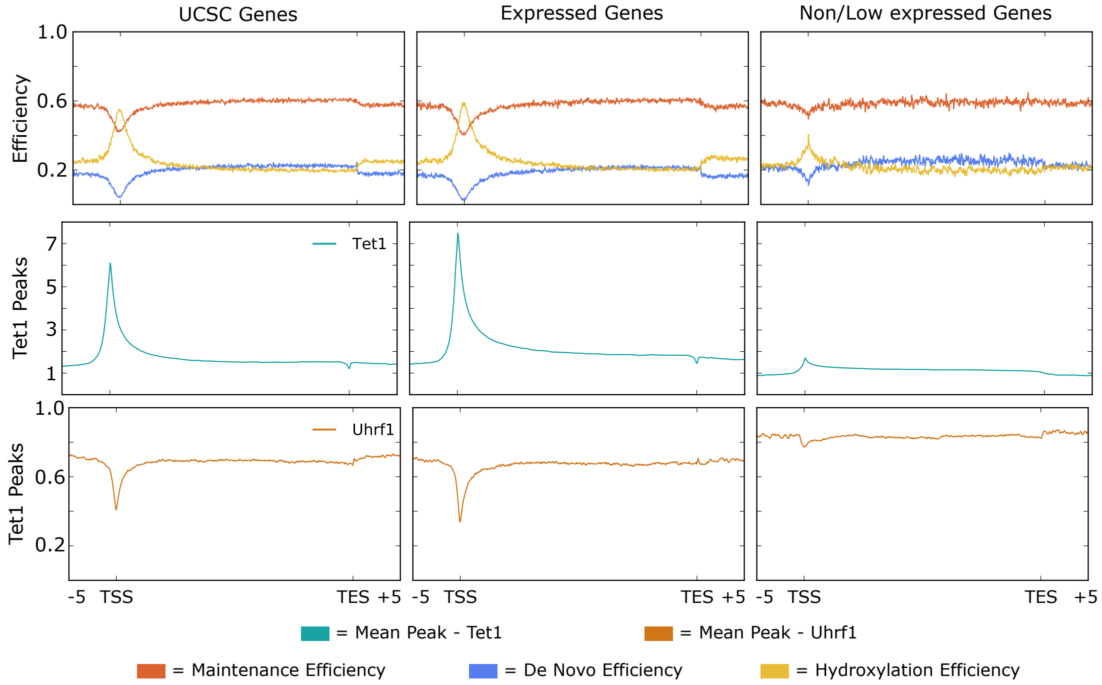
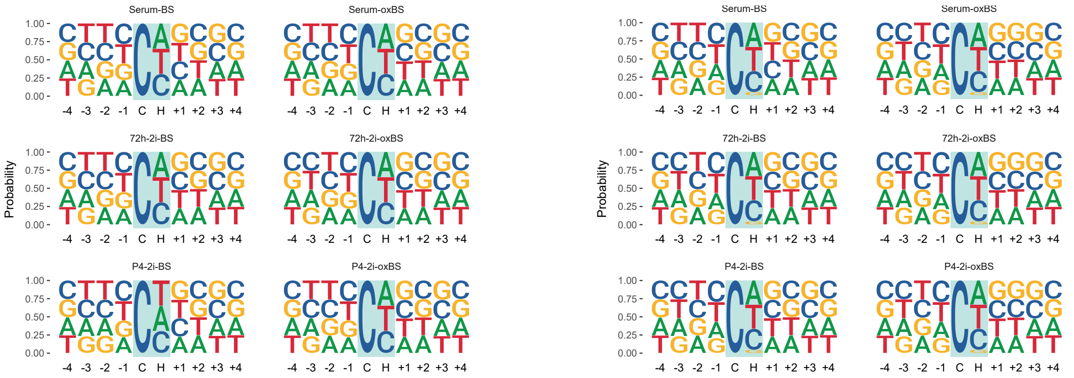


Fig. 5.42: Estimated Efficiencies of Dnmts and Tets in form of maintenance methylation(red), *de novo* methylation and hydroxylation as well as binding of Tet1 and Uhrf1.

enrichment at TSS matching the strongly increased hyroxylation efficiency observed by our model.

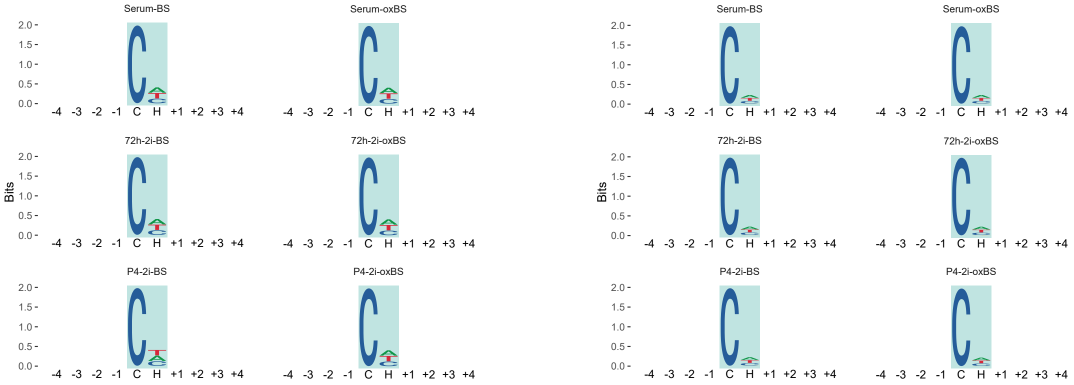
nonCpG Methylation

Frequently, DNA methylation occurs outside of a CpG context [5, 6]. Hence, we determined the sequence occurrence of nonCpG methylation in our WT samples. For our analysis, we considered only nonCpG positions which are (i) methylated above the conversion error, (ii) show at least three methylated reads and (iii) a coverage of $\geq$ 10. In accordance with literature, we find that CpA is the most common methylated sequence after CpG on both DNA strands (Figure 5.43 and Figure 5.44). Furthermore, we see that the majority of all nonCpG in our data set correspond to FMRs and PMDs, whereas only a small fraction can be found in LMRs and UMRs (Figure 5.45). This observation nicely matches our model's prediction according to which FMRs and PMDs exhibit higher *de novo* methylation activity (main manuscript figure 9) mainly caused by Dnmt3a and 3b.



(a) nonCpG methylation +Strand (b) nonCpG methylation −Strand

Fig. 5.43: Occurrences of nonCpG methylation in Serum and 2i cultivated WT ES cells. Size of bases indicate the probability at a given position. nonCpG with 4 bases up- and downstream are shown.



(a) nonCpG methylation +Strand (b) nonCpG methylation −Strand

Fig. 5.44: Occurrences of nonCpG methylation in Serum and 2i cultivated WT ES cells. Size of bases indicate the probability at a given position. nonCpG with 4 bases up- and downstream are shown.

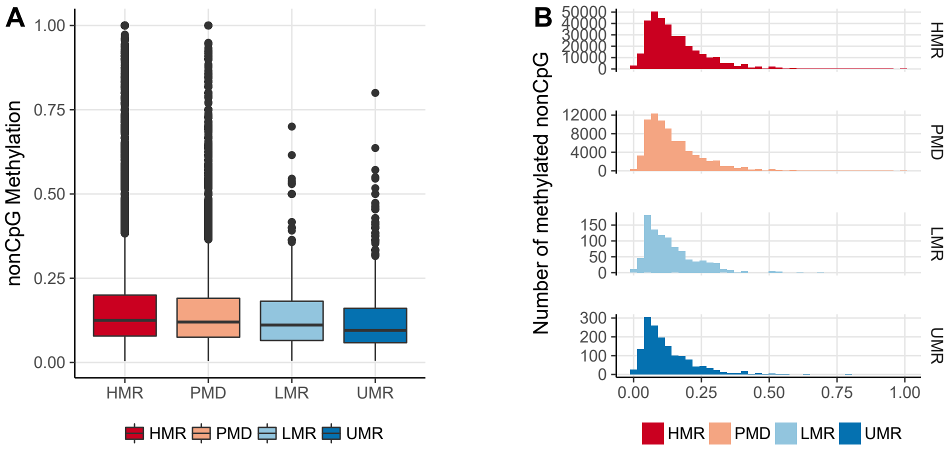


Fig. 5.45: Methylation level and distribution of methylated nonCpG in FMRs, PMDs LMRs and UMRs. Methylation level (A). nonCpG methylation distribution in FMRs, PMDs, LMRs and UMRs (B)

.

Bibliography

[1] Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[2] Arthur D Riggs. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[3] Déborah Bourc'his and Timothy H Bestor. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking dnmt3l. *Nature*, 431(7004):96, 2004.

[4] En Li, Timothy H Bestor, and Rudolf Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.

[5] Bernard H Ramsahoye, Detlev Biniszkiewicz, Frank Lyko, Victoria Clark, Adrian P Bird, and Rudolf Jaenisch. Non-cpg methylation is prevalent in embryonic stem cells and may be mediated by dna methyltransferase 3a. *Proceedings of the National Academy of Sciences*, 97(10):5237–5242, 2000.

[6] Michael J Ziller, Fabian Müller, Jing Liao, Yingying Zhang, Hongcang Gu, Christoph Bock, Patrick Boyle, Charles B Epstein, Bradley E Bernstein, Thomas Lengauer, et al. Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. *PLoS genetics*, 7(12):e1002389, 2011.

[7] Ryan Lister, Mattia Pelizzola, Robert H Dowen, R David Hawkins, Gary Hon, Julian Tonti-Filippini, Joseph R Nery, Leonard Lee, Zhen Ye, Que-Minh Ngo, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *nature*, 462(7271):315, 2009.

[8] Heinrich Leonhardt, Andrea W Page, Heinz-Ulrich Weier, and Timothy H Bestor. A targeting sequence directs dna methyltransferase to sites of dna replication in mammalian nuclei. *Cell*, 71(5):865–873, 1992.

[9] Linda S-H Chuang, Hang-In Ian, Tong-Wey Koh, Huck-Hui Ng, Guoliang Xu, and Benjamin FL Li. Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1. *Science*, 277(5334):1996–2000, 1997.

[10] Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.

[11] Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908–912, 2007.

[12] Kyohei Arita, Mariko Ariyoshi, Hidehito Tochio, Yusuke Nakamura, and Masahiro Shirakawa. Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. *Nature*, 455(7214):818, 2008.

[13] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 279(46):48350–48359, 2004.

[14] Masaki Okano, Shaoping Xie, and En Li. Cloning and characterization of a family of novel mammalian dna (cytosine-5) methyltransferases. *Nature genetics*, 19(3):219, 1998.

[15] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[16] Daniela Meilinger, Karin Fellinger, Sebastian Bultmann, Ulrich Rothbauer, Ian Marc Bonapace, Wolfgang EF Klinkert, Fabio Spada, and Heinrich Leonhardt. Np95 interacts with de novo dna methyltransferases, dnmt3a and dnmt3b, and mediates epigenetic silencing of the viral cmv promoter in embryonic stem cells. *EMBO reports*, 10(11):1259–1264, 2009.

[17] Gangning Liang, Matilda F Chan, Yoshitaka Tomigahara, Yvonne C Tsai, Felicidad A Gonzales, En Li, Peter W Laird, and Peter A Jones. Cooperativity between dna methyltransferases in the maintenance methylation of repetitive elements. *Molecular and cellular biology*, 22(2):480–491, 2002.

[18] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[19] Ryoichi Ono, Tomohiko Taki, Takeshi Taketani, Masafumi Taniwaki, Hajime Kobayashi, and Yasuhide Hayashi. Lcx, leukemia-associated protein with a cxxc domain, is fused to mll in acute myeloid leukemia with trilineage dysplasia having t (10; 11)(q22; q23). *Cancer research*, 62(14):4075–4080, 2002.

[20] RB Lorsbach, J Moore, S Mathew, SC Raimondi, ST Mukatira, and JR Downing. Tet1, a member of a novel protein family, is fused to mll in acute myeloid leukemia containing the t (10; 11)(q22; q23). *Leukemia*, 17(3):637, 2003.

[21] Lakshminarayan M Iyer, Mamta Tahiliani, Anjana Rao, and L Aravind. Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. *Cell cycle*, 8(11):1698–1710, 2009.

[22] Mamta Tahiliani, Kian Peng Koh, Yinghua Shen, William A Pastor, Hozefa Bandukwala, Yevgeny Brudno, Suneet Agarwal, Lakshminarayan M Iyer, David R Liu, L Aravind, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, 324(5929):930–935, 2009.

[23] Shinsuke Ito, Li Shen, Qing Dai, Susan C Wu, Leonard B Collins, James A Swenberg, Chuan He, and Yi Zhang. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, 333(6047):1300–1303, 2011.

[24] Daniel Globisch, Martin Münzel, Markus Müller, Stylianos Michalakis, Mirko Wagner, Susanne Koch, Tobias Brückl, Martin Biel, and Thomas Carell. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one*, 5(12):e15367, 2010.

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

[26] Aleksandra Szwagierczak, Sebastian Bultmann, Christine S Schmidt, Fabio Spada, and Heinrich Leonhardt. Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic dna. *Nucleic acids research*, 38(19):e181–e181, 2010.

[27] Martin Bachman, Santiago Uribe-Lewis, Xiaoping Yang, Michael Williams, Adele Murrell, and Shankar Balasubramanian. 5-hydroxymethylcytosine is a predominantly stable dna modification. *Nature chemistry*, 6(12):1049, 2014.

[28] Eun-Ang Raiber, Pierre Murat, Dimitri Y Chirgadze, Dario Beraldi, Ben F Luisi, and Shankar Balasubramanian. 5-formylcytosine alters the structure of the dna double helix. *Nature Structural and Molecular Biology*, 22(1):44, 2015.

[29] Matthew W Kellinger, Chun-Xiao Song, Jenny Chong, Xing-Yu Lu, Chuan He, and Dong Wang. 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of rna polymerase ii transcription. *Nature Structural and Molecular Biology*, 19(8):831, 2012.

[30] Hideharu Hashimoto, Yiwei Liu, Anup K Upadhyay, Yanqi Chang, Shelley B Howerton, Paula M Vertino, Xing Zhang, and Xiaodong Cheng. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research*, 40(11):4841–4849, 2012.

[31] Victoria Valinluck and Lawrence C Sowers. Endogenous cytosine damage products alter the site selectivity of human dna maintenance methyltransferase dnmt1. *Cancer research*, 67(3):946–950, 2007.

[32] Debin Ji, Krystal Lin, Jikui Song, and Yinsheng Wang. Effects of tet-induced oxidation products of 5-methylcytosine on dnmt1-and dnmt3a-mediated cytosine methylation. *Molecular bioSystems*, 10(7):1749–1752, 2014.

[33] Yu-Fei He, Bin-Zhong Li, Zheng Li, Peng Liu, Yang Wang, Qingyu Tang, Jianping Ding, Yingying Jia, Zhangcheng Chen, Lin Li, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, 333(6047):1303–1307, 2011.

[34] Atanu Maiti and Alexander C Drohat. Thymine dna glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine potential implications for active demethylation of cpg sites. *Journal of Biological Chemistry*, 286(41):35334–35338, 2011.

[35] Zachary D Smith, Michelle M Chan, Tarjei S Mikkelsen, Hongcang Gu, Andreas Gnirke, Aviv Regev, and Alexander Meissner. A unique regulatory phase of dna methylation in the early mammalian embryo. *Nature*, 484(7394):339, 2012.

[36] J Oswald, S Engemann, N Lane, W Mayer, A Olek, R Fundele, W Dean, W Reik, and J Walter. Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478, 2000.

[37] Petra Hajkova, Sean J Jeffries, Caroline Lee, Nigel Miller, Stephen P Jackson, and M Azim Surani. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science*, 329(5987):78–82, 2010.

[38] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[39] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[40] Marius Walter, Aurélie Teissandier, Raquel Pérez-Palacios, and Déborah Bourc'his. An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. *Elife*, 5, 2016.

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

[42] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[43] Pascal Giehr and Jörn Walter. Hairpin bisulfite sequencing: Synchronous methylation analysis on complementary dna strands of individual chromosomes. In *DNA Methylation Protocols*, pages 573–586. Springer, 2018.

[44] Lei Zhao, Ming-an Sun, Zejuan Li, Xue Bai, Miao Yu, Min Wang, Liji Liang, Xiaojian Shao, Stephen Arnovitz, Qianfei Wang, et al. The dynamics of dna methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome research*, 24(8):1296–1307, 2014.

[45] Michael J Booth, Miguel R Branco, Gabriella Ficz, David Oxley, Felix Krueger, Wolf Reik, and Shankar Balasubramanian. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science*, 336(6083):934–937, 2012.

[46] Charalampos Kyriakopoulos, Pascal Giehr, and Verena Wolf. H (o) ta: estimation of dna methylation and hydroxylation levels and efficiencies from time course data. *Bioinformatics*, 33(11):1733–1734, 2017.

[47] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

[48] Jacob Porter, Ming-an Sun, Hehuang Xie, and Liqing Zhang. Investigating bisulfite short-read mapping failure with hairpin bisulfite sequencing data. *BMC genomics*, 16(11):S2, 2015.

[49] Babraham Bioinformatics - Trim Galore!

[50] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[51] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The gem mapper: fast, accurate and versatile alignment by filtration. *Nature methods*, 9(12):1185, 2012.

[52] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[53] Leo Goodstadt. Ruffus: a lightweight python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779, 2010.

[54] Samuel L Braunstein. How large a sample is needed for the maximum likelihood estimator to be approximately gaussian? *Journal of Physics A: Mathematical and General*, 25(13):3813, 1992.

[55] Scott J Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[56] Ronald Schoenberg. Constrained maximum likelihood. *Computational Economics*, 10(3):251–266, 1997.

[57] Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. Identification of active regulatory regions from dna methylation data. *Nucleic acids research*, 41(16):e155–e155, 2013.

[58] Cath Tyner, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Christopher Eisenhart, Clayton M Fischer, David Gibson, Jairo Navarro Gonzalez, Luvina Guruvadoo, et al. The ucsc genome browser database: 2017 update. *Nucleic acids research*, 45(D1):D626–D634, 2016.

[59] Nathan C Sheffield and Christoph Bock. Lola: enrichment analysis for genomic region sets and regulatory elements in r and bioconductor. *Bioinformatics*, 32(4):587–589, 2015.

[60] Ferdinand von Meyenn, Mario Iurlaro, Ehsan Habibi, Ning Qing Liu, Ali Salehzadeh-Yazdi, Fátima Santos, Edoardo Petrini, Inês Milagre, Miao Yu, Zhenqing Xie, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861, 2016.

[61] Xiaoli Liu, Qinqin Gao, Pishun Li, Qian Zhao, Jiqin Zhang, Jiwen Li, Haruhiko Koseki, and Jiemin Wong. Uhrf1 targets dnmt1 for dna methylation through cooperative binding of hemi-methylated dna and methylated h3k9. *Nature communications*, 4:1563, 2013.

[62] Hao Wu, Ana C D'alessio, Shinsuke Ito, Kai Xia, Zhibin Wang, Kairong Cui, Keji Zhao, Yi Eve Sun, and Yi Zhang. Dual functions of tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473(7347):389, 2011.

[63] Yufei Xu, Feizhen Wu, Li Tan, Lingchun Kong, Lijun Xiong, Jie Deng, Andrew J Barbera, Lijuan Zheng, Haikuo Zhang, Stephen Huang, et al. Genome-wide regulation of 5hmc, 5mc, and gene expression by tet1 hydroxylase in mouse embryonic stem cells. *Molecular cell*, 42(4):451–464, 2011.

[64] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768, 2011.

[65] Christoph Bock, Isabel Beerman, Wen-Hui Lien, Zachary D Smith, Hongcang Gu, Patrick Boyle, Andreas Gnirke, Elaine Fuchs, Derrick J Rossi, and Alexander Meissner. Dna methylation dynamics during in vivo differentiation of blood and skin stem cells. *Molecular cell*, 47(4):633–647, 2012.

[66] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412, 2016.

[67] Michael B Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J Oakeley, Dimos Gaidatzis, et al. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 2011.

[68] Ichiro Hiratani, Tyrone Ryba, Mari Itoh, Tomoki Yokochi, Michaela Schwaiger, Chia-Wei Chang, Yung Lyou, Tim M Townes, Dirk Schübeler, and David M Gilbert. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS biology*, 6(10):e245, 2008.

[69] Mahesh Kumar and Nitin R Patel. Clustering data with measurement errors. *Computational Statistics & Data Analysis*, 51(12):6084–6101, 2007.

[70] Hitoshi Niwa, Tom Burdon, Ian Chambers, and Austin Smith. Self-renewal of pluripotent embryonic stem cells is mediated via activation of stat3. *Genes & development*, 12(13):2048–2060, 1998.

[71] Michiel Vermeulen, Klaas W Mulder, Sergei Denissov, WWM Pim Pijnappel, Frederik MA van Schaik, Radhika A Varier, Marijke PA Baltissen, Henk G Stunnenberg, Matthias Mann, and H Th Marc Timmers. Selective anchoring of tfiid to nucleosomes by trimethylation of histone h3 lysine 4. *Cell*, 131(1):58–69, 2007.

[72] Hao Jiang, Abhijit Shukla, Xiaoling Wang, Wei-yi Chen, Bradley E Bernstein, and Robert G Roeder. Role for dpy-30 in es cell-fate specification by regulation of h3k4 methylation within bivalent domains. *Cell*, 144(4):513–525, 2011.

[73] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520(7546):243, 2015.

[74] Tianpeng Gu, Xueqiu Lin, Sean M Cullen, Min Luo, Mira Jeong, Marcos Estecio, Jianjun Shen, Swanand Hardikar, Deqiang Sun, Jianzhong Su, et al. Dnmt3a and tet1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome biology*, 19(1):88, 2018.

[75] Hao Wu, Ana C D'Alessio, Shinsuke Ito, Zhibin Wang, Kairong Cui, Keji Zhao, Yi Eve Sun, and Yi Zhang. Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Genes & development*, 25(7):679–684, 2011.

[76] Hume Stroud, Suhua Feng, Shannon Morey Kinney, Sriharsa Pradhan, and Steven E Jacobsen. 5-hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome biology*, 12(6):R54, 2011.

[77] Kevin C Johnson, E Andres Houseman, Jessica E King, Katharine M Von Herrmann, Camilo E Fadul, and Brock C Christensen. 5-hydroxymethylcytosine localizes to enhancer elements and is associated with survival in glioblastoma patients. *Nature communications*, 7:13177, 2016.

[78] Michael J Booth, Giovanni Marsico, Martin Bachman, Dario Beraldi, and Shankar Balasubramanian. Quantitative sequencing of 5-formylcytosine in dna at single-base resolution. *Nature chemistry*, 6(5):435, 2014.

[79] Chenxu Zhu, Yun Gao, Hongshan Guo, Bo Xia, Jinghui Song, Xinglong Wu, Hu Zeng, Kehkooi Kee, Fuchou Tang, and Chengqi Yi. Single-cell 5-formylcytosine landscapes of mammalian early embryos and escs at single-base resolution. *Cell Stem Cell*, 20(5):720–731, 2017.

[80] Peter Beerli. Comparison of bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3):341–345, 2005.

[81] Daniel McNeish. On using bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5):750–773, 2016.

[82] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

[83] David Shen, Zaizai Lu, et al. Computation of correlation coefficient and its confidence interval in sas. *SUGI: Paper*, pages 170–31, 2006.

6. A STOCHASTIC MODEL FOR THE FORMATION OF SPATIAL METHYLATION PATTERNS

The content of chapter 6 has been published as:

Lück, A.*, Giehr, P.†, Walter, J.†, & Wolf, V.* (2017, September). A Stochastic Model for the Formation of Spatial Methylation Patterns. In International Conference on Computational Methods in Systems Biology (pp. 160-178). Springer, Cham.

AUTHOR CONTRIBUTIONS

Alexander Lück: Design and development of the model. Statistical analysis of model output. Authoring of the manuscript including abstract, as well as the sections 6.1 introduction, 6.2 preliminaries (including the generation of figure 6.2), 6.3 model (including the generation of figure 6.3 to 6.4, as well as all formulas), 6.4 results (including the generation of figures 6.5 to 6.7, as well as tables 6.1 and 6.2), 6.5 related work and 6.6 conclusion.

Pascal Giehr: Hairpin sequencing data preparation. Consulting on model design. Revision of the manuscript i.e. changes in formulation/wording and structure of the text mainly restricted to the biological background and interpretation of the presented results. Rewriting parts of section 6.1 introduction, and generation of figure 6.1, as well as section 6.4 results, mainly paragraphs 1 and 3, an section 6.6 concerning biological background and interpretation.

Prof. Dr. Jörn Walter: Supervision. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Prof. Dr. Verena Wolf: Supervision. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

*Department of Computer Science, Saarland University, D-66123 Saarbrücken, Germany

†Department of Biological Sciences, Saarland University, D-66123 Saarbrücken, Germany

Abstract

DNA methylation is an epigenetic mechanism whose important role in development has been widely recognized. This epigenetic modification results in heritable changes in gene expression not encoded by the DNA sequence. The underlying mechanisms controlling DNA methylation are only partly understood and recently different mechanistic models of enzyme activities responsible for DNA methylation have been proposed. Here we extend existing Hidden Markov Models (HMMs) for DNA methylation by describing the occurrence of spatial methylation patterns over time and propose several models with different neighborhood dependencies. We perform numerical analysis of the HMMs applied to bisulfite sequencing measurements and accurately predict wild type data. In addition, we find evidence that the enzymes' activities depend on the left 5' neighborhood but not on the right 3' neighborhood.

6.1 Introduction

The DNA code of an organism determines its appearance and behavior by encoding protein sequences. In addition, there is a multitude of additional mechanisms to control and regulate the ways in which the DNA is packed and processed in the cell and thus determine the fate of a cell. One of these mechanisms in cells is DNA methylation, which is an epigenetic modification that occurs at the cytosine (C) bases of eukaryotic DNA. Cytosines are converted to 5-methylcytosine (5mC) by DNA methyltransferase (Dnmt) enzymes. The neighboring nucleotide of a methylated cytosine is usually guanine (G) and together with the GC-pair on the opposite strand, a common pattern is that two methylated cytosines are located diagonally to each other on opposing DNA strands. DNA methylation at CpG dinucleotides is known to control and mediate gene expression and is therefore essential for cell differentiation and embryonic development. In human somatic cells, approximately 70-80% of the cytosine nucleotides in CpG dyads are methylated on both strands and methylation near gene promoters varies considerably depending on the cell type. Methylation of promoters often correlates with low or no transcription [1] and can be used as a predictor of gene expression [2]. Also significant differences in overall and specific methylation levels exist between different tissue types and between normal cells and cancer cells from the same tissue. However, the exact mechanism which leads to a methylation of a specific CpG and the formation of distinct methylation patterns at certain genomic regions is still not fully understood. Recently proposed measurement techniques based on hairpin bisulfite sequencing (BS-seq) allow to determine on both DNA strands the level of 5mC at individual CpGs dyads [3]. Based on a small hidden Markov model, the probabilities of the different states of a CpG can be accurately estimated (assuming that enough samples per CpG are provided) [4, 5].

Mechanistic models for the activity of the different Dnmts usually distinguish de novo activities, i.e., adding methyl groups at cytosines independent of the methylation state of the opposite strand, and maintenance activities, which refers to the copying of methylation from an existing DNA strand to its newly synthesized partner (containing no methylation) after replication [6, 7]. Hence, maintenance methylation is responsible for re-establishment of the same DNA methylation pattern before and after cell replication. A common hypothesis is that the copying of DNA methylation patterns after replication is performed by Dnmt1, an enzyme that shows a preference for hemimethylated CpG sites (only one strand is methylated) as they appear after DNA replication. Moreover, studies have shown that Dnmt1 is highly processive and able to methylate long sequences of hemimethylated CpGs without dissociation from the target DNA strand [6]. However, an exact transmission of the methylation information to the next cellular generation is not guaranteed. The enzymes Dnmt3a and Dnmt3b show equal activities on hemi- and unmethylated DNA and are mainly responsible for de novo methylation, i.e., methylation without any specific preference for the current state of the CpG (hemi- or unmethylated) [7]. However, by now evidence exists that the activity of the different enzymes is not that exclusive, i.e., Dnmt1 shows to a certain degree also de novo and Dnmt3a/b maintenance methylation activity [8]. The way how methyltransferases interact with the DNA and introduce CpG methylation was investigated in many *in vitro* studies. Basically, one can distinguish between two mechanisms. A distributive one, where the enzyme periodically binds and dissociates from the DNA, leaping more or less randomly from one CpG to another and a processive one in which the enzyme migrates along the DNA without detachment from the DNA [9, 10, 11], as illustrated in Fig. 6.1. Note that for Dnmt1, for instance, it is reasonable to assume that it is processive in 5' to 3' direction since it is linked to the DNA replication machinery. In particular for the Dnmt3's different hypotheses about the processivity and neighborhood dependence exist [12, 13], but the detailed mechanisms remain elusive.

Several models that describe the dynamics of the formation of methylation patterns have been proposed. In the seminal paper of Otto and Walbot, a dynamical model was proposed that assumed independent methylation events for a single CpG. The main idea was to track the frequencies of fully, hemi- and unmethylated CpGs during several cell generations [14]. Later, refined models allowed to distinguish between maintenance and de novo methylation on the parent and daughter strands [15, 16]. More sophisticated extensions of the original model of Otto and Walbot models have been successfully used to predict *in vivo* data still assuming a neighbor-independent methylation process for a single CpG site [8, 17]. However, measurements indicate that methylation events at a single CpG may depend on the methylation state of neighboring CpGs, which is not captured by these models.

Here, we follow the dynamical HMM approach proposed in [8] where knockout data

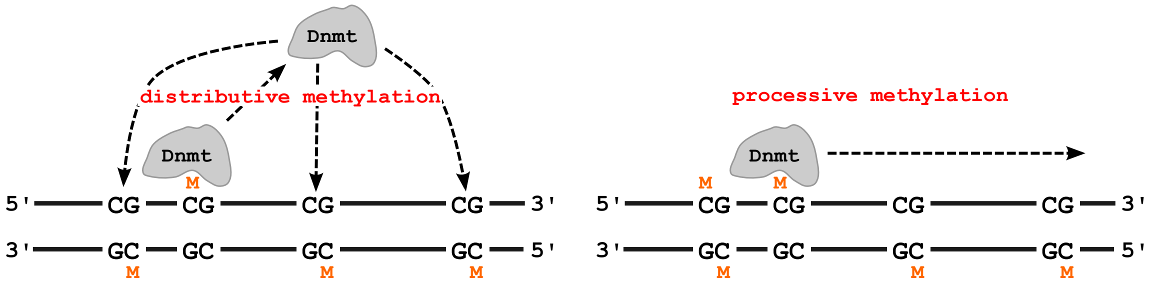


Fig. 6.1: Dnmts can methylate DNA in a distributive manner, "jumping" randomly from one CpG to another or in a processive way where the enzyme starts at one CpG and slides in 5' to 3' direction over the DNA.

was used to train a model that accurately predicts wild-type methylation levels for BS-seq data of repetitive elements from mouse embryonic stem cells. We extend this model by describing the methylation state of several CpGs instead of a single CpG and use similar dependency parameters as introduced in [18]. More specifically, we design different models by combining the activities of the two types of Dnmts and test for both, maintenance and de novo methylation the hypotheses illustrated in Fig 6.1. The models vary according to the order in which the enzymes act, whether they perform methylation in a processive manner or not, and how much their action depends on the left/right CpG neighbor. We use the same BS-seq data as in [8], i.e. data where Dnmt1 or Dnmt3a/b was knocked out (KO) and learn the parameters of the different models. Then, similar as in [8], we predict the behavior of the measured wild-type (WT), in which both types of enzymes are active, by designing a combined model that describes the activity of both enzymes and compare the results to the WT data.

We found that all proposed models show a similar behavior in terms of prediction quality such that no model can be declared as the best fit. However, our results indicate that Dnmt1 works independently of the methylation state of its neighborhood, which is in accordance to the current hypothesis that Dnmt1 is linked to the replication machinery and copies the methylation state on the opposite strand. On the other hand, Dnmt3a/b shows a dependency to the left but no dependency to the right, which supports hypotheses of processive or cooperative behavior.

6.2 Preliminaries

Consider a sequence of L neighboring CpG dyads[*], which is represented as a lattice of length L and width two (for the two strands). Each cytosine in the lattice can either be methylated or not, leading to four possible states at each position l:

[*] The exact nucleotide distance between two neighboring dyads is not considered here, but we assume that this distance is small. For the BS-seq data that we consider, the average distance between two CpGs is 14 bp and the maximal distance is 46 bp.

- *State 0*: Both sites are not methylated.
- *State 1*: The cytosine on the upper strand is methylated, the lower one not.
- *State 2*: The cytosine on the lower strand is methylated, the upper one not.
- *State 3*: Both cytosines are methylated.

A sequence of four CpGs, each of which is in one of the four possible states, is shown in Fig. 6.2.

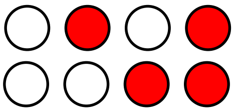


Fig. 6.2: A lattice of length $L = 4$ containing all possible states 0, 1, 2 and 3, forming the pattern 0123.

For a system of length L there are in total 4^L possibilities to combine the states of individual CpGs. These combinations are called *patterns* in the following. A pattern is denoted by a concatenation of states, e.g. 321, 0123 or 33221.

In order to represent the pattern distribution as a vector it is necessary to uniquely assign a reference number to each pattern. A pattern can be perceived as a number in the tetral system, such that converting to the decimal system leads to a unique reference number. After the conversion an additional 1 is added in order to start the referencing at 1 instead of 0.

Examples for $L = 3$:

$$\begin{aligned} 000 &\quad\longrightarrow\quad 1\ (= 0 + 1) \\ 123 &\quad\longrightarrow\quad 28\ (= 27 + 1) \\ 333 &\quad\longrightarrow\quad 64\ (= 63 + 1) \end{aligned}$$

This reference number then corresponds to the position of the pattern in the respective distribution vector.

6.3 Model

We describe the state of a sequence of L CpGs by a discrete-time Markov chain with pattern distribution $\pi(t)$, i.e., the probability of each of the 4^L patterns after t cell divisions. For the initial distribution $\pi(0)$, we use the distribution measured in the wild-type when the cells are in equilibrium. Note that other initial conditions gave very similar results, i.e., the choice of the initial distribution does not significantly affect the results. The reason is that also the KO data is measured after a relatively high number of cell divisions where the cells are almost in equilibrium. Transitions between patterns are triggered by different processes: First due to *cell division* the methylation on one strand is kept as it is (e.g. the upper strand), whereas the newly synthesized strand (the new lower strand) does not contain any methyl group. Afterwards, methylation is added due to different mechanisms. On the newly synthesized strand a site can be methylated if the cytosine at the opposite strand is already methylated (*maintenance*). It is widely accepted that maintenance in form of Dnmt1 is linked to the replication machinery and thus occurs during/directly after the synthesis of the new strand. Furthermore, CpGs on both strands can be methylated independent of the methylation state of the opposite site (*de novo*).

The transition matrix P is defined by composition of matrices for cell division, maintenance and de novo methylation of each site.

6.3.1 Cell Division

Depending on which daughter cell is considered after cell replication, the upper ($s = 1$) or lower ($s = 2$) strand is the parental one after cell division. Then, the new pattern can be obtained by applying the following state replacements:

$$s = 1 : \begin{cases} 0 & \longrightarrow & 0 \\ 1 & \longrightarrow & 1 \\ 2 & \longrightarrow & 0 \\ 3 & \longrightarrow & 1 \end{cases} \qquad s = 2 : \begin{cases} 0 & \longrightarrow & 0 \\ 1 & \longrightarrow & 0 \\ 2 & \longrightarrow & 2 \\ 3 & \longrightarrow & 2 \end{cases} \tag{6.1}$$

Given some initial pattern with reference number i, applying the transformation (6.1) to each of the L positions leads to a new pattern with reference number j (notation: $i \overset{(6.1)}{\rightsquigarrow} j$). The corresponding transition matrix $D_s \in \{0,1\}^{4^L \times 4^L}$ has the form

$$D_s(i,j) = \begin{cases} 1, & \text{if } i \overset{(6.1)}{\rightsquigarrow} j, \\ 0, & \text{else.} \end{cases} \tag{6.2}$$

6.3.2 Maintenance and De Novo Methylation

For maintenance and de novo methylation, the single site transition matrices are built according to the following rules:

Consider at first the (non-boundary) site $l = 2, \ldots, L-1$ and its left and right neighbor $l-1$ and $l+1$ respectively. The remaining sites do not change and do not affect the transition. The probabilities of the different types of transitions in Fig. 6.3 have the form

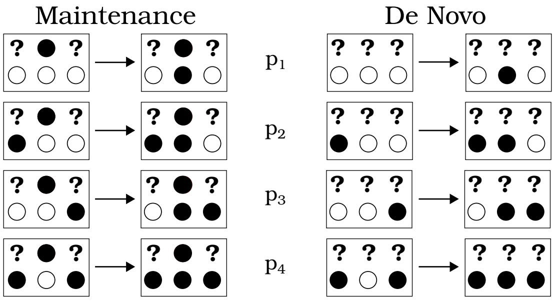


Fig. 6.3: Possible maintenance and de novo transitions depicted for the lower strand, where ∘ denotes an unmethylated, • a methylated site and **?** a site where the methylation state does not matter. Note that the same transitions can occur on the upper strand.

$$p_1 = 0.5\cdot(\psi_L + \psi_R)x, \tag{6.3}$$

$$p_2 = 0.5\cdot(\psi_L + \psi_R)x + 0.5\cdot(1 - \psi_L), \tag{6.4}$$

$$p_3 = 0.5\cdot(\psi_L + \psi_R)x + 0.5\cdot(1 - \psi_R), \tag{6.5}$$

$$p_4 = 1 - 0.5\cdot(\psi_L + \psi_R)(1 - x), \tag{6.6}$$

where $x = \mu$ is the maintenance probability, $x = \tau$ is the de novo probability and $\psi_L,\ \psi_R \in [0, 1]$ the dependency parameters for the left and right neighbor.
A dependency value of $\psi_i = 1$ corresponds to a total independence on the neighbor whereas $\psi_i = 0$ leads to a total dependence. Hence, μ and τ can be interpreted as the probability of maintenance and de novo methylation of a single cytosine between two cell divisions assuming independence from neighboring CpGs. Moreover, all CpGs that are part of the considered window of the DNA have the same value for the parameters μ, τ, ψ_L, and ψ_R, since in earlier experiments only very small differences have been found between the methylation efficiencies of nearby CpGs [8].

In order to understand the form of the transition probabilities consider at first a case with only one neighbor. The probabilities then have the form ψx if the neighbor is unmethylated and $1 - \psi(1 - x)$ if the neighbor is methylated. Note that both forms evaluate to x for $\psi = 1$, meaning that a site is methylated with probability x, independent of its neighbor. For $\psi = 0$ the probabilities become 0 and 1, meaning that if there is no methylated neighbor the site cannot be methylated or will be methylated for sure if there is a methylated neighbor respectively.
The probabilities for two neighbors are obtained by a linear combination of the one neighbor cases, with ψ_L for the left and ψ_R for the right neighbor, and an additional weight of 0.5 to normalize the probability.
The same considerations also apply to the boundary sites however there is no way of knowing the methylation states outside the boundaries (denoted by ?). Therefore instead

of a concrete methylation state (○ for unmethylated, ● for methylated site) the average methylation density ρ is used to compute the transition probabilities at the boundaries (depicted here for de novo):

$$?\circ\circ \rightarrow ?\bullet\circ \qquad \tilde{p}_1 = (1-\rho)\cdot p_1 + \rho\cdot p_2, \tag{6.7}$$

$$?\circ\bullet \rightarrow ?\bullet\bullet \qquad \tilde{p}_2 = (1-\rho)\cdot p_3 + \rho\cdot p_4, \tag{6.8}$$

$$\circ\circ ? \rightarrow \circ\bullet ? \qquad \tilde{p}_3 = (1-\rho)\cdot p_1 + \rho\cdot p_3, \tag{6.9}$$

$$\bullet\circ ? \rightarrow \bullet\bullet ? \qquad \tilde{p}_4 = (1-\rho)\cdot p_2 + \rho\cdot p_4. \tag{6.10}$$

Note that the same considerations hold for maintenance at the boundaries if the opposite site of the boundary site is already methylated.

For each position l, there are four transition matrices: two for maintenance and two for de novo, namely one for the upper and one for the lower strand in each process. In order to construct these matrices consider the three positions $l-1$, l and $l+1$, where the transition happens at position l. Only the transitions depicted in Fig. 6.3 can occur. Furthermore the transitions are unique, i.e. for a given reference number i the new reference number j is uniquely determined. For patterns not depicted in Fig. 6.3 no transition can occur, i.e. the reference number does not change.

The matrix describing a maintenance event at position l and strand s has the form

$$M_s^{(l)}(i,j) = \begin{cases} 1, & \text{if } i = j \text{ and } \nexists j' : \ i \rightsquigarrow j', \\ 1-p, & \text{if } i = j \text{ and } \exists j' : \ i \rightsquigarrow j', \\ p, & \text{if } i \neq j \text{ and } i \rightsquigarrow j, \\ 0, & \text{else}, \end{cases} \tag{6.11}$$

where the probability p is given by one of the Eqs. (6.3)-(6.10) that describes the corresponding case and $x = \mu$. Note that $M_s^{(l)}$ depends on s and l since it describes a single transition from pattern i to pattern j, which occurs on a particular strand and at a particular location with probability p. We define matrices $T_s^{(l)}$ for de novo methylation according to the same rules except that $x = \tau$ and the possible transitions are as in Fig. 6.3, right.

The advantage of defining the matrices position- and process-wise is that different models can be realized by changing the order of multiplication of these matrices.

It is important to note that 5mC can be further modified by oxidation to 5-hydroxymethyl- (5hmC), 5-formyl- (5fC) and 5-carboxyl cytosine(5caC) by Tet enzymes. These modifications are involved in the removal of 5mC from the DNA and can potentially interfere with methylation events. However, our data does not capture these modifications and therefore we are not able to consider these modifications in our model.

6.3.3 Combination of Transition Matrices

For all subsequent models it is assumed that first of all cell division happens and maintenance methylation only occurs on the newly synthesized strand given by s, whereas de novo methylation happens on both strands. Given the mechanisms in Fig. 6.1, the two different kinds of methylation events, and the two types of enzymes, there are several possibilities to combine the transition matrices. We consider the following four models, which we found most reasonable based on the current state of research in DNA methylation:

a. first processive maintenance and then processive de novo methylation

$$P_s = \prod_{l_1=1}^{L} M_s^{(l_1)} \prod_{l_2=1}^{L} T_1^{(l_2)} \prod_{l_3=1}^{L} T_2^{(l_3)}, \tag{6.12}$$

b. first processive maintenance and then de novo in arbitrary order

$$P_s = \frac{1}{(L!)^2} \prod_{l_1=1}^{L} M_s^{(l_1)} \left(\sum_{\sigma_1 \in S_L} \prod_{l_2=1}^{L} T_1^{(\sigma_1(l_2))} \right) \left(\sum_{\sigma_2 \in S_L} \prod_{l_3=1}^{L} T_2^{(\sigma_2(l_3))} \right), \tag{6.13}$$

c. maintenance and de novo at one position, processive

$$P_s = \prod_{l=1}^{L} M_s^{(l)} T_1^{(l)} T_2^{(l)}, \tag{6.14}$$

d. maintenance and de novo at one position, arbitrary order

$$P_s = \frac{1}{L!} \sum_{\sigma \in S_L} \prod_{l=1}^{L} M_s^{(\sigma(l))} T_1^{(\sigma(l))} T_2^{(\sigma(l))}, \tag{6.15}$$

where S_L is the set of all possible permutations for the numbers $1, \ldots, L$.
Note that the de novo events on both strands are independent, i.e. the de novo events on the upper strand do not influence the de novo events on the lower strand and vice versa, such that $[T_1^{(l)}, T_2^{(l')}] = 0$ independent of ψ_i[†]. Obviously it is important whether maintenance or de novo happens first, since the transition probabilities and the transitions themselves depend on the actual pattern. Furthermore in the case $\psi_i < 1$ (dependency on right and/or left neighbor) the order of the transitions on a strand matters, i.e. $[M_s^{(l)}, M_s^{(l')}] \neq 0$ and $[T_s^{(l)}, T_s^{(l')}] \neq 0$ for $l \neq l'$. The total transition matrix is then given by a combination of the cell division and maintenance/de novo matrices.

[†] $[A, B] = AB - BA$ is the commutator of the matrices A and B.

Recall that we consider two different types of Dnmts, i.e., Dnmt1 and Dnmt3a/b. If only one type of Dnmt is active (KO data) the matrix has the form

$$P = 0.5\cdot(D_1\cdot P_1 + D_2\cdot P_2) \tag{6.16}$$

and if all Dnmts are active (WT data)

$$P = 0.5\cdot(D_1\cdot P_1\cdot\tilde{P}_1 + D_2\cdot P_2\cdot\tilde{P}_2), \tag{6.17}$$

where P_s and $\tilde{P}_s$ have one of the forms (6.12)-(6.15). This leads to four different models for one active enzyme or 16 models for all active enzymes respectively. In the second case P_s represents the transitions caused by Dnmt1 and $\tilde{P}_s$ the transitions caused by Dnmt3a/b. Note that if $\psi_L = \psi_R = 1$ all models are the same within each case.

6.3.4 Conversion Errors

$$\begin{tikzcd}
u \arrow[d, "c"'] \arrow[dr, dashed, "1-c"' near start] & m \arrow[d, "d"] \arrow[dl, dashed, "1-d" near start] \\
T & C
\end{tikzcd}$$

Fig. 6.4: Conversions of the unobservable states u, m to observable states T, C with respective rates.

The actual methylation state of a C cannot be directly observed. During BS-seq, with high probability every unmethylated C (denoted by u) is converted into Thymine (T) and every 5mC (denoted by m) into C. However, conversion errors may occur and we define their probability as $1-c$ and $1-d$, respectively, as shown by the dashed arrows in Fig. 6.4. It is reasonable that these conversion errors occur independently and with approximately identical probability at each site and thus the error matrix for a single CpG takes the form

$$\Delta_1 = \begin{pmatrix} c^2 & c(1-c) & c(1-c) & (1-c)^2 \\ c(1-d) & cd & (1-c)(1-d) & d(1-c) \\ c(1-d) & (1-c)(1-d) & cd & d(1-c) \\ (1-d)^2 & d(1-d) & d(1-d) & d^2 \end{pmatrix}. \tag{6.18}$$

Due to the independency of the events this matrix can easily be generalized for systems with $L > 1$ by recursively using the Kronecker-product

$$\Delta_L = \Delta_1 \otimes \Delta_{L-1} \qquad \text{for } L \geq 2. \tag{6.19}$$

Hence, Δ_L gives the probability of observing a certain sequence of C and T nucleotides for each given unobservable methylation pattern. In order to compute the likelihood $\hat{\pi}$ of the observed BS-seq data, we therefore first compute the transient distribution $\pi(t)$ of the underlying Markov chain at the corresponding time instant[‡] t by solving

$$\pi(t) = \pi(0) \cdot P^t \tag{6.20}$$

and then multiply the distribution of the unobservable patterns with the error matrix.

$$\hat{\pi} = \pi(t) \cdot \Delta_L. \tag{6.21}$$

Note that this yields a hidden Markov model with emission probabilities Δ_L. In the following the values for c were chosen according to [8]. Since the value for d was not determined in [8], we measured the conversion rate $d = 0.94$ in an independent experiment under comparable conditions (data not shown).

6.3.5 Maximum Likelihood Estimator

In order to estimate the parameters $\theta = (\mu, \psi_L, \psi_R, \tau)$, we employ a Maximum (Log)Likelihood Estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta), \quad \ell(\theta) = \sum_{j=1}^{4^L} \log(\hat{\pi}_j(\theta)) \cdot N_j, \tag{6.22}$$

where $\hat{\pi}$ is the pattern distribution obtained from the numerical solution of (6.20) and (6.21) for a given time t and N_j is the number of occurrences of pattern j in the measured data. The parameters $\theta = \hat{\theta}$ are chosen in such a way that ℓ is maximized. Visual inspection of all two dimensional cuts of the likelihood landscapes showed only a single local maximum.

We employ the MLE twice in order to estimate the parameter vector $\hat{\theta}_1$ for Dnmt1 from the 3a/b DKO (double knockout) data and the vector $\hat{\theta}_{3a/b}$ for Dnmt3a/b from the Dnmt1 KO data, where transition matrix (6.16) is used. The corresponding time instants are $t = 26$ for the 3a/b DKO data and $t = 41$ for the 1KO data.

We approximate the standard deviations of the estimated parameters $\hat{\theta}$ as follows: Let $\mathcal{I}(\hat{\theta}) = \mathbb{E}[-\mathcal{H}(\hat{\theta})]$ be the expected Fisher information, with the Hessian $\mathcal{H}(\hat{\theta}) = \nabla\nabla^{\intercal}\ell(\hat{\theta})$. The inverse of the expected Fisher information is a lower bound for the covariance matrix of the MLE such that we can use the approximation $\sigma(\hat{\theta}) \approx \sqrt{\text{diag}(-\mathcal{H}(\hat{\theta}))}$.

A prediction for the wild-type can be computed by combining the estimated vectors such that in the model both types of enzymes are active. For this, we insert $\hat{\theta}_1$ in P_s and $\hat{\theta}_{3a/b}$

[‡] The number of cell divisions is estimated from the time of the measurement since these cells divide once every 24 hours.

in $\tilde{P}_s$ in (6.17) to obtain the transition matrix for the wild-type.

6.4 Results

For our analysis we focused at the single copy genes Afp (5 CpGs) and Tex13 (10 CpGs) as well as the repetitive elements IAP (intracisternal A particle) (6 CpGs), L1 (Long interspersed nuclear elements) (7 CpGs) and mSat (major satellite) (3 CpGs). Repetitive elements occur in multiple copies and are dispersed over the entire genome. Therefore they allow capturing an averaged, more general behavior of methylation dynamics. If a locus contains more than three CpGs, the analysis is done for all sets of three adjacent sites independently, in order to keep computation times short and memory requirements low. In the sequel, we mainly focus on the estimated dependency parameters ψ_L and ψ_R and on the prediction quality of the different models.

The estimates for all the available KO data and all suggested models obtained using the transition matrix in Eq. (6.16) are summarized as histograms in Fig. 6.5. Because of the different possibilities to combine the four different models in Eq. (6.12)-(6.15) and because of the different loci considered, in total there are 84 estimates for each KO data set. We plot the number of occurrences N of ψ_L (left) and ψ_R (right) in different ranges for both sorts of KO data (Dnmt1KO and Dnmt3a/b DKO).

The estimates of ψ_L spread over the whole interval $[0, 1]$ while in the case of ψ_R, nearly all estimates are larger than 0.99 and only in a few cases the dependency parameter is significantly smaller than 1. Hence, in most cases the methylation probabilities are independent of the right neighbor for both Dnmt1KO and Dnmt3a/b DKO. For ψ_L the dependency parameter in the Dnmt3a/b DKO case occurs more often close to 1, meaning that the transitions induced by Dnmt1 have little to no dependency on the left neighbor. On the other hand for Dnmt1KO the dependency parameter occurs more often at smaller values giving evidence that there is a dependency on the left neighbor for the activity of Dnmt3a/b. Note that all models show a similar behavior in terms of the dependency parameters for a given locus or position within a locus respectively, i.e. either $\psi_i \approx 1$ or $\psi_i < 1$ for all models. The difference between the behaviors at different loci and positions may be explained by explicitly including the distances between the CpGs and is planned as future work.

Since ψ_R is usually close to 1 a smaller model with only three parameters $\theta = (\mu, \psi, \tau)$ can be proposed, where ψ is a dependency parameter for the left neighbor. This model can either be obtained by fixing $\psi_R = 1$ in the original model and setting $\psi = \psi_L$ or by redefining the transition probabilities to ψx if the left neighbor is unmethylated and $1 - \psi(1 - x)$ if the left neighbor is methylated. In that case ψ and ψ_L are related via $\psi = 0.5(\psi_L + 1)$. Note that both versions yield the same results.

In order to check whether there is a significant difference in the original and the smaller model, we performed a Likelihood-ratio test with the null hypothesis that the smaller model is a special case of the original model. Since the original model with more parameters is always as least as good as the smaller model, our goal is to check in which cases the smaller model is sufficient. Indeed if ψ_R was estimated to be approximately 1 the Likelihood-ratio test indicates that the smaller model is sufficient (p-value ≈ 1). On the other hand, for the few cases where ψ_R differs significantly from 1 the original model has to be used (p-value < 0.01).

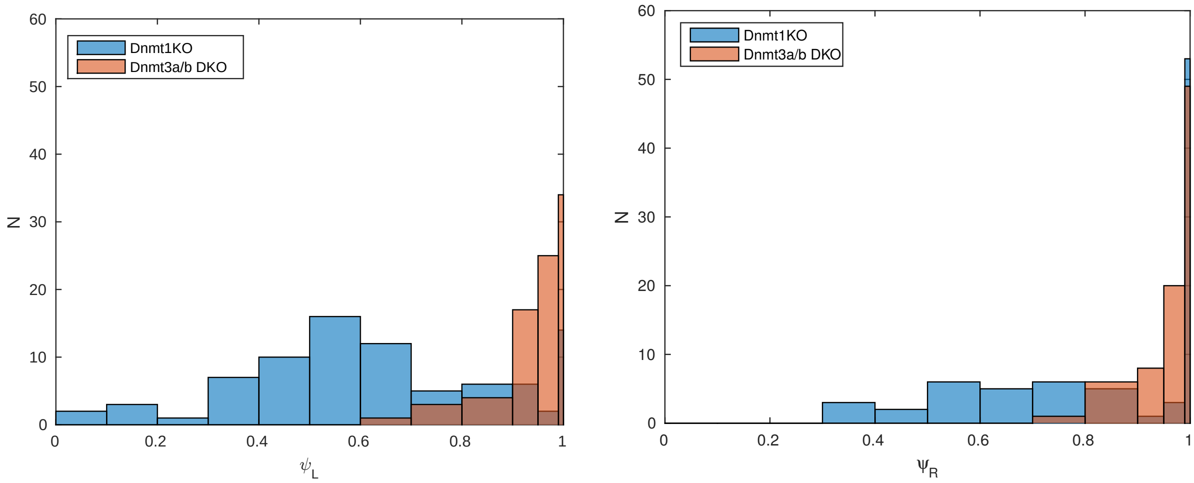


Fig. 6.5: Histograms for the estimated dependency parameters ψ_L and ψ_R for all sets of three adjacent CpGs in all loci and for all suggested models.

As a next step we used the estimated parameters from the KO data to predict the WT data. The models from Eqs.(6.12) - (6.15) are referred to as Models *1-4*. For the prediction, the notation (x, y) is used to refer to Model x for the Dnmt3a/b DKO (only Dnmt1 active) and Model y for the Dnmt1KO case (only Dnmt3a/b active). One instance of the prediction, for which Model 1 was used for both Dnmt1KO and Dnmt3a/b DKO, i.e. (1,1), are shown in Fig. 6.6. Note that all wild-type predictions yielded a very similar accuracy (see also Appendix). We list the corresponding estimations for the parameters for an example of a single copy gene (Afp) and a repetitive element (L1) in Tab. 6.1. While the standard deviation of the estimated parameters for µis always of the order 10^{-2} and for τ of order 10^{-3}, it is usually of oder 10^{-2} for ψ_i. Depending on the model, locus and position, standard deviations up to order 10^{-1} may occur for the dependency parameters in a few cases.

In Fig. 6.6 the predictions for the pattern distribution together with the WT pattern distribution and a prediction from the neighborhood independent model ($\psi_L = \psi_R = 1$) for all loci are shown in the main plot. As an inset the distributions are shown on a smaller scale to display small deviations. With the exception of patterns 0 and 64 (which corresponds to no methylation/full methylation of all sites) in L1 and pattern 64 in all loci, where the difference between WT and the numerical solution is about 10%, the difference is always small ($< 5\%$) as seen in the insets.

Tab. 6.1: Estimated parameters for the KO data and model based on Eq. (6.12) for the loci L1 and Afp with sample size n.

KO	μ	ψ_L	ψ_R	τ	n	Locus
Dnmt1	0.334 ± 0.051	0.576 ± 0.067	1.000 ± 0.122	0.038 ± 0.004	1047	L1
Dnmt3a/b	0.789 ± 0.037	1.000 ± 0.038	0.984 ± 0.045	$10^{-10} \pm 0.002$	805	L1
Dnmt1	0.452 ± 0.062	0.383 ± 0.076	1.000 ± 0.094	0.091 ± 0.016	134	Afp
Dnmt3a/b	0.990 ± 0.003	0.984 ± 0.011	1.000 ± 0.006	$10^{-10} \pm 0.011$	186	Afp

In general all 16 models show a similar performance for all loci and positions in terms of accuracy of the prediction. On the large scale the differences are not visible and even for the smaller scale the differences are small, as shown for mSat in Fig. 6.7. This is in accordance to the corresponding Kullback-Leibler divergences

$$KL = \sum_{j=1}^{4^L} \pi_j(\mathrm{WT}) \log \left(\frac{\pi_j(\mathrm{WT})}{\pi_j(\mathrm{pred})} \right)$$

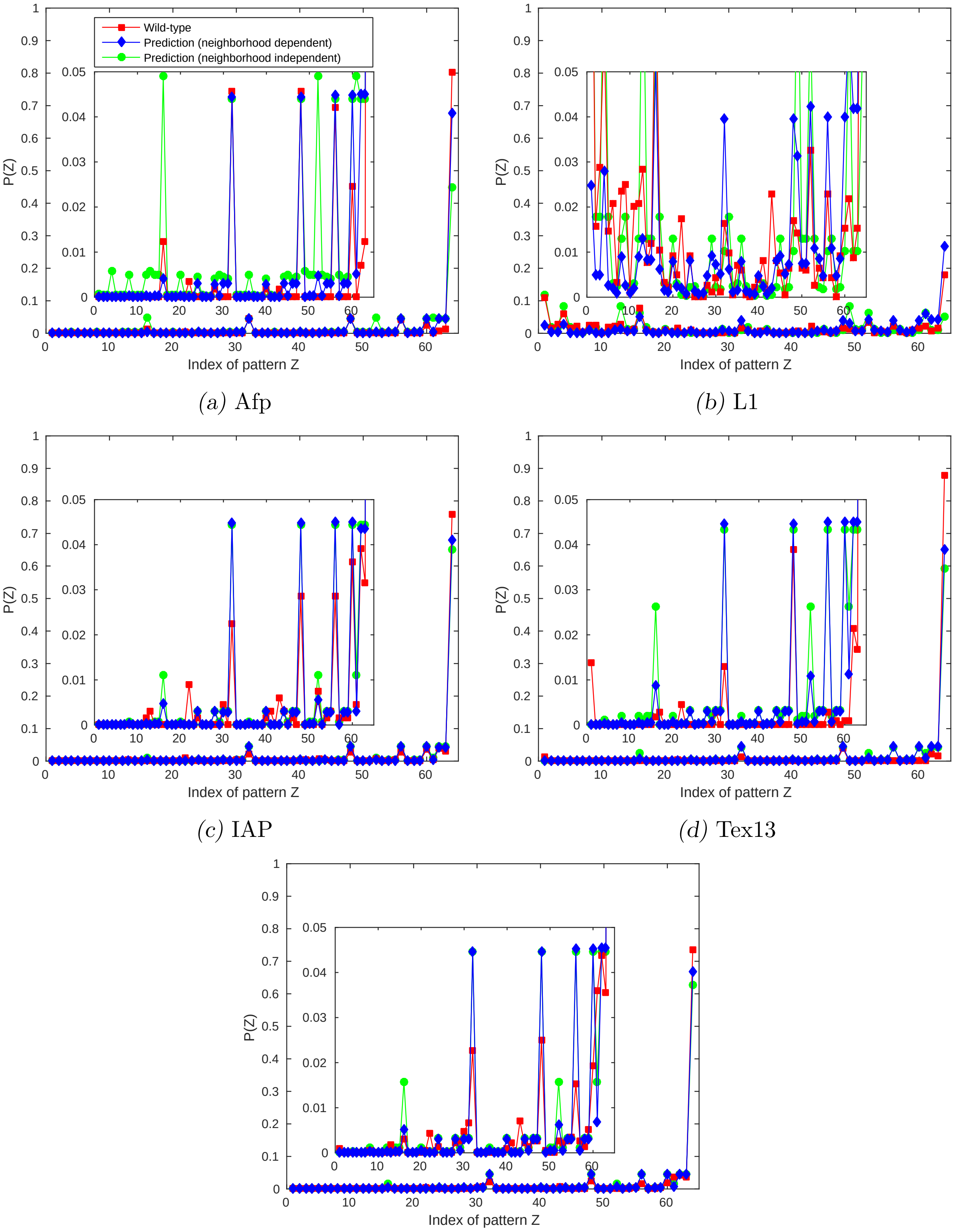
that we list in Table 6.2. The difference in KL between the "best" and the "worst" case is about 0.01. The mean and standard deviation for KL was obtained via bootstrapping of the wild-type data (10.000 bootstrap samples for each model). Since no confidence intervals of the parameters are included, this standard deviation can be regarded as a lower bound. However, even with these lower bounds the intervals of KL overlap for all models, such that no model can be favorized.

Tab. 6.2: Kullback-Leibler divergence KL for the 16 models.

Model	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$
KL	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1337 ± 0.0127
Model	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$
KL	0.1438 ± 0.0137	0.1439 ± 0.0136	0.1439 ± 0.0137	0.1374 ± 0.0133
Model	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$
KL	0.1399 ± 0.0134	0.1399 ± 0.0134	0.1398 ± 0.0133	0.1337 ± 0.0127
Model	$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$
KL	0.1410 ± 0.0137	0.1411 ± 0.0136	0.1409 ± 0.0135	0.1349 ± 0.0130

6.5 Related Work

In [18] location- and neighbor-dependent models are proposed for single-stranded DNA methylation data in blood and tumor cells. The (de-)methylation rates depend on the



(a) Afp

(b) L1

(c) IAP

(d) Tex13

(e) mSat

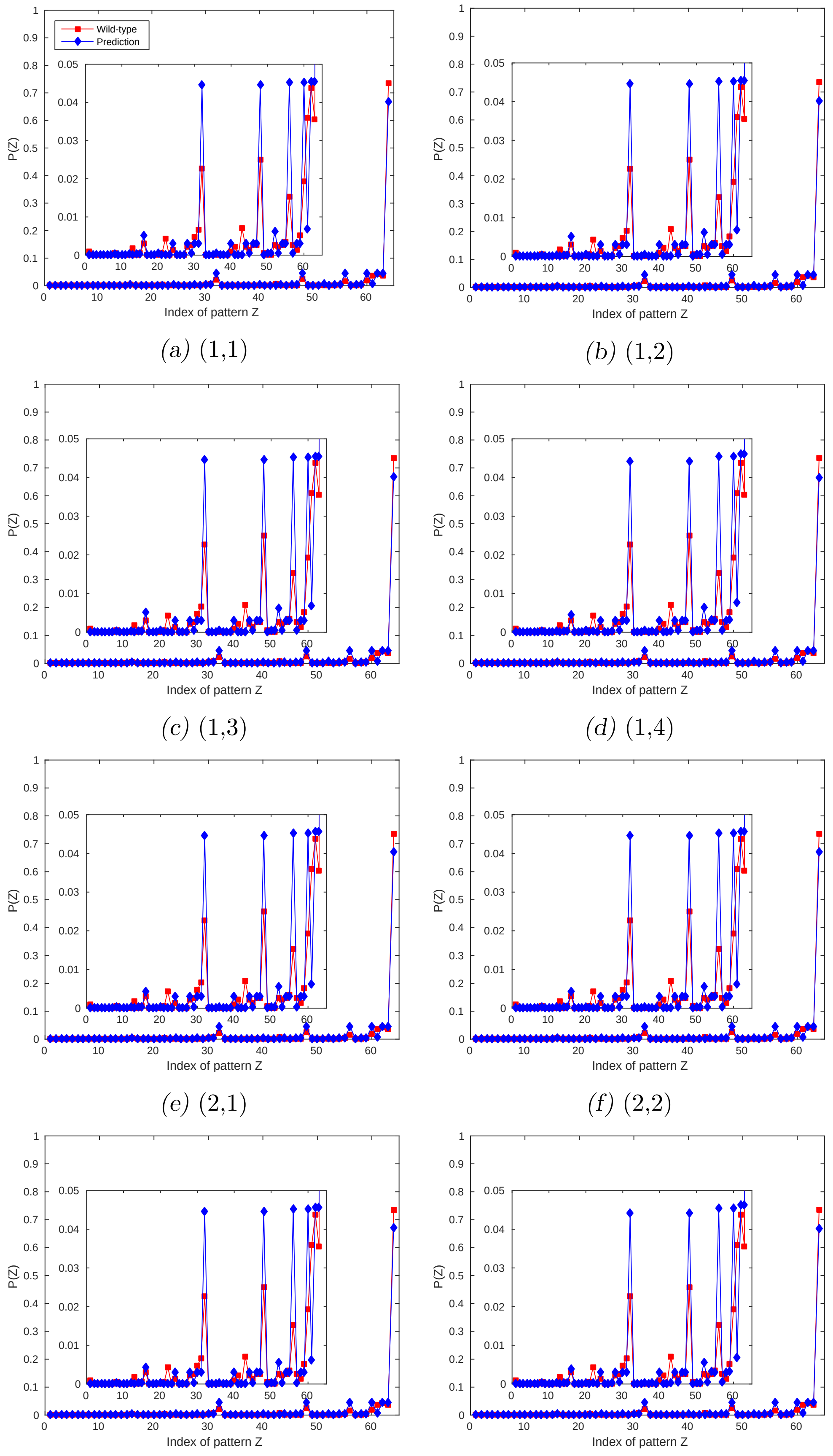
Fig. 6.6: The figures show an example for the predicted (neighbor dependent and neighbor independent) and the measured pattern distribution for each locus. The inset shows a zoomed in version of the distribution.

position of the CpG relative to the 3' or 5' end and/or on the methylation state of the left neighbor only. The dependency is realized by the introduction of an additional parameter. In our proposed models we use double-stranded DNA and can therefore include hemimethylated sites and even distinguish on which strand the site is methylated. Furthermore we allow dependencies on both neighbors by introducing two different dependency parameters. In contrast [19] copes with the neighborhood dependency indirectly by allowing different parameter values for different sites. In order to reduce the dimensionality of the parameter vector, a hierarchical model based on beta distributions is proposed. Another difference to our model is the distinction between de novo rates for parent and daughter strand. However, this can easily be included in future work. A density-dependent Markov model was proposed [20]. In this model, the probabilities of (de-)methylation events may depend on the methylation density in the CpG neighborhood. In addition, a neighboring sites model has been developed, in which the probabilities for a given site are directly influenced by the states of neighboring sites to the left and right [20]. When these models were tested on double-stranded methylation patterns from two distinct tandem repeat regions in a collection of ovarian carcinomas, the density-dependent and neighboring sites models were superior to independent models in generating statistically similar samples. Although this model also includes the dependence on the methylation state on the left and right neighbor for double-stranded DNA the approach is different. The transition probabilities of the neighbor-independent model are transformed into a transition probability of a neighbor-dependent model by introducing only one additional parameter. The state of the left and right neighbor are taken into account by exponentiating this parameter by some norm. In addition, this approach does not allow the intuitive interpretation of the dependency parameter.

6.6 Conclusion

We proposed a set of stochastic models for the formation and modification of methylation patterns over time. These models take into account the state of the CpG sites in the spatial neighborhood and allow to describe different hypotheses about the underlying mechanisms of methyltransferases adding methyl groups at CpG sites. We used knockout data from bisulfite sequencing at several loci to learn the efficiencies at which these enzymes perform methylation. By combining these efficiencies, we accurately predicted the probability distribution of the patterns in the wild-type. Moreover, we found that in all cases the models predict values for the dependency parameters ψ_L and ψ_R close to 1 and therefore independence of methylation for the Dnmt3a/b DKO meaning that Dnmt1 methylates CpGs independent of the methylation of neighboring CpGs. For Dnmt3a/b on the other hand we could identify dependencies on the neighboring CpGs. Both findings are in accordance with current existing mechanistic models: Dnmt1 reliably copies the methylation from the template strand to maintain the distinct methylation patterns, whereas Dnmt3a/b try to establish and keep a certain amount of CpG methylation at a given loci. Interestingly, our models only suggest dependencies of de novo methylation activity on the CpGs in the 5' neighborhood. This indicates that Dnmt3a and Dnmt3b show a preference to methylate CpGs in a 5' to 3' direction and could point towards a processive or cooperative behavior of these enzymes like recently described in *in vitro* experiments [13, 10]. Compared to a neighborhood independent model with $\psi_L = \psi_R = 1$, a neighborhood dependent model shows better predictions and furthermore allows to investigate (possible) connections of adjacent CpGs and their methylation states.

As future work, we plan to investigate models in which we distinguish between the actions of Dnmt3a and Dnmt3b and in which we allow a diagonal dependency for de novo methylation, i.e., a dependency on the state of neighboring CpGs on the opposite strand. Moreover, we will design models that take into account the number of base pairs between adjacent CpG sites. To investigate a potential impact of oxidized cytosine forms on the methylation at neighboring CpG sites we further plan to include the CpG states 5hmC, 5fC and 5caC in our model.



(a) (1,1)

(b) (1,2)

(c) (1,3)

(d) (1,4)

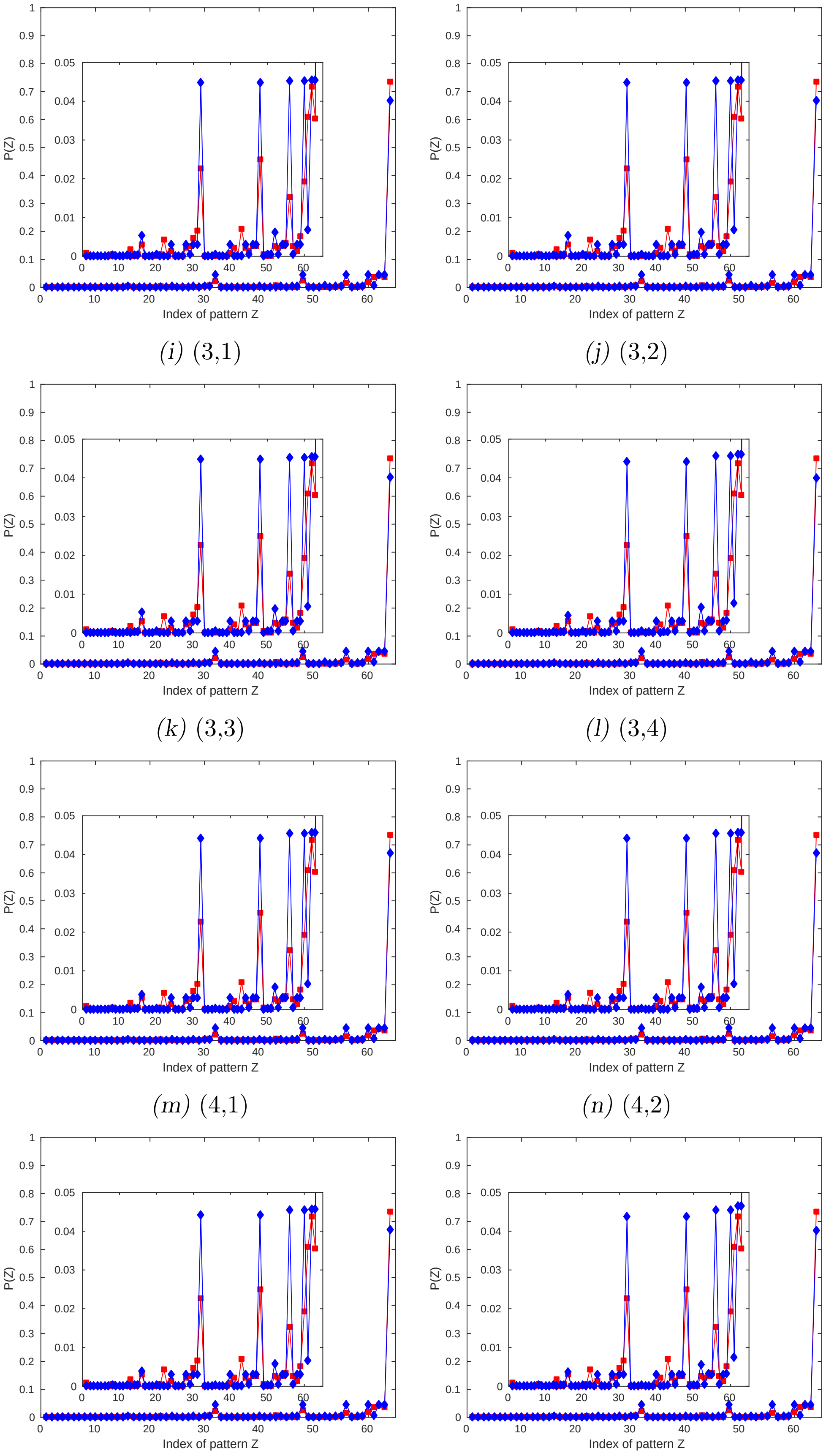
(e) (2,1)

(f) (2,2)

(g) (2,3)

(h) (2,4)

Fig. 6.7: The figures show the predicted and the measured pattern distribution for all 16 models for mSat. The inset shows a zoomed in version of the distribution.



(i) (3,1)

(j) (3,2)

(k) (3,3)

(l) (3,4)

(m) (4,1)

(n) (4,2)

(o) (4,3)

(p) (4,4)

Fig. 6.7: (continued)

Bibliography

[1] Miho M Suzuki and Adrian Bird. Dna methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465, 2008.

[2] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412, 2016.

[3] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[4] Tarmo Äijö, Yun Huang, Henrik Mannerström, Lukas Chavez, Ageliki Tsagaratou, Anjana Rao, and Harri Lähdesmäki. A probabilistic generative model for quantification of dna modifications enables analysis of demethylation pathways. *Genome biology*, 17(1):49, 2016.

[5] Charalampos Kyriakopoulos, Pascal Giehr, and Verena Wolf. H (o) ta: estimation of dna methylation and hydroxylation levels and efficiencies from time course data. *Bioinformatics*, 33(11):1733–1734, 2017.

[6] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 2004.

[7] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[8] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[9] Humaira Gowher and Albert Jeltsch. Molecular enzymology of the catalytic domains of the dnmt3a and dnmt3b dna methyltransferases. *Journal of Biological Chemistry*, 2002.

[10] Celeste Holz-Schietinger and Norbert O Reich. The inherent processivity of the human de novo methyltransferase 3a (dnmt3a) is enhanced by dnmt3l. *Journal of Biological Chemistry*, 285(38):29091–29100, 2010.

[11] Allison B Norvil, Christopher J Petell, Lama Alabdi, Lanchen Wu, Sandra Rossie, and Humaira Gowher. Dnmt3b methylates dna by a noncooperative mechanism, and its activity is unaffected by manipulations at the predicted dimer interface. *Biochemistry*, 2016.

[12] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520(7546):243, 2015.

[13] Max Emperle, Arumugam Rajavelu, Richard Reinhardt, Renata Z Jurkowska, and Albert Jeltsch. Cooperative dna binding and protein/dna fiber formation increases the activity of the dnmt3a dna methyltransferase. *Journal of Biological Chemistry*, pages jbc–M114, 2014.

[14] Sarah P Otto and Virginia Walbot. Dna methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. *Genetics*, 124(2):429–437, 1990.

[15] Diane P Genereux, Brooks E Miner, Carl T Bergstrom, and Charles D Laird. A population-epigenetic model to infer site-specific methylation rates from double-stranded dna methylation patterns. *Proceedings of the National Academy of Sciences*, 102(16):5802–5807, 2005.

[16] Laura B Sontag, Matthew C Lorincz, and E Georg Luebeck. Dynamics, stability and inheritance of somatic dna methylation imprints. *Journal of theoretical biology*, 242(4):890–899, 2006.

[17] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

[18] Nicolas Bonello, James Sampson, John Burn, Ian J Wilson, Gail McGrown, Geoff P Margison, Mary Thorncroft, Philip Crossbie, Andrew C Povey, Mauro Santibanez-Koref, et al. Bayesian inference supports a location and neighbour-dependent model of dna methylation propagation at the mgmt gene promoter in lung tumours. *Journal of theoretical biology*, 336:87–95, 2013.

[19] Audrey Qiuyan Fu, Diane P Genereux, Reinhard Stöger, Charles D Laird, and Matthew Stephens. Statistical inference of transmission fidelity of dna methylation patterns over somatic cell divisions in mammals. *The annals of applied statistics*, 4(2):871, 2010.

[20] Michelle R Lacey and Melanie Ehrlich. Modeling dependence in methylation patterns with application to ovarian carcinomas. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

7. HIDDEN MARKOV MODELLING REVEALS NEIGHBORHOOD DEPENDENCE OF DNMT3A AND 3B ACTIVITY

The content of Chapter 7 presents a preprint version of a manuscript which as by been submitted under:

Lück, A.[*‡], Giehr, P.[†‡], Nordström, K., [†] Walter, J.[†], & Wolf, V.[*] (2018, Jun). Hidden Markov modelling reveals neighborhood dependence of Dnmt3a and 3b activity." IEEE/ACM transactions on computational biology and bioinformatics (2019).

AUTHOR CONTRIBUTIONS

Alexander Lück: Design and development of the model. Design and development of the statistical and computational tools for genome wide application. Clustering of dependency parameter. Authoring of the manuscript including abstract, as well as the sections 7.1 introduction (including the generation of figure 7.2), 7.2 model (including the generation of figure 7.3 and 7.4, as well as all equations), 7.3 results (including the generation of figures 7.5 - 7.11, as well as tables 7.1 - 7.3), 7.4 related work and 7.5 conclusion.

Pascal Giehr: Conduction of reduced representation hairpin bisulfite sequencing. Meta data analysis including LOLA enrichment analysis and data comparison to genomic sequence context. Revision of the manuscript i.e. changes in formulation/wording and structure of the text mainly restricted to the biological background and interpretation of the presented results. Rewriting parts of section 7.1 introduction including the generation of figure 7.1, as well as section 7.3 results (mainly section 7.3.1 paragraph 1 and 3 and section 7.3.5 paragraph 4) including figure 7.12 and 7.5 conclusion.

Dr. Karl Nordström: Processing of sequencing data and generation of primary hairpin sequencing output, meta data analysis in form of genetic annotation of CpGs.

[*] Department of Computer Science, Saarland University, D-66123 Saarbrücken, Germany

[‡] These Authors contributed euqally to this work

[†] Department of Biological Sciences, Saarland University, D-66123 Saarbrücken, Germany

[‡] These Authors contributed euqally to this work

Prof. Dr. Jörn Walter: Supervision. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Prof. Dr. Verena Wolf: Supervision. Financing. Revision of the manuscript i.e. changes in formulation/wording and structure of the text.

Abstract

DNA methylation is an epigenetic mark whose important role in development has been widely recognized. This epigenetic modification results in heritable information not encoded by the DNA sequence. The underlying mechanisms controlling DNA methylation are only partly understood. Several mechanistic models of enzyme activities responsible for DNA methylation have been proposed. Here we extend existing Hidden Markov Models (HMMs) for DNA methylation by describing the occurrence of spatial methylation patterns over time and propose several models with different neighborhood dependencies. Furthermore we investigate correlations between the neighborhood dependence and other genomic information. We perform numerical analysis of the HMMs applied to comprehensive hairpin and non-hairpin bisulfite sequencing measurements and accurately predict wild-type data. We find evidence that the activities of Dnmt3a and Dnmt3b responsible for *de novo* methylation depend on 5' (left) but not on 3' (right) neighboring CpGs in a sequencing string.

7.1 Introduction

The DNA code of an organism determines its appearance and behavior by encoding protein sequences. In addition, there is a multitude of additional mechanisms to control and regulate the ways in which the DNA is packed and processed in the cell and thus determine the fate of a cell. One of these mechanisms in cells is DNA methylation, which is an epigenetic modification that occurs at cytosine (C) bases of eukaryotic DNA. Cytosines are converted to 5-methylcytosine (5mC) by DNA methyltransferase (Dnmt) enzymes. The neighboring nucleotide of a methylated cytosine is usually guanine (G) and together with the CG-pair on the opposite strand, a common pattern is that two methylated cytosines are located diagonally to each other on opposing DNA strands. DNA methylation at CpG dinucleotides is known to control and mediate gene expression and is therefore essential for cell differentiation and embryonic development. In human somatic cells, approximately 70-80% of the cytosine nucleotides in CpG dyads are methylated on both strands and methylation near gene promoters varies considerably depending on the cell type. Methylation of promoters often correlates with low or no transcription [1] and can be used as a predictor of gene expression [2]. Also, significant differences in overall and specific methylation levels exist between different tissue types and between normal cells and cancer cells from the same tissue. However, the exact mechanism which leads to a methylation of a specific CpG and the formation of distinct methylation patterns at certain genomic regions is still not fully understood. Recently proposed measurement techniques based on hairpin bisulfite sequencing (BS-seq) allow to determine the level of 5mC at individual CpGs dyads on both DNA strands [3]. Based on a small hidden Markov model, the probabilities of the different states of a CpG can be accurately estimated (assuming that enough samples per CpG are provided) [4, 5, 6].

Mechanistic models for the activity of the different Dnmts usually distinguish *de novo* activities, i.e., adding methyl groups at cytosines independent of the methylation state of the opposite strand, and maintenance activities, which refers to the copying of methylation from an existing DNA strand to its newly synthesized partner (containing no methylation) after replication [7, 8]. Hence, maintenance methylation is responsible for re-establishment of the same DNA methylation pattern before and after cell replication. A common hypothesis is that the copying of DNA methylation patterns after replication is performed by Dnmt1, an enzyme that shows a preference for hemimethylated CpG sites (only one strand is methylated) as they appear after DNA replication. Moreover, studies have shown that Dnmt1 is highly processive and able to methylate long sequences of hemimethylated CpGs without dissociation from the target DNA strand [7]. However, an exact transmission of the methylation information to the next cellular generation is not guaranteed. The enzymes Dnmt3a and Dnmt3b show equal activities on hemi- and unmethylated DNA and are mainly responsible for *de novo* methylation, i.e., methylation without any specific preference for the current state of the CpG (hemi- or unmethylated) [8]. However, by now evidence exists that the activity of the different enzymes is not that exclusive, i.e., Dnmt1 shows to a certain degree also *de novo* and Dnmt3a/b maintenance methylation activity [9]. The way how methyltransferases interact with the DNA and introduce CpG methylation was investigated in many *in vitro* studies. Basically, one can distinguish between two mechanisms. A distributive one, where the enzyme periodically binds and dissociates from the DNA, leaping more or less randomly from one CpG to another and a processive one in which the enzyme migrates along the DNA without detachment from the DNA [10, 11, 12], as illustrated in Fig. 7.1. Note that for Dnmt1, for instance, it is reasonable to assume that it is processive in 5' to 3' direction since it is linked to the DNA replication machinery. In particular for the Dnmt3's different hypotheses about the processivity and neighborhood dependence exist [13, 14], but the detailed mechanisms remain elusive.

Several models that describe the dynamics of the formation of methylation patterns have been proposed. In the seminal paper of Otto and Walbot, a dynamical model was proposed that assumed independent methylation events for a single CpG. The main idea was to track the frequencies of fully, hemi- and unmethylated CpGs during several cell

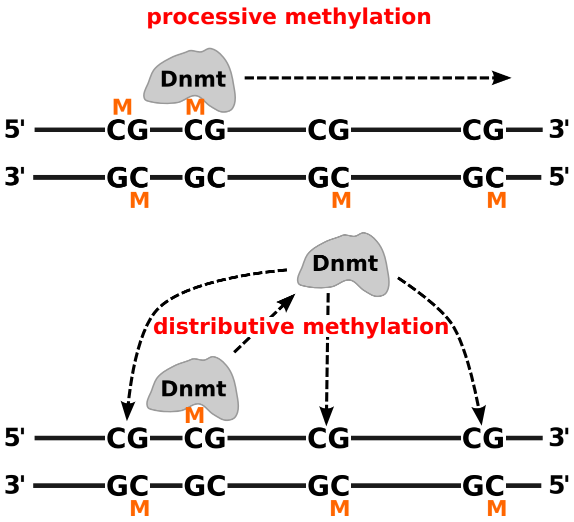


Fig. 7.1: Dnmts can methylate DNA in a processive way where the enzyme starts at one CpG and slides in 5' to 3' direction over the DNA or in a distributive manner, "jumping" randomly from one CpG to another.

generations [15]. Later, refined models allowed to distinguish between maintenance and *de novo* methylation on the parent and daughter strands [16, 17]. More sophisticated extensions of the original model of Otto and Walbot models have been successfully used to predict *in vivo* data still assuming a neighbor-independent methylation process for a single CpG site [9, 5]. However, measurements indicate that methylation events at a single CpG may depend on the methylation state of neighboring CpGs, which is not captured by these models.

Here, we follow the dynamical HMM approach proposed in [9] where knockout data was used to train a model that accurately predicts wild-type methylation levels for BS-seq data of repetitive elements from mouse embryonic stem cells. We extend this model by describing the methylation state of several CpGs instead of a single CpG and use similar dependency parameters as introduced in Bonello et al. [18]. More specifically, we design different models by combining the activities of the two types of Dnmts and test for both, maintenance and *de novo* methylation the hypotheses illustrated in Fig 7.1. The models vary according to the order in which the enzymes act, whether they perform methylation in a processive manner or not, and how much their action depends on the left/right CpG neighbor. We use the same BS-seq hairpin data as in [9], i.e. data where Dnmt1 or Dnmt3a/b was knocked out (KO) and learn the parameters of the different models. We also relate the estimated dependency parameters to the distance between the respective adjacent CpGs in order to investigate their possible influence. Then, similar as in [9], we predict the behavior of the measured wild-type (WT), in which both types of enzymes are active, by designing a combined model that describes the activity of both enzymes and compare the results to the WT data. Finally, we apply our model to non-hairpin data.

We found that all proposed models show a similar behavior in terms of prediction quality such that no model can be declared as the best fit. However, our results indicate that Dnmt1 works independently of the methylation state of its neighborhood, which is in accordance to the current hypothesis that Dnmt1 is linked to the replication machinery and copies the methylation state on the opposite strand. On the other hand, Dnmt3a/b shows a dependency to the left but no dependency to the right, which supports hypotheses of processive or cooperative behavior. Furthermore, we find evidence that at least for small distances rather the genetic region than the distance determines the dependence on the neighbors. Applying our model to a genome-wide data set we find three distinct clusters based on the dependency parameters and distances between adjacent CpGs. These clusters also show different methylation levels and reveal that hypomethylated CpGs in promoter regions behave independent of their neighborhood. Finally our results show that our model can also be used for non-hairpin data as long as no information from the opposite strand is needed as for example in Dnmt1KO data.

This paper is organized is as follows: Our model is introduced in section II and the results are presented in section III. In section IV we discuss the related work. We conclude the paper in section V and give a brief outline on future work.

7.2 Model

7.2.1 Notation

Consider a sequence of L neighboring CpG dyads*, which is represented as a lattice of length L and width two (for the two strands). Each cytosine in the lattice can either be methylated or not, leading to four possible states at each position l:

- *State 0*: Both cytosines are not methylated.
- *State 1*: The cytosine on the upper strand is methylated, the lower one not.
- *State 2*: The cytosine on the lower strand is methylated, the upper one not.
- *State 3*: Both cytosines are methylated.

A sequence of four CpGs, each of which is in one of the four possible states, is shown in Fig. 7.2.

For a system of length L there are in total 4^L possibilities to combine the states of individual CpGs. These combinations are called *patterns* in the following. A pattern is denoted by a concatenation of states, e.g. 321, 0123 or 33221.

In order to represent the pattern distribution as a vector it is necessary to uniquely assign

*The exact nucleotide distance between two neighboring dyads is not considered here explicitly, but we assume that this distance is small. For the BS-seq data that we consider, the average distance between two CpGs is 14 bps (base pairs) and the maximal distance is 46 bps.

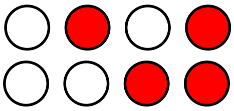


Fig. 7.2: A lattice of length $L = 4$ containing all possible states 0, 1, 2 and 3, forming the pattern 0123.

a reference number to each pattern. A pattern can be perceived as a number in the tetral system, such that converting to the decimal system leads to a unique reference number. After the conversion an additional 1 is added in order to start the referencing at 1 instead of 0.
Examples for $L = 3$:

$$\begin{aligned} 000 &\longrightarrow 1\ (= 0 + 1) \\ 123 &\longrightarrow 28\ (= 27 + 1) \\ 333 &\longrightarrow 64\ (= 63 + 1) \end{aligned}$$

This reference number then corresponds to the position of the pattern in the respective distribution vector.

We describe the state of a sequence of L CpGs by a discrete-time Markov chain with pattern distribution $\pi(t)$, i.e., the probability of each of the 4^L patterns after t cell divisions. For the initial distribution $\pi(0)$, we use the distribution measured in the wild-type when the cells are in equilibrium. Note, that other initial conditions gave very similar results, i.e., the choice of the initial distribution does not significantly affect the results. The reason is that also the KO data is measured after a relatively high number of cell divisions where the cells are almost in equilibrium. Transitions between patterns are triggered by different processes: First due to *cell division* the methylation on one strand is kept as it is (e.g. the upper strand), whereas the newly synthesized strand (the new lower strand) does not contain any methyl group. Afterwards, methylation is added due to different mechanisms. On the newly synthesized strand a site can be methylated if the cytosine at the opposite strand is already methylated (*maintenance*). It is widely accepted that maintenance in form of Dnmt1 is linked to the replication machinery and thus occurs during/directly after the synthesis of the new strand. Furthermore, CpGs on both strands can be methylated independent of the methylation state of the opposite site (*de novo*). The transition matrix P is defined by composition of matrices for cell division, maintenance and *de novo* methylation of each site.

7.2.2 Cell Division

Depending on which daughter cell is considered after cell replication, the upper ($s = 1$) or lower ($s = 2$) strand is the parental one after cell division. Then, the new pattern can

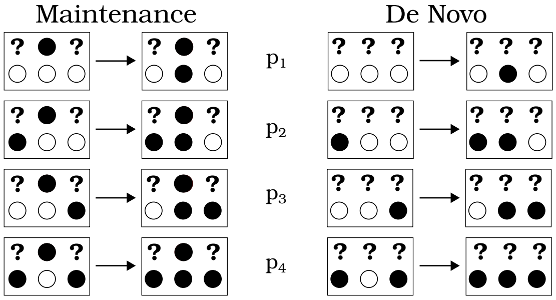


Fig. 7.3: Possible maintenance and *de novo* transitions depicted for the lower strand, where ○ denotes an unmethylated, • a methylated site and **?** a site where the methylation state does not matter. Note that the same transitions can occur on the upper strand.

be obtained by applying the following state replacements:

$$s = 1: \begin{cases} 0 & \longrightarrow & 0 \\ 1 & \longrightarrow & 1 \\ 2 & \longrightarrow & 0 \\ 3 & \longrightarrow & 1 \end{cases} \qquad s = 2: \begin{cases} 0 & \longrightarrow & 0 \\ 1 & \longrightarrow & 0 \\ 2 & \longrightarrow & 2 \\ 3 & \longrightarrow & 2 \end{cases} \tag{7.1}$$

Given some initial pattern with reference number i, applying the transformation (7.1) to each of the L positions leads to a new pattern with reference number j (notation: $i \overset{(7.1)}{\rightsquigarrow} j$). The corresponding transition matrix $D_s \in \{0,1\}^{4^L \times 4^L}$ has the form

$$D_s(i,j) = \begin{cases} 1, & \text{if } i \overset{(7.1)}{\rightsquigarrow} j, \\ 0, & \text{else.} \end{cases} \tag{7.2}$$

7.2.3 *Maintenance and* De Novo *Methylation*

For maintenance and *de novo* methylation, the single site transition matrices are built according to the following rules:

Consider at first the (non-boundary) site $l = 2, \dots, L-1$ and its left and right neighbor $l-1$ and $l+1$ respectively. The remaining sites do not change and do not affect the transition. The probabilities of the different types of transitions in Fig. 7.3 have the form

$$p_1 = 0.5 \cdot (\psi_L + \psi_R)x, \tag{7.3}$$

$$p_2 = 0.5 \cdot (\psi_L + \psi_R)x + 0.5 \cdot (1 - \psi_L), \tag{7.4}$$

$$p_3 = 0.5 \cdot (\psi_L + \psi_R)x + 0.5 \cdot (1 - \psi_R), \tag{7.5}$$

$$p_4 = 1 - 0.5 \cdot (\psi_L + \psi_R)(1 - x), \tag{7.6}$$

where we set the probability x to $x = \mu$ in case of maintenance and to $x = \tau$ in case of *de novo* methylation. $\psi_L,\ \psi_R \in [0,1]$ are the dependency parameters for the left and right

neighbor.
A dependency value of $\psi_i = 1$ corresponds to a total independence on the neighbor whereas $\psi_i = 0$ leads to a total dependence. Hence, μ and τ can be interpreted as the probability of maintenance and *de novo* methylation of a single cytosine between two cell divisions assuming independence from neighboring CpGs. Moreover, all CpGs that are part of the considered window of the DNA have the same value for the parameters μ, τ, ψ_L, and ψ_R, since in earlier experiments only very small differences have been found between the methylation efficiencies of nearby CpGs [9].

In order to understand the form of the transition probabilities consider at first a case with only one neighbor. The probabilities then have the form ψx if the neighbor is unmethylated and $1-\psi(1-x)$ if the neighbor is methylated. Note that both forms evaluate to x for $\psi = 1$, meaning that a site is methylated with probability x, independent of its neighbor. For $\psi = 0$ the probabilities become 0 and 1, meaning that if there is no methylated neighbor the site cannot be methylated or will be methylated for sure if there is a methylated neighbor respectively.
The probabilities for two neighbors are obtained by a linear combination of the one neighbor cases, with ψ_L for the left and ψ_R for the right neighbor, and an additional weight of 0.5 to normalize the probability.
The same considerations also apply to the boundary sites however there is no way of knowing the methylation states outside the boundaries (denoted by ?). Therefore instead of a concrete methylation state (○ for unmethylated, ● for methylated site) the average methylation density ρ is used to compute the transition probabilities at the boundaries (depicted here for *de novo*):

$$? \circ \circ \rightarrow ? \bullet \circ \qquad \tilde{p}_1 = (1-\rho)\cdot p_1 + \rho\cdot p_2, \tag{7.7}$$

$$? \circ \bullet \rightarrow ? \bullet \bullet \qquad \tilde{p}_2 = (1-\rho)\cdot p_3 + \rho\cdot p_4, \tag{7.8}$$

$$\circ \circ ? \rightarrow \circ \bullet ? \qquad \tilde{p}_3 = (1-\rho)\cdot p_1 + \rho\cdot p_3, \tag{7.9}$$

$$\bullet \circ ? \rightarrow \bullet \bullet ? \qquad \tilde{p}_4 = (1-\rho)\cdot p_2 + \rho\cdot p_4. \tag{7.10}$$

Note that the same considerations hold for maintenance at the boundaries if the opposite site of the boundary site is already methylated.
For each position l, there are four transition matrices: two for maintenance and two for *de novo*, namely one for the upper and one for the lower strand in each process. In order to construct these matrices consider the three positions $l-1$, l and $l+1$, where the transition happens at position l. Only the transitions depicted in Fig. 7.3 can occur. Furthermore the transitions are unique, i.e. for a given reference number i the new reference number j is uniquely determined. For patterns not depicted in Fig. 7.3 no transition can occur, i.e. the reference number does not change.

The matrix describing a maintenance event at position l and strand s has the form

$$M_s^{(l)}(i,j) = \begin{cases} 1, & \text{if } i = j \text{ and } \nexists j' : \ i \rightsquigarrow j', \\ 1-p, & \text{if } i = j \text{ and } \exists j' : \ i \rightsquigarrow j', \\ p, & \text{if } i \neq j \text{ and } i \rightsquigarrow j, \\ 0, & \text{else,} \end{cases} \tag{7.11}$$

where the probability p is given by one of the Eqs. (7.3)-(7.10) that describes the corresponding case and $x = \mu$. Note that $M_s^{(l)}$ depends on s and l since it describes a single transition from pattern i to pattern j, which occurs on a particular strand and at a particular location with probability p. We define matrices $T_s^{(l)}$ for *de novo* methylation according to the same rules except that $x = \tau$ and the possible transitions are as in Fig. 7.3, right. All matrices are of size $4^L \times 4^L$.

The advantage of defining the matrices position- and process-wise is that different models can be realized by changing the order of multiplication of these matrices.

It is important to note that 5mC can be further modified by oxidation to 5-hydroxymethyl- (5hmC), 5-formyl- (5fC) and 5-carboxyl cytosine(5caC) by Tet enzymes. These modifications are involved in the removal of 5mC from the DNA and can potentially interfere with methylation events. However, our data does not capture these modifications and therefore we are not able to consider these modifications in our model.

7.2.4 Combination of Transition Matrices

For all subsequent models it is assumed that first of all cell division happens and maintenance methylation only occurs on the newly synthesized strand given by s, whereas *de novo* methylation happens on both strands. Given the mechanisms in Fig. 7.1, the two different kinds of methylation events, and the two types of enzymes, there are several possibilities to combine the transition matrices. We consider the following four models, which we found most reasonable based on the current state of research in DNA methylation:

a. first processive maintenance and then processive *de novo* methylation

$$P_s = \prod_{l_1=1}^{L} M_s^{(l_1)} \prod_{l_2=1}^{L} T_1^{(l_2)} \prod_{l_3=1}^{L} T_2^{(l_3)}, \tag{7.12}$$

b. first processive maintenance and then *de novo* in arbitrary order

$$P_s = \frac{1}{(L!)^2} \prod_{l_1=1}^{L} M_s^{(l_1)} \left(\sum_{\sigma_1 \in S_L} \prod_{l_2=1}^{L} T_1^{(\sigma_1(l_2))} \right) \cdot \left(\sum_{\sigma_2 \in S_L} \prod_{l_3=1}^{L} T_2^{(\sigma_2(l_3))} \right), \tag{7.13}$$

c. maintenance and *de novo* at one position, processive

$$P_s = \prod_{l=1}^{L} M_s^{(l)} T_1^{(l)} T_2^{(l)}, \tag{7.14}$$

d. maintenance and *de novo* at one position, arbitrary order

$$P_s = \frac{1}{L!} \sum_{\sigma \in S_L} \prod_{l=1}^{L} M_s^{(\sigma(l))} T_1^{(\sigma(l))} T_2^{(\sigma(l))}, \tag{7.15}$$

where S_L is the set of all possible permutations for the numbers $1, \dots, L$.

Note that the *de novo* events on both strands are independent, i.e. the *de novo* events on the upper strand do not influence the *de novo* events on the lower strand and vice versa, such that $[T_1^{(l)}, T_2^{(l')}] = 0$ independent of ψ_i[†]. Obviously it is important whether maintenance or *de novo* happens first, since the transition probabilities and the transitions themselves depend on the actual pattern. Furthermore in the case $\psi_i < 1$ (dependency on right and/or left neighbor) the order of the transitions on a strand matters, i.e. $[M_s^{(l)}, M_s^{(l')}] \neq 0$ and $[T_s^{(l)}, T_s^{(l')}] \neq 0$ for $l \neq l'$. Note that this definition of models in principle allows to consider an arbitrary number of CpGs. However, at least three CpGs are needed to properly include the influence of the left and right neighbor in the transitions. It is also important to note that independent of the number of considered CpGs the window size of the influential CpGs for the transition rates is always kept at size three. However, treating more than three CpGs at once has two major drawbacks: First of all the number of possible patterns grows rapidly (recall 4^L possible patterns for L CpGs) and hence the transition matrices become very large as well ($4^L \times 4^L$). This may lead to memory issues while calculating the distributions, which can however be circumvented by sampling approaches, i.e. stochastic simulation of the underlying Markov chain. Another problem with the large number of possible patterns is that more data is required in order to ensure a good coverage, i.e. the number of measurements should be larger than the number of patterns.

The second main problem is that using the same dependency parameters for all pairs of

[†] $[A, B] = AB - BA$ is the commutator of the matrices A and B.

$$\begin{tikzcd} u \arrow[d, "c"'] \arrow[dr, dashed, "1-c"' near start] & m \arrow[d, "d"] \arrow[dl, dashed, "1-d" near start] \\ T & C \end{tikzcd}$$

Fig. 7.4: Conversions of the unobservable states u, m to observable states T, C with respective rates.

adjacent CpGs is a rather strong assumption. Note that this assumption becomes more problematic for larger windows, due to e.g. different distances between the CpGs. One solution would be to introduce extra dependency parameters for each pair, however this may lead to difficulties in the parameter identification.

The total transition matrix is then given by a combination of the cell division and maintenance/*de novo* matrices. Recall that we consider two different types of Dnmts, i.e., Dnmt1 and Dnmt3a/b. If only one type of Dnmt is active (KO data) the matrix has the form

$$P = 0.5 \cdot (D_1 \cdot P_1 + D_2 \cdot P_2) \tag{7.16}$$

and if all Dnmts are active (WT data)

$$P = 0.5 \cdot (D_1 \cdot P_1 \cdot \tilde{P}_1 + D_2 \cdot P_2 \cdot \tilde{P}_2), \tag{7.17}$$

where P_s and $\tilde{P}_s$ have one of the forms (7.12)-(7.15). This leads to four different models for one active enzyme or 16 models for all active enzymes respectively. In the second case P_s represents the transitions caused by Dnmt1 and $\tilde{P}_s$ the transitions caused by Dnmt3a/b. Note that if $\psi_L = \psi_R = 1$ all models are the same within each case since they reduce to the neighborhood independent model from [9]. Furthermore, the cell division, maintenance, and *de novo* transition matrices for a single CpG at a given position are sparse. However, upon combining them to the full transition matrices in Eqs. (7.16) or (7.17), the final matrices become dense and therefore have higher memory requirements.

7.2.5 Conversion Errors

The actual methylation state of a C cannot be directly observed. During BS-seq, with high probability every unmethylated C (denoted by u) is converted into Thymine (T) and every 5mC (denoted by m) into C. However, conversion errors may occur and we define their probability as $1-c$ and $1-d$, respectively, as shown by the dashed arrows in Fig. 7.4. It is reasonable that these conversion errors occur independently and with approximately identical probability at each site and thus the error matrix for a single CpG takes the form

$$\Delta_1 = \begin{pmatrix} c^2 & c\bar{c} & c\bar{c} & \bar{c}^2 \\ c\bar{d} & cd & \bar{c}\bar{d} & d\bar{c} \\ c\bar{d} & \bar{c}\bar{d} & cd & d\bar{c} \\ \bar{d}^2 & d\bar{d} & d\bar{d} & d^2 \end{pmatrix}, \tag{7.18}$$

with $\bar{c} = 1 - c$ and $\bar{d} = 1 - d$. Due to the independency of the events this matrix can easily be generalized for systems with $L > 1$ by recursively using the Kronecker-product

$$\Delta_L = \Delta_1 \otimes \Delta_{L-1} \qquad \text{for } L \geq 2. \tag{7.19}$$

Hence, Δ_L gives the probability of observing a certain sequence of C and T nucleotides for each given unobservable methylation pattern. In order to compute the likelihood $\hat{\pi}$ of the observed BS-seq data, we therefore first compute the transient distribution $\pi(t)$ of the underlying Markov chain at the corresponding time instant[‡] t by solving

$$\pi(t) = \pi(0) \cdot P^t \tag{7.20}$$

and then multiply the distribution of the unobservable patterns with the error matrix.

$$\hat{\pi} = \pi(t) \cdot \Delta_L. \tag{7.21}$$

Note that this yields a hidden Markov model with emission probabilities Δ_L. In the following the values for c were chosen according to [9]. Since the value for d was not determined in [9], we measured the conversion rate $d = 0.94$ in an independent experiment under comparable conditions (data not shown).

7.2.6 Maximum Likelihood Estimator

In order to estimate the parameters $\theta = (\mu, \psi_L, \psi_R, \tau) \in [0,1]^4$, we employ a Maximum (Log)Likelihood Estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta), \quad \ell(\theta) = \sum_{j=1}^{4^L} \log(\hat{\pi}_j(\theta)) \cdot N_j, \tag{7.22}$$

where $\hat{\pi}$ is the pattern distribution obtained from the numerical solution of (7.20) and (7.21) for a given time t and N_j is the number of occurrences of pattern j in the measured data. The parameters $\theta = \hat{\theta}$ are chosen in such a way that ℓ is maximized. In order to ensure that the global maximum in $[0,1]^4$ is found during the optimization, we ran the estimation several times with different random starting points. In all cases the estimation

[‡] The number of cell divisions is estimated from the time of the measurement since these cells divide once every 24 hours.

yielded the same results, such that we can conclude that indeed the global optimum was found.

We employ the MLE twice in order to estimate the parameter vector $\hat{\theta}_1$ for Dnmt1 from the 3a/b DKO (double knockout) data and the vector $\hat{\theta}_{3a/b}$ for Dnmt3a/b from the Dnmt1 KO data, where transition matrix (7.16) is used. The corresponding time instants are $t = 26$ for the 3a/b DKO data and $t = 41$ for the 1KO data.

We approximate the standard deviations of the estimated parameters $\hat{\theta}$ as follows: Let $\mathcal{I}(\hat{\theta}) = \mathbb{E}[-\mathcal{H}(\hat{\theta})]$ be the expected Fisher information, with the Hessian $\mathcal{H}(\hat{\theta}) = \nabla\nabla^\intercal \ell(\hat{\theta})$. The inverse of the expected Fisher information is a lower bound for the covariance matrix of the MLE such that we can use the approximation $\sigma(\hat{\theta}) \approx \sqrt{\mathrm{diag}(-\mathcal{H}(\hat{\theta}))}$.

A prediction for the wild-type can be computed by combining the estimated vectors such that in the model both types of enzymes are active. For this, we insert $\hat{\theta}_1$ in P_s and $\hat{\theta}_{3a/b}$ in $\tilde{P}_s$ in (7.17) to obtain the transition matrix for the wild-type.

7.2.7 Data

For our analysis we focused on hairpin data of the single copy genes Afp (5 CpGs) and Tex13 (10 CpGs) as well as the repetitive elements IAP (intracisternal A particle) (6 CpGs), L1 (Long interspersed nuclear elements) (7 CpGs) and mSat (major satellite) (3 CpGs). During the workflow of hairpin bisulfite sequencing, the two DNA strands are linked together covalently, i.e. the methylation status of both strands from an individual chromosome (DNA molecule) is known. Repetitive elements occur in multiple copies and are dispersed over the entire genome. Therefore they allow capturing an averaged, more general behavior of methylation dynamics. Typical data sets are shown in Fig. 7.5. Note that the WT data is almost always fully methylated, while the Dnmt1KO data is mostly un- or hemimethylated. The Dnmt3ab DKO data is somewhat in between.

7.3 Results

7.3.1 Parameter Estimation

In the following we focus on the hairpin data for the single copy genes and repetitive elements as introduced in the previous section. If a locus contains more than three CpGs, the analysis is done for all sets of three adjacent sites independently, in order to keep computation times short and memory requirements low. In the sequel, we mainly focus on the estimated dependency parameters ψ_L and ψ_R and on the prediction quality of the different models.

The estimates for all the available KO data and all suggested models obtained using the transition matrix in Eq. (7.16) are summarized as histograms in Fig. 7.6. Because of the different possibilities to combine the four different models in Eq. (7.12)-(7.15) and

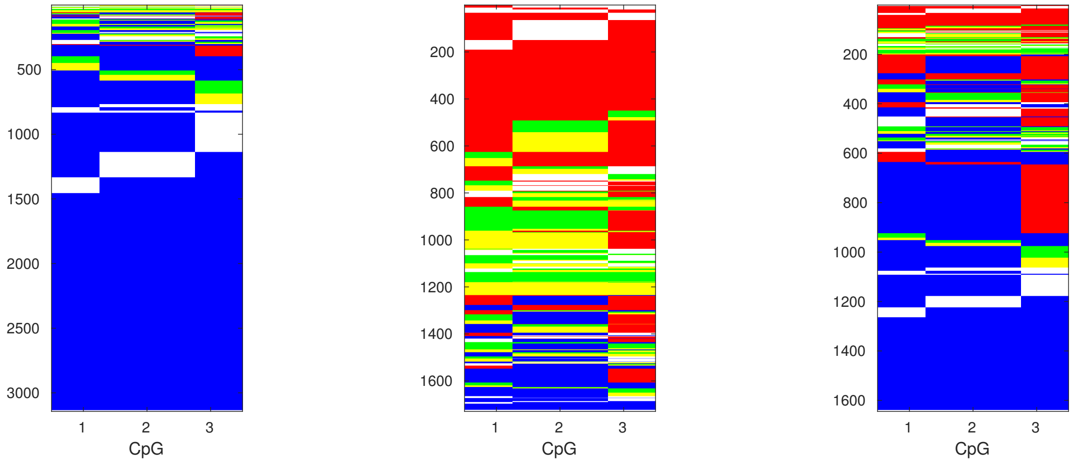


Fig. 7.5: Representations of WT (left), Dnmt1KO (middle) and Dnmt3a/b DKO (right) data for mSat. On the X axis the CpGs and on the Y axis the measured cells are shown. The different colors encode the states as follows: Red: 0, green: 1, yellow: 2, blue:3, and white: “no measurement”.

because of the different loci considered, in total there are 84 estimates for each KO data set. We plot the number of occurrences N of ψ_L (top) and ψ_R (bottom) in different ranges for both sorts of KO data (Dnmt1KO and Dnmt3a/b DKO).

The estimates of ψ_L spread over the whole interval $[0, 1]$ while in the case of ψ_R, nearly all estimates are larger than 0.99 and only in a few cases the dependency parameter is significantly smaller than 1. Hence, in most cases the methylation probabilities are independent of the right neighbor for both Dnmt1KO and Dnmt3a/b DKO. For ψ_L the dependency parameter in the Dnmt3a/b DKO case occurs more often close to 1, meaning that the transitions induced by Dnmt1 have little to no dependency on the left neighbor. On the other hand for Dnmt1KO the dependency parameter occurs more often at smaller values giving evidence that there is a dependency on the left neighbor for the activity of Dnmt3a/b. Note that all models show a similar behavior in terms of the dependency parameters for a given locus or position within a locus respectively, i.e. either $\psi_i \approx 1$ or $\psi_i < 1$ for all models. Since the histograms for Dnmt3a/b DKO look very similar for ψ_L and ψ_R, we used a two-sample Kolmogorov-Smirnov test to assess if they differ significantly. The resulting p-value of 1 indicates that there is no significant difference in this case. Note that we also get quite high p-values (0.786 and 0.433) when applying the test to the Dnmt1KO histogram for ψ_R and the two Dnmt3a/b DKO histograms. On the other hand, the p-values are significantly smaller for Dnmt1KO ψ_L histogram, with a minimum of 0.019, indicating a different behavior for the dependency on the left neighbor for Dnmt3a/b.

Since ψ_R is usually close to 1 a smaller model with only three parameters $\theta = (\mu, \psi, \tau)$ can be proposed, where ψ is a dependency parameter for the left neighbor. This model can either be obtained by fixing $\psi_R = 1$ in the original model and setting $\psi = \psi_L$ or by redefining the transition probabilities to ψx if the left neighbor is unmethylated and

$1 - \psi(1-x)$ if the left neighbor is methylated. In that case ψ and ψ_L are related via $\psi = 0.5(\psi_L + 1)$. Note that both versions yield the same results. In order to check whether there is a significant difference in the original and the smaller model, we performed a Likelihood-ratio test with the null hypothesis that the smaller model is a special case of the original model. Since the original model with more parameters is always as least as good as the smaller model, our goal is to check in which cases the smaller model is sufficient. Indeed, if ψ_R was estimated to be approximately 1 the Likelihood-ratio test indicates that the smaller model is sufficient (p-value ≈ 1). On the other hand, for the few cases where ψ_R differs significantly from 1 the original model has to be used (p-value < 0.01).

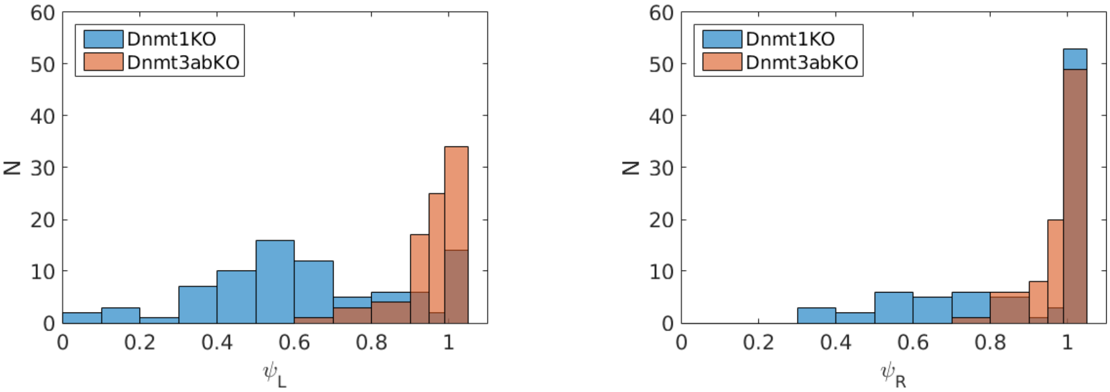


Fig. 7.6: Histograms for the estimated dependency parameters ψ_L and ψ_R for all sets of three adjacent CpGs in all loci and for all suggested models.

7.3.2 CpG Distances

We now take a closer look at the estimated dependency parameters shown in the histograms in Fig. 7.6 and link the parameters to their respective loci and distances between adjacent CpGs. The results for the estimation of the left and right dependency parameter for both Dnmt3a/b DKO and Dnmt1KO data, based on the transition matrix in Eq. (7.12) are shown in Fig. 7.7. The results based on the other transition matrices yielded similar results and are therefore not presented here. The coloring of the symbols for the different loci is as follows: mSat (red), Afp (blue), IAP (green), L1 (pink) and Tex13 (black). As already seen before, in all cases, except for the dependency of the activity of Dnmt3a/b on the left neighbor, the dependency parameter is always close to 1, independent of the distance between the CpGs, i.e. the majority of the estimates for the dependency parameters fall into the interval $0.9 < \psi < 1$. Only Dnmt3a/b shows a stronger dependency on the left neighbor, i.e. in most cases $\psi < 0.9$, but no simple relation to the distance is visible. Another observation from Fig. 7.7 (c) is that the depencency parameters show very similar behaviors within the same locus. However, it is impossible to draw reliable conclusions due to the small sample size within each loci.

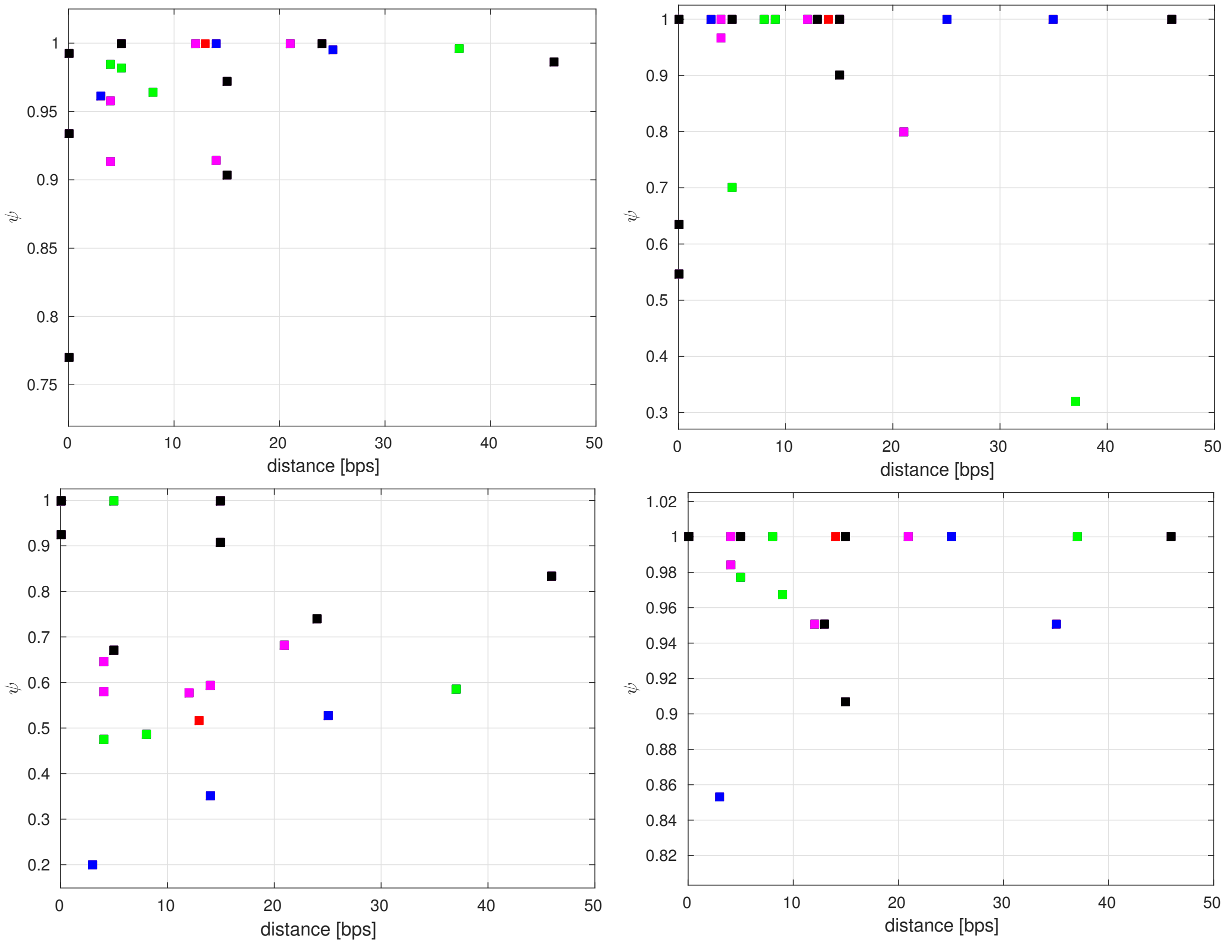


Fig. 7.7: Dependency parameter versus distance between CpGs measured in bps. The top row shows the results for the Dnmt3a/b DKO data, the bottom row for Dnmt1KO. The left (right) column shows results for the dependency parameter to the left (right). The right column shows results for the dependency parameter to the right. The different colors of the symbols represent the different loci and are explained in the main text. Note the different ranges on the Y axes. Red dots = mSat, blue dots = Afp, green dots = IAP, pink dots = L1 and black dots = Tex13.

7.3.3 Wild-Type Prediction

As a next step we used the estimated parameters from the KO data to predict the WT data. The models from Eq.(7.12)-(7.15) are referred to as *Models 1-4*. For the prediction, the notation (x, y) is used to refer to Model x for the Dnmt3a/b DKO (only Dnmt1 active) and Model y for the Dnmt1KO case (only Dnmt3a/b active). One instance of the prediction, for which Model 1 was used for both Dnmt1KO and Dnmt3a/b DKO, i.e. $(1, 1)$, are shown in Fig. 7.8. Note that all wild-type predictions yielded a very similar accuracy. We list the corresponding estimations for the parameters for an example of a single copy gene (Afp) and a repetitive element (L1) in Tab. 7.1. While the standard deviation of the estimated parameters for μ is always of the order 10^{-2} and for τ of order 10^{-3}, it is usually of order 10^{-2} for ψ_i. Depending on the model, locus and position, standard deviations up to order 10^{-1} may occur for the dependency parameters in a few

Tab. 7.1: Estimated parameters for the KO data and model $(1, 1)$ based on Eq. (7.12) for the loci Afp and L1 with sample size n.

KO	μ	ψ_L	ψ_R	τ	n	Locus
Dnmt1	0.452 ± 0.062	0.383 ± 0.076	1.000 ± 0.094	0.091 ± 0.016	134	Afp
Dnmt3a/b	0.990 ± 0.003	0.984 ± 0.011	1.000 ± 0.006	$10^{-10} \pm 0.011$	186	Afp
Dnmt1	0.334 ± 0.051	0.576 ± 0.067	1.000 ± 0.122	0.038 ± 0.004	1047	L1
Dnmt3a/b	0.789 ± 0.037	1.000 ± 0.038	0.984 ± 0.045	$10^{-10} \pm 0.002$	805	L1

cases.

In Fig. 7.8 the predictions for the pattern distribution together with the WT pattern distribution and a prediction from the neighborhood independent model ($\psi_L = \psi_R = 1$) for all loci are shown in the main plot. As an inset the distributions are shown on a smaller scale to display small deviations. With the exception of patterns 1 and 64 (which corresponds to no methylation/full methylation of all sites) in L1 and pattern 64 in all loci, where the difference between WT and the numerical solution is about 10%, the difference is always small ($< 5\%$) as seen in the insets. In order to compare the performance of the neighborhood dependent and neighborhood independent model, we compute Kullback-Leibler divergence

$$KL = \sum_{j=1}^{4^L} \pi_j(\mathrm{WT}) \log\left(\frac{\pi_j(\mathrm{WT})}{\pi_j(\mathrm{pred})}\right) \tag{7.23}$$

for both cases and each locus and list the results in Tab. 7.2. The mean and standard deviation were obtained via bootstrapping of the wild-type data (10.000 bootstrap samples). The results show that the mean of KL as well as its standard deviation are always smaller for the neighborhood dependent model, i.e. the neighborhood dependent model yields the more accurate predictions.

For the 16 proposed models from Eq. (7.17) we observe a similar performance for all loci and positions in terms of accuracy of the prediction. On the large scale the differences are not visible and even for the smaller scale the differences are small. We therefore only show two examples for mSat in Fig. 7.9. By comparing KL that we list in Tab. 7.3, the similar performance of all 16 models can clearly be seen. The difference in KL between the "best" and the "worst" case is about 0.01. Again, the mean and standard deviation for KL were obtained via bootstrapping of the wild-type data (10.000 bootstrap samples for each model). Since no confidence intervals of the parameters are included, this standard deviation can be regarded as a lower bound. However, even with these lower bounds the intervals of KL overlap for all models, such that no model can be favorized.

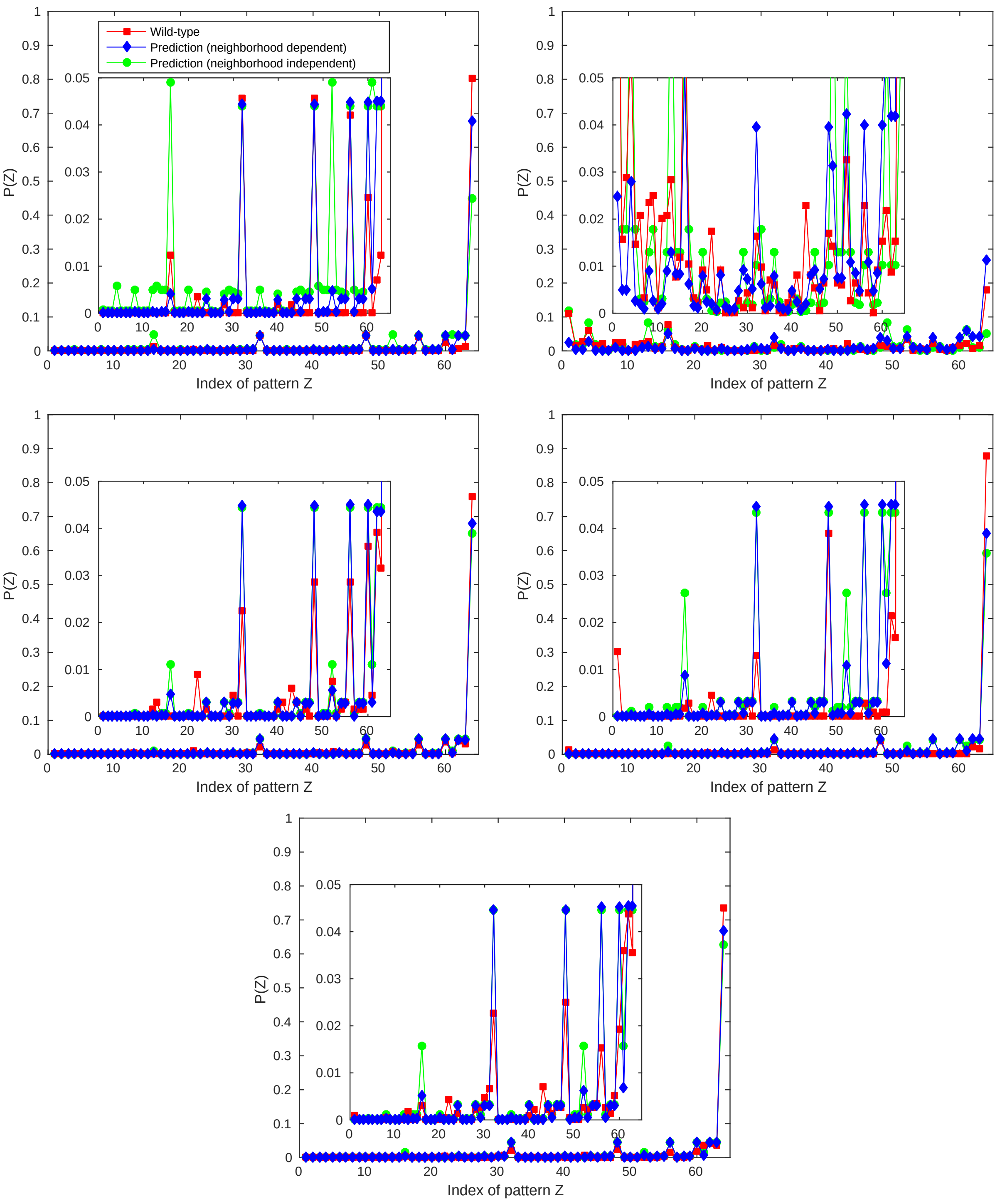
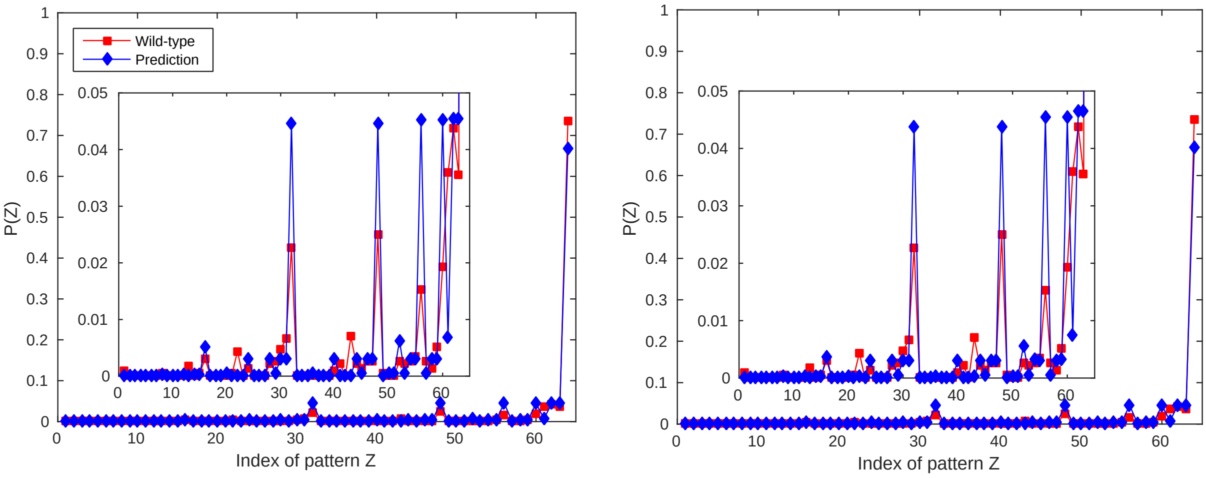


Fig. 7.8: The figures show an example for the predicted (neighborhood dependent and neighborhood independent) and the measured pattern distribution for each locus. The inset shows a zoomed in version of the distribution.

7.3.4 Non-Hairpin Data

So far we restricted the usage of the model to hairpin data, i.e. for one DNA molecule the methylation state of both strands is measured. For non-hairpin data there is only knowledge available for each strand independently. The information which strands stem



(a) (1,1)

(b) (4,4)

Fig. 7.9: The figures show the predicted and the measured pattern distribution for two, (1,1) and (4,4), of the 16 models for mSat. The inset shows a zoomed in version of the distribution. The red WT distribution is the same in both plots. Note the slight differences in both predictions for example in pattern 16, 62 and 63.

Tab. 7.2: Kullback-Leibler divergence *KL* for the neighborhood dependent and independent predictions at all loci.

Locus	Afp	L1	IAP	Tex13	mSat
KL_{dep}	0.6820 ± 0.0914	0.5342 ± 0.0638	0.3615 ± 0.0482	1.3364 ± 0.3235	0.1398 ± 0.0134
KL_{ind}	3.3557 ± 0.0979	0.5639 ± 0.0771	0.5390 ± 0.0602	2.0120 ± 0.3637	0.2582 ± 0.0286

from the same chromosome is not known. However, it is possible to compute the product of the likelihood of the individual strand patterns, which resembles the likelihood of real hairpin data (assuming independence). Our results show that this approach works well as long as the states of the opposite strand do not determine the transition probabilities, which is the case for Dnmt1KO data, since Dnmt3a/b shows only little maintenance activity. Since Dnmt1's main activity is maintenance, we indeed found that the WT and Dnmt3a/b DKO data does not yield good results (results not shown).

To compare the performance of the model for hairpin and non-hairpin data, we split the original hairpin data in upper and lower strand and computed the product of likelihoods for the patterns using the independence assumption. We then estimated the parameters via MLE with our model and the computed distributions. We found that for Dnmt3a/b the results are very close to the original hairpin data in terms of dependency parameter ψ_L and ψ_R, since in the model definition these parameters rely only on information on the same strand. No information from the opposite strand influences the dependency parameters. The ratio $R = \mu/\tau$ is usually smaller, i.e. the maintenance is under- and the *de novo* activity overestimated, for the non-hairpin data as shown in Fig. 7.10. However, this does not lead to contradictory results since maintenance and *de novo* methylation can not be distinguished by the model if the CpG on the opposite strand is methylated.

Tab. 7.3: Kullback-Leibler divergence KL for all 16 models.

Model	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$
KL	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1337 ± 0.0127
Model	$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$
KL	0.1438 ± 0.0137	0.1439 ± 0.0136	0.1439 ± 0.0137	0.1374 ± 0.0133
Model	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$
KL	0.1399 ± 0.0134	0.1399 ± 0.0134	0.1398 ± 0.0133	0.1337 ± 0.0127
Model	$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$
KL	0.1410 ± 0.0137	0.1411 ± 0.0136	0.1409 ± 0.0135	0.1349 ± 0.0130

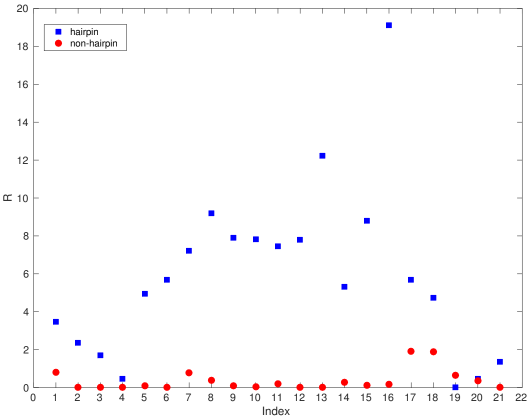


Fig. 7.10: Ratio $R = \mu/\tau$ between maintenance and *de novo* rate for hairpin (blue) and non-hairpin data (red) for all loci. The loci are mapped to the indices as follows: mSat:1, Afp:2–4, IAP:5–8, L1:9–13, Tex13:14–21.

7.3.5 Genome-wide Data

Due to the limited amount of CpGs for the experiments in the previous sections, we also considered genome-wide hairpin data obtained from mouse embryonic stem cells to substantially increase the number of measured CpGs and hence also the number of possible distances between adjacent CpGs. In the genome-wide data the methylation state of the CpGs were recorded in windows of approximately 150 bps for a subset of CpGs, such that there is information available for about 4 million CpGs of the entire genome. The data contains the methylation state of each CpG and the position on the DNA, from which the distance between adjacent CpGs can be derived. For our analysis, we only consider CpGs within the sames read i.e. in the 150 bp window. This last information is of great importance since we want to investigate the neighborhood dependency and have to ensure that the three adjacent CpGs stem from the same DNA molecule. Therefore the data is filtered such that we omit all CpGs which do not form a sequence of at least three consecutive CpGs within one read. Note that we do not consider all cases where either only one or two CpGs were covered in the measurement window or because of missing CpGs the consecutive sequence is split in chunks of two CpGs or smaller. Furthermore we only considered CpG triples for which at least 64 (i.e. the number of possible patterns) measurements were taken. After applying these constrains there are 3,489 CpG triples left.

Since only WT data (and no KO data) was available for the whole genome, we had to use a modified version of the parameter estimation based on Eq. (7.17), which contains eight parameters (four for each enzyme). In order to reduce the model complexity we use the observations from the previous experiments, namely that only Dnmt3a/b shows a dependency to the left, and we therefore set the remaining dependency parameters $\psi_L^{(1)}$, $\psi_R^{(1)}$ and $\psi_R^{(3a/b)}$ to 1. The conversion errors for the data set are $c = 0.996$ and $d = 0.93$. The conversion rates are derived from short synthetic DNA fragments containing different cytosine forms at definite positions. These oligos become part of the hairpin bisulfite library and therefore undergo the same treatment as the stem cell DNA. Thus, after sequencing, we can determine the conversion rate of C and 5mC independently of our biological sample.

Despite considering only CpG triples with a coverage of at least 64, in general the coverage is pretty low compared to the hairpin data used for the parameter estimation in the previous section. We therefore employ Bayesian inference rather than MLE for the parameter estimation in the genome-wide data. We use a Metropolis Hastings algorithm with the estimations from ML as starting points and a Gaussian proposal distribution with mean 0 and a standard deviation of 0.01 such that on average 40% of the 5000 total trials per CpG triplet are accepted for the posterior distribution. Afterwards a variant of the k-means algorithm is applied, which also considers standard deviations of the quantities that should be clustered [19]. Note that in order to avoid a domination by the much larger distances in the clustering, the distance is normalized before the algorithm is applied. The ideal number of clusters is chosen by minimizing the Davis-Bouldin index [20], which is defined as the ratio between cluster separation and similarity within the clusters. The results of the parameter estimation and the clustering is shown in Fig. 7.11. Note that the clustering is based on dependency parameter and distance only. The methylation state is not an input of the clustering algorithm.

In our results the methylation state of a CpG shows a strong dependency on the methylation state of the left neighbor even for distances up to 70 bps. We therefore conclude that the independence starts at much larger distances. Note that due to the restriction that the three CpGs have to be within the same 150 bps window during the measurement, even for the genome-wide data the distances between the CpGs are rather short. It is therefore not possible with the current data and measurement techniques to check hypotheses such as the independency of neighboring CpGs for large distances. Nevertheless, we see distinct methylation profiles for the three individual clusters as shown in Fig. 7.12. First, we analysed the frequencies of the four methylation states of each cluster as displayed in Fig 7.12 (a). CpGs in cluster 0 show low frequencies of fully- or hemimethylated

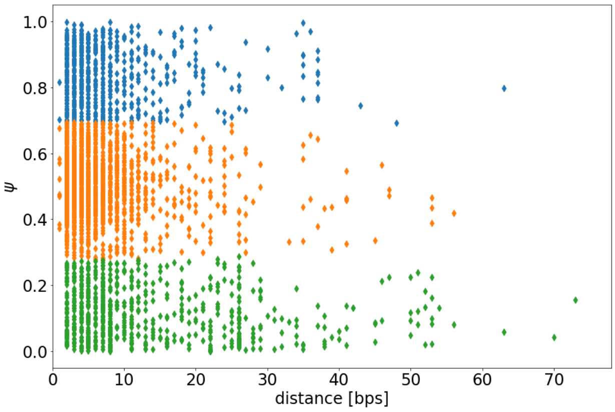
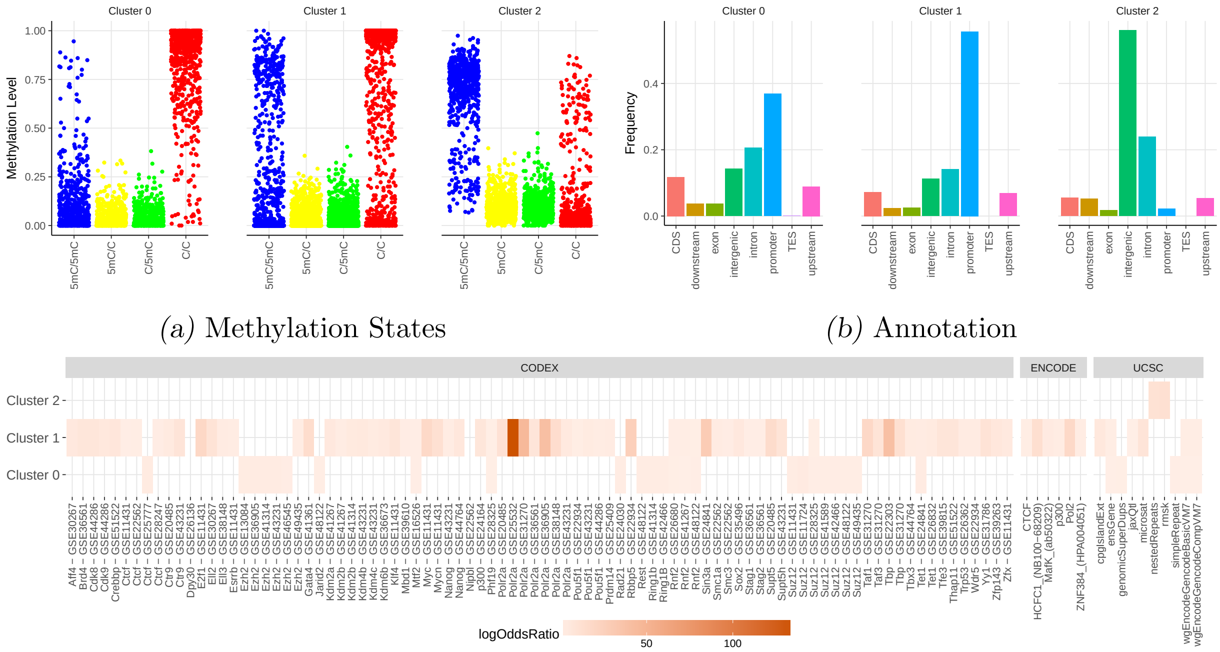


Fig. 7.11: Dependency parameter versus distance between CpGs for the genome-wide data. The three colors represent three clusters. Cluster 0: blue, cluster 1: orange, cluster 2: green.

states and in general appear to be unmethylated. Cluster 2 exhibits an inverse behavior compared to cluster 0, meaning that CpGs are more often found in a fully methylated state. Lastly, cluster 1 displays a bimodal distribution of fully- and unmethylated states but similar frequencies in 5mC/C and C/5mC. In other words, unmethylated CpGs seem to show less dependency compared to methylated ones. Furthermore, the CpG of these three individual clusters differ also in their genomic localization. Whereas most of the CpGs in cluster 2 are located in introns or intergenic regions, the majority of CpGs in cluster 0 and cluster 1 are found at promoters (Fig. 7.12 (b)). In addition, we conducted an enrichment analysis of transcription factors using the recently developed R package LOLA (Fig. 7.12 (c)) [21]. We found strong enrichment of cluster 2 CpGs at transcription factor binding sites (TFBS) including Pol2 and Polr2a pointing towards a relation of active transcription. Taken together, our findings suggest that hypomethylated CpGs at promoters and TFBS behave more independently. One possible explanation would be the constant setting (most likely by Dnmt3a/b) and removal of CpG methylation at these regions, which would point towards a constant turn over of 5mC. However, a more detailed analysis is needed to address this question.

7.4 Related Work

In [18] location- and neighbor-dependent models are proposed for single-stranded DNA methylation data in blood and tumor cells. The (de-)methylation rates depend on the position of the CpG relative to the 3' or 5' end and/or on the methylation state of the left neighbor only. The dependency is realized by the introduction of an additional parameter. In our proposed models we use double-stranded DNA and can therefore include hemimethylated sites and even distinguish on which strand the site is methylated. Furthermore we allow dependencies on both neighbors by introducing two different dependency param-



(a) Methylation States

(b) Annotation

(c) LOLA

Fig. 7.12: Biological context of CpG clustering. **(a)** Frequency CpG methylation state; states are indicated as follows: state 0 = C/C - red, state 1 = 5mC/C - yellow, state 2 = c/5mC - green, state 3 = 5mC/5mC - blue. **(b)** Frequency of annotated genomic features within the individual clusters. **(c)** Result of LOLA enrichment analysis against transcription factors from CODEX or ENCODE and UCSC features. All depicted enrichments possess a q-value above 0.05.

eters. In contrast [22] copes with the neighborhood dependency indirectly by allowing different parameter values for different sites. In order to reduce the dimensionality of the parameter vector, a hierarchical model based on beta distributions is proposed. Another difference to our model is the distinction between *de novo* rates for parent and daughter strand. However, this can easily be included in future work. A density-dependent Markov model was proposed [23]. In this model, the probabilities of (de-)methylation events may depend on the methylation density in the CpG neighborhood. In addition, a neighboring sites model has been developed, in which the probabilities for a given site are directly influenced by the states of neighboring sites to the left and right [23]. When these models were tested on double-stranded methylation patterns from two distinct tandem repeat regions in a collection of ovarian carcinomas, the density-dependent and neighboring sites models were superior to independent models in generating statistically similar samples. Although this model also includes the dependence on the methylation state on the left and right neighbor for double-stranded DNA the approach is different. The transition probabilities of the neighbor-independent model are transformed into a transition probability of a neighbor-dependent model by introducing only one additional parameter. The state of the left and right neighbor are taken into account by exponentiating this parameter by some norm. In addition, this approach does not allow the intuitive interpretation of the dependency parameter. Recently the model from [23] was extended to include the influence of different distances between the CpGs [24]. However this model is still restricted to single-stranded methylation data. In [25] it has been shown that the collaboration between CpG sites is required to obtain stable fractions of methylation states over time in CpG islands. In this model another nearby CpG serves as a mediator such that its state influences the possible reactions. In a more recent version of this model the distance to the mediator CpG is taken into account [26]. However, both models feature active demethylation, have no explicit dependency parameter and do not distinguish between the two different hemimethylated states.

7.5 Conclusion

We proposed a set of stochastic models for the formation and modification of methylation patterns over time. These models take into account the state of the CpG sites in the spatial neighborhood and allow to describe different hypotheses about the underlying mechanisms of methyltransferases adding methyl groups at CpG sites. We used knockout data from bisulfite sequencing at several loci to learn the efficiencies at which these enzymes perform methylation. By combining these efficiencies, we accurately predicted the probability distribution of the patterns in the wild-type. Moreover, we found that in all cases the models predict values for the dependency parameters ψ_L and ψ_R close to 1 and therefore independence of methylation for the Dnmt3a/b DKO meaning that Dnmt1 methylates CpGs independent of the methylation of neighboring CpGs. For Dnmt3a/b on the other hand we could identify dependencies on the neighboring CpGs. Both findings are in accordance with current existing mechanistic models: Dnmt1 reliably copies the methylation from the template strand to maintain the distinct methylation patterns, whereas Dnmt3a/b try to establish and keep a certain amount of CpG methylation at a given loci. Interestingly, our models only suggest dependencies of *de novo* methylation activity on the CpGs in the 5' neighborhood. This indicates that Dnmt3a and Dnmt3b show a preference to methylate CpGs in a 5' to 3' direction and could point towards a processive or cooperative behavior of these enzymes like recently described in *in vitro* experiments [14, 11]. Our results indicate that, at least for small distances, rather the genetic region than the distance determines the dependence on the neighbors. Compared to a neighborhood independent model with $\psi_L = \psi_R = 1$, a neighborhood dependent model shows better predictions and furthermore allows to investigate (possible) connections of adjacent CpGs and their methylation states. As long as no information from the opposite strand is needed, i.e. if maintenance activity is not too high, as in the Dnmt1KO data, our model can also be used for non-hairpin data. Applying our model at genome-wide data reveals distinct dependency clusters with individual methylation patterns. We finde, that hypomethylated CpGs at promoter and TFBS are more likely to behave independent of their neighborhood compared to hypermethylated CpGs.

As future work, we plan to investigate models in which we distinguish between the actions of Dnmt3a and Dnmt3b and in which we allow a diagonal dependency for *de novo* methylation, i.e., a dependency on the state of neighboring CpGs on the opposite strand. Furthermore, we intend to explicitely include the actual distance of neighboring CpGs in our model by making the dependency parameters distance dependent. This also eases the modelling of more than three CpGs since we then do not longer assume the same dependency parameters for all CpGs and therefore make the model more flexible. To investigate a potential impact of oxidized cytosine forms on the methylation at neighboring CpG sites we further plan to include the CpG states 5hmC, 5fC and 5caC in our model.

Acknowledgements

This research has been partially funded by the German Research Council (DFG) as part of the Cluster of Excellence on Multimodal Computing and Interaction and as part of the Collaborative Research Centre (SFB) 1027 at Saarland University.

Author Contributions

Conceived and designed the experiments: VW JW. Performed the experiments: PG. Analyzed the data: AL PG. Wrote the paper: AL PG. Designed/implemented the software used in analysis: AL.

Bibliography

[1] Miho M Suzuki and Adrian Bird. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6):465–476, 2008.

[2] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412, 2016.

[3] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *PNAS*, 101(1):204–209, 2004.

[4] Tarmo Äijö, Yun Huang, Henrik Mannerström, Lukas Chavez, Ageliki Tsagaratou, Anjana Rao, and Harri Lähdesmäki. A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biology*, 17(1):49, 2016.

[5] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The Influence of Hydroxylation on Maintaining CpG Methylation Patterns: A Hidden Markov Model Approach. *PLoS Comput Biol*, 12(5):e1004905, 2016.

[6] Charalampos Kyriakopoulos, Pascal Giehr, and Verena Wolf. H(O)TA: estimation of DNA methylation and hydroxylation levels and efficiencies from time course data. *Bioinformatics*, 33(11):1733–1734, 2017.

[7] Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The Dnmt1 DNA-(cytosine-c5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 279(46):48350–48359, 2004.

[8] Masaki Okano, Daphne W Bell, Daniel A Haber, and En Li. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257, 1999.

[9] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet*, 8(6):e1002750, 2012.

[10] Humaira Gowher and Albert Jeltsch. Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. *Journal of Biological Chemistry*, 277(23):20409–20414, 2002.

[11] Celeste Holz-Schietinger and Norbert O Reich. The inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L. *Journal of Biological Chemistry*, 285(38):29091–29100, 2010.

[12] Allison B Norvil, Christopher J Petell, Lama Alabdi, Lanchen Wu, Sandra Rossie, and Humaira Gowher. Dnmt3b Methylates DNA by a Noncooperative Mechanism, and Its Activity Is Unaffected by Manipulations at the Predicted Dimer Interface. *Biochemistry*, 2016.

[13] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546):243–247, 2015.

[14] Max Emperle, Arumugam Rajavelu, Richard Reinhardt, Renata Z Jurkowska, and Albert Jeltsch. Cooperative DNA binding and protein/DNA fiber formation increases the activity of the Dnmt3a DNA methyltransferase. *Journal of Biological Chemistry*, 289(43):29602–29613, 2014.

[15] Sarah P Otto and Virginia Walbot. DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. *Genetics*, 124(2):429–437, 1990.

[16] Diane P Genereux, Brooks E Miner, Carl T Bergstrom, and Charles D Laird. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *PNAS*, 102(16):5802–5807, 2005.

[17] Laura B Sontag, Matthew C Lorincz, and E Georg Luebeck. Dynamics, stability and inheritance of somatic DNA methylation imprints. *Journal of Theoretical Biology*, 242(4):890–899, 2006.

[18] Nicolas Bonello, James Sampson, John Burn, Ian J Wilson, Gail McGrown, Geoff P Margison, Mary Thorncroft, Philip Crossbie, Andrew C Povey, Mauro Santibanez-Koref, et al. Bayesian inference supports a location and neighbour-dependent model of DNA methylation propagation at the MGMT gene promoter in lung tumours. *Journal of Theoretical Biology*, 336:87–95, 2013.

[19] Mahesh Kumar and Nitin R. Patel. Clustering data with measurement errors. *Computational Statistics & Data Analysis*, 51(12):6084 – 6101, 2007.

[20] D. L. Davies and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[21] Nathan C Sheffield and Christoph Bock. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*, 32(4):587–589, 2015.

[22] Audrey Qiuyan Fu, Diane P Genereux, Reinhard Stöger, Charles D Laird, and Matthew Stephens. Statistical inference of transmission fidelity of DNA methylation patterns over somatic cell divisions in mammals. *The Annals of Applied Statistics*, 4(2):871, 2010.

[23] Michelle R Lacey, Melanie Ehrlich, et al. Modeling dependence in methylation patterns with application to ovarian carcinomas. *Stat Appl Genet Mol Biol*, 8(1):40, 2009.

[24] Karlene Nicole Meyer and Michelle Lacey. Modeling Methylation Patterns with Long Read Sequencing Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2017.

[25] Jan O Haerter, Cecilia Lövkvist, Ian B Dodd, and Kim Sneppen. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. *Nucleic Acids Research*, 42(4):2235–2244, 2013.

[26] Cecilia Lövkvist, Ian B Dodd, Kim Sneppen, and Jan O Haerter. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Research*, 44(11):5123–5132, 2016.

8. DISCUSSION

The here summarised studies approach the activity and cooperation of Dnmts and Tets in generating, sustaining and changing distinct DNA methylation patterns. Hence, these studies include experimental and computational strategies to match the requirements in addressing complex mechanistic questions. Generally, this thesis can be divided into two main parts. First, the assessment of Tet and 5hmC influence on Dnmt methylation processes and second, the investigation of how existing methylation patterns modulate Dnmt activity. The following chapter will highlight the essential aspects of this cumulative work, discuss biological findings and eventually provide an outlook for possible follow-up studies.

8.1 Combining Hairpin Sequencing and Hidden Markov Models

The presented studies are based on HPBS and the subsequent application of HMMs, which allows to estimate the efficiency of DNA modifying enzymes [1, 2, 3, 4]. Previously, this combinatorial approach has been used by Arand *et al.* to determine the contribution of distinct Dnmts in generating and sustaining stable methylation patterns [3]. Compared to classical sequencing strategies, the advantage of HPBS is the detection of double strand DNA methylation patterns of individual DNA molecules, which is used to distinguish between *de novo* and maintenance methylation events in stochastic modelling.

8.1.1 Detecting 5hmC and Tet Hydroxylation Efficiency

Following the protocols described in Arand *et al.* and Ficz *et al.*, HPBS was combined with oxBS for the simultaneous, strand specific detection of 5mC and 5hmC (see Chapter 3 and Chapter 4) [5, 6]. The usage of particular hairpin linker containing C, 5mC and 5hmC provides a per molecule control sequence and permits the accurate estimation of conversion rates after BS and oxBS. Accordingly, two linked HMMs were constructed describing the evolution of BS and oxBS methylation patterns in the context DNA replication (cell division), DNA methylation and hydroxylation, respectively. Several studies, applied oxBS or TAB-Seq for the investigation of 5hmC, however, none provides a strand specific detection of 5hmC. Furthermore, even though these studies determine the conversion rates during the chemical treatment, these rates were not used in later estimation of 5mC and 5hmC levels [7, 8, 9, 10]. Thus, the pipeline presented in Chapter 3 and Chapter 4, provides a much more accurate estimation of methylation, or hydroxymethylation patterns.

Next, in order to capture 5mC and 5hmC patterns in a more global context, a genome wide hairpin sequencing protocol was developed (Chapter 5). A first genome wide approach has been published by Zhao *et al.* [11]. However, the pipeline published by Zhao *et al.* demands large amounts of DNA (10µg), which in general makes the application challenging, particularly in the case of primary biological or clinical samples. Furthermore, the method covers the entire genome, which raises the sequencing costs substantially and, furthermore, reduces the coverage for each CpG position obtained after sequencing and lastly, is not able to capture 5hmC. Thus, as part of this thesis, ‘reduced representation hairpin oxidaive bisulfite sequencing’ (RRHPoxBS) has been developed Chapter 5. RRHPoxBS requires only 1.2µg input material and covers a subset of about 4 million CpGs equally distributed across the genome. Hence, the method provides the possibility for global detection of 5mC and 5hmC with reduced sequencing costs and a higher coverage compared to the previously described protocol, suited for stochastic modelling.

8.1.2 A Pipeline for the Analysis of Spatial Methylation Patterns

In a second approach, presented in Chapter 6 and 7, this thesis investigates the formation of spatial methylation patterns. Compared to the previously described model, the spatial HMM makes use of the methylation patterns derived from HPBS and describes the modification state of three neighbouring CpG positions instead of just only one CpG. The additional information is used to determine a possible dependency of adjacent CpGs in relation to their methylation state.

Previous investigations of spatial methylation patterns were performed on very few genomic regions and often rely on single strand methylation patterns derived from classical BS [12, 13, 14]. In contrast, the here presented studies in Chapter 6 and 7 are based solely on complementary methylation patterns obtained from HPBS. Therefore, the model can distinguish more accurately between maintenance and *de novo* methylation events compared to previous published models. Moreover, in a first instance, the model is applied to data derived from WT, Dnmt1 KO, as well as Dnmt3a/3b DKO ES cells published by Arand *et al.*, which allows for the first time the prediction of enzyme specific dependencies for distinct Dnmts [3]. The genome wide HPBS data is then used to investigate the dependency of neighbouring CpG methylation in a global manner and provides more diversity in terms of sequence context, including CpG density, distance of CpGs and methylation state. Lastly, the use of two dependency parameters permits to distinguish between dependency towards the ‘left’ and ‘right’ neighbouring CpG (Chapter 6).

8.2 Further Applications of RRHPBS

Classical BS approaches identify DNA methylation of CpGs and nonCpGs by aligning the obtained sequencing read to a reference sequence. Hence, the comparison of C positions within the reference sequence and the corresponding read position will determine the methylation state of each read. A methylated C will be displayed as a C, also in the read sequence, while an unmethylated C will be seen as T (Figure 8.1 A).

In contrast, HPBS does not require a reference sequence for methylation calling. After sequencing, the obtained sequence of the lower DNA strand is aligned against the sequence of the upper strand. The comparison of both sequences can be used to reconstruct the genomic sequence, as well as for direct identification of the methylation state (8.1 B).

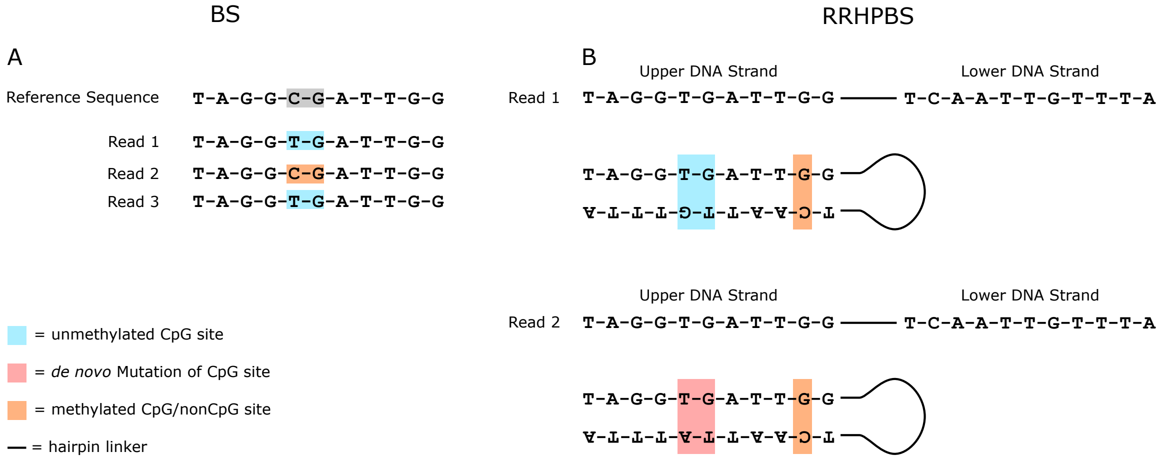

BS
RRHPBS
A
Reference Sequence T-A-G-G-C-G-A-T-T-G-G
Read 1 T-A-G-G-T-G-A-T-T-G-G
Read 2 T-A-G-G-C-G-A-T-T-G-G
Read 3 T-A-G-G-T-G-A-T-T-G-G
B
Upper DNA Strand
Lower DNA Strand
Read 1 T-A-G-G-T-G-A-T-T-G-G T-C-A-A-T-T-G-T-T-T-A
T-A-G-G-T-G-A-T-T-G-G
T-C-A-A-T-T-G-T-T-T-A
Upper DNA Strand
Lower DNA Strand
Read 2 T-A-G-G-T-G-A-T-T-G-G T-C-A-A-T-T-G-T-T-T-A
T-A-G-G-T-G-A-T-T-G-G
T-C-A-A-T-T-A-T-T-T-A
= unmethylated CpG site
= de novo Mutation of CpG site
= methylated CpG/nonCpG site
= hairpin linker

Fig. 8.1: Schematic representation of DNA methylation calling of common BS, as well as HPBS pipelines. **(A)** BS approach, methylation state of CpGs are identified by aligning the sequencing reads against the reference sequence. **(B)** RRHPoxBS, upper and lower DNA strand are first aligned against each other, which permits to derive the genomic sequence as well as the methylation state of CpGs and nonCpG without the use of a reference sequence. Read 1, identification of an unmethylated CpG site (blue), as well as a methylated nonCpG position (orange). Read 2, identification of SNP or *de novo* mutation at a former CpG site (red), as well as a methylated nonCpG position (orange).

The reference sequence free methylation calling of RRHPBS represents a clear advantages compared to common BS techniques. First, an impact of genetic variation due to distinct genetic backgrounds of reference genome and sample DNA are minimised, which permits a more precise estimation of DNA methylation and, furthermore, an explicit discrimination between CpG and nonCpG positions. Second, having the sequence information of both complementary DNA strands permits a more secure identification of SNPs at a given position. Moreover, deriving the genomic sequence directly from the sample DNA prevents the impact of genetic variations within the reference sequence and permits the identification of rare *de novo* mutation events.

8.3 Opposed Efficiencies of Dnmts and Tets form the Methylome

A cell's methylome consists of alternating methylated and un (UMRs) or lowly (LMRs) methylated domains [15, 16, 17]. In this context, the methylation state often corresponds to the binding of Dnmts and Tets, i.e. unmethylated domains are occupied by Tets, while methylated domains are bound by Dnmts [18, 19] (Chapter 5, Figure 5.4 and Chapter 5, Supplement, Figure S5.41 and S5.42). Yet, the here presented studies describe the first approach to accurately estimate the activity of Dnmts and Tets for single CpGs also in a genome wide manner.

Generally, the estimated efficiencies show that binding and activity of Dnmts and Tets are not exclusive. In other words, Dnmts, as well as Tets bind and act notably at both unmethylated and methylated regions. Nevertheless, Dnmts and Tets display opposed efficiency profiles (Chapter 5 Figure 5.2 and Figure 5.6). UMRs and LMRs exhibit high hydroxylation paired with low maintenance and *de novo* methylation efficiencies, while PMDs and HMRs display a reversed behaviour (Chapter 5 Figure 5.8 and 5.10). Hence, these observations correspond essentially to previously described ChIP profiles of Dnmts and Tets and demonstrate the accuracy of our approach in estimating enzyme efficiencies (Chapter Figure 5-5.2 and Chapter 5, Supplement, S5.41 and S5.42).

8.4 Tets Enhance DNA Demethylation During the Serum-to-2i Shift

The epigenetic adaptation of mouse ES cells to 2i containing medium after long term cultivation under Serum/LIF conditions has been the subject of several studies [20, 21, 22, 23]. In this context, the impairment of maintenance methylation, due to the reduction of Uhrf1, as well as H3K9 di-methylation, has been identified as the main cause of DNA demethylation [23]. Additionally, Dnmt3a and 3b undergo gradual reduction in ES cells after incubation in 2i medium, which further reduces the methylation activity [20]. Besides, contribution of Tet enzymes in DNA demethylation appears to be restricted to particular genomic regions [20, 23].

The data presented in Chapters 3 - 5 mostly agree with the previous observations. Upon cultivation of ES cells in 2i medium, the model estimates a reduced maintenance activity comparable to the description by von Meyenn *et al.* [23]. Furthermore, the reduced maintenance efficiency shows no further decrease and remains stable in 2i. This indicates that the reduction of Uhrf1 and H3K9me2 is an early event and in addition suggests that the level of Uhrf1 and H3K4me3 attain a new steady state within the first 24 hours. Moreover, the pipeline also identifies a gradual decrease of *de novo* methylation efficiency which is in agreement with the successive loss of Dnmt3a and 3b [20]. Thus, these observations indicate once more the accuracy of the pipeline in describing the complex underlying molecular changes.

However, the presented data indicate a notable contribution of Tet enzymes in DNA demethylation. The sequence specific analysis suggest that regions with high 5hmC levels are more likely to lose 5mC. More evidence comes from the parameter p, which determines if 5hmC will be recognised by the maintenance machinery (Chapter 3, Section 3.3). In both cases, local and genome wide estimations, the model clearly favours a scenario in which 5hmC leads to an inhibition of maintenance methylation. Furthermore, comparison of DNA demethylation rates of WT and Tet TKO ES cells show strongly reduced demethyltion rates in the absence of Tets, demonstrating the importance of 5hmC and Tets for efficient demethylation Chapter 5, Supplement Figure S5.35.

8.5 Tet Hydroxylation Supports ES Cell Self Renewal and Differentiation

Based on their enzyme efficiency signature, CpGs can be divided into two main categories (Chapter 5, Figure 5.10): CpGs with high methylation efficiency (both maintenance and *de novo*) and low hydroxylation efficiency, as well as CpGs displaying low methylation efficiency and high hydroxylation efficiency. The latter show considerably lower levels of DNA methylation and are predominately located at promoters and transcription factor binding sites (TFBS) (Chapter 5, Figure 5.10 D). Identified TFBS include binding sites of known pluripotency markers, such as Nanog, Oct4 (Pou5f1), as well as Sox2 and furthermore, regions of proteins involved in ES cell self renewal, e.g. Stat3 or Stat5b. The strongest enrichment of CpGs exhibiting high hydroxylation efficiency can be found at Dpy30 binding sites. As a subunit of the SET1/MLL histone methyltransferase complex, Dpy30 is involved in the generation of H3K4me3 [24, 25]. Dpy30 is not involved in stem cell self renewal, instead it appears to promote ES cell differentiation by regulation of bivalent promoters [26]. Previous studies already express a dual role of Tet enzymes in ES cells. Tet1, for example, binds to active as well as bivalent promoters and its knock down results in partial loss of stem cell identity [27, 28, 29]. In contrast, KO of multiple Tet enzymes prevents ES cell differentiation [30, 31]. The above discussed results support a dual function of Tet enzymes in ES cells and demonstrate that catalytic active Tet enzymes on the one hand, support stem cell self renewal and on the other hand prepare stem cell differentiation by modulating bivalent promoters.

8.6 Tet Enzymes Protect from DNA Methylation Spreading

Comparison of WT and Tet TKO efficiency profiles demonstrates a notable misregulation of both maintenance and *de novo* methylation efficiency in the absence of Tets. Explicitly, the model describes an increase of maintenance efficiency across UMRs and LMRs, as well as, a global increase of *de novo* methylation efficiency at late 2i stages (Chapter 5, Figure

5.11, Figure 5.6 and Figure 5.5).

The increased maintenance efficiency can not simply be explained by the absence of 5hmC. Any inhibitory effect of 5hmC towards maintenance efficiency is captured by the parameter p. Moreover, notable levels of 5hmC in WT ES cells are not only detected at promoters and enhancers, but also across the gene body. Accordingly, loss of 5hmC would rather cause a global increase in maintenance methylation efficiency. Hence, the observed increase in Tet TKO cells must be a result of missing Tet proteins. One possible explanation is that in WT ES cells, the high occupancy of Tets at UMRs/LMRs shields the DNA from binding of the maintenance machinery, while the absence of Tet provides more accessibility for Dnmts and proteins involved in maintenance methylation.

The almost stable *de novo* methylation efficiency is surprising, considering the down regulation of Dnmt3a and 3b in WT ES cells cultivated in 2i containing medium. However, the increased nonCpG methylation levels in Tet TKO cells further support the concept of remaining Dnmt3a/3b activity. Once again, this increase in *de novo* methylation could be the result of higher DNA accessibility for Dnmt3a and 3b due to the absence of Tets. However, it is not intuitive that the absence of Tet proteins would cause a global increase in DNA accessibility for Dnmt3a and 3b, while in the case of proteins involved in maintenance methylation this effect remains restricted to UMRS/LMRs.

The global increase in *de novo* methylation efficiency in Tet TKO ES cells could also be explained by an elevated expression of Dnmt3a or 3b. A study by Freudenberg *et al.* describes an increased expression of Dnmt3b in ES cells after knock down of Tet1 [29]. Yet, RNA or protein measurements would be required to confirm up-regulation of Dnmt3a or 3b in Tet TKO cells.

Tet1 protection against DNA methylation events, in particular at CpG islands, has been previously described [32]. However, the underlying mechanisms remained elusive. Taken together, the results obtained from RRBS of WT and Tet TKO ES cells in Chapter 5 suggest that Tet enzymes protect UMRs/LMRs against methylation spreading in ES cells in three ways. Firstly, the high hydroxylation efficiency at UMRs/LMRs ensures an instant conversion of 5mC, resulting in passive or active removal DNA methylation. Secondly, Tets inhibit the effectiveness of maintenance methylation across UMRs/LMRs and prevents the inheritance of ectopic methylation. Lastly, Tet enzymes ensure an efficient downregulation of *de novo* methyltransferases, at least in naive ES cells.

8.7 The Activity of Dnmt3a and 3b Depends on Neighbouring CpG Methylation

Novel epigenetic studies show that depending on the genomic location, the stable inheritance of methylation patterns relies on the combined activity of Dnmt1, as well as Dnmt3a and 3b [3, 33, 34]. In this context, the activity of Dnmts often relies on pre-existing DNA methylation. Dnmt1 for example is driven by the presence of hemi-methylated DNA after replication. However, there is evidence that the methylation state of CpGs influences the activity of Dnmts also at later stages [13, 14].

The application of a spatial HMM on HPBS data presented in Chapter 6 and 7 indeed identifies a dependence of CpGs in relation to their methylation state. However, the analysis of methylation patterns derived from Dnmt KO ES cells reveals that this dependency is enzyme specific. In the case of Dnmt1, the model finds no dependence of the methylation activity towards neighbouring CpGs. This observation can be explained by the biological function of Dnmt1 to reproduce methylation patterns after replication at the newly synthesised DNA strand. The influence of neighbouring CpGs would cause changes in methylation patterns with each replications. In contrast, Dnmt3a and 3b display a dependency towards CpG methylation states, however, the activity is only affected by the left (5') CpG position. In other words, Dnmt3a and 3b tend to methylate CpGs with higher efficiency when the previous CpG is methylated.

The application of the spatial HMM on genome wide HPBS data shows that unmethylated CpG positions exhibit less dependency on the neighbouring CpGs compared to methylated once. Moreover, genomic annotations of CpGs reveal that CpGs, which behave independent of each other, are located mainly at promoters. Note, that the model considers dependency in the case of both, a sequence of methylated, but also unmethylated CpGs. Such independence of unmethylated CpGs could for instance be explained by an infrequent *de novo* activity of Dnmt3a or 3b. In this context, the sporadic generation of 5mC might point towards a distributive, rather than a processive activity of Dnmt3a and 3b. A second possible scenario is a constant setting and removal of CpG methylation. Indeed, a recent publication by Rulands *et al.* describe oscillations of DNA methylation at promoters of mouse ES cells [35] The high hydroxylation efficiency at unmethylated regions observed in Chapter 5 further supports a constant turnover of generated DNA methylation. From the biological point of view, such a dynamic system can quickly respond to extracellular signals and switch from active to repressed gene expression or *vice versa.*

The dependency of CpGs might also be influenced by the distance of adjacent CpG sites. However, neither local HPBS nor RRHPBS provide enough variability to address this question. Detected distances between CpGs in RRHPBS lie within a range of 1 to 70bp. Within this range, no definite impact on the dependence of CpG position in relation to their methylation state can be identified.

8.8 Conclusion

The studies presented in this thesis provide robust experimental and computational strategies for the investigation of methylation patterns, as well as the underlying enzymatic activities of Dnmts and Tets.

Examination of complementary DNA methylation patterns from WT and Tet TKO ES cells demonstrate that Dnmts and Tets cooperatively create alternating methylated and unmethylated domains with clear boundaries across the genome. In this context, Tet enzymes protect unmethylated domains due to efficient conversion of 5mC to 5hmC, the inhibition of maintenance methylation, as well as the effective downregulation of *de novo* methylation. Moreover, regions exhibiting high Tet activity mainly present enhancer and gene promoters, which are essential for maintaining the stem cell phenotype.

The application of a spatial HMM on methylation patterns reveals that the *de novo* methylation efficiency of Dnmt3a and 3b is affected by the methylation state of the 5' neighbouring CpG position. However, a dependency of CpGs with respect to their methylation state is mainly restricted to highly methylated domains, while CpGs at unmethylated regions, particularly at promoters, behave independently of each other. This behaviour suggests either sporadic *de novo* methylation events at promoters or processive methylation followed by Tet oxidation and potential removal of 5mC.

8.9 Outlook

Besides the demethylating effect of 5hmC due to inhibition of maintenance methylation efficiency, higher oxidised cytosine forms, i.e. 5fC and 5caC, may also contribute to either active or passive DNA demethylation. In the present studies, 5fC and 5caC will be detected as T in BS and oxBS and are therefore not considered by the HMMs. However, several methods have been described, which permit the selective identification of oxCs, such as MAB-Seq, fCAB, CLEVER-Seq and combination of these methods with HPBS is straightforward [36, 37]. First experiments combining HPBS and MAB-Seq, as well as the expansion of the HMMs have already been performed (data not shown) and in the future will allow to determine the impact of oxCs towards methylation pattern formation.

In this respect, three distinct di-oxygenases, Tet1, Tet2 and Tet3 are responsible for the generation of oxCs and in ES cells, mainly Tet1 and Tet2 are expressed [38, 20, 23]. Similarly to the study by Arand *et al.*, in which Dnmt KO systems were analysed, the investigation of individual Tet KO ES cell lines would allow to separate the contribution of Tet1 and Tet2 in 5mC oxidation, as well as their impact on Dnmt methylation activity.

Recently, Xu & Corces published a computational pipeline, which reconstructs complementary DNA methylation patterns based on common whole genome bisulfite sequencing [39]. A combinatorial approach of this pipeline and the presented HMM in Chapter 5, would permit the usability on non hairpin approaches for the investigation of Dnmt and Tet efficiencies. During the last years, several consortia generated reference methylomes for multiple mouse and human cell types based on whole genome bisulfite sequencing (WGBS). Similarly, several WGBS data sets of Dnmt and Tet KO ES cells lines have been published in the past. Re-analysis of such existing data sets using a synergistic approach of the method by Xu & Corces and the HMM from Chapter 5 presents a great opportunity for large scale mechanistic studies regarding Dnmt and Tet activity [39]. Furthermore, common genome wide approaches often require less amounts of DNA and have even been applied to single cells. Thus, a synergistic approach of the method by Xu & Corces and the HMM in Chapter 5 would also permit the analysis of demanding samples with limited amount of DNA.

Lastly, novel sequencing techniques such as Nanopore or SMRT sequencing, provide the opportunity of bisulfite free detection of oxidised cytosine forms. Both technologies process single native DNA molecules without the need of prior amplification and are able to directly detect modified bases [40, 41, 42]. In addition, both methods provide much larger reads of more than 20kb. Such long reads are best suited for spatial methylation analysis and would provide the needed diversity to investigate the impact of CpG distance towards methylation dependencies of neighbouring CpGs. SMRT sequencing relies on the analysis of large circular DNA fragments, i.e. hairpin molecules and therefore naturally provides the methylation patterns of complementary DNA equally to RRHPBS. Exploiting this technique in the future will greatly benefit the investigation of oxidised cytosine variants and the understanding of molecular enzymatic mechanisms of Dnmts and Tets.

Bibliography

[1] Charles D Laird, Nicole D Pleasant, Aaron D Clark, Jessica L Sneeden, KM Anwarul Hassan, Nathan C Manley, Jay C Vary, Todd Morgan, R Scott Hansen, and Reinhard Stöger. Hairpin-bisulfite pcr: assessing epigenetic methylation patterns on complementary strands of individual dna molecules. *Proceedings of the National Academy of Sciences*, 101(1):204–209, 2004.

[2] Brooks E Miner, Reinhard J Stöger, Alice F Burden, Charles D Laird, and R Scott Hansen. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite pcr. *Nucleic acids research*, 32(17):e135–e135, 2004.

[3] Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8(6):e1002750, 2012.

[4] Pascal Giehr and Jörn Walter. Hairpin bisulfite sequencing: Synchronous methylation analysis on complementary dna strands of individual chromosomes. In *DNA Methylation Protocols*, pages 573–586. Springer, 2018.

[5] Pascal Giehr, Charalampos Kyriakopoulos, Gabriella Ficz, Verena Wolf, and Jörn Walter. The influence of hydroxylation on maintaining cpg methylation patterns: a hidden markov model approach. *PLoS computational biology*, 12(5):e1004905, 2016.

[6] Pascal Giehr, Charalampos Kyriakopoulos, Konstantin Lepikhov, Stefan Wallner, Verena Wolf, and Jörn Walter. Two are better than one: Hpoxbs-hairpin oxidative bisulfite sequencing. *Nucleic acids research*, 2018.

[7] Jianghan Qu, Meng Zhou, Qiang Song, Elizabeth E Hong, and Andrew D Smith. Mlml: consistent simultaneous estimates of dna methylation and hydroxymethylation. *Bioinformatics*, 29(20):2645–2646, 2013.

[8] Zongli Xu, Jack A Taylor, Yuet-Kin Leung, Shuk-Mei Ho, and Liang Niu. oxbsmle: an efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated dna. *Bioinformatics*, 32(23):3667–3669, 2016.

[9] Tarmo Äijö, Xiaojing Yue, Anjana Rao, and Harri Lähdesmäki. Luxglm: a probabilistic covariate model for quantification of dna methylation modifications with complex experimental designs. *Bioinformatics*, 32(17):i511–i519, 2016.

[10] Samara F Kiihl, Maria Jose Martinez-Garrido, Arce Domingo-Relloso, Jose Bermudez, and Maria Tellez-Plaza. Mlml2r: an r package for maximum likelihood

estimation of dna methylation and hydroxymethylation proportions. *Statistical applications in genetics and molecular biology*, 2019.

[11] Lei Zhao, Ming-an Sun, Zejuan Li, Xue Bai, Miao Yu, Min Wang, Liji Liang, Xiaojian Shao, Stephen Arnovitz, Qianfei Wang, et al. The dynamics of dna methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. *Genome research*, pages gr–163147, 2014.

[12] Diane P Genereux, Brooks E Miner, Carl T Bergstrom, and Charles D Laird. A population-epigenetic model to infer site-specific methylation rates from double-stranded dna methylation patterns. *Proceedings of the National Academy of Sciences*, 102(16):5802–5807, 2005.

[13] Audrey Qiuyan Fu, Diane P Genereux, Reinhard Stöger, Charles D Laird, and Matthew Stephens. Statistical inference of transmission fidelity of dna methylation patterns over somatic cell divisions in mammals. *The annals of applied statistics*, 4(2):871, 2010.

[14] Nicolas Bonello, James Sampson, John Burn, Ian J Wilson, Gail McGrown, Geoff P Margison, Mary Thorncroft, Philip Crossbie, Andrew C Povey, Mauro Santibanez-Koref, et al. Bayesian inference supports a location and neighbour-dependent model of dna methylation propagation at the mgmt gene promoter in lung tumours. *Journal of theoretical biology*, 336:87–95, 2013.

[15] Michael B Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J Oakeley, Dimos Gaidatzis, et al. Dna-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 2011.

[16] Lukas Burger, Dimos Gaidatzis, Dirk Schübeler, and Michael B Stadler. Identification of active regulatory regions from dna methylation data. *Nucleic acids research*, 41(16):e155–e155, 2013.

[17] Abdulrahman Salhab, Karl Nordström, Gilles Gasparoni, Kathrin Kattler, Peter Ebert, Fidel Ramirez, Laura Arrigoni, Fabian Müller, Julia K Polansky, Cristina Cadenas, et al. A comprehensive analysis of 195 dna methylomes reveals shared and cell-specific features of partially methylated domains. *Genome biology*, 19(1):150, 2018.

[18] Tuncay Baubec, Daniele F Colombo, Christiane Wirbelauer, Juliane Schmidt, Lukas Burger, Arnaud R Krebs, Altuna Akalin, and Dirk Schübeler. Genomic profiling of dna methyltransferases reveals a role for dnmt3b in genic methylation. *Nature*, 520(7546):243, 2015.

[19] Tianpeng Gu, Xueqiu Lin, Sean M Cullen, Min Luo, Mira Jeong, Marcos Estecio, Jianjun Shen, Swanand Hardikar, Deqiang Sun, Jianzhong Su, et al. Dnmt3a and tet1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome biology*, 19(1):88, 2018.

[20] Gabriella Ficz, Timothy A Hore, Fatima Santos, Heather J Lee, Wendy Dean, Julia Arand, Felix Krueger, David Oxley, Yu-Lee Paul, Jörn Walter, et al. Fgf signaling inhibition in escs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell stem cell*, 13(3):351–359, 2013.

[21] Ehsan Habibi, Arie B Brinkman, Julia Arand, Leonie I Kroeze, Hindrik HD Kerstens, Filomena Matarese, Konstantin Lepikhov, Marta Gut, Isabelle Brun-Heath, Nina C Hubner, et al. Whole-genome bisulfite sequencing of two distinct interconvertible dna methylomes of mouse embryonic stem cells. *Cell stem cell*, 13(3):360–369, 2013.

[22] Marius Walter, Aurelie Teissandier, Raquel Perez-Palacios, and Deborah Bourc'his. An epigenetic switch ensures transposon repression upon dynamic loss of dna methylation in embryonic stem cells. *Elife*, 5:e11418, 2016.

[23] Ferdinand von Meyenn, Mario Iurlaro, Ehsan Habibi, Ning Qing Liu, Ali Salehzadeh-Yazdi, Fátima Santos, Edoardo Petrini, Inês Milagre, Miao Yu, Zhenqing Xie, et al. Impairment of dna methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Molecular cell*, 62(6):848–861, 2016.

[24] Pengfei Wang, Chengqi Lin, Edwin R Smith, Hong Guo, Brian W Sanderson, Min Wu, Madelaine Gogol, Tara Alexander, Christopher Seidel, Leanne M Wiedemann, et al. Global analysis of h3k4 methylation defines mll family member targets and points to a role for mll1-mediated h3k4 methylation in the regulation of transcriptional initiation by rna polymerase ii. *Molecular and cellular biology*, 29(22):6074–6085, 2009.

[25] Yali Dou, Thomas A Milne, Alexander J Ruthenburg, Seunghee Lee, Jae Woon Lee, Gregory L Verdine, C David Allis, and Robert G Roeder. Regulation of mll1 h3k4 methyltransferase activity by its core components. *Nature Structural and Molecular Biology*, 13(8):713, 2006.

[26] Hao Jiang, Abhijit Shukla, Xiaoling Wang, Wei-yi Chen, Bradley E Bernstein, and Robert G Roeder. Role for dpy-30 in es cell-fate specification by regulation of h3k4 methylation within bivalent domains. *Cell*, 144(4):513–525, 2011.

[27] Hao Wu, Ana C D'alessio, Shinsuke Ito, Kai Xia, Zhibin Wang, Kairong Cui, Keji Zhao, Yi Eve Sun, and Yi Zhang. Dual functions of tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473(7347):389, 2011.

[28] Shinsuke Ito, Ana C D'alessio, Olena V Taranova, Kwonho Hong, Lawrence C Sowers, and Yi Zhang. Role of tet proteins in 5mc to 5hmc conversion, es-cell self-renewal and inner cell mass specification. *nature*, 466(7310):1129, 2010.

[29] Johannes M Freudenberg, Swati Ghosh, Brad L Lackford, Sailu Yellaboina, Xiaofeng Zheng, Ruifang Li, Suresh Cuddapah, Paul A Wade, Guang Hu, and Raja Jothi. Acute depletion of tet1-dependent 5-hydroxymethylcytosine levels impairs lif/stat3 signaling and results in loss of embryonic stem cell identity. *Nucleic acids research*, 40(8):3364–3377, 2011.

[30] Meelad M Dawlaty, Achim Breiling, Thuc Le, Günter Raddatz, M Inmaculada Barrasa, Albert W Cheng, Qing Gao, Benjamin E Powell, Zhe Li, Mingjiang Xu, et al. Combined deficiency of tet1 and tet2 causes epigenetic abnormalities but is compatible with postnatal development. *Developmental cell*, 24(3):310–323, 2013.

[31] Meelad M Dawlaty, Achim Breiling, Thuc Le, M Inmaculada Barrasa, Günter Raddatz, Qing Gao, Benjamin E Powell, Albert W Cheng, Kym F Faull, Frank Lyko, et al. Loss of tet enzymes compromises proper differentiation of embryonic stem cells. *Developmental cell*, 29(1):102–111, 2014.

[32] Chunlei Jin, Yue Lu, Jaroslav Jelinek, Shoudan Liang, Marcos RH Estecio, Michelle Craig Barton, and Jean-Pierre J Issa. Tet1 is a maintenance dna demethylase that prevents methylation spreading in differentiated cells. *Nucleic acids research*, 42(11):6956–6971, 2014.

[33] Taiping Chen, Yoshihide Ueda, Jonathan E Dodge, Zhenjuan Wang, and En Li. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by dnmt3a and dnmt3b. *Molecular and cellular biology*, 23(16):5594–5605, 2003.

[34] Jian Feng, Yu Zhou, Susan L Campbell, Thuc Le, En Li, J David Sweatt, Alcino J Silva, and Guoping Fan. Dnmt1 and dnmt3a maintain dna methylation and regulate synaptic function in adult forebrain neurons. *Nature neuroscience*, 13(4):423, 2010.

[35] Steffen Rulands, Heather J Lee, Stephen J Clark, Christof Angermueller, Sebastien A Smallwood, Felix Krueger, Hisham Mohammed, Wendy Dean, Jennifer Nichols, Peter Rugg-Gunn, et al. Genome-scale oscillations in dna methylation during exit from pluripotency. *bioRxiv*, page 338822, 2018.

[36] Hao Wu, Xiaoji Wu, and Yi Zhang. Base-resolution profiling of active dna demethylation using mab-seq and camab-seq. *nature protocols*, 11(6):1081, 2016.

[37] Chenxu Zhu, Yun Gao, Hongshan Guo, Bo Xia, Jinghui Song, Xinglong Wu, Hu Zeng, Kehkooi Kee, Fuchou Tang, and Chengqi Yi. Single-cell 5-formylcytosine landscapes of mammalian early embryos and escs at single-base resolution. *Cell Stem Cell*, 20(5):720–731, 2017.

[38] Kian Peng Koh, Akiko Yabuuchi, Sridhar Rao, Yun Huang, Kerrianne Cunniff, Julie Nardone, Asta Laiho, Mamta Tahiliani, Cesar A Sommer, Gustavo Mostoslavsky, et al. Tet1 and tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell stem cell*, 8(2):200–213, 2011.

[39] Chenhuan Xu and Victor G Corces. Resolution of the dna methylation state of single cpg dyads using in silico strand annealing and wgbs data. *Nature protocols*, 14(1):202, 2019.

[40] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[41] Jonas Korlach, Keith P Bjornson, Bidhan P Chaudhuri, Ronald L Cicero, Benjamin A Flusberg, Jeremy J Gray, David Holden, Ravi Saxena, Jeffrey Wegener, and Stephen W Turner. Real-time dna sequencing from single polymerase molecules. In *Methods in enzymology*, volume 472, pages 431–455. Elsevier, 2010.

[42] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265, 2009.

LIST OF ABBREVIATIONS

°C	Degree Celsius
%	Percent
µg	Microgram
µl	Microlitre
2i	Two inhibitors (PD0325901, CHIR99021)
5caC	5-carboxyluracil
5fC	5-formylcytosine
5fU	5-formyluracil
5hmC	5-hydroxymethylcytosine
5hmU	5-hydroxymethyluracil
5mC	5-methylcytosine
A	Adenine
Afp	Alpha fetoprotein
AID	Activation Induced Deaminase
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
ATP	Adenosine triphosphate
BER	Base excision repair
BI	Bayesian inference
bp	Base pair
bps	Base pairs
BS	Bisulfite treatment or bisulfite sequencing
C	Cytosine
CEGX	Cambridge Epigenetix
CGI	CpG island
CpG	Cytosine-phosphatidyl-Guanine
CpH	Cytosine-phosphatidyl-H (H = A,C,T)
Cys	Cysteine
d	Day
ddH$_2$O	Double-distilled water
DKO	Double knockout
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid

Dnmt	DNA methyltransferase
Dpy30	Protein dpy-30 homolog
ds-DNA	Double strand DNA
e.g.	lat. *exempli gratia*, eng. ‘for example’
E10.5	Embryonic day 10.5
E10.5	Embryonic day 11.5
ENCODE	Encyclopaedia of DNA elements
EQ	Equation
ES cell	Embryonic stem cell
fCAB-Seq	5fC chemically assisted bisulfite sequencing
FCS	Foetal calf serum
Fig	Figure
G	Guanine
Gsk3	Glycogen synthase kinase 3
h	Hour
H(O)TA	Hairpin (Oxidative) bisulfite sequencing Time course Analyzer
HDAC	Histone deacetylase
HEK	Human embryonic kidney
HMM	Hidden Markov model
HMR	Highly methylated regions
HP	hairpin
HPBS	Hairpin bisulfite sequencing
HPoxBS	Hairpin oxidative bisulfite sequencing
i.e.	lat. *id est*, eng. ‘that is’
IAP	Intracisternal A particle
KD	Knockdown
KO	Knockout
L1	Long interspersed nuclear elements 1
LIF	leukemia inhibiting factor
LMR	Low methylated regions
lncRNA	Long non coding RNA
M	Molar
MAB-Seq	*M.Sss*I assisted bisulfte sequencing
MeCP2	Methyl-CpG-binding protein 2
Mek	Mitogen activated protein kinase
min	Minutes
ml	Millilitre
MLE	Maximum likelihood estimation
mM	Millimolar

mSat	Major satellite
MuERVL	Murine endogenous retrovirus
Nanog	Homeobox protein NANOG
NaOH	Sodium hydroxide
NEB	New England Biolabs
ng	Nanogram
NGS	Next generation sequencing
oxBS	Oxidative bisulfite treatment or oxidative bisulfite sequencing
PCR	Polymerase chain reaction
PGC	Primordial germ cell
Pho	Phosphate
piRNA	PIWI associated RNAs
PIWI	P-element Induced WImpy testis in Drosophila
PMD	Partially methylated domains
pmol	Picomol
Pou5f1	POU domain, class 5, transcription factor 1 (Oct4)
PTM	Post-translation modification
qPCR	Quantitative real-time PCR
RE	Restriction enzyme
REST	RE1-Silencing Transcription factor
RNA	Ribonucleic acid
RRBS	Reduced representation bisulfite sequencing
RRHPBS	Reduced representation hairpin bisulfite sequencing
RRHPoxBS	Reduced representation hairpin oxidative bisulfite sequencing
RT	Room temperature
s	Seconds
SAH	S-adenosyl-homocysteine
SAM	S-adenosyl methionine
SDS	Sodium dodecyl sulfate
SMRT-Seq	Single molecule real-time sequencing
SMUG1	Single-strand selective monofunctional uracil DNA glycosylase
SNP	Single nucleotide polymorphism
Sox2	Transcription factor SOX-2
T	Thymine
TDG	Thymine DNA glycosylase
TE buffer	Tris EDTA buffer
Tet	Ten-eleven translocation
Tex13	Testis exprimiertes Gen 13
TFBS	Transcription factor binding site

tiRNA	Transcription initiation RNA
TKO	Triple knockout
TpG	Thymine-phosphatidyl-Guanine
Ttc25	Tetratricopeptide Repeat Domain 25
U	Uracil or Units
UCSC	University of California, Santa Cruz (UCSC genome browser)
Uhrf1	Ubiquitin-Like PHD And RING Finger Domain-Containing Protein 1
UMI	Unique molecular identifier
UMR	Unmethylated region
WT	Wild type
Zim3	Zinc Finger Imprinted 3

LIST OF FIGURES

LIST OF TABLES

EIDESSTATTLICHE VERSICHERUNG

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 21.01.2020

Peter Pascal Giehr