Stochastic Modeling of DNA Demethylation Dynamics in ESCs

Charalampos Kyriakopoulos

Dissertation

submitted towards

the degree Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science

of Saarland University

Saarbrücken

2019

Tag des Kolloquiums: 30.04.2019
Dekan: Prof. Dr. Sebastian Hack

Prüfungsausschuss
Vorsitzender: Prof. Dr. Wolfgang Paul
Berichterstatter: Prof. Dr. Verena Wolf,
Prof. Dr. Luca Bortolussi,
Prof. Dr. Guido Sanguinetti
Akademischer Mitarbeiter: Dr. Daniel Stan

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 in replication dependent and independent demethylation processes has been suggested to be influenced by Tet mediated oxidation of a methylated cytosine, 5mC, to a hydroxylated cytosine, 5hmC. Based only on in vitro experiments, several alternative mechanisms have been proposed on how 5hmC influences replication dependent maintenance of DNA methylation and replication independent processes of active demethylation. In this thesis we design an extended and easily generalizable hidden Markov model that uses as input hairpin (oxidative-)bisulfite sequencing data to precisely determine the over time dynamics of 5mC and 5hmC, as well as to infer the activities of the involved enzymes at a single CpG resolution. Developing the appropriate statistical and computational tools, we apply the model to discrete high-depth sequenced genomic loci, and on a whole genome scale with a much smaller sequencing depth. Performing the analysis of the model's output on mESCs data, we show that the presence of Tet enzymes and 5hmC has a very strong impact on replication dependent demethylation by establishing a passive demethylation mechanism, implicitly impairing methylation maintenance, but also down-regulating the *de novo* methylation activity.

Zusammenfassung

DNA-Methylierung und Demethylierung sind gegenläufige Prozesse, die im Gleichgewicht stabile Muster des epigenetischen Gedächtnisses erzeugen. Es wird angenommen, dass die Kontrolle der DNA-Methylierungsmusterbildung in replikationsabhängige und unabhängige Demethylierungsprozesse durch Tet-regulierte Oxidation eines methylierten Zytosins (5mC) zu einem hydroxylierten Zytosin (5hmC) beeinflusst wird. Aufgrund von In-Vitro-Experimenten, wurden verschiedene Mechanismen vorgeschlagen wie 5hmC die replikationsabhängige Aufrechterhaltung der DNA-Methylierung und die replikationsunabhängigen Prozesse der aktiven Demethylierung beeinflusst. In dieser Arbeit entwerfen wir ein erweitertes und leicht verallgemeinertes Hidden Markov Modell, das mit Hilfe von Hairpin (oxidative-)Bisulfit Sequenzierung gewonnener Daten die Zeitdynamik von 5mC und 5hmC genau bestimmt und die Aktivitäten der beteiligten Enzyme auf der Ebene einzelner CpGs schätzt. Wir entwickeln geeignete statistische Methoden, um das Modell sowohl auf der Ebene der sequenzspezifischen Tiefensequenzierung einzelner Loci, als auch auf genomweiter Ebene mit stark verringerter Sequenzierungstiefe anzuwenden. Wir zeigen, dass die Anwesenheit von Tet-Enzymen und 5hmC einen sehr starken Einfluss auf die replikationsabhängige Demethylierung hat, indem sie einen passiven Demethylierungsmechanismus etabliert, der die Methylierungserhaltung implizit beeinträchtigt, aber auch die *de novo*-Methylierung herunterreguliert.

Acknowledgements

I would like to thank my advisor Prof. Verena Wolf for introducing me to the fascinating world of stochastic modeling, giving me the chance to perform high-quality research, and for always providing me with a peaceful and friendly environment for doing that. I thank a lot my colleague from the Epigenetics Department Pascal Giehr with whom we worked closely together for more than four years. I can not omit to mention Prof. Luca Bortolussi and his bright ideas and comments during my doctoral time. Last, many thanks go to my ex-colleagues Alexander, David, Michael and Gerrit. I feel more than fortunate to have been collaborated with such highly talented individuals.

On a more personal note, there is a long list of people that have been next to me during all these years. Aggeliki, Eirini, Fabrizio, Nikos, Pedro and Alina are the guys that sparked up my Saarbrücken life. I am genuinely grateful for all the precious and different things that each of you gave me. Vaso, Michalis, Apostolos, Dimitris, Nikos and Thodoris are my invaluable friends already from the Diploma years in Patras. I am connected to them via a deep and lifelong relationship and I suspect their contribution in my thinking is a bit more than they imagine. Finally, I guess no words are enough to express my gratitude to my sister Nikoletta, my mother Vasiliki and my father Tasos. I will simply say that I owe you everything that I am today.

To my parents

for the depth they gave me

Contents

Abstract . . . iii

Zusammenfassung . . . iv

Acknowledgements . . . v

List of Tables . . . xi

List of Figures . . . xii

1 Introduction 1

1.1 Related Work . . . 5

1.2 Contributions . . . 7

1.3 Structure . . . 9

2 Preliminaries 11

2.1 Mathematical Formalism . . . 11

2.1.1 Discrete-time Markov chains . . . 11

2.1.2 Continuous-time Markov chains . . . 14

2.1.3 Hidden Markov Models . . . 16

2.2 Experimental Methods . . . 22

2.2.1 Protocol HPoxBS for individual loci . . . 22

2.2.2 Protocol RRHPoxBS for the whole genome . . . 25

3 The Modeling Approach 27

3.1 A Hidden Markov Model for Hydroxylation . . . 27

3.1.1 Hidden state space 29
3.1.2 Observable state space and conversion errors 33
3.2 Estimation of Model Parameters 35
3.2.1 Initial hidden states' distribution 37
3.2.2 Estimation of the efficiency functions 38
3.2.3 Model validation 45
3.3 A Hybrid Generalization of the Model 46
3.3.1 Estimation of model parameters 48

4 Individual Genomic Loci Analysis **59**
4.1 Results 59
4.1.1 Estimation of the enzymatic activities 60
4.1.2 Goodness of fit and validation 64
4.1.3 Prediction of hidden states' levels 64
4.2 Discussion 68
4.2.1 Biological findings 69

5 H(O)TA Tool **73**
5.1 H(O)TA Description 73
5.2 Run H(O)TA 76
5.2.1 Installation 76
5.2.2 Input 77
5.2.3 Output 79
5.3 H(O)TA v.2 - Hybrid model, further extensions 80

6 Whole Genome Analysis **84**
6.1 Computational Methods 86
6.1.1 Bayesian inference 86
6.1.2 Clustering of enzymatic efficiencies 89

6.1.3 Spatial and non-spatial cross-correlations 95
6.2 Results 97
6.2.1 BI vs MLE 97
6.2.2 *k*-error vs *k*-mean 100
6.2.3 Enzymatic activities' correlations 104
6.3 Output Analysis 106
6.3.1 Spatial analysis 108
6.3.2 Temporal analysis 122
6.4 Discussion 125
6.4.1 Biological findings 127

7 Conclusions & Future Work 131

7.1 Conclusions 131
7.2 Future Work 133

A Individual Genomic Loci Analysis 135

A.1 HPoxBS Protocol Details 135
A.1.1 Primer- and reference sequences 136
A.2 Input Data 137
A.3 Single CpGs' Results 139

B Whole Genome Analysis 144

B.1 RRHPoxBS Protocol Details 144
B.1.1 Read mapping and methylation calling 145
B.1.2 Spike-in analysis 145
B.2 Implementation Details 147
B.3 BI vs MLE 149
B.4 ESCs Chromosomes' Results 149

Bibliography 160

List of Tables

3.1 Emission matrices of BS and oxBS 35

3.2 Emission matrices for BS, oxBS and mabBS 55

4.1 Estimated coefficients of the enzymatic efficiency functions 61

4.2 Estimated coefficients of the total methylation efficiency 61

4.3 Kullback Leibler divergence and Bhattacharya distance by LOOCV . 66

6.1 Kullback-Leibler divergence for BI and MLE 100

A.1 Hairpin linker sequences 137

A.2 Primer for amplification after BS and oxBS 137

A.3 Reference sequences 138

A.4 Input BS and oxBS data and conversion errors (repetitive elements) . 139

A.5 Input BS and oxBS data and conversion errors (single copy genes) . . 140

B.1 Conversion rates in the TrueMethyl Spike of BS treatment 145

B.2 Conversion rates in the TrueMethyl Spike of oxBS treatment 146

List of Figures

1.1 DNA Maintenance methylation 2

1.2 Hydroxylation of DNA 4

1.3 Chapter dependencies. 9

2.1 Schematic outline of HPoxBS 23

2.2 Schematic outline of RRHPoxBS. 26

3.1 DNA demethylation 28

3.2 Cell division DTMC graph 30

3.3 Methylation DTMC graph 31

3.4 Hydroxylation DTMC graph 33

3.5 Conversion errors during BS and oxBS experiments 34

3.6 Active demethylation cycle of a CpG site 47

3.7 Conversion errors during BS, oxBS and mabBS experiments 54

4.1 Enzymatic efficiencies and standard deviations over time 63

4.2 Comparison between data and model prediction for BS and oxBS . . 65

4.3 Hidden states' probabilities 67

4.4 The effect of passive demethylation 68

5.1 Conversion errors' illustration 74

5.2 The main window of the graphical user interface of H(O)TA. 76

5.3 The input panel and example input files of H(O)TA 78
5.4 H(O)TA's output file in .txt format 80
5.5 H(O)TA v.2 81
5.6 MLE vs subtraction estimator for hydroxylation levels 83
5.7 H(O)TA v.2: Model input file example 83

6.1 Metropolis-Hastings' new sample generation 88
6.2 Ellipse of the bivariate normal 90
6.3 Illustration of k-means vs k-error clustering 92
6.4 Elbow method 95
6.5 Genome-wide model fit for MLE and BI estimators 99
6.6 Bar plots for estimated enzymes' efficiencies by MLE and BI method 101
6.7 Optimal number of clusters for k-means and k-error algorithms 102
6.8 Optimal k-means clustering 103
6.9 Optimal k-error clustering 103
6.10 Spatial auto- and cross-correlation of the efficiencies across the genome 105
6.11 Pearson correlation between efficiencies and modification levels 107
6.12 Hemimethylated CpGs across expressed and not/low expressed genes 109
6.13 Genome-wide CpG and nonCpG methylation levels 109
6.14 NonCpG methylation in Serum and 2i cultivated WT ESCs 110
6.15 Gene average enzymatic activity and methylation pattern in WT ESCs 112
6.16 Gene average enzymatic activity in WT vs Tet TKO ESCs 113
6.17 Expressed and non-expressed genes' profiles 114
6.18 Enzymatic activity at TFBS and across histone marks 116
6.19 Enzymatic activity at TFBS and across histone marks in WT vs Tet TKO ESCs 117
6.20 Profile of ENCODE histone modifications across segmentation regions 118
6.21 NonCpG methylation for different segments 119

6.22 Methylome Segmentation, modification levels and efficiencies per segment 120
6.23 Clustering of single CpGs based on enzymatic activity, profiles of annotated genomic features . . . 121
6.24 Demethylation rate in WT and Tet TKO ESCs . . . 123
6.25 Clustered CpGs in WT vs Tet TKO ESCs . . . 126

A.1 Single CpGs: Estimated efficiencies and standard deviations . . . 142
A.2 Single CpGs: (Hydroxy-)methylation levels prediction . . . 143

B.1 CpGs per observation days and average day coverage for BS, oxBS . . . 146
B.2 Genome-wide modification levels' prediction . . . 149
B.3 Number of CpGs per number of observation days for each chromosome 151
B.4 Average coverage per day for each chromosome . . . 152
B.5 Efficiency profiles for each chromosome in WT and Tet TKO ESCs . . . 153
B.6 Modification levels for each chromosome in WT and Tet KO ESCs . . . 154
B.7 25 most frequent repetitive elements: Efficiency profiles in WT . . . 155
B.8 25 most frequent repetitive elements: Modification levels in WT . . . 156
B.9 25 most frequent repetitive elements: 5hmC distribution in WT . . . 157
B.10 25 most frequent repetitive elements: Efficiency profiles in Tet TKO . . . 158
B.11 25 most frequent repetitive elements: Modification levels in Tet TKO 159

Chapter 1

Introduction

DNA methylation is an essential epigenetic modification that regulates the transcriptional access to genetic information and controls the genome stability and genome function [11, 53, 37, 78]. During development the distribution of DNA methylation is under strict control to maintain a temporal and cell type specific persistence of epigenetic information [68]. In mammals DNA methylation is restricted to the C-5 position of cytosine and it is predominantly found in a palindromic CpG di-nucleotide context [6, 17, 77, 100, 56]. The generation of a 5-methyl cytosine (5mC) is controlled by the DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b that catalyze the transfer of a methyl group from s-adenosyl methionine to the fifth carbon atom of cytosine.

Methyltransferase Dnmt1 is responsible for maintaining an existing methylation pattern after cell replication. Via interaction with Uhrf1 and PCNA, Dnmt1 is tightly associated with the replication machinery [51, 14, 10]. The palindromic nature of a CpG sequence in which 5mC occurs allows the recognition of a 5mC hemimethylated state after semi-conservative replication and the copying of the parental methylation pattern to the newly synthesized DNA strand [35, 10, 81]. This process is known as maintenance methylation and it is described in Figure 1.1. In contrast to Dnmt1, methyltransferases Dnmt3a and Dnmt3b act without any preference on both hemi-

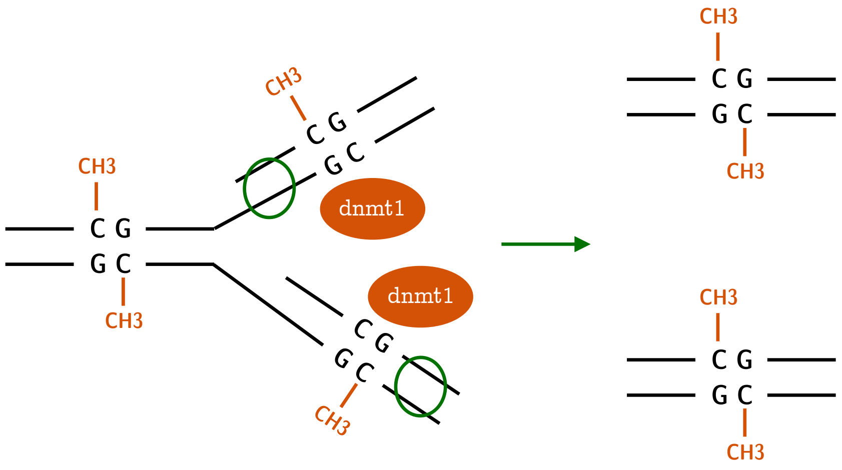


Figure 1.1: DNA Maintenance methylation. After DNA replication Dnmt1 is recruited to the replication fork via its association with Uhrf1 [10] and PCNA [14], that confer processivity to polymerases (green ring). Dnmt1 recognizes a hemimethylated site [73], i.e., a pair of CpG dinucleotides with a 5-methylcytosine on the original strand and a normal cytosine on the newly synthesized strand. Once Dnmt1 binds to the hemimethylated site, it transfers a methyl group onto the newly synthesized cytosine base.

as well as unmethylated CpGs, and their activity is not coupled to DNA replication [69, 68, 25]. Both of these enzymes are highly regulated and regarded as the main enzymes to establish new methylation patterns[1]. For this reason they are classified as *de novo* DNA methyltransferases.

Besides the establishment and the persistence of methylation, its removal, called demethylation, is also of major biological importance. Demethylation events can occur on a local scale in case of individual gene activation, but also on a genome-wide scale in the early zygote, on embryonic stem cells (ESCs), as well as during the maturation of primordial germ cells (PGCs) [84, 71, 30], where genomes are reprogrammed for new developmental functions [28, 52]. In all these types of cells the stability of DNA demethylation is influenced by the oxidation of a 5-methylcytosine

[1] There is evidence that a strict separation of Dnmt1 and Dnmt3a/b activity is not coherent and that under certain conditions these enzymes exhibit cooperation, the details of which are not yet well understood [65, 55, 46]. For instance, Dnmt1, at least in the absence of Dnmt3a/b, may also *de novo* methylate unmethylated dyads [58, 1]. Nevertheless, the observed Dnmt1 *de novo* activity is usually very small and hence it seems fair to categorize Dmnts in the way we did above [25, 40].

(5mC) into 5-hydroxymethylcytosine (5hmC) and, potentially, into further oxidized forms (oxCs) as 5-formylcytosine (5fC) [3, 76, 45] and 5-carboxylcytosine (5caC) [90, 38]. The subsequent conversions to oxC modifications happen via the activity of a family of di-oxigenases, the ten-eleven translocation enzymes Tet1, Tet2 and Tet3 [70, 59, 39] in a sequence of steps, that change the chemical properties and the biological function of each base. As a consequence it has been observed that disturbances or depletion of Tet enzymes in the previously mentioned cells result in massive changes of 5hmC and lead to developmental consequences [95, 26, 15].

Based solely on in vitro experiments, several possible explanations have been proposed regarding the connection of the oxidative cytosine derivatives and the Tet enzymes with DNA demethylation. According to these, oxCs might serve as an intermediate during the course of either active, or passive demethylation [32, 92, 41, 34, 61, 98], which consequently results to the impairment of replication dependent copying of 5mC. Active demethylation is defined as the demethylation that happens within a cell-replication cycle, by the gradual transformation of 5mC to 5hmC, 5fC and 5caC and then to simple cytosine (Figure 1.2). On the contrary, passive demethylation is the decrease of methylation levels among two consecutive cell replications due to the impairment of maintenance methylation machinery, or possibly the interference of oxC derivatives with this machinery. In vitro observations hint towards such interference for instance, stating that Dnmt1 binds to 5hmC with a much lower affinity than to 5mC [32]. The truthfulness of the above theories, however, has never been tested in vivo, and the detailed underlying mechanisms that demethylation processes take place are still under debate. Questions such as how oxCs are inherited across cell replication, or what is really their impact on maintaining or changing an existing methylome, remain still elusive.

The goal of this doctoral thesis is to shed light on the in vivo dynamics of DNA demethylation in ESCs by studying mainly the influence of the most abundant ox-

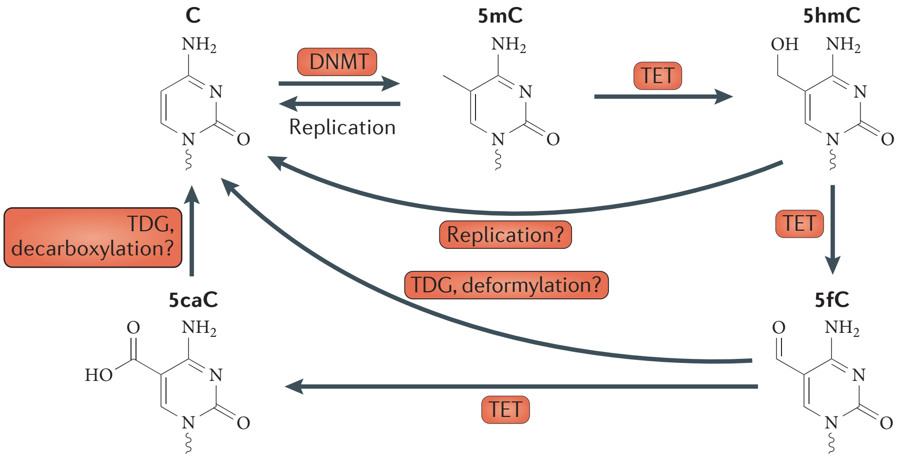


Figure 1.2: Potential pathways for DNA demethylation. Genomic 5-methylcytosine (5mC) can be removed passively during replication, but several pathways for active demethylation have also been proposed, including those in which 5-hydroxymethylcytosine (5hmC) is an intermediate. Alternatively, 5hmC may be further oxidized to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) by Tet enzymes [34, 38]. Although it is possible that the deformylation of 5fC and decarboxylation of 5caC convert these intermediates directly back to cytosine, no enzymatic activities have been described to date. Instead, thymine DNA glyco-sylase (TDG) has been shown to cleave 5fC and 5caC [34, 61], producing an abasic site that is then repaired by the base excision repair (BER) machinery. Adapted from "Uncovering the role of 5hydroxymethylcytosine in the epigenome," by Miguel R. Branco, 2012, Nature Reviews Genetics, 13, p.7. Copyright 2011 by the Nature Publishing Group.

idative variant [23, 48, 88], 5hmC, and of Tet enzymes on DNA demethylation over subsequent cell-replications. For this purpose a completely novel stochastic model is designed. Based on reasonable assumptions regarding the current knowledge about the function of Dnmt and Tet enzymes, the model is able i) to accurately determine the evolution of methylation and hydroxylation patterns for both complementary DNA strands over time, as well as ii) to estimate the efficiencies of maintenance, *de novo* and hydroxylation enzymes. The sequencing data given as input is from mouse embryonic stem cells (mESCs) which are transfered from Serum medium containing LIF (primed state) to a synthetic 2i medium and show, therefore, a gradual genome-wide loss of methylation [18, 29]. The model is applied to single CpGs of specific loci with ultra-deep sequencing, as well as on a genome-wide scale with less deep sequencing. Its output analysis offers particularly valuable insights into DNA demethylation mechanisms in ESCs.

1.1 Related Work

There are several approaches in the literature for estimating 5mC levels and modeling methylation dynamics on various cell types. In one of the earliest models Genereux et al. [21] derived a system of simple non-linear equations in order to estimate site specific maintenance and to investigate different assumptions regarding the *de novo* methylation of upper and lower strand. They used for this purpose hairpin bisulfite PCR data at a single CpG resolution. Sontag et al. [85] developed a simple Markov Chain in order to explain how the methylation levels of hypo or hyper-methylated regions remain stable despite a cell division process as well as possible sporadic transitions between the two equilibria. In the same direction Fu et al. [19] apply a hidden Markov model (HMM) to hairpin bisulfite PCR data to study the association and processivity, as well as, the substrate specificity of Dnmts on three different loci of human genome. Similar to Sontag, Arand et al. use a simple HMM to identify each of the Dnmt enzymes' static activities and specificities in a sequence of wild-type (WT) and KO experiments.

Regarding capturing 5hmC and estimating hydroxylation levels of single CpGs, the MLML tool [74] was the first to produce consistent estimates for methylation and hydroxylation levels getting as input any two of bisulfite sequencing (BS-seq), oxidative bisulfite sequencing (oxBS-seq) [9], or Tet-assisted bisulfite sequencing (TAB-seq) data. The tool makes use of an expectation maximization (EM) algorithm, but it does not take into account any of the possible conversion errors during the sequencing and can not cope with time course data. An extended version of this work was done by Äijö et al., where Bayesian Inference is used to compute the levels of 5mC, 5hmC and also of further oxidative modifications 5fC and 5caC, given that data from other sequencing methods as MAB-seq, fCAB-seq, redBS-seq, or CAB-seq is available. In [94] von Meyenn et al. the authors use genome-wide data to study the demethylation process in mESCs. Their modeling approach consists of solving the mean field equations of a 6-state Markov chain. Since they do not have hairpin TAB-seq or oxBS-seq data they assume 5hmC is uniformly distributed to all possible double stranded states with at least one 5hmC. This unavoidably leads to an inaccurate model. In addition, a possible mis-recognition of 5hmC by the maintenance methyltransferase that can be a potential mechanism for passive demethylation has also not been investigated in this paper, due to the unavailability of double stranded oxBS or TAB data. As a result, in a contradicting, to our opinion, finding the authors report that the presence of 5hmC plays a very minor role in the demethylation dynamics in ESCs.

We follow here a different approach that is more complete in capturing 5mC/5hmC dynamics compared to the methods used in the work described above. The strength of our approach is based on the neat combination of two innovative parts. The first part is the production of single CpG double strand time course data available via the development of two new protocols that are characterized by the original application of a hairpin setup [50] on both BS-seq and oxBS-seq. The second part is the design of the stochastic model described in this thesis that is able to take advantage of this sequencing data to accurately estimate the changes of 5mC/5hmC levels for a single CpG over time. This is achieved via the novel combination of two HMMs, one for each sequencing method, that incorporate all possible conversion errors that can happen during each experiment. Based on realistic biological assumptions about the underlying epigenetic subprocesses, the presented model is additionally capable of inferring the dynamic activity of certain enzymes, and studying potential mechanisms involved in the creation and the loss of 5mC/5hmC. As a result, the model results provide the opportunity to investigate several hypotheses about the behavior of factors relevant to loss of methylation, and thus greatly deepen our understanding on what causes DNA demethylation.

1.2 Contributions

The contributions of this doctoral thesis consist of:

- Design of a novel stochastic model that gives (hydroxy-)methylation levels' estimates for single CpGs over time, while simultaneously provides estimates about the activities of Dnmt and Tet enzymes over time.
- Development of a numerical optimization algorithm to efficiently identify the unknown model parameters along with their respective confidence intervals and performance of hypothesis tests regarding the behavior of the individual enzymes.
- Generalization of the above model to cope with generalized enzymatic efficiency functions, a larger number of experiments and cell types, and predict estimations for further oxidative forms other than hydroxylation.
- Implementation of the above functions in a software called H(O)TA with a very simple user-interface especially targeted to be used by biologists.
- Model's application on a genome-wide scale via the development of inference methods that cope with medium or small coverage and an efficient parallel implementation for the model's execution on a large cluster of machines.
- Development of a sophisticated clustering approach for identifying distinct profiles in the genome-wide data, spatial and temporal analysis of the model's genome-wide output, and detailed biological interpretation of the results.

These contributions have been published in parts or in whole in the following research papers:

P. Giehr*, **C. Kyriakopoulos***, G. Fisz, V. Wolf, J. Walter: *The Influence of Hydroxylation on Maintaining CpG Methylation Patterns: a HMM approach*, **PLOS Computational Biology, 2016**

M. A. Eckersley-Maslin, V. Svensson, C. Krüger, T. Stubbs, P. Giehr, F. Krüger, R. Miragaia, **C. Kyriakopoulos**, et al. *MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs*, **Cell Reports, 2016**

C. Kyriakopoulos, P. Giehr, V. Wolf: *H(O)TA: estimation of DNA methylation and hydroxylation levels and efficiencies from time course data*, **Bioinformatics, 2017**

P. Giehr, **C. Kyriakopoulos**, V. Wolf, J. Walter: *Two are better than one: HPoxBS - Hairpin oxidative Bisulfite Sequencing*, **Nucleic Acids Research, 2018**

C. Kyriakopoulos, P. Giehr, A. Lück, J. Walter, V. Wolf, : *A Hybrid HMM Approach for the Dynamics of DNA Methylation*, **HSB 2019**

C. Kyriakopoulos*, P. Giehr*, K. Nordström, S. Abdulrahman, F. Müller, F. von Meyenn, G. Ficz, W. Reik, V. Wolf, J. Walter: *Genome-wide single-base resolution efficiency profiling reveals modulation of maintenance and de novo methylation by Tet di-oxygenases*, **(under review)**

*These authors have contributed equally to this work.

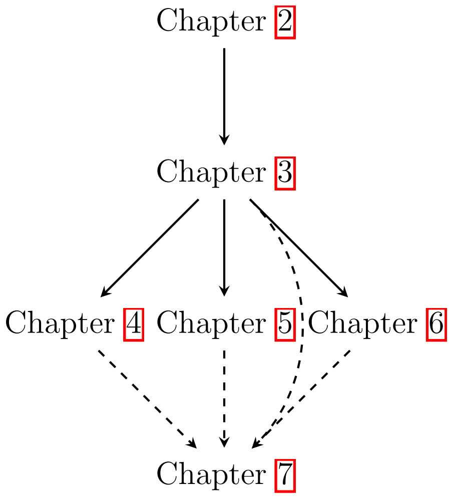


Figure 1.3: Chapter dependencies.

1.3 Structure

We sketch below the contents of the chapters of this thesis, while in Figure 1.3 we point at the dependencies of the chapters on each other. A solid line indicates an explicit relation, and a dashed line an implicit relation between two chapters.

Chapter 2: Definition of the basic mathematical concepts that are used in this thesis. Formal introduction of discrete- and continuous-time Markov Chains, as well as Hidden Markov Models. Brief description of the two biological protocols from which the data used in this thesis has been produced.

Chapter 3: In detail presentation of the core hidden Markov model we build in order to study the demethylation process of ESCs. Derivation of an efficient gradient descent numerical optimization approach to estimate the values of the unknown parameters and to approximate the confidence intervals of these estimators. Validation of the model output. In detail presentation of a hybrid generalization of the core model.

Chapter 4: Report and analysis of the results from the application of the core-model to individual genome regions of high-depth sequencing. Validation of the model's behavior in known phenomena via control-regions. Examination of concrete biological hypotheses regarding possible mechanisms of demethylation in mESCs.

Chapter 5: Presentation of the H(O)TA tool that incorporates all previous analysis in a user-friendly software specifically targeted to biologists interested in studying a (de-)methylation process at certain DNA loci.

Chapter 6: Description of the inference methods developed to cope with genome-wide single CpG resolution data of medium coverage and of the parallel implementation on a cluster of computing machines. Presentation of a clustering method to identify regions of similar enzymatic activity in the genome. Report and in depth biological interpretation of the model results on the genome-wide scale.

Chapter 7: Summary of models, computational methods and tools developed in this thesis, as well as short outline of the major biological findings. List of possible future extensions of the current work towards the better understanding of the various epigenetic processes.

Chapter 2

Preliminaries

2.1 Mathematical Formalism

In this Section we give the formal definitions of discrete- and continuous-time Markov chains and we distinguish between time-homogeneous and time-inhomogeneous Markov chains. We introduce then the main stochastic model that is being used along this thesis, the hidden Markov models (HMMs).

A Markov chain is a countable family of random variables that take values in a discrete set S (called *state space*) obeying the *Markov property*. They typically describe the temporal dynamics of a process (or system) over time and can be separated into chains that act in continuous and chains that act in discrete time.

2.1.1 Discrete-time Markov chains

Definition 1: Discrete-time Markov chain

▷**Definition 1**

A discrete-time Markov chain (DTMC) X is a sequence of random variables (RVs) $X_n : \Omega \to S, n \in \mathbb{N}_0$, on a countable state space S s.t. for all $n \in \mathbb{N}_0$ and for all $x_0, \ldots, x_n, x_{n+1} \in S$

$$\begin{aligned} P(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) \\ = P(X_{n+1} = x_{n+1} \mid X_n = x_n). \end{aligned} \tag{2.1}$$

Eq. (2.1) is known as Markov (or memoryless) property and expresses the assumption that the future state of the system (X_{n+1}) is dependent only on the information about the present state, i.e., $X_n = x_n$, whereas additional information about the past states $X_j = x_j,\ j \leq n-1$ is irrelevant.

W.l.o.g. we can assume that $S \subseteq \mathbb{N}$[1] and we subsequently write $p_{ij} = P(X_{n+1} = j \mid X_n = i)$ for the transition probabilities from state i to state j, for $i, j \in S$. We can arrange the transition probabilities in a matrix $\mathbf{P} = (p_{ij})_{i,j \in \{1,2,\ldots\}}$. The matrix $\mathbf{P}$ is called the transition probability matrix of the DTMC X. In case that all the entries of $\mathbf{P}$ are independent of the time unit n we call the Markov Chain time-homogeneous, whereas we call the Markov Chain time-inhomogeneous if there is at least one entry of transition matrix $\mathbf{P}$ that is dependent on n.

Transient distribution

As transient distribution we name the discrete probability distribution of the states of the Markov chain at a certain time step. Let $\pi(0)$ be the row vector that contains the *initial distribution* of the Markov chain at time zero, i.e., the entries $\pi(i, 0) = P(X_0 = i)$ for all states $i \in S$. From the law of total probability it holds

$$P(X_1 = j) = \sum_{i \in S} P(X_1 = j \mid X_0 = i) \cdot P(X_0 = j), \quad \forall i \in S \tag{2.2}$$

[1] This we do by simply enumerating all states in S such that $S = \{x_1, x_2, \ldots\}$ and then just write $X = i$ instead of $X = x_i$.

and, hence, writing Eq. (2.2) as a vector-matrix multiplication we get the *transient distribution* of the DTMC after one time unit (step) as $\pi(1) = \pi(0) \cdot \mathbf{P}$. In the same fashion, successive application of this argument, i.e., n multiplications with transition matrix $\mathbf{P}$, give us the vector $\pi(n)$ of transient probabilities after n steps

$$\pi(n) = \pi(n-1) \cdot \mathbf{P} = \ldots = \pi(0) \cdot \mathbf{P}^n.$$

Note that $\mathbf{P}$ is a stochastic matrix, i.e., the row sums are one and it has only non-negative entries. The n-th power of the transition matrix, $\mathbf{P}^n$, is also stochastic and essentially contains for each state the probability of reaching any other state after n steps.

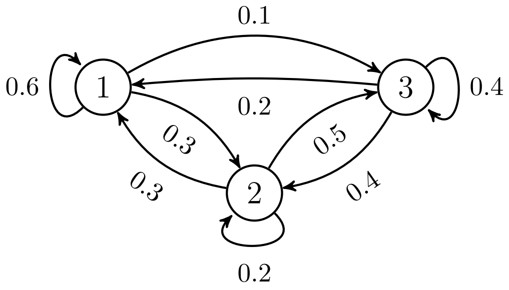
Consequently, given $\mathbf{P}$ and $\pi(0)$, we have two ways of computing the vector $\pi(n)$ of state probabilities after n steps. Either we successively compute $\pi(1), \pi(2), \ldots, \pi(n)$ by multiplying the vector with $\mathbf{P}$, or we compute the matrix product $\mathbf{P}^n$ and multiply it with $\pi(0)$. The former approach is usually preferable since it requires significantly less computational work and it also needs less memory[2]. Note that even if $\mathbf{P}$ is a sparse matrix the powers of $\mathbf{P}$ will typically not be sparse. Finally, note that both of the previous methods are possible if the DTMC is time-inhomogeneous. The difference is that at every update step, whether computing the matrix or the vector-matrix product, one has to multiply with matrix $\mathbf{P}(n)$, containing the transition probabilities at time unit n.

Example 1: A simple DTMC

▷**Example 1**

[2] In case $\mathbf{P}$ is large matrix multiplication needs $\mathcal{O}(|S|^3)$ time, and $\mathcal{O}(|S|^2)$ space, while matrix-vector multiplication is of $\mathcal{O}(|S|^2)$ time and of $\mathcal{O}(|S|)$ space complexity.

Consider the following simple DTMC that makes a weather forecast based on the weather of the current day. The states are 1:=sunny, 2:=cloudy, 3:=rainy.



The matrix

$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.2 & 0.5 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

contains the transition probabilities of the DTMC. Let $\pi_0 = (1, 0, 0)$ be the initial distribution of the chain. Then the transient distribution after three steps is

$$\pi_3 = \pi_0 \cdot \mathbf{P}^3 = (0.4160,\ 0.2970,\ 0.2870)$$

that is, for instance, $P(X_3 = 2) = 0.2970$.

2.1.2 Continuous-time Markov chains

Definition 2: Continuous-time Markov chain

▷**Definition 2**

A continuous-time Markov Chain (CTMC) is a sequence of random variables $X_t : \Omega \to S,\ \ t \geq 0$, on a countable state space S s.t. for all $x_0, \ldots, x_k, i, j \in S$, all $\Delta > 0, t \geq 0$ and all $t_0, \ldots, t_k \in [0, t)$ with $t_0 <, \ldots, < t_k$

$$\begin{aligned} &P(X_{t+\Delta} = j \mid X_t = i, X_{t_k} = x_k, \ldots, X_{t_0} = x_0) \\ &\qquad = P(X_{t+\Delta} = j \mid X_t = i). \end{aligned} \tag{2.3}$$

Eq. (2.3) is the Markov property defined for continuous time. Using the time derivatives of the transition probabilities of Eq. (2.3) one defines the infinitesimal generator matrix $\mathbf{Q}$ as shown in Eq. (2.4).

$$\mathbf{Q}_{ij} = \begin{cases} \lim_{h\to 0} \frac{P(X_h = j \mid X_0 = i)}{h} & \text{if } i \neq j, \text{and} \\ -\sum_{k\neq i} \mathbf{Q}_{ik} & \text{otherwise.} \end{cases} \tag{2.4}$$

For $i \neq j$ the entry $\mathbf{Q}_{ij}$ is the transition rate of state i to state j, while the negative diagonal entries $-\mathbf{Q}_{ii}$ contain the total *exit rate* to take any transition from state i. The infinitesimal generator matrix $\mathbf{Q}$ uniquely determines the behavior of a CTMC.

Let, similar to DTMCs, $\pi(t)$ be the transient distribution of the chain at time point $t \in \mathbb{R}$, and $\mathbf{P}(\Delta)$ the matrix that contains the transition probabilities for time interval $\Delta \in \mathbb{R}$, i.e., $\mathbf{P}(\Delta) = (\mathbf{P}(\Delta)_{ij})_{i,j\in\mathcal{S}} = (P(X_\Delta = j \mid X_0 = i))_{i,j\in\mathcal{S}}$. From this definitions we get then $\pi(t) = \pi(0)\mathbf{P}(t)$, and combining with the Kolmogorov forward equation $\frac{d}{dt}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{Q}$ we can prove that the time derivative of the transient distribution $\frac{d}{dt}\pi(t)$ is described by Eq. (2.5) as a set of ordinary differential equations (ODEs) that depend on the infinitesimal generator matrix $\mathbf{Q}$ [47].

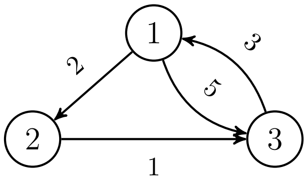
$$\frac{d}{dt}\pi(t) = \pi(t) \cdot \mathbf{Q}. \tag{2.5}$$

For a finite state space CTMC the solution of these ODEs (Eq. (2.6)) gives us the transient probability distribution over time.

$$\pi(t) = \pi(0) \cdot e^{\mathbf{Q}\cdot t}. \tag{2.6}$$

Example 2: A simple CTMC

▷**Example 2**



The Q matrix of the above CTMC is

$$\mathbf{Q} = \begin{pmatrix} -7 & 2 & 5 \\ 0 & -1 & 1 \\ 3 & 0 & -3 \end{pmatrix}.$$

Assuming the initial distribution is $\pi(0) = (0.5, 0.5, 0)$, the transient distribution after 2.5 time units is going to be $\pi(2.5) = \pi(0) \cdot e^{Q \cdot 2.5} = (0.1864,\ 0.3779,\ 0.4356)$

2.1.3 Hidden Markov Models

A hidden Markov model (HMM) is a Markov chain of finite state space and of discrete- or continuous-time in which the system being modeled is assumed to have unobserved (hidden) states. An HMM extends a Markov chain in that besides the state space of the Markov chain, i.e., the set $S = \{1, \ldots, |S|\}$ of hidden states, there is also a set $O = \{1, \ldots, |O|\}$ of observable states. This intuitively expresses the fact that an observed output might not be the true (hidden state) of the system, but it does not necessarily mean that the sets S and O are disjoint. For instance, in case the observed states are the result of some observation error over the hidden states, the two sets coincide. W.l.o.g. a formal definition of an HMM of discrete time follows.

Definition 3: Hidden Markov model

▷**Definition 3**

A hidden Markov model (HMM) is a combination of two stochastic processes X and Y. X is a DTMC with $X_n : \Omega \to S, n \in \mathbb{N}_0$, on a hidden state space S, while Y, with $(Y_n) : \Omega' \to O, n \in \mathbb{N}_0$, is dependent on X and it is defined on a countable state space O which can be observed.

In order to define an HMM of discrete time we need the transition matrix $\mathbf{P}$ of the underlying DTMC[3] and the emission probability matrix $\mathbf{E}$. The $\mathbf{E}_{i,j}$ entry of the emission matrix equals the probability that we observe state $j \in O$, given that the hidden state of the system is $i \in S$, i.e, $\mathbf{E}_{i,j} = P(Y_n = j \mid X_n = i)$ for all n.

Given for an HMM the initial distribution $\pi(0)$, the transition probabilities (or rates), the emission matrix $\mathbf{E}$, and a sequence of observed states, we state some typical problems that can arise.

i Compute the probability distribution over the last hidden state of the observed sequence (*filtering*).

ii Compute the probability distribution over a hidden state in the middle of the observed sequence (*smoothing*).

iii Find the most likely sequence of hidden states that produced the given sequence of observed states (*most likely explanation*).

In order to solve the problem of *filtering* we use the forward algorithm, while for solving *smoothing* the forward-backward algorithm needs to be used [75]. To find the *most likely explanation* for a given sequence of observed states, we use the Viterbi algorithm [93]. Since problem (i) is a subcase of problem (ii), we present in the sequel the forward-backward and the Viterbi algorithm.

Forward-backward algorithm

Given a sequence of $k+1$ observations $o = \{o_0, \ldots, o_k\} = o_{0:k}$ at discrete time points, we want to compute the probability $P(X_\ell = i \mid o)$ of every possible hidden state i

[3] For an HMM of continuous time we would need the infinitesimal generator matrix $\mathbf{Q}$.

after ℓ steps. From Bayes law and the conditional independence of $o\ :_{0:\ell}$ and $o_{\ell+1:k}$ given X_ℓ we get

$$\begin{aligned} P(o \mid X_\ell = i) &= \frac{P(o, X_\ell = i)}{P(o)} \\ &= \frac{P(o_{0:\ell}, X_\ell = i) P(o_{\ell+1:k} \mid o_{0:\ell}, X_\ell = i)}{P(o)} \\ &= \frac{P(o_{0:\ell}, X_\ell = i) P(o_{\ell+1:k} \mid X_\ell = i)}{P(o)} \end{aligned} \tag{2.7}$$

for all $i \in S$. We compute the probability $P(o_{0:\ell}, X_\ell = i) = f_\ell(i)$ of (2.7) by the ℓ-th step of the forward algorithm, while we can compute the probability $P(o_{\ell+1:k} \mid X_\ell = i) = b_\ell(i)$ by the $(k-\ell)$-th step of the backward algorithm for all $i \in S$. We write then

$$P(X_\ell = i \mid o) = \frac{f_\ell(i) b_\ell(i)}{\sum_i f_\ell(i) b_\ell(i)}.$$

For the probability of the forward algorithm it holds:

$$\begin{aligned} f_\ell(i) &= P(o_{0:\ell}, X_\ell = i) \\ &= \sum_{x_0} \ldots \sum_{x_{\ell-1}} P(o_{0,\ell} \mid X_0 = x_0, \ldots, X_{\ell-1} = x_{\ell-1}, X_\ell = i) \\ &\quad \cdot P(X_0 = x_0, \ldots, X_{\ell-1} = x_{\ell-1}, X_\ell = i) \\ &= \sum_{x_0} \ldots \sum_{x_{\ell-1}} P(X_0 = x_0, \ldots, X_{\ell-1} = x_{\ell-1}, X_\ell = i) \cdot \prod_{n=0}^{\ell-1} \mathbf{E}_{x_n,o_n} \mathbf{E}_{n,o_\ell} \\ &= \sum_{x_0} \ldots \sum_{x_{\ell-1}} P(X_\ell = i \mid X_{\ell-1} = x_{\ell-1}) \cdot \ldots \cdot P(X_1 = x_1 \mid X_0 = x_0) P(X_0 = x_0) \\ &\quad \prod_{n=0}^{\ell-1} \mathbf{E}_{x_n,o_n} \mathbf{E}_{i,o_\ell} \\ &= \sum_{x_0} P(X_0 = x_0) \mathbf{E}_{x_0,o_0} \sum_{x_1} P(X_1 = x_1 \mid X_0 = x_0) \mathbf{E}_{x_1,o_1} \ldots \\ &\quad \sum_{x_{\ell-1}} P(X_{\ell-1} = x_{\ell-1} \mid X_{\ell-2} = x_{\ell-2}) \mathbf{E}_{x_{\ell-1},o_{\ell-1}} P(X_\ell = i \mid X_{\ell-1} = x_{\ell-1}) \mathbf{E}_{i,o_\ell} \end{aligned} \tag{2.8}$$

Given an initial distribution $\pi(0)$ for the hidden states we can write the above sum as the following vector-matrix product to get the row vector containing the forward probabilities of Eq. (2.8) for every hidden state $i \in S$, i.e.,

$$f_\ell = \pi(0)\mathbf{T}_0\mathbf{P}\mathbf{T}_1\mathbf{P}\ldots\mathbf{T}_{\ell-1}\mathbf{P}\mathbf{T}_\ell,$$

where $\mathbf{T}_n,\ n = \{0, \ldots, k\}$, is a diagonal matrix of the same size as $\mathbf{P}$ whose diagonal entry of i-th row is equal to $\mathbf{E}_{i,o_n}$. Exploiting in a very similar way the probability $b_\ell(i) = P(o_{\ell+1:k} \mid X_\ell = i)$ it arises that we can get the conditional probability (column) vector of the backward algorithm as:

$$b_\ell^\intercal = \mathbf{P}\mathbf{T}_{\ell+1}\ldots\mathbf{P}\mathbf{T}_k\phi_k^\intercal,$$

where $\phi_k = 1^{|S|}$.

Note that the combined use of the forward-backward algorithm makes only sense in case we are interested in finding the most possible hidden state for an o_ℓ in the middle of the observation sequence. If $\ell = k$, then it is enough to compute only the forward probability vectors.

Baum-Welch Finally, we mention that it is possible to estimate the initial distribution π_0, as well as the transition and the emission probabilities of an HMM, given an observation sequence o. If θ is a vector that contains all unknown parameters, the Baum-Welch algorithm iteratively alternates between computing the probability distribution of the hidden states $P(X \mid o, \theta)$, $X = \{X_0, \ldots, X_k\}$ using the forward-backward algorithm, and estimating the parameter vector value θ^* that maximizes the conditional expected log-likelihood $\mathbb{E}_{X|o,\theta}[\log P(X \mid \theta)]$ until a predefined convergence.

Viterbi algorithm

For a given sequence of observations $o = o_{0:k}$ we would like to find the most likely sequence of hidden states $x^* = x_0^*, x_1^*, \dots, x_k^*$. The Viterbi algorithm determines $x^* = \operatorname{argmax}_x P(x \mid o)$, which is due to the Bayes law equivalent to $x^* = \operatorname{argmax}_x P(x, o)$, by iteratively computing for all $i \in S$ the value

$$w_\ell(i) = \max_{x_0, x_1 \dots x_{\ell-1}} P(X_0 = x_0, \dots, X_{\ell-1} = x_{\ell-1}, X_\ell = i, o_{0:\ell})$$

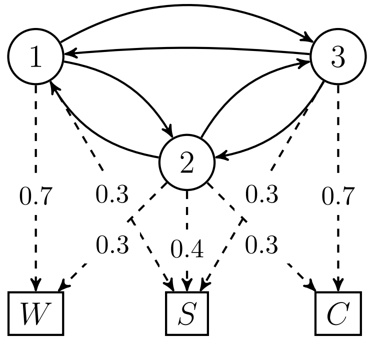
for $\ell \in \{0, \dots, k\}$. The value $w_\ell(i)$ gives the maximal probability of all paths that produce the observed sequence $o_{0:\ell}$, and end up in state i after ℓ steps. Note, then, that $\max_i w_k(i) = P(x^*, o)$ gives us the maximal probability among all paths that produce the whole observed sequence o. By induction it is not hard to show that for all $i, j \in S$ and $\ell = \{0 \dots, k-1\}$ it holds

$$w_{\ell+1}(j) = \max_i \left(w_\ell(i)\ p_{ij}\right) \mathbf{E}_{j, o_{\ell+1}}.$$

Hence, we can iteratively compute the vectors $w_0, w_1, \dots, w_k$ and track the sequence of states that corresponds to the maximum as described in Algorithm 1.

Example 3: A simple HMM

▷**Example 3**



Algorithm 1: Viterbi algorithm

```
Input: $\pi(0), \mathbf{P}, \mathbf{E}, o$
Output: $P(x^*, o), x^*$
Initialize $w_0$ with entries $w_0(i) = \pi(i, 0)\mathbf{E}_{i,o_0}$.
For all $\ell \in \{1, 2, \dots, k\}$ compute the vector $w_\ell$ with entries
    $$w_\ell(j) = \max_i \left(w_{\ell-1}(i)\ p_{ij}\right) \mathbf{E}_{j,o_\ell}, \qquad j \in S$$
  and store the states that correspond to the maximum in the vectors $s_\ell$ with
  entries
    $$s_\ell(j) = \arg\max_i (w_{\ell-1}(i)\ p_{ij}), \qquad j \in S.$$
Return the probability $P(x^*, o) = \max_i w_k(i)$ and the last maximal state
  $x^*_k = \text{argmax}_i\, w_k(i)$.
Backtrack the remaining states of the optimal sequence $x^*$ by setting
    $$x^*_\ell = s_{\ell+1}(x^*_{\ell+1}), \qquad \ell = k-1, \dots, 1, 0$$
```

Consider the DTMC of Example 1 where the DTMC's transition and the self-loops probabilities remain the same (not shown here). Imagine, though, the Markov process describing each day's weather can not be directly observed and the only thing that we, as external observers, know is a certain person's activity: whether she went for a walk (W), she went shopping (S), or she cleaned (C). The transition probability matrix $\mathbf{P}$ and the emission probability matrix $\mathbf{E}$ follow.

$$\mathbf{P} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.3 & 0.2 & 0.5 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}, \quad \mathbf{E} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} W & S & C \end{array} \\ \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.3 & 0.7 \end{pmatrix} \end{array}.$$

Given $\pi(0) = (0, 0.5, 0.5)$ the probability of observing the sequence $o = S, C, W$ is $P(o) = \pi(0)\mathbf{T}_0\mathbf{P}\mathbf{T}_1\mathbf{P}\mathbf{T}_2\mathbf{e} = 0.0372$, where

$$\mathbf{T}_0 = \begin{pmatrix} 0.3 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0 & 0 & 0.3 \end{pmatrix}, \mathbf{T}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.7 \end{pmatrix}, \mathbf{T}_2 = \begin{pmatrix} 0.7 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In addition, using the steps of Viterbi algorithm described above we compute that the most likely sequence of hidden states is $x^* = \operatorname{argmax}_x P(x, o) = \{3, 2, 1\}$ and the joint probability of this sequence and observation o is $P(x, o) = 0.00378$.

2.2 Experimental Methods

In this section we describe in short the two main biological experiments from which the data to produce the results of this thesis was extracted. The two experimental protocols developed and executed in the Epigenetics Lab of Saarland University by P. Giehr. The first protocol, named HPoxBS, is described in detail in Giehr et al. 2018 [22] and provides us with double strand (hydroxy-)methylation data for individual genomic loci, whereas the second protocol RRHPoxBS is its genome-wide analog.

2.2.1 Protocol HPoxBS for individual loci

The experimental protocol hairpin oxidative bisulfite sequencing (HPoxBS) is the first protocol that produces data that allows for the tracking of 5hmC fate at a single base resolution level. To get measurements that provide double strand (hydroxy-)methylation information we applied for the first time a hairpin linker setup on DNA samples taken from both bisulfite sequencing (BS-seq) and oxidative bisulfite sequencing (oxBS-seq) methods. The combination of this novel application of the hairpin linker [50] on both BS-seq and oxBS-seq with our specific modeling approach presented in Chapter 3 is what gives us the possibility to determine the levels of each modification status at both sites of any individual CpG dyad within a sequenced loci. In addition, the incorporation of the hairpin linker is what also allows us to accurately measure possible conversion errors during the two treatments.

To achieve a sufficient coverage (>1000x) per CpG we used a very deep NGS based sequencing on selected DNA loci referred as “single-copy genes”. To additionally cover larger parts of the genome we also included the analysis of mobile elements which occur in multiple identical copies across the genome and to which we refer as “repetitive elements”. 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 7% of mSat (N=44) elements. In the case of these multi-copy regions the >1000x coverage has to be seen as the aggregate coverage of all individual copies of a given repetitive element.

Protocol 1: HPoxBS

▷Protocol 1

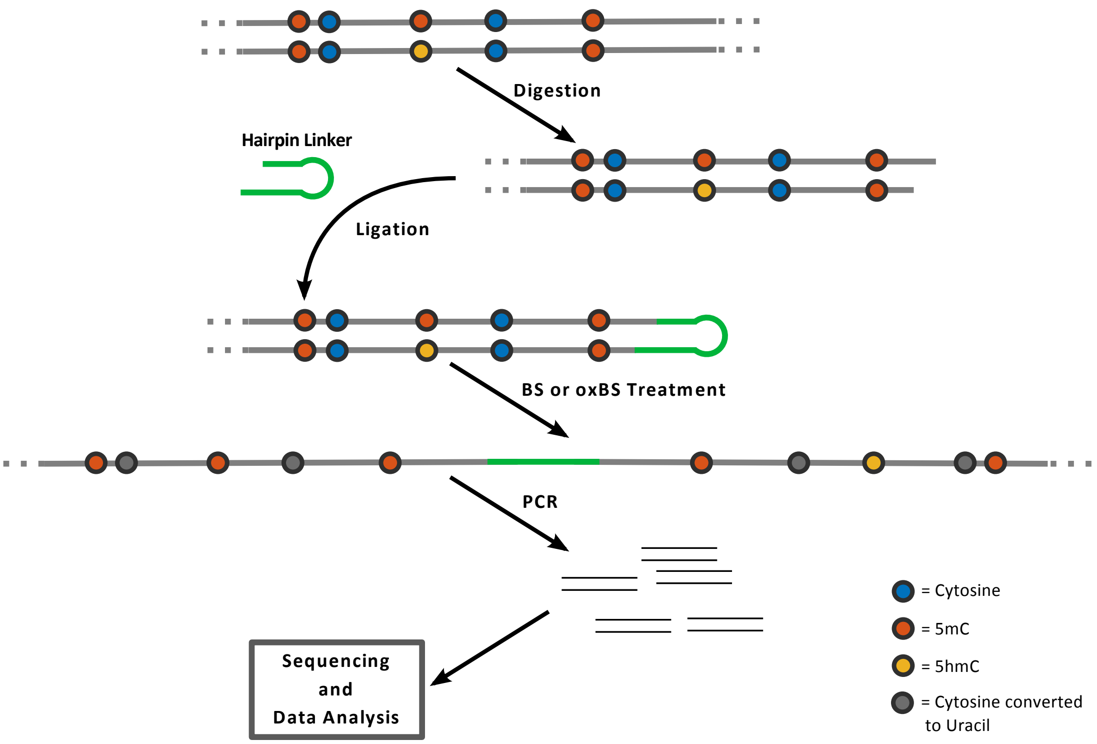


Figure 2.1: Schematic outline of HPoxBS. The hairpin BS and hairpin oxBS sequencing methods are based on enzymatic digestion of genomic DNA and the covalent connection of upper and lower DNA strands by ligating a hairpin oligonucleotide. PCR enrichment of the BS/oxBS treated sample is used for amplicon generation followed by sequencing and data analysis.

Figure 2.1 outlines the main experimental steps of the HPoxBS protocol. In the first step the region to be analyzed is located via the reference sequence, and the 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 platform (Illumina®) compatible sequencing adapters to perform deep (paired end 2x300bp) sequencing (up to 10K/product). Sequencing data are processed using the in-house software BiQ-HT and a Python script. In the bisulfite only reaction 5mC (red) and 5hmC (yellow) remain as cytosines, whereas the unmodified C (gray) is converted to uracil and finally to thymine. In the oxidative bisulfite reaction only 5mC remains as cytosine, and 5hmC, C are both converted to uracil/thymine.

Coversion errors Each individual sequence includes 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 and calculate for each modification accurate conversion rates by dividing the number of wrongly converted bases by the total number of them.

Additional information about the protocol is given in Section A.1 together with more details about the reference-, primer- and linker-sequences that have been used.

2.2.2 Protocol RRHPoxBS for the whole genome

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 [99] but come together with a very high sequencing cost, and demands large amounts of DNA. The RRHPoxBS protocol restricts the analysis to around 3 million CpGs equally distributed across the genome, and as a result it reduces the sequencing cost and provides a higher coverage for the informative CpGs. In addition, its pipeline only uses about one tenth of the DNA amounts formerly needed and can probably be scaled down further. Below we highlight the differences of the RRHPoxBS protocol compared to HPoxBS.

Protocol 2: RRHPoxBS

▷Protocol 2

Figure 2.2 outlines the main steps of the RRHPoxBS protocol. First DNA is divided into three equal parts, each digested with a distinct restriction enzyme (AluI, HaeII and HpyCH4V) that cuts a distinct sequence of bases (1). As a result, short DNA fragments are formed (2) and be subjected to ligation of the hairpin linker and the sequencing adapter (3). As ligation is a stochastic process, it might, besides fragments with sequencing adapter on one side and hairpin on the other, also create unwanted fragments with sequencing adapters on both sides[a] (4). However, because hairpins carry biotin molecules which are exclusively used to bind to streptavidin beads, the non hairpin fragments, i.e., sequencing adapter on both sides, will not bind to the beads and subsequently not processed for sequencing. Similar to HPoxBS, the hairpin fragments are split then in two parts, with the first part subjected to BS and the second to oxBS treatment, respectively. The treated DNA will finally be amplified and prepared for sequencing by PCR using primers specific for the sequencing adapters (5).

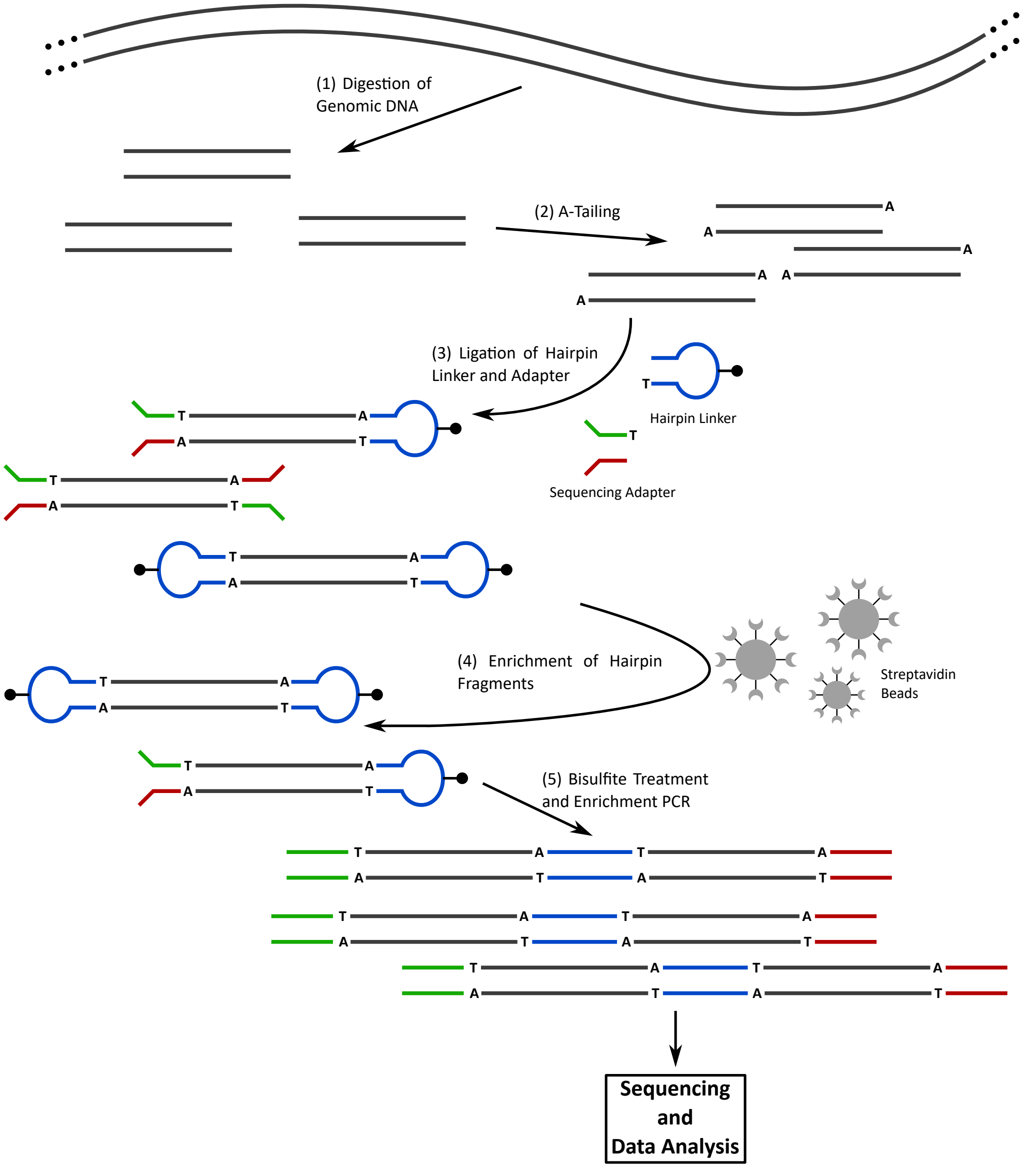

(1) Digestion of Genomic DNA
(2) A-Tailing
(3) Ligation of Hairpin Linker and Adapter
Hairpin Linker
Sequencing Adapter
(4) Enrichment of Hairpin Fragments
Streptavidin Beads
(5) Bisulfite Treatment and Enrichment PCR
Sequencing and Data Analysis

Figure 2.2: Schematic outline of RRHPoxBS. Ligation can create fragments with sequencing adapters on both sides, but only hairpin fragments carrying biotin molecules are going to bind to streptavidin beads and processed for PCR amplification and sequencing.

[a] Ligation might create fragments with hairpin on both sides as well, but these will anyway not bind to the sequencer later on.

Chapter 3

The Modeling Approach

3.1 A Hidden Markov Model for Hydroxylation

The main 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 and describes the evolution of both methylation and hydroxylation patterns over time. The suggested model allows that methylation can be lost only as a result of cell replication while methyl groups can be added due to either maintenance or *de novo* enzyme activity [1, 85]. In addition, all methylated sites can potentially be hydroxylated within the cycle of one cell replication.

Since every epigenetic modification, such as 5mC or 5hmC, can be observed only implicitly in both bisulfite and oxidative bisulfite experiments we first define a DTMC which models the hidden state of a cytosine and then for each of the two sequencing methods we define a hidden Markov model (HMM) that represents the output of the method. For each experiment we construct a likelihood function based on a time series observational data for several different days, where the time span of one day corresponds to a cell replication cycle.

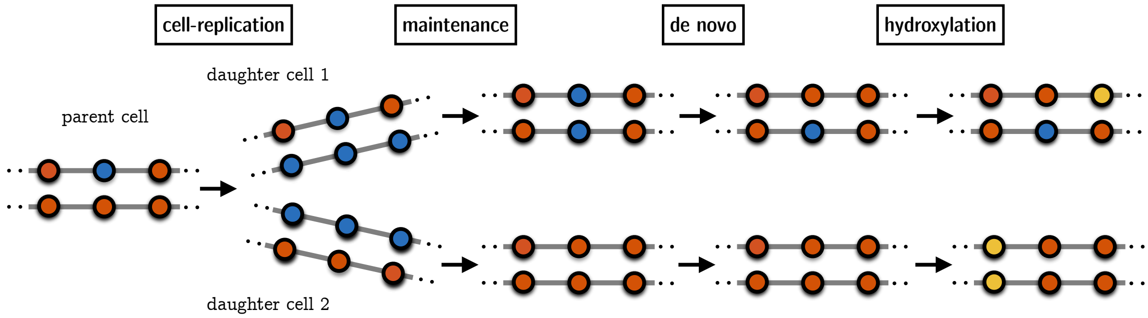


Figure 3.1: DNA demethylation: When a cell divides each of the daughter cells keeps the epigenetic information only of either the upper or the lower parental strand. The complementary strand of the new cell has only unmethylated cytosines (blue color). maintenance machinery which is associated with cell replication methylates a hemimethylated CpG. *De novo* methylation is known to act independently of the status of the opposite CpG site and follows maintenance. At last an already methylated cytosine (5mC, red color) can be oxidized by Tet enzymes to a hydroxy-cytosine (5hmC, yellow color).

If we are interested only in the (hydroxy)-methylation levels we can combine the two likelihoods, in an approach similar to [74] and derive estimations for (hydroxy-)methylation levels at each time point. In order to exploit further the mechanisms behind DNA demethylation and be able to shed light on how the different enzymes are involved in a demethylation process we increase the complexity of our model. We relate the unknown model parameters to the activities of these enzymes over time and not to the levels, i.e, probabilities, of the HMM's hidden states. We automatically get predictions for the hidden states' probabilities during the process of estimating the HMM's unknown parameters via a numerical optimization approach.

Hence, based on observations at many different time points the combination of the two likelihoods allows us to first determine the initial distribution of the hidden states, and next the methylation and hydroxylation efficiencies over time. As a consequence, our model accurately predicts the evolution of the (hydroxy-)methylation patterns, while it enables us to test different assumptions about the activities of the involved enzymes.

3.1.1 Hidden state space

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 the set $\mathcal{S} = \{u, m, h\}^2$. 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, in state $(s_1, s_2) = (u, h)$ the upper strand is unmethylated and the lower strand is hydroxylated. For the simplicity of notation we will often in this thesis 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 (or subprocesses): 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 always 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. This makes the Markov chain time-inhomogeneous and the whole analysis quite more complicated. In the sequel we give a detailed description of each subprocess and the corresponding matrices $\mathbf{D}(t)$, $\mathbf{M}(t)$, and $\mathbf{H}(t)$.

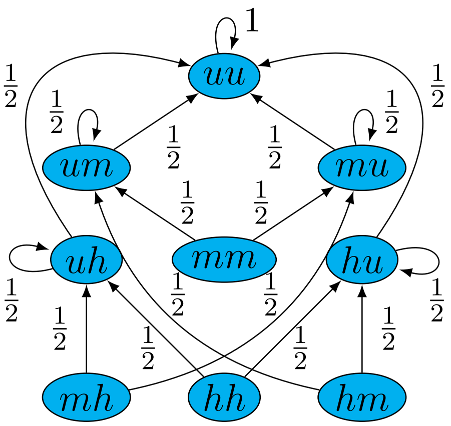


Figure 3.2: Cell Division: Possible transitions of the 9 different states of a CpG site, where 5mC or 5hmC marks of either the lower or the upper strand are removed after a cell divides.

Demethylation through cell division

DNA replication/cell division temporarily results in a direct loss of methyl or hydroxyl groups. During a DNA replication each of the two daughter cells is created by one parental strand and a newly synthesized strand. While the epigenetic pattern of the parental strand remains unchanged, the newly formed strand initially consists only by unmodified cytosines. Hence, a previously methylated CpG site in a daughter cell keeps only half of its (hydroxy-)methylated state. By averaging over all daughter cells, if the current state is (mm) then immediately after cell division the new state is either (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). The transition probabilities of the remaining states are defined in a similar way and we illustrate the corresponding matrix $\mathbf{D}(t)$ in Figure 3.2.

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

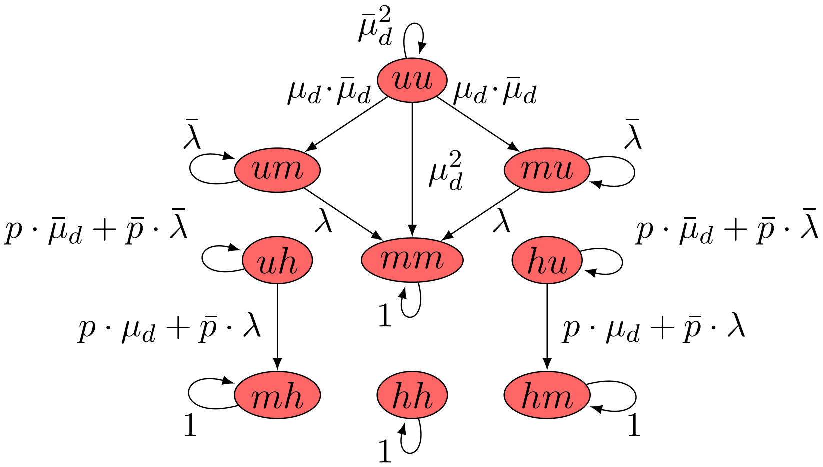

$\bar{\mu}_d^2$
uu
$\mu_d \cdot \bar{\mu}_d$
$\mu_d \cdot \bar{\mu}_d$
$\bar{\lambda}$
um
μ_d^2
mu
$\bar{\lambda}$
$p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda}$
λ
λ
$p \cdot \bar{\mu}_d + \bar{p} \cdot \bar{\lambda}$
uh
mm
hu
1
$p \cdot \mu_d + \bar{p} \cdot \lambda$
$p \cdot \mu_d + \bar{p} \cdot \lambda$
mh
hh
hm
1
1
1

Figure 3.3: Methylation: 5mC marks are added due to maintenance (μ_m) or *de novo* methylation (μ_d). For the methylation on a hemimethylated dyad we allow both maintenance and *de novo* methylation, while we make the assumption that *de novo* enzymes' efficiency is the same on hemi- and unmethylated sites.

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 [68] and the in vivo results of [1], we assume that Dnmt3a/b act on hemimethylated sites with the same efficiency as on unmethylated sites.

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 Figure 3.3. 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 Figure 3.3). 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 Figure 3.4. 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.

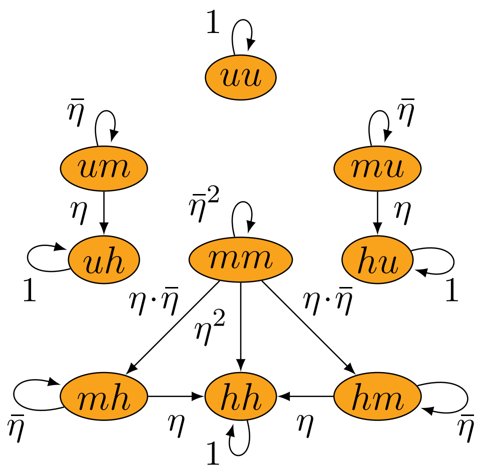


Figure 3.4: Hydroxylation: Probabilities that 5mC marks are hydroxylated (η) by Tet enzymes under the assumption that the hydroxylation of the two sites of a CpG dyad happens independently.

3.1.2 Observable state space and conversion errors

Before defining the observable states and the corresponding emission probabilities, we have to remind the reader some details of the hairpin (oxidative) bisulfite sequencing. After the DNA is cut by a restriction enzyme, the DNA fragments are linked covalent to a hairpin linker resulting in the connection of upper and lower strand. The resulting hairpin fragments are divided into two halves, from which one is treated with a standard bisulfite reaction and the other is subjected to an oxidation followed by bisulfite treatment.

As a result of the above procedure, in the ideal case of no conversion errors, an unmethylated cytosine (C) will first transform to uracil (U) and it will appear after sequencing as thymine (T) in both oxidative (oxBS) and non-oxidative bisulfite (BS) treatment. A methylated cytosine (5mC) will stay unaffected and present itself as a cytosine (C) after sequencing, again during both treatments. On the contrary, a hydroxylated cytosine (5hmC) will remain unaffected and appear as C at the end of a simple BS treatment, whereas it will get oxidized to 5fC, converted during bisulfite to 5fU, and appear as T at the end of the oxBS experiment.

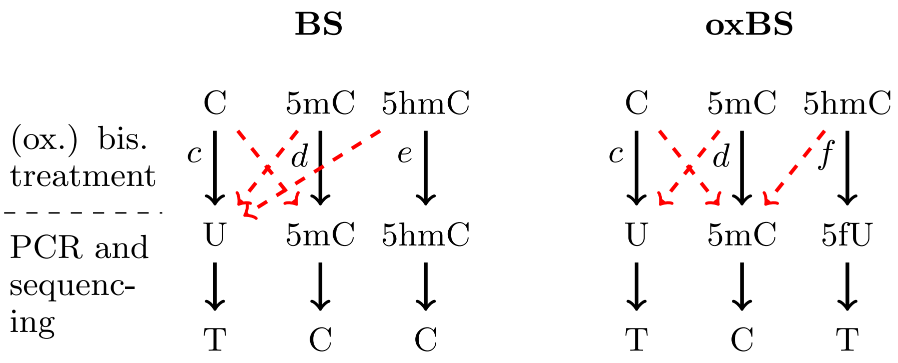


Figure 3.5: Schematic outline of the conversion of Cytosine, 5mC and 5hmC during BS and oxBS treatment and after sequencing: In the bisulfite reaction a cytosine (C) is converted to uracil (U), whereas 5mC and 5hmC remain untouched. In the oxidative bisulfite sequencing only 5mC remains untouched and 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.

Since every base will eventually transform into a thymine (T) or a cytosine (C), the set of the observable states for a CpG dyad with two cytosines is $\mathcal{S}_{obs} = \{\mathrm{T}, \mathrm{C}\}^2$. All transitions from a site's possible hidden states to the observable ones are shown in Figure 3.5, where the blacked arrows indicate the conversion probabilities c, d, e, f and the red dashed arrows correspond to possible conversion errors. The entries of the complete emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ for the transitions from hidden to all observable states in each treatment can be found in Table 3.1.

Measuring the conversions errors To accurately measure the conversion probabilities c, d, e, f, we artificially incorporated an unmodified cytosine, as well as a 5mC and a 5hmC into the hairpin linker (see A.1.1 for more technical details) and computed the ratio of right conversions in the individual measurements (around $5 \cdot 10^3$x on average). Note that the values of c and d can differ between the two treatments and that the conversion probabilities can also differ over time.

	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	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot d$	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot d$
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	$\bar{d} \cdot \bar{e}$	$d \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot e$	$\bar{d} \cdot f$	$d \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot \bar{f}$
mh	$\bar{d} \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot \bar{e}$	$d \cdot e$	$\bar{d} \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot f$	$d \cdot \bar{f}$
mm	$\bar{d}^2$	$d \cdot \bar{d}$	$d \cdot \bar{d}$	d^2	$\bar{d}^2$	$d \cdot \bar{d}$	$d \cdot \bar{d}$	d^2
hh	$\bar{e}^2$	$e \cdot \bar{e}$	$e \cdot \bar{e}$	e^2	f^2	$f \cdot \bar{f}$	$f \cdot \bar{f}$	$\bar{f}^2$

Table 3.1: Transition probabilities from hidden to the observable states in BS and in oxBS. The conversion errors of single bases $\bar{c}, \bar{d}, \bar{e}, \bar{f}$ represent the complements $1-c, 1-d, 1-e, 1-f$ of a right conversion probability.

3.2 Estimation of Model Parameters

Given the number of times $n_{bs}(j,t)$ and $n_{ox}(j,t)$ that state $j \in \mathcal{S}_{obs} = \{\mathrm{T},\mathrm{C}\}^2$ has been observed during independent BS and oxBS measurements at time t we use a maximum likelihood approach to estimate the unknown parameters of the HMMs. These are, the initial distribution of the hidden states, $\mathcal{S} = \{u,m,h\}^2$, the unknown functions $\mu_d(t), \mu_m(t)$ and $\eta(t)$ of the enzymatic efficiencies, as well as the probability p at which a 5hmC site is not considered during maintenance.

Formally, let $\pi(t)$ be the row vector of the (hidden) state probabilities of DTMC $\mathcal{X}$ after t cell divisions, where $\pi(0)$ is the initial distribution of the hidden states in serum conditions. Both HMMs corresponding to BS and oxBS have the same distribution $\pi(t)$ for the hidden states as for both experiments the same cell population is used and when the parameter values are fixed, for a time-inhomogeneous Markov chain $\pi(t)$ is given by the equation

$$\pi(t) = \pi(0) \cdot \prod_{k=1}^{t} \mathbf{P}(k). \tag{3.1}$$

For $i \in \mathcal{S}$ let $\pi(i,t) = P(\mathcal{X}(t) = i)$ denote the entry of $\pi(t)$ that corresponds to state i. In general, the probability of observing state $\mathcal{O}(t) = j \in \mathcal{S}_{obs}$ at time t is given according to the law of total probability 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 $P(\mathcal{O}(t) = j \mid \mathcal{X}(t) = i)$ is the emission probability of observing j, when the hidden state is $i \in \mathcal{S}$. For the two treatments BS and oxBS this yields in matrix-vector form

$$\pi_{bs}(t) = \pi(t) \cdot \mathbf{E}_{bs}(t) \quad \text{and} \quad \pi_{ox}(t) = \pi(t) \cdot \mathbf{E}_{ox}(t), \tag{3.2}$$

respectively, where $\pi_{bs}(t)$ and $\pi_{ox}(t)$ are the vectors with the distribution over the observable states at time t.

Our strategy to estimate the “optimal” values for the unknown parameters is to first estimate an “optimal” initial distribution for the hidden states, and then based on this estimation and Eq. (3.1), (3.2) to apply a numerical optimization algorithm to find the values of the enzymatic efficiencies that maximize the likelihood (MLE) of our data. Note that due to the high coverage, the information for the observable states given at $t = 0$ suffices for an accurate estimation of the initial hidden states' distribution. We choose, hence, to split the parameter inference into two (global) optimization problems in order to avoid the curse of dimensionality of one high-dimensional parameter space. Each optimization problem involves the maximization of a likelihood for the solution of which it is convenient to minimize its negative logarithm. Deriving expressions for the first and second derivatives of the log-likelihood is in both cases possible and it guarantees a fast convergence of a gradient descent optimization routine (interior-point method), even with multiple starting values. A formal description of the two optimization problems follows in Section 3.2.1, 3.2.2.

3.2.1 Initial hidden states' distribution

Let $\mathcal{L}_1$ be the combined likelihood of the observed data in conventional serum conditions. Due to the small number of cells measured compared to the size of the pool that they are sampled from around 10^7, it is very unlikely that we pick two cells with a common descendant and it is consequently meaningful to assume independence between the measurements. Hence, we get

$$\mathcal{L}_1(\pi(0)) = \prod_{j \in \mathcal{S}_{obs}} \pi_{bs}(j,0)^{n_{bs}(j,0)} \cdot \pi_{ox}(j,0)^{n_{ox}(j,0)} \tag{3.3}$$

We estimate the initial distribution $\pi(0)^*$ based on the initial independent BS and oxBS measurements by solving the following maximization problem:

$$\begin{aligned} \underset{\pi(0)}{\text{maximize}} \quad & \mathcal{L}_1(\pi(0)) \\ \text{subject to} \quad & \sum_i \pi(i,0) = 1. \end{aligned} \tag{3.4}$$

Since $\mathcal{L}_1$ depends only on the unknown vector $\pi(0)$ and on the known emission matrices we can determine the initial distribution of the hidden states, if we maximize the likelihood over all vectors $\pi(0)$ that sum up to one.

To solve (3.4) we consider the log-likelihood

$$\log \mathcal{L}_1(\pi(0)) = \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,0) \cdot \log \pi_{bs}(j,0) + n_{ox}(j,0) \cdot \log \pi_{ox}(j,0)). \tag{3.5}$$

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

$$\frac{d}{d\pi(0)} \log \mathcal{L}_1(\pi(0)) = \sum_{j \in \mathcal{S}_{obs}} n_{bs}(j,0) \cdot \frac{\frac{d}{d\pi(0)}\pi_{bs}(j,0)}{\pi_{bs}(j,0)} + n_{ox}(j,0) \cdot \frac{\frac{d}{d\pi(0)}\pi_{ox}(j,0)}{\pi_{ox}(j,0)}. \tag{3.6}$$

Letting, now, $\pi_{bs}(t), \pi_{ox}(t)$ be the vectors with entries $\pi_{bs}(j,t), \pi_{ox}(j,t)$, $\forall j \in \mathcal{S}_{obs}, \forall t \in T_{obs}$, we can write the 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

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

which after insertion into Eq. (3.6) gives us the gradient of the log-likelihood function w.r.t. the initial distribution of the hidden states.

Finally, note that the above MLE procedure for estimating the initial distribution over the hidden states can be applied, exactly the same way, at any time point $t \in T_{obs}$ for which we have measurements, in order to estimate the distribution of the hidden states at this time point.

3.2.2 Estimation of the efficiency functions

Given an estimate for $\pi(0)$, and a fixed value for the parameter vector $\mathbf{v}$, we can compute for $t \in \{1, 2, \ldots\}$ the state probabilities $\pi(t)$ of the hidden states and consider the common likelihood

$$\mathcal{L}_2(\mathbf{v}) = \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j} \pi_{bs}(j,t)^{n_{bs}(j,t)} \cdot \pi_{ox}(j,t)^{n_{ox}(j,t)} \tag{3.7}$$

for the observations at all remaining observation time points $t \in T_{obs} \setminus \{0\}$. Here we assume that the cells divide every 24 hours, hence t ranges over all days at which measurements were made, and as above we assume independence between all $n_{tot} = \sum_{t \in T_{obs}} \sum_j n_{bs(j,t)} + n_{ox(j,t)}$ cell measurements.

The parameter vector $\mathbf{v}$ in Eq.(3.7) consists of the unknown functions $\mu_m(t), \mu_d(t)$, $\eta(t)$ and the unknown probability p, where μ_m stands for maintenance, μ_d for *de novo*, η for hydroxylation efficiency, and p is the probability that 5hmC is not considered during maintenance. Hence the value of $\mathbf{v}$ uniquely defines the DTMC's transition matrix $\mathbf{P}(t) = \mathbf{D}(t) \cdot \mathbf{M}(t) \cdot \mathbf{H}(t)$, where[1]

$$\mathbf{D} = \begin{array}{c|ccccccccc} & uu & um & mu & uh & hu & hm & mh & mm & hh \\ uu & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ um & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ mu & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ uh & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ hu & 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ hm & 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ mh & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ mm & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ hh & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \end{array},$$

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

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

and

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

In more detail, 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 seven unknown parameters, containing the coefficients of each efficiency linear time function, e.g., $\mu_m(t) = \beta_0^{\mu_m} + t \cdot \beta_1^{\mu_m}$, and the constant over time probability p. Given $\pi(0)^*$ as the solution of Eq.(3.4) we want to compute the MLE $\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \mathcal{L}_2(\mathbf{v})$. The constraints of this optimization problem arise from the requirement that the efficiencies should be probabilities for all the considered time points, and the same constraint should hold for p. Hence, the maximization problem we solve is:

$$\begin{aligned} \underset{\pi(0)}{\text{maximize}} \quad & \mathcal{L}_2(\mathbf{v}) \\ \text{subject to} \quad & 0 \leq p \leq 1 \text{ and} \\ & 0 \leq \beta_0^x + \beta_1^x \cdot t \leq 1, \ \forall x \in \{\mu_m, \mu_d, \eta\}, \ \forall t \in T_{obs}. \end{aligned} \tag{3.8}$$

Note that the above constraints are linear with straightforward algebra they can be deduced to $0 \leq \beta_0^x \leq 1$ and $\frac{-1}{t_{\max}} \leq \beta_1^x \leq \frac{1-\beta_0^x}{t_{\max}}$.

To ease the computations of Eq. (3.8) we derive the log-likelihood

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

we get the score-vector w.r.t. to $\mathbf{v}$

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

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

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

where the entries of the emission matrices $\mathbf{E}_{bs}(t)$ and $\mathbf{E}_{ox}(t)$ are given in Table 3.1.

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

$$\begin{aligned} \pi(t) &= \pi(t-1) \cdot \mathbf{P}(t) \\ \frac{d}{d\mathbf{v}} \pi(t) &= \frac{d}{d\mathbf{v}} \pi(t-1) \cdot \mathbf{P}(t) + \pi(t-1) \frac{d}{d\mathbf{v}} \mathbf{P}(t), \ \ \forall t \geq 1 \end{aligned} \tag{3.9}$$

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 the parameter vector $\mathbf{v}$. Applying once more the product rule in Eq. (3.9), and using similar arguments as above we, additionally, 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 $\mathcal{H} = \nabla\nabla^{\mathrm{T}}$ $\log \mathcal{L}_2(\mathbf{v})$.

Confidence intervals of the estimators

Due to the large number of samples (Table A.4, A.5) 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 [16].

In particular, 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 for a linear function $f(t) = \beta_0 + \beta_1 \cdot t$ it holds that $\mathrm{Var}(f(t)) = \mathrm{Var}(\beta_0 + \beta_1 \cdot t) = \mathrm{Var}(\beta_0) + t^2\,\mathrm{Var}(\beta_1) + 2t\,\mathrm{Cov}(\beta_0, \beta_1)$.

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

$$\begin{aligned} \mathrm{Var}(\lambda) = {} & \mathrm{Var}(\mu_m) + \mathrm{Var}(\mu_d) + 2\,\mathrm{Cov}(\mu_m, \mu_d) + \mathrm{Var}(\mu_m\mu_d) \\ & - 2\,\mathrm{Cov}(\mu_m, \mu_m\mu_d) - 2\,\mathrm{Cov}(\mu_d, \mu_m\mu_d), \end{aligned} \tag{3.10}$$

For Eq. (3.10) we compute

$$\mathrm{Cov}(\mu_m, \mu_d) = \mathrm{Cov}(\beta_0^{\mu_m}, \beta_0^{\mu_d}) + t\,\mathrm{Cov}(\beta_0^{\mu_m}, \beta_1^{\mu_d}) + t\,\mathrm{Cov}(\beta_1^{\mu_m}, \beta_0^{\mu_d}) + t^2\,\mathrm{Cov}(\beta_1^{\mu_m}, \beta_1^{\mu_d}),$$

using basic properties of the covariance and we exploit the last three terms

$$\mathrm{Var}(\mu_m\mu_d) = \mathbb{E}[\mu_m^2\mu_d^2] - \mathbb{E}[\mu_m\mu_d]^2 \tag{3.11}$$

$$\mathrm{Cov}(\mu_m, \mu_m\mu_d) = \mathbb{E}[\mu_m^2\mu_d] - \mathbb{E}[\mu_m]\,\mathbb{E}[\mu_m\mu_d] \tag{3.12}$$

$$\mathrm{Cov}(\mu_d, \mu_m\mu_d) = \mathbb{E}[\mu_d^2\mu_m] - \mathbb{E}[\mu_d]\,\mathbb{E}[\mu_m\mu_d] \tag{3.13}$$

using the identity $\mathbb{E}[XY] = \mathrm{Cov}(X, Y) + \mathbb{E}[X]\,\mathbb{E}[Y]$ for random variables X, Y. Since the MLEs are approximately normally distributed, every linear combination of them such as μ_m, μ_d and η follows also a normal distribution. From the raw third and fourth moments of the bivariate normal distribution we get then

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

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

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

and hence all terms in Eq. (3.11) - (3.13) 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 $z = F^{-1}\big(\frac{\beta+1}{2}\big)$ and F is the cumulative distribution function (cdf) of the standard normal distribution. Similarly, we get the confidence intervals for all remaining parameters.

Hypothesis testing

For estimates taken from maximum likelihood a number of hypotheses tests, such as likelihood ratio, score test, or Wald test are possible. Here we describe the details of the Wald test which is the one we mainly used to validate our results, because of the easiness it offers in testing multiple hypotheses in parallel. We mention that the use of all the alternative tests mentioned before returned similar p-values for our estimates and did not lead to a different result regarding the cases that one rejects H_0.

Wald test 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 asymptotic covariance matrix, here equal to the inverse of 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 distribution of $W_{\mathbf{v}^*}$ converges to a Chi-square distribution with r degrees of freedom [89].

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

3.2.3 Model validation

Leave One Out Cross Validation

To further exclude the possibility of over-fitting for our model due to the imposed linear assumption for the efficiencies, we perform leave-one-out cross-validation (LOOCV) to estimate the test error of our model with constant efficiencies versus a linear model. More precisely, we compute the test error of the model by performing LOOCV for the two following competing assumptions:

1. "The enzymes' efficiencies are constant over time"
2. "The enzymes' efficiencies are linear over time".

We test the prediction of the model for each single CpG, having trained it on the data of the other CpGs and finally we average 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 data distribution P and the distribution Q predicted by the model. Kullback-Leibler (KL) divergence is defined as $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$ and the Bhattacharyya (BC) distance as $BC(P,Q) = -\log\big(\sum_i \sqrt{P(i)Q(i)}\,\big)$, where i goes here over the observable states.

Sensitivity Analysis

To validate the robustness of the model, a sensitivity analysis of the parameters has been conducted. Perturbing one parameter at a time (OAT) by ±1% we examined the absolute change of the model's output, i.e., in this case the predicted levels of the hidden states.

3.3 A Hybrid Generalization of the Model

For the core-model of this thesis described in Section 3.1 we made the following two main assumptions:

1. There is no *active demethylation* (Figure 3.6 right) in the system. Once a 5hmC is formed, the only way to fade out is via cell replication and *passive demethylation* (Figure 3.6 left).

2. All methylation and hydroxylation events are discrete time events that happen once after each cell division.

Assumption 1 was obligatory since according to the given information from BS and oxBS experiments there is no way to distinguish between a simple cytosine C and a formylcytosine (5fC) or a carboxylcytosine (5caC). Assumption 2 was meaningful because we were interested only in the total effect of each subprocess within a cell replication cycle.

Nevertheless, in reality the formation of 5fC/5caC and active demethylation can potentially happen in both kinds of cells; cells that undergo, but also cells that do not undergo mitosis (cell division). In dividing ESCs active demethylation might contribute to the demethylation dynamics and to the effect of passive demethylation to a larger or smaller extent. In systems, on the other hand, where no cell division happens we know with certainty that active demethylation is the exclusive reason of

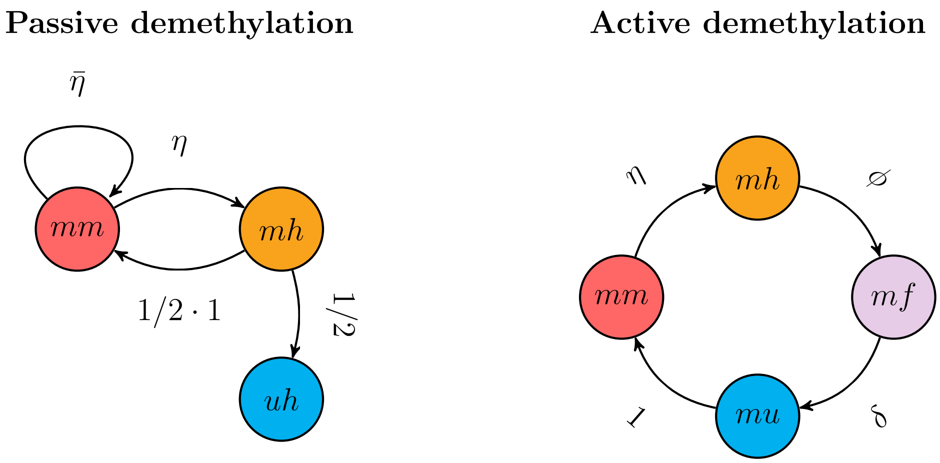


Figure 3.6: Let for simplicity the enzymes of the system act only on the lower strand, maintenance methylation to be perfect, i.e, $\mu_m = 1$, and assume 5hmC is not recognized by the maintenance machinery. Left: The DTMC of a fully methylated CpG dyad during passive demethylation. If a methylated site gets hydroxylated, the dyad will return to mm state only in one from two daughter cells. In any case, though, we are not going to observe a loss of 5mC before the next cell division. Right: The CTMC of a fully methylated CpG dyad during active demethylation. An unmethylated (u) site gets methylated (m) by the maintenance machinery, then it becomes hydroxylated (h), this turns into (formyl-)/(carboxyl-)cytosine (f) and then it goes back to unmethylated cytosine (u) within one cell division period.

the methylation loss over time. Examples of such cells are monocyte-macrophages, cardiac muscle cells, neurons, osteocytes and the majority of mature blood cells.

Hence, given data from a sequencing method such as mabBS-seq, fCAB-seq, redBS-seq, or CAB-seq, that is able to identify further oxidative forms 5fC/5caC, it might be of interest to compute how many times a particular subprocess, e.g., *de novo*, or hydroxylation, has happened within a certain time interval or to answer questions such as “what is the average time of a whole DNA demethylation cycle” (Figure 3.6 right). In this case the only subprocesses that needs to be described as events of discrete time are (only in case of dividing cells) the cell division and the subsequent maintenance methylation. All other events that follow, namely *de novo*, hydroxylation, formylation and active demethylation, can theoretically happen many times within a cell cycle and thus a continuous-time assumption is appropriate.

In the sequel we present an extension of the core-model in the following aspects:

- Given data from mabBS experiments in addition to BS and oxBS, we consider a third HMM that captures the observational states of mabBS and gives us the possibility to reveal the status of a further oxidized modification of 5hmC such as 5fC or 5caC. For this, we, naturally, extend the previous state space to $\{u, m, h, f\}^2$, where f accounts for both of the further oxidized forms 5fC and 5caC.

- We define a hybrid model that consists of both discrete- and continuous-time events. The discrete time events are cell replication and maintenance methylation. The continuous part of the model includes the subsequent *de novo*, hydroxylation, formylation, and active demethylation events, that happen within a cell replication cycle, possibly multiple times.

- Motivated by preliminary results from the application of the core-model in different cell types from ESCs, such as monocyte-macrophages, we generalize the behavior of the system's enzymes by loosing the previously imposed linear time constraints. Using instead piecewise polynomials of a certain degree, we can capture more complicated activity patterns over time and potentially apply the model to various cell types.

3.3.1 Estimation of model parameters

Initial distribution of the hidden states

Similarly to Section 3.2.1 let $\pi(0)$ be the unknown initial distribution of the 16, here, hidden states and let $\pi(i, t) = P(\mathcal{X}(t) = i)$ represent the entry of $\pi(t)$ that corresponds to state $i \in S$. To compress the notation we define, now, the set of the different sequencing methods $E = \{bs, ox, mab\}$, where bs stands for bisulfite, ox for oxidative bisulfite and mab for mab bisulfite sequencing. We denote as $n_e(j, t)$ the number of times that state $j \in \mathcal{S}_{obs}$ has been observed during independent measurements of sequencing method $e \in E$, as $\mathbf{E}_e(t)$ its emission matrix of at time point t and as $\pi_e(t)$ the probability distribution over the observable states of this method.

We 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 the likelihood to be maximized is

$$\mathcal{L}_1(\pi(0)) = \prod_{e \in E} \prod_{j \in \mathcal{S}_{obs}} \pi_e(j,0)^{n_e(j,0)}.$$

In a similar fashion to Section 3.2.1 we consider the log-likelihood

$$\log \mathcal{L}_1(\pi(0)) = \sum_{e \in E} \sum_{j \in \mathcal{S}_{obs}} n_e(j,0) \cdot \log \pi_e(j,0).$$

and deriving its derivative we apply a gradient descent numerical optimization procedure to compute ML estimator for the hidden states' distribution.

Estimation of the efficiencies of the hybrid HMM using cubic splines

In order to model a more complicated behavior of the enzymes we assume, here, that the efficiencies of the enzymes (probabilities and rates since we have a hybrid model) are polynomial functions over time. More concretely, in order to cope with possible over-fitting, we define the efficiencies of the model as piecewise cubic polynomials. Additionally, to impose smoothness on the efficiency functions, we require that the rate function, as well as its first and second derivative are continuous functions over time. The choice that simultaneously fulfills all previous criteria is cubic splines.

For an enzymatic efficiency function $r \in \{\mu_m, \mu_d, \eta, \phi, \delta\}$, where μ_m stands for maintenance, μ_d for *de novo*, η for hydroxylation, ϕ for formylation, and δ for active demethylation efficiency a cubic spline has the following form:

$$r(t) = \beta_0^r + \ldots + \beta_3^r t^3 + \beta_4^r h(t, \xi_1) + \ldots + \beta_{K+3}^r h(t, \xi_K), \tag{3.14}$$

where ξ_i, is the i-th knot, $i = \{1, \ldots, K\}$, and h is the truncated power basis function defined as

$$h(t,\xi) = (t,\xi)_+ = \begin{cases} (t-\xi)^3 & t \geq \xi \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{v}_r = \{\beta_0^r, \beta_1^r, \ldots \beta_{K+3}^r\}$ be the vector of the unknown parameters, i.e., the coefficients of a qubic spline function with K knots. Then $\mathbf{v} = (\mathbf{v}_{\mu_m}, \mathbf{v}_{\mu_d}, \mathbf{v}_\eta, \mathbf{v}_\phi, \mathbf{v}_\delta, p)$, is the vector of all unknown parameters of all efficiencies, and the probability p that 5hmC is not considered during maintenance.

We consider a continuous-time Markov chain (CTMC) with a discrete update step at every time unit that a cell division happens. Since cell division is always being followed by a maintenance event we define the transition matrix of the discrete Markov chain at time unit t is $\mathbf{P}(t) = \mathbf{D} \cdot \mathbf{M}_m(t)$, where

$$\mathbf{D} = \begin{array}{c} \\ uu \\ um \\ uh \\ uf \\ mu \\ mm \\ mh \\ mf \\ hu \\ hm \\ hh \\ hf \\ fu \\ fm \\ fh \\ ff \end{array} \begin{array}{c} \begin{array}{cccccccccccccccc} uu & um & uh & uf & mu & mm & mh & mf & hu & hm & hh & hf & fu & fm & fh & ff \end{array} \\ \left(\begin{array}{cccccccccccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0
\end{array} \right) \end{array},$$

$$\mathbf{M}_m(t) = \begin{array}{cc} & \begin{array}{cccccccccccccccc} uu & um & uh & uf & mu & mm & mh & mf & hu & hm & hh & hf & fu & fm & fh & ff \end{array} \\ \begin{array}{c} uu \\ um \\ uh \\ uf \\ mu \\ mm \\ mh \\ mf \\ hu \\ hm \\ hh \\ hf \\ fu \\ fm \\ fh \\ ff \end{array} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1-\mu_m & 0 & 0 & 0 & 0 & \mu_m & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-\mu_m+p\mu_m & 0 & 0 & 0 & \bar{p}\mu_m & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-\mu_m & \mu_m & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1-\mu_m+p\mu_m & \bar{p}\mu_m & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} .$$

After applying the discrete step we let the time run “continuously” multiplying with the $\mathbf{Q}(t)$ matrix

$$\mathbf{Q}(t) = \begin{array}{cc} & \begin{array}{cccccccccccccccc} uu & um & uh & uf & mu & mm & mh & mf & hu & hm & hh & hf & fu & fm & fh & ff \end{array} \\ \begin{array}{c} uu \\ um \\ uh \\ uf \\ mu \\ mm \\ mh \\ mf \\ hu \\ hm \\ hh \\ hf \\ fu \\ fm \\ fh \\ ff \end{array} & \begin{pmatrix} -2\mu_d & \mu_d & 0 & 0 & \mu_d & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\mu_d-\eta & \eta & 0 & 0 & \mu_d & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\mu_d-\phi & \phi & 0 & 0 & \mu_d & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \delta & 0 & 0 & -\mu_d-\delta & 0 & 0 & 0 & \mu_d & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\eta-\mu_d & \mu_d & 0 & 0 & \eta & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2\eta & \eta & 0 & 0 & \eta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\eta-\phi & \phi & 0 & 0 & \eta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta & 0 & 0 & -\eta-\delta & 0 & 0 & 0 & \eta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\phi-\mu_d & \mu_d & 0 & 0 & \phi & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\phi-\eta & \eta & 0 & 0 & \phi & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2\phi & \phi & 0 & 0 & \phi & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & -\phi-\delta & 0 & 0 & 0 & \phi \\ \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\mu_d-\delta & \mu_d & 0 & 0 \\ 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\delta-\eta & \eta & 0 \\ 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\delta-\phi & \phi \\ 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & -2\delta \end{pmatrix} \end{array} .$$

that incorporates the processes of *de novo* (μ_d), hydroxylation (η), formation of 5fC-5caC from $5hmC$, let us name it formylation, (ϕ), and active demethylation (δ), i.e., the transition from 5fC and 5caC back to unmethylated cytosine. The biological interpretation of the rates of the above processes is the number of events of the process happening within one cell replication.

Given, now, $\pi(0)^*$, we seek for the maximum likelihood estimator (MLE)

$$\mathbf{v}^* = \text{argmax}_{\mathbf{v}} \, \log\mathcal{L}_2(\mathbf{v}), \tag{3.15}$$

where

$$\mathcal{L}_2(\mathbf{v}) = \prod_{e \in E} \prod_{t \in T_{obs} \setminus \{0\}} \prod_{j \in \mathcal{S}_{obs}} \pi_e(j,t)^{n_e(j,t)}. \tag{3.16}$$

It holds

$$\log \mathcal{L}_2(\mathbf{v}) = \sum_{e \in E} \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_e(j,t) \cdot \log \pi_e(j,t)$$

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

$$\frac{d}{d\mathbf{v}} \log \mathcal{L}_2(\mathbf{v}) = \sum_{e \in E} \sum_{t \in T_{obs} \setminus \{0\}} \sum_{j \in \mathcal{S}_{obs}} n_e(j,t) \cdot \frac{\frac{d}{d\mathbf{v}} \pi_e(j,t)}{\pi_e(j,t)}.$$

To compute the above likelihood and its score vector we need the transient probabilities vector and its derivative w.r.t. the parameters. Let us assume we have L time points t_i for $i \in \{0, \ldots, L-1\}$, at which cell-division happens. Computing the transient distribution of the hidden states in our hybrid model means solving the Chapman-Kolmogorov Eq. (3.1) of the continuous Markov chain along with the derivatives equation system within every interval $[t_i, t_{i+1}]$. Hence, it suffices to simultaneously solve the following two differential equation systems.

$$\begin{aligned} \frac{d}{dt}\pi(t) &= \pi(t) \cdot \mathbf{Q}(t) \\ \frac{d}{d\mathbf{v}}\pi(t) &= \frac{d}{d\mathbf{v}}\pi(t) \cdot \mathbf{Q}(t) + \pi(t)\frac{d}{d\mathbf{v}}\mathbf{Q}(t), \quad \forall t \in [t_i, \ldots t_{i+1}] \end{aligned} \tag{3.17}$$

with $\pi(t_0) = \pi(0)^*$ and $\frac{d}{d\mathbf{v}}\pi(t_0) = 0$ updating in parallel the transient distribution $\pi(t_i) := \pi(t_i) \cdot P(t_i)$ and the derivatives vector

$$\frac{d}{d\mathbf{v}}\pi(t_i) := \frac{d}{d\mathbf{v}}\pi(t_i) \cdot \mathbf{P}(t_i) + \pi(t_i) \cdot \frac{d}{d\mathbf{v}}\mathbf{P}(t_i) \tag{3.18}$$

at every cell division time point, where the derivative of the transition matrix is $\frac{d}{d\mathbf{v}}\mathbf{P}(t) = \frac{d}{d\mathbf{v}}(\mathbf{D} \cdot \mathbf{M}_m(t)) = \mathbf{D} \cdot \frac{d}{d\mathbf{v}}\mathbf{M}_m(t)$. To compute the derivatives $\frac{d}{d\mathbf{v}}\mathbf{Q}(t)$ and $\frac{d}{d\mathbf{v}}\mathbf{M}_m(t)$ we have to apply the chain rule and writing $r(t)$ instead of $r(\mathbf{v}_r, t)$ we get $\forall\ \beta_i^r \in \mathbf{v}_r$:

$$\frac{d}{d\beta_i^r}\mathbf{Q}(t) = \frac{d}{dr(t)}\mathbf{Q}(t) \cdot \frac{d}{d\beta_i^r}r(t),$$

where r denotes one of $\{\mu_m, \mu_d, \eta, \phi, \delta\}$. Since the parameter of non-recognition probability p is not a function of time we can directly compute $\frac{d}{dp}\mathbf{M}_m(t)$.

Having computed the distribution of the hidden states and its derivative for all $t \in T_{obs}$, we get the the observable states' distribution

$$\pi_e(t) = \pi(t) \cdot \mathbf{E}_e(t) \quad \text{and} \quad \frac{d}{d\mathbf{v}}\pi_e(t) = \frac{d}{d\mathbf{v}}\pi(t) \cdot \mathbf{E}_e(t),$$

for a sequencing method $e \in E$. The entries of the emission matrices $\mathbf{E}_e(t)$ for all $e \in E$ are given in Table 3.2. A schematic outline of the conversion errors of each experiment which define the entries of the experiment's emission matrix is shown in Figure 3.7.

Applying once more the product rule in Eq. (3.17) and (3.18) and using similar arguments as above we compute the second partial derivatives $\frac{d^2}{d\mathbf{v}_i d\mathbf{v}_j}\log \mathcal{L}_2(\mathbf{v})$ which will give us the (i,j)-th entry of the Hessian matrix $\mathcal{H} = \nabla\nabla^{\mathrm{T}} \log \mathcal{L}_2(\mathbf{v})$.

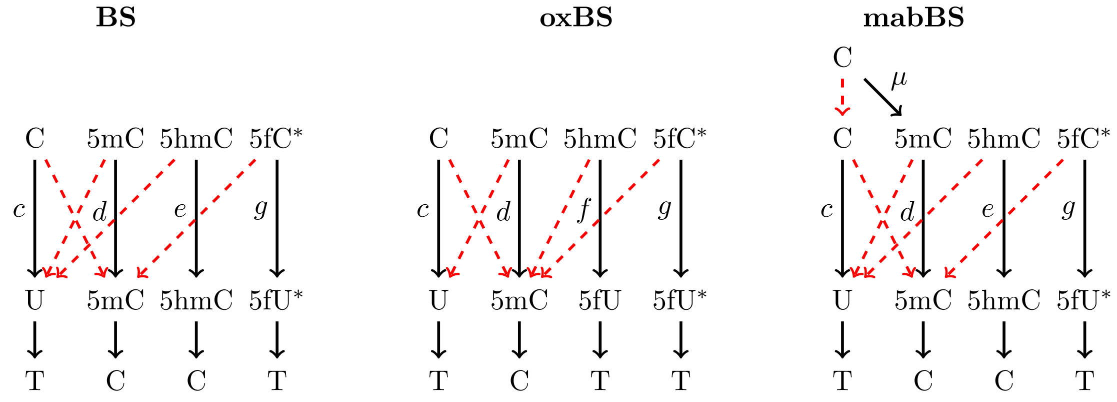


Figure 3.7: Schematic outline of the conversion of Cytosine, 5mC, 5hmC and 5fC* (5fC or 5caC) during BS, oxBS and mabBS sequencing. In both BS and oxBS 5fC* can not be distinguished from C. Hence, in mabBS we apply a preliminary step that transforms C to 5mC and then we apply bisulfite sequencing. As a result in the ideal case of no conversions errors only 5fC* is seen at the end as T. The conversion errors are illustrated as dashed red arrows and the conversion probabilities c, d, e, f, g, μ as black arrows.

Non-linear constraints

First note that in order the parameters of the hybrid model of Figure 3.6 to be identifiable we need to have observation time points within the period of an active demethylation cycle. The constraints for the optimization problem of Eq. (3.15), (3.16) are that the efficiencies at all considered time points are either between 0 and 1, in case they are probabilities (this holds only for maintenance methylation), or between 0 and some positive upper bound ub, in case they are rates. I.e. $0 \leq \mu_m(t) \leq 1$, and $0 \leq \mu_d(t) \leq ub, \ \forall t \in \{0, t_{\max}\}$ for all other efficiencies. For the $5hmC$ non-recognition probability p, it should also hold $0 \leq p \leq 1$.

Note that if we don't provide an upper bound ub, to the optimization algorithm, we can again run into unidentifiability problems since, no matter how many data points we get within one period, the demethylation cycle can happen arbitrarily fast. We set a value for the upper bound $ub = 12$ for each rate by assuming that the total time of a subprocess r can not be less than 1/12 hours, i.e. $\frac{1}{r(t)} \geq \frac{1}{12} \ \forall t \in [0, t_{\max}]$,

	bisulfite sequencing				ox. bisulfite sequencing				mab. bisulfite sequencing			
	TT	TC	CT	CC	TT	TC	CT	CC	TT	TC	CT	CC
uu	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$	c^2	$c \cdot \bar{c}$	$c \cdot \bar{c}$	$\bar{c}^2$	$\bar{j}^2$	$j \cdot \bar{j}$	$j \cdot \bar{j}$	j^2
um	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot d$	$c \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot \bar{d}$	$\bar{c} \cdot d$	$\bar{d} \cdot \bar{j}$	$d \cdot \bar{j}$	$\bar{d} \cdot j$	$d \cdot j$
uh	$c \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot \bar{e}$	$\bar{c} \cdot e$	$c \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot f$	$\bar{c} \cdot \bar{f}$	$\bar{e} \cdot \bar{j}$	$e \cdot \bar{j}$	$\bar{e} \cdot j$	$e \cdot j$
uf	$c \cdot g$	$c \cdot \bar{g}$	$\bar{c} \cdot g$	$\bar{c} \cdot \bar{g}$	$c \cdot g$	$c \cdot \bar{g}$	$\bar{c} \cdot g$	$\bar{c} \cdot \bar{g}$	$g \cdot j$	$\bar{g} \cdot j$	$g \cdot \bar{j}$	$\bar{g} \cdot \bar{j}$
mu	$c \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$	$c \cdot \bar{d}$	$\bar{c} \cdot \bar{d}$	$c \cdot d$	$\bar{c} \cdot d$	$\bar{d} \cdot j$	$\bar{d} \cdot \bar{j}$	$d \cdot j$	$d \cdot \bar{j}$
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	$\bar{d}^2$	$\bar{d} \cdot d$	$d \cdot \bar{d}$	d^2
mh	$\bar{d} \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot \bar{e}$	$d \cdot e$	$\bar{d} \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot f$	$d \cdot \bar{f}$	$\bar{d} \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot \bar{e}$	$d \cdot e$
mf	$\bar{d} \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot g$	$\bar{d} \cdot g$	$\bar{d} \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot g$	$d \cdot \bar{g}$	$\bar{d} \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot g$	$d \cdot \bar{g}$
hu	$c \cdot \bar{e}$	$\bar{c} \cdot \bar{e}$	$c \cdot e$	$\bar{c} \cdot e$	$c \cdot f$	$\bar{c} \cdot f$	$c \cdot \bar{f}$	$\bar{c} \cdot \bar{f}$	$\bar{e} \cdot \bar{j}$	$\bar{e} \cdot j$	$e \cdot \bar{j}$	$e \cdot j$
hm	$\bar{d} \cdot \bar{e}$	$d \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot e$	$\bar{d} \cdot f$	$d \cdot f$	$\bar{d} \cdot \bar{f}$	$d \cdot \bar{f}$	$\bar{d} \cdot \bar{e}$	$d \cdot \bar{e}$	$\bar{d} \cdot e$	$d \cdot e$
hh	$\bar{e}^2$	$\bar{e} \cdot e$	$e \cdot \bar{e}$	e^2	f^2	$f \cdot \bar{f}$	$f \cdot \bar{f}$	$\bar{f}^2$	$\bar{e}^2$	$\bar{e} \cdot e$	$e \cdot \bar{e}$	e^2
hf	$\bar{e} \cdot g$	$\bar{e} \cdot \bar{g}$	$e \cdot \bar{g}$	$e \cdot g$	$f \cdot g$	$f \cdot \bar{g}$	$\bar{f} \cdot \bar{g}$	$\bar{f} \cdot g$	$\bar{e} \cdot g$	$\bar{e} \cdot \bar{g}$	$e \cdot \bar{g}$	$e \cdot g$
fu	$c \cdot g$	$\bar{c} \cdot \bar{g}$	$c \cdot g$	$\bar{c} \cdot \bar{g}$	$c \cdot g$	$\bar{c} \cdot \bar{g}$	$c \cdot g$	$\bar{c} \cdot \bar{g}$	$g \cdot \bar{j}$	$\bar{g} \cdot j$	$g \cdot \bar{j}$	$\bar{g} \cdot j$
fm	$\bar{d} \cdot g$	$d \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot \bar{g}$	$\bar{d} \cdot g$	$d \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot \bar{g}$	$\bar{d} \cdot g$	$d \cdot g$	$\bar{d} \cdot \bar{g}$	$d \cdot \bar{g}$
fh	$\bar{e} \cdot g$	$e \cdot g$	$\bar{e} \cdot \bar{g}$	$e \cdot \bar{g}$	$f \cdot g$	$\bar{f} \cdot g$	$f \cdot \bar{g}$	$\bar{f} \cdot \bar{g}$	$\bar{e} \cdot g$	$e \cdot g$	$\bar{e} \cdot \bar{g}$	$e \cdot \bar{g}$
ff	g^2	$g \cdot \bar{g}$	$g \cdot \bar{g}$	$\bar{g}^2$	g^2	$g \cdot \bar{g}$	$g \cdot \bar{g}$	$\bar{g}^2$	g^2	$g \cdot \bar{g}$	$g \cdot \bar{g}$	$\bar{g}^2$

Table 3.2: Transition probabilities from hidden to the observable states in BS, oxBS and mabBS. For mabBS we write the probability of a desired conversion from C→C as $j = \mathrm{P}(\mathrm{C}\rightarrow\mathrm{C}) = \mu d + (1-\mu)(1-c)$ and the complementary probability $\bar{j} = \mathrm{P}(\mathrm{C}\rightarrow\mathrm{T}) = (1-\mu)c + \mu(1-d)$.

based on the fact that the average turnover time $\mathbb{E}[T_{\text{turnover}}] = \frac{1}{\mu_d} + \frac{1}{\eta} + \frac{1}{\phi} + \frac{1}{\alpha}$ in certain promoters of human cells is between 75 and 120 minutes [43, 66].

The above condition gets simplified to the following two non-linear constraints:

$$0 \leq r(t) \leq ub \quad \Longleftrightarrow \quad 0 \leq r(t_{\arg\min}) \text{ and } r(t_{\arg\max}) \leq ub \tag{3.19}$$

for $t_{\arg\min}, t_{\arg\max} \in [0, t_{\max}]$.

Hence, we only have to compute the critical points of the spline function by considering the points where the first time derivative of $r(t)$ becomes zero, i.e., $\frac{d}{dt}r(t) = \beta_1^r + \beta_2^r t + 2\beta_3^r h(t-\xi_1)^{1/2} + \ldots + 2\beta_{K+3}^r h(t-\xi_K)^{1/2} = 0$, and deciding whether each critical point is a maximum or a minimum by evaluating the second derivative $\frac{d^2}{dt^2}r(t)$ at this point.

Defining the constraint function $g_r(\mathbf{v}_r) : \mathbb{R}^{|\mathbf{v}_r|} \to \mathbb{R}^2$ of each rate r as $g_r(\mathbf{v}_r) = [g_{1,r}(\mathbf{v}_r),\ g_{2,r}(\mathbf{v}_r)]$, where $g_{1,r}(\mathbf{v}_r) = ub - r(t_{\arg\max})$, and $g_{2,r}(\mathbf{v}_r) = r(t_{\arg\min})$, we get $V_s = \bigcup V_r$ as the feasible parameter space of the optimization problem, where $V_r = \{\mathbf{v}_r : g_{1,r}(\mathbf{v}_r), g_{2,r}(\mathbf{v}_r) \geq 0\}$ is the feasible space for efficiency r. Computing the Jacobian matrix $J_{g_r}(\mathbf{v}_r)$ of the constraint function g_r we can steer the optimization algorithm towards parameters' feasible regions.

Standard deviations and confidence intervals

Similarly to Section 3.2.2 we approximate the covariance matrix of $\mathbf{v}^*$ as the inverse of the Hessian matrix $\mathcal{H}(\mathbf{v}^*)$ at the point of the maximum likelihood estimate and we get the standard deviation estimates as $\sigma(\mathbf{v}^*) = \sqrt{\mathrm{diag}(-\mathcal{H}^{-1}(\mathbf{v}^*))}$. In order to get the standard deviations of the efficiency functions over time, i.e., $\sigma(\mu_m(t)), \sigma(\mu_d(t))$ and $\sigma(\eta(t))$, we exploit the fact that if

$$r(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 h(t - \xi_1) + \ldots + \beta_{K+2} h(t - \xi_K),$$

where ξ_i, is the i-th knot, $i = \{1, \ldots, K\}$ then

$$\begin{aligned}
& \mathrm{Var(r(t))} \\
= \ & \mathrm{Var}(\beta_0 + \beta_1 \mathrm{t} + \beta_2 \mathrm{t}^2 + \beta_3 \mathrm{h}(\mathrm{t} - \xi_1) + \ldots + \beta_{\mathrm{K}+2} \mathrm{h}(\mathrm{t} - \xi_{\mathrm{K}})) \\
= \ & \mathrm{Var}(\beta_0) + \mathrm{t}^2 \mathrm{Var}(\beta_1) + \mathrm{t}^4 \mathrm{Var}(\beta_2) + \mathrm{h}(\mathrm{t} - \xi_1)^2 \mathrm{Var}(\beta_3) + \ldots + \mathrm{h}(\mathrm{t} - \xi_{\mathrm{K}})^2 \mathrm{Var}(\beta_{\mathrm{K}+2}) \\
& +2t \operatorname{Cov}(\beta_0, \beta_1) + 2t^2 \operatorname{Cov}(\beta_0, \beta_2) + 2h(t - \xi_1) \operatorname{Cov}(\beta_0, \beta_3) + \ldots + 2h(t - \xi_K) \operatorname{Cov}(\beta_0, \beta_{K+2}) \\
& +2t^3 \operatorname{Cov}(\beta_1, \beta_2) + 2h(t - \xi_1) t \operatorname{Cov}(\beta_1, \beta_3) + \ldots + 2h(t - \xi_K) t \operatorname{Cov}(\beta_1, \beta_K) \\
& +2 \operatorname{Cov}(\beta_2, \beta_3) h(t - \xi_1) + \ldots + 2 \operatorname{Cov}(\beta_2, \beta_K) h(t - \xi_2) \\
& + \ldots, \qquad (3.20)
\end{aligned}$$

where ξ_i is the i-th knot, $i = \{1 \dots K\}$. Eq. (3.20) can be written in a much simpler form if we define the vector $t_p = [0, t, t^2, h(t-\xi_1), \dots, h(t-\xi_K)]$. Then we can write $(t_p^\top \cdot t_p) \otimes \mathrm{Cov}_{\mathbf{v}_r}$, where $\mathrm{Cov}_{\mathbf{v}_r}$ is the submatrix of Cov that corresponds to $\mathbf{v}_r$ parameters.

Obtaining the standard deviations of all the efficiencies over time one can create the corresponding confidence intervals for a fixed confidence level, i.e., $\beta = 95\%$, as

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

where $z = F^{-1}\left(\frac{\beta+1}{2}\right)$ and F is the cumulative distribution function (cdf) of the standard normal distribution.

Model selection - choosing the knots' number and location

The only thing that remains regarding the use of the cubic spline functions is to decide the maximum number of knots which is necessary for our model so that we avoid overfitting, and also define the points that the knots should be placed. We do it by applying one of the following strategies.

- Decide a priori the number and / or the location of the knots by observing the “specifications” of the input data patterns for a certain locus.
- Apply a LOOCV similarly to Section 3.2.3 training the data in all but one CpGs of a region every time. Choose the model that provides the smallest test error based on KL divergence.
- Compute for every model either Akaike Information Criterion AIC $= 2d - 2\log \mathcal{L}_2(\mathbf{v}^*)$ or Bayesian Information Criterion BIC $= \log(n_{tot})d - 2\log \mathcal{L}_2(\mathbf{v}^*)$, where n_{tot} is the total number of observations over all time points and experiments, and d is the number of estimable parameters (degrees of freedom)[2]. Both

[2] In case of the cubic spline with K knots we get $4(K+1) - 3K = K + 4$ actual degrees of freedom due to the continuity up to the second derivative in each knot.

of these terms provide a compromise between a very good fit of the data and model's simplicity by correcting for the number of unknown parameters. We choose as 'optimal' the model that minimizes either of the above criteria.

Chapter 4

Individual Genomic Loci Analysis

4.1 Results

We run the core-model presented in Section 3.1 and the inference and validation methods presented in Section 3.2 on an ultra-deep generated DNA methylation data set of selected loci in mouse ES cells (mESCs) collected at defined time points after cultivation in 2i. For our analysis we choose four multi-copy, repetitive elements, IAPs (intracisternal A particle), L1mdA and L1mdT (both Long interspersed nuclear elements) and mSat (major satellite), as well as four single-copy loci in the genes Afp, Snrpn, Ttc25 and Zim3. Most of the above repetitive were already known to be 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 [18], whereas 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 Section 2.2.1). 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). Then, the hairpin refolding of sequences of Section 2.2.1 in combination with the HMMs described in Section 3.1 let us determine the accurate double strand CpG methylation levels in a given locus (hemi-, fully- or unmethylated).

4.1.1 Estimation of the enzymatic activities

We used our model on the above data 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. Single CpG modeling largely returned similar results with the average data of the same locus (with slightly increased confidence intervals) and therefore in our further analysis and model's interpretation we refer to the average data. The results for the single CpG modeling of all examined regions can be found in Section A.3.

In Table 4.1 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 Table A.4, A.5 for each of the eight genome loci. Table 4.2 shows the computed coefficients of the total methylation $\lambda(t)$, which can be implicitly taken from the maintenance and *de novo* estimated coefficients (see Section 3.1.1, Maintenance).

In Figure 4.1 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 Table 4.3,

Table 4.1: 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	0.2577	-	-	-

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	-	-	-

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	-	-	-

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	-	-	-

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	0.0208	-	-	-

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	0.0670	-	-	-

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	0.0255	-	-	-

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	-	-	-

Table 4.2: 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	$1.21 \cdot 10^{-4}$	$< 10^{-5}$
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

4.2. From the above efficiencies one 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., $\lambda(t) - \mu_m(t) = \bar{\mu}_m(t) \cdot \mu_d(t)$. Our results indicate that persistence of DNA methylation in IAP, Afp and mSat elements clearly depends as well on *de novo* enzymes acting on hemimethylated CpGs.

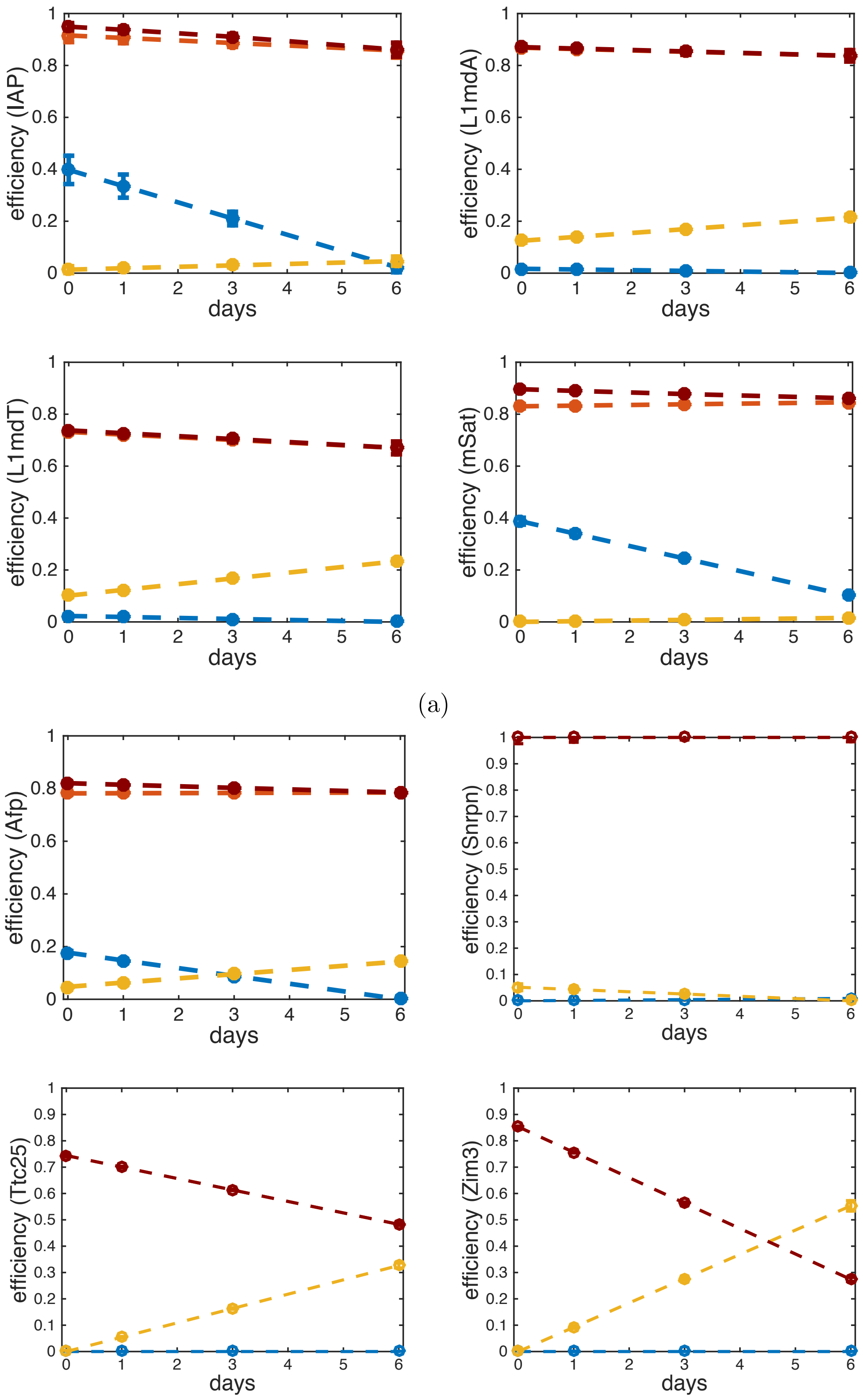
Wald Test

For each efficiency, we performed a statistical test with a confidence level of 1% for the null hypothesis that the slope of the corresponding function is zero, i.e., that the efficiencies are constant functions of time. Hence, 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 $H_0 : \beta_1^\lambda = 0 \wedge \beta_2^\lambda = 0$, since λ is a quadratic function of time.

From the performed Wald test we found a statistically significant decrease for the *de novo*, and the total methylation efficiencies in all eight loci (besides *de novo* at Ttc25, Zim3 and Snrpn where it is absent and total methylation in L1mdA). Similarly, the increase of hydroxylation for five out of eight loci is statistically significant. However, for the maintenance function we have to accept the null hypothesis in most of the loci (namely all repetitive elements and Afp), that is, we cannot exclude the possibility that for these loci maintenance is constant over time.

Sensitivity Analysis

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 loci, time points and parameters) absolute change of 0.0053 for the total hydroxylation level and 0.0198 for the total methylation level. This ensures that the model is sufficiently robust.



(a)

(b)

Figure 4.1: The diagrams show the predicted by the model enzymatic efficiencies and their standard deviations for maintenance (red), *de novo* (blue), hydroxylation (yellow) and total efficiency on a hemimethylated CpG (dark red) for all (a) multi- and (b) single-copy studied loci.

4.1.2 Goodness of fit and validation

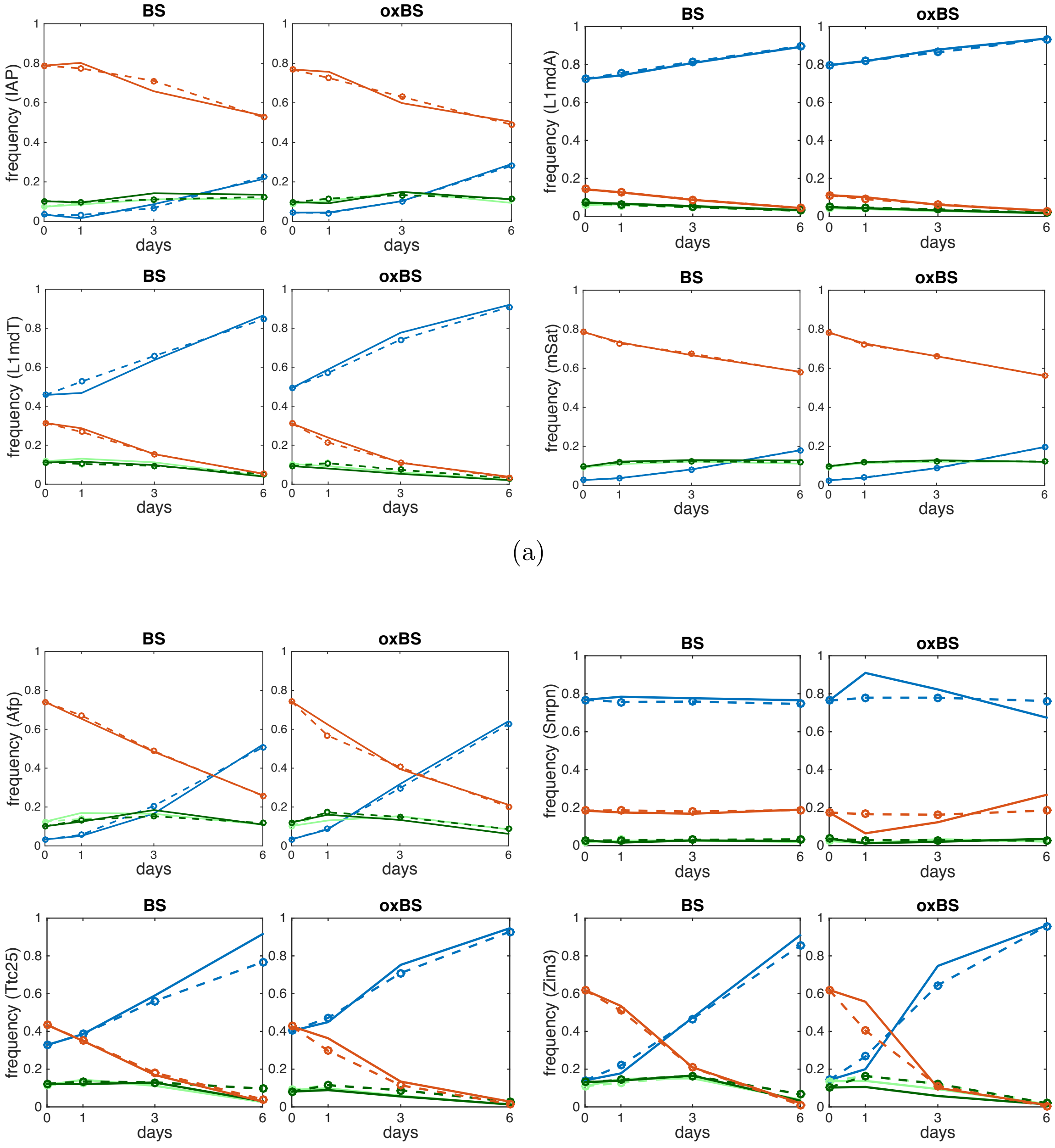
Using the estimated values of the model's unknown parameters for each region we could predict the probabilities of the observable states and compare them to the frequencies of the measured data (Table A.4, A.5) at various time points. As we observe in Figure 4.2 the model accurately describes the dynamics for all loci except for some underestimations of two states CC and TT for oxBs in Ttc25 and Zim3, respectively.

Leave One Out Cross Validation

To validate the linear assumption for the efficiencies in our model we performed LOOCV (see Section 3.2.3 for details). The results in Table 4.3 show that the linear assumption for the enzymes' efficiencies improves the prediction up to 38.3%, compared with constant efficiencies. For all loci the test error becomes evidently smaller for the case where we allow efficiencies to be linear over time using either Kullback Leibler (KL) divergence or Bhattacharyya (BC) distance as a distance measure. The improvement ("gain") $\frac{KL_{\text{const}}-KL_{\text{linear}}}{KL_{\text{const}}}$ 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, and from 0.3% to 38.1% for the BC distance. Note that the predictive potential gain of the model described by the above ratio depends on the available number of CpGs for the training data and also on how much the efficiencies deviate from constant behavior over time. Clearly, though, it is in general very well supported by the data that the linear assumption for the efficiencies evidently increases the model's predictive ability.

4.1.3 Prediction of hidden states' levels

Figure 4.3 shows the probabilities of the hidden states in all single and multi-copy gene loci, where the parameters are chosen according to the results of the maximum



(a)

(b)

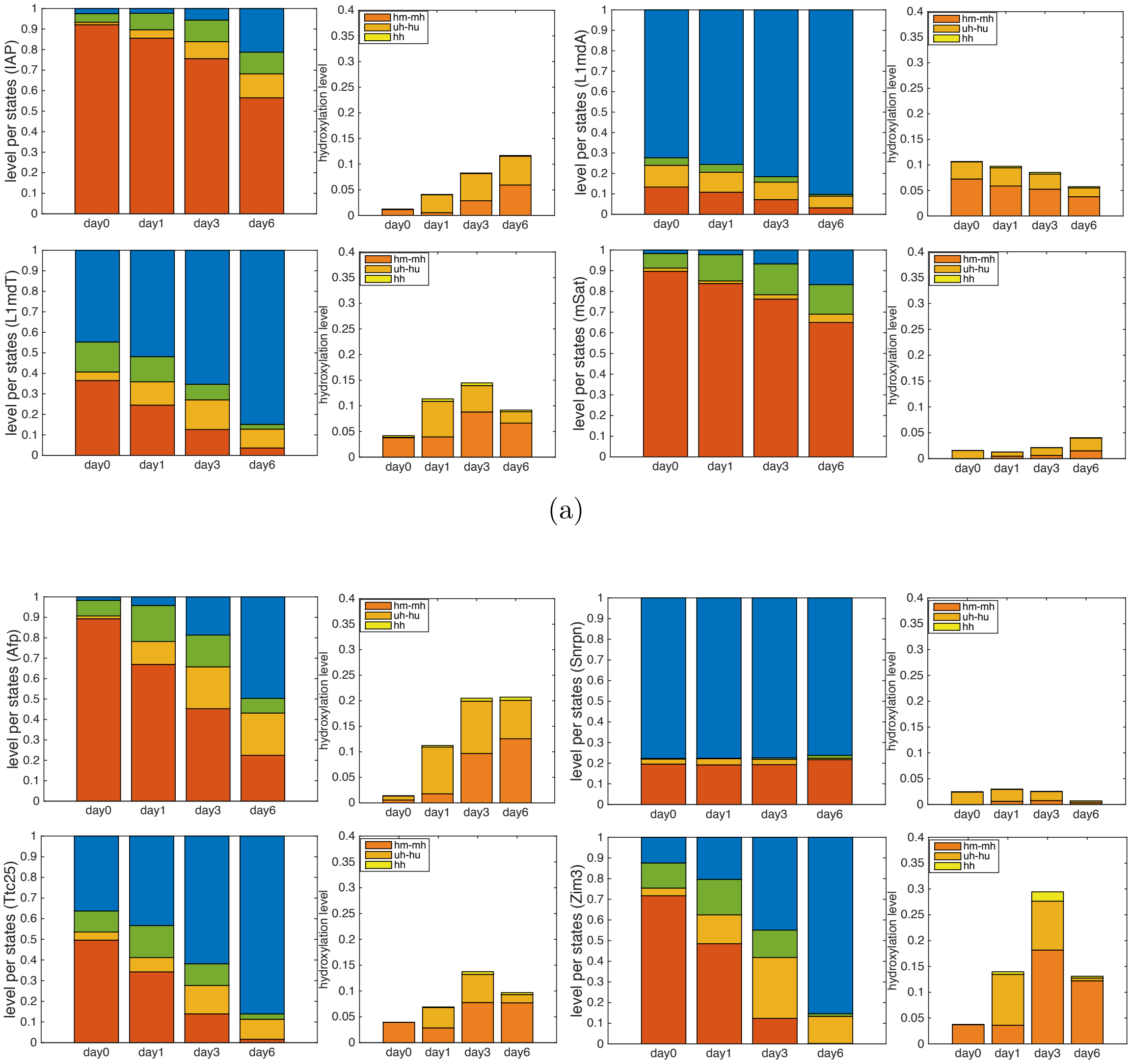
Figure 4.2: Comparison of predicted modification levels and the obtained sequencing data for BS and oxBS for the loci L1mdT, mSat, Afp, Zim3, IAP, L1mdA, Ttc25 and Snrpn. Probabilities of the observable states TT (blue), TC (light green), CT (dark green), CC (red). The solid lines show the experimentally measured frequencies states and the dashed lines correspond to the values predicted by the two HMMs.

Table 4.3: 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 %
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 %

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 [18, 29]. However, precise levels had not been determined per locus. 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. 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, and IAP (see Figure 4.3), 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 Figure 4.3) relative to the total modification level (5hmC + 5mC) (red, orange and green bars), becomes maximal on



(a)

(b)

Figure 4.3: Probabilities of the hidden states for (a) multi-copy loci IAP, L1mdA, L1mdT, mSat, and (b) single-copy loci Afp, Snrpn, Ttc25 and Zim3: The left diagram depicts the amount of fully methylated (*mm*) sites in red color, hemimethylated (*um* and *mu*) sites in green color, and unmethylated (*uu*) sites in blue color. 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.

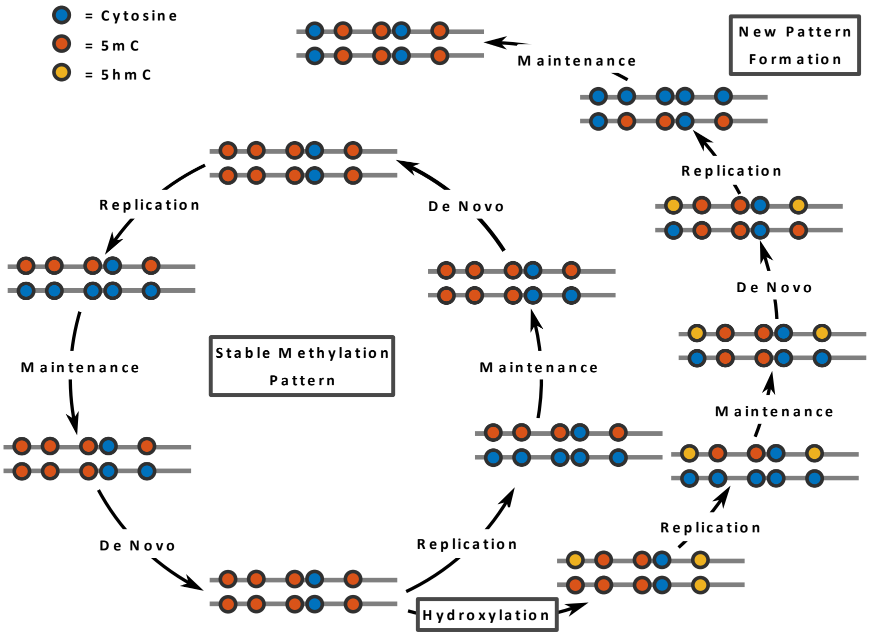


Figure 4.4: Maintenance and *de novo* methylation are usually cooperating to maintain a stable methylation pattern (inner circle). The oxidation of 5mC to 5hmC may interfere with the maintenance machinery causing a (partial) loss of CpG methylation after DNA replication. DNA strands are indicated by lines whereas the CpG are shown as colored circles.

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.

This is, indeed, further stressed by the fact that 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 (Table 4.1). We estimated smaller values for p only for those loci where hydroxylation is nearly absent (mSat, Snrpn).

4.2 Discussion

The goal of this chapter was the application of our, previously described, model at both single- multi-copy loci across the genome, in order to investigate the role of

5hmC in the process of progressive DNA demethylation in these regions. As a model system we used the DNA of mESCs grown under conditions where the cells experience a genome wide reduction of DNA methylation [29, 18].

Using time dependent comparative bisulfite and oxidative bisulfite hairpin sequencing data we applied two HMMs: one that captures the dynamics of total modifications 5mC and 5hmC in BS, and another only representing the 5mC levels in oxBS. The combination allowed us to accurately determine the amount and changes of 5hmC for each single CpG dyad 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.

4.2.1 Biological findings

A careful analysis of the model's output reveals the following key biological findings.

Time-variant hydroxylation levels

First, we observe that 5hmC levels and distribution change 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 and Tets' dynamics, which is congruent with the finding that several Tet enzymes are up-regulated in 2i medium [18, 29]. 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 from locus to locus. However, the dynamics of the (hydroxy-)methylation levels, as well as the efficiency profiles, show a certain homogeneity for the CpGs of the same locus (see Figure A.1).

Mainly constant maintenance, decreasing *de novo*

Maintenance methylation shows an impaired behavior in 2i (it is on average 0.82 in non-control regions) in comparison with its Serum perfect function. Within 2i, though, it is predicted from our model to remain stable within in 6 of 8 examined loci and this nicely agrees with the previously shown unchanged mRNA expression of Dnmt1 and Uhrf1 in 2i [18, 29]. Interestingly, for the single copy genes Ttc25 and Zim3 we predict a clear decrease of maintenance function (see Figure 4.1, red line), which is independent of the high 5hmC levels at these loci, as the influence of 5hmC on the maintenance mechanism is captured by the non-recognition probability p. This might indicate an additional impairment or absence of the maintenance machinery at these loci, which would hint towards the existence of genomic loci with specific enzymatic profiles. However, since we cannot capture further oxidized cytosine forms with the current experimental/ model design, we can also not exclude the possibility that with the strong decrease in maintenance efficiency our model, at least to some extent, compensates for active demethylation.

Being able to estimate the *de novo* methylation impact of Dnmt3a/b on hemimethylated sites, a 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 *de novo* enzymes' activities follows their relative expression in 2i medium, in which both Dnmt3a and Dnmt3b are clearly down regulated [18, 29]. 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.

Passive demethylation mechanism

The most important finding of our modeling is that it 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. As it is illustrated in Figure 4.4 the model suggests that the accumulation of 5hmC causes a passive dilution mechanism of CpG methylation with each DNA replication/cell cycle, despite the fact that the maintenance activity is predicted to remain stable in most of the analyzed loci. In mESCs maintained in 2i medium this passive demethylation mechanism appears to be, together with the initial maintenance impairment, the main driving force for a rapid and linear DNA demethylation. As a result, loci with an enrichment of 5hmC are more likely to lose DNA methylation over time. Indeed, data supports that 5hmC containing DNA strands such as IAP, L1mdT, Afp and TTc25 show higher demethylation rates compared to low 5hmC regions as mSat or Snrpn.

Summary

In this chapter we presented the results of a novel HMM method that allows to accurately measure and describe effects related to the influence of 5hmC on the persistence of DNA methylation in the mammalian genome. Taking advantage of BS and oxBS hairpin data over different time points our modeling is the first one that allows to accurate double strand CpG (hydroxy-)methylation levels in a given locus, enables us to accurately infer enzymatic activities and thus allows us to decipher complex DNA methylation patterns.

We reveal a strong passive loss of 5mC that happens due to its transformation to 5hmC and the non-recognition of it by the maintenance machinery. This passive demethylation mechanism is enhanced by a non-perfect maintenance, an increasing Tet enzymes activity and in some regions by a decreasing *de novo* efficiency as well.

In its current form the model already captures a fraction of possible demethylation dynamics and scenarios most likely reflecting many loci in the genome. However, our approach could also be used to accurately model 5hmC dependent methylation dynamics in diseases, e.g., certain cancers and in aging processes of long lived cells. Finally, as it has been shown in detail in Section 3.3, the model can be enhanced to additionally capture active demethylation by integrating additional sequencing methods that are able of detecting further oxidized modifications such as 5fC and 5caC. In this case the mechanisms behind DNA demethylation will be described in an even greater resolution.

Chapter 5

H(O)TA Tool

5.1 H(O)TA Description

Hairpin (Oxidative) bisulfite sequencing Time course Analyzer (H(O)TA) is a tool that accurately infers (hydroxy-)methylation levels and efficiencies of the involved enzymes at a certain DNA locus. The tool gets as input time course measurements from hairpin BS-seq and oxBS-seq and it is based on the construction of two coupled hidden Markov models (HMMs) which take into account all relevant conversion errors. The underlying stochastic model and the core estimation procedure of the unknown parameters are being described in Chapter 3, Section 3.1, 3.2, respectively.

The interested user can download the tool and run the input data of Section A.2 or upload his own BS-seq oxBS-seq epigenomic data. In this chapter we describe the functions that one can operate using H(O)TA and provide a short and easy installation guide.

H(O)TA has been mainly developed in MATLAB with some routines written in C++. Its execution requires the installation of the free MATLAB runtime environment (MRE). The tool and the MRE can be downloaded as a single installation file available for Linux, MacOS, and Windows operating systems.

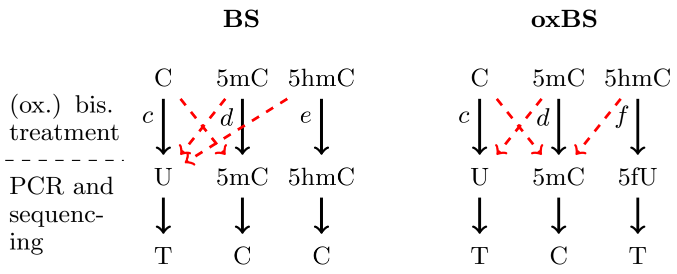


Figure 5.1: Conversion scheme for BS-seq and oxBS-seq. The first (last) line determines the hidden (observable) states. Conversion errors (illustrated by the red dashed arrows) are taken into account by adjusting the emission probabilities of the HMMs by the conversion probabilities c, d, e, and f.

Its graphical user interface consists of two windows: a **dialogue window** for loading the input files of a DNA locus and running the analysis and the **main window** (Figure 5.2) for visualizing the output. The tool can automatically aggregate data of different CpGs of a locus and compute average (hydroxy-) methylation levels as well as average efficiencies. In addition, the same analysis can be performed for each CpG individually. The model can be applied to both, clean cell populations and cell mixtures. However, when dealing with convolutions of cells individual methylation patterns might be hidden and the results will only reflect the average behavior of all cell types. Users can provide three input .txt files. The first file contains BS-seq time course data, the second one oxBS-seq time course data and the third file should contain the conversion errors of the two experiments (see dashed arrows in Figure 5.1), as well as a string that describes how many cell divisions take place between two observation time points. Conversion errors can be obtained either by including unmodified cytosine, 5mC and 5hmC into the hairpin linker or by the inclusion of a spike-in sequence into the sample containing the different cytosine variants (see A.1.1). If only BS-seq data is given, then the tool will predict only the methylation levels and efficiencies (merged with the corresponding unknown hydroxylation values) of the given region. A detailed documentation of the input files is given in Section 5.2.2.

The main window consists of two panels. The left panel of the main window is an overview in small resolution of the detailed output that is shown on the right. The (hydroxy-)methylation levels and the efficiencies of all individual CpGs are plotted such that they can be compared with each other and with the corresponding plots of the right panel. Based on the selection made by the user in the upper left corner, the right panel shows the output of the analysis either for the aggregated data or for each of the previously chosen CpG sites. The observable states reflect the possible outcomes of hairpin BS-seq and hairpin oxBS-seq, respectively, that is, $\{\mathrm{T}, \mathrm{C}\}^2$ (cf. last line in Figure 5.1), where T stands for thymine and C for cytosine. The upper left and middle plots of the right panel (Figure 5.2) show the fit between the data (dense line) and the model prediction (dashed line) for the observable states TT, TC, CT, CC in each of the two experiments. As opposed to methods for single time point data, H(O)TA performs an analysis that considers the transient probability distribution over the set $\{u, m, h\}^2$ of nine hidden states of the two cytosines of a CpG dyad, where u, m and h describe C, 5mC, and 5hmC, respectively. Thus, besides the states uu and mm, which correspond to the blue and red bars in the bar plots of the hidden states' probabilities in Figure 5.2, lower left plot, the model's output also includes the time evolution of the levels of hemimethylated sites (states um, mu, green bars) as well as those of hydroxylated sites (states uh, hu, hm, mh, hh, orange bars). The lower right plot of the main window shows the detailed distribution of the different hydroxylation states. For each observation time point, estimations of the enzymes' efficiencies, i.e., the probabilities of a methylation or a hydroxylation event between two cell divisions, are made in the upper right plot for the maintenance methylation (red), de novo methylation (blue) and hydroxylation (orange) as well as the total methylation (dark red) on hemimethylated CpGs. In addition, an estimation is provided for the probability that no maintenance is performed when the current state is mh or hm, which hints on the existence of a passive demethylation mechanism

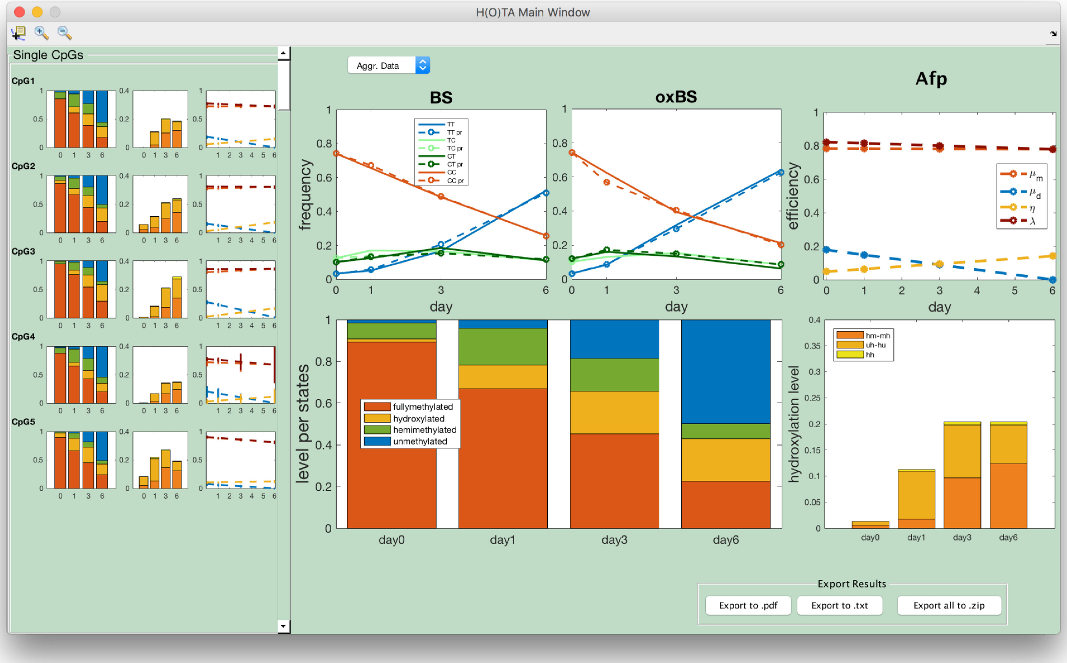


Figure 5.2: The main window of the graphical user interface of H(O)TA.

induced by hydroxylation. H(O)TA provides the user with several options (lower right corner) for exporting the estimation results in a desirable format. For all the estimated parameters confidence intervals are computed and a statistical test is carried out in order to verify certain hypotheses about the efficiencies. For a complete description of the underlying model and details about the optimization as well as the statistical validation of the results, the reader can refer to Chapter 3.

5.2 Run H(O)TA

5.2.1 Installation

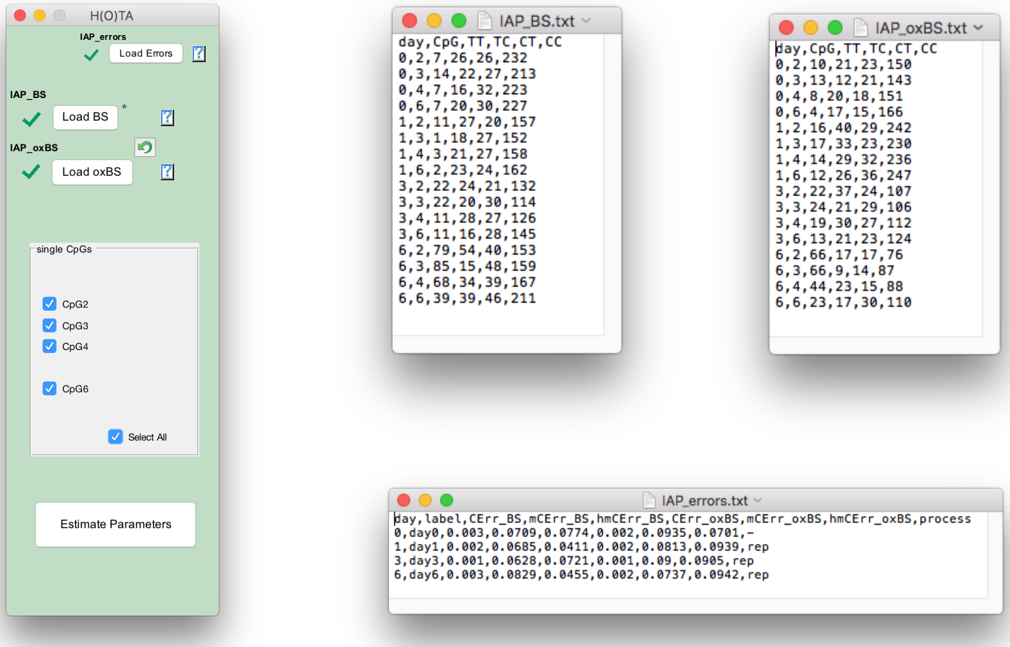
Downloading the following installation files provides the user with detailed instructions on how to install H(O)TA and Matlab Runtime which is necessary for the executable to run.

- **H(O)TA, Linux 64bit**
- **H(O)TA, MacOS 64bit**
- **H(O)TA, Windows 64bit**

> **Note**: For Linux machines unzip the downloaded file and then execute
>
> ```
> $sudo ./HOTA_Installer_web.install
> ```
>
> in the extracted directory. Once H(O)TA and Runtime are both installed go to H(O)TA's installation directory and run HOTA script by giving as an argument the path of the installation directory of the Runtime. E.g. if the version of the runtime is v901 and the installation path is `/usr/local/MATLAB` then type
>
> ```
> $./run_HOTA.sh/usr/local/MATLAB/MATLAB_Runtime/v901,
> ```
>
> while being in the directory of `run_HOTA.sh`.

5.2.2 Input

The names of the files containing the bisulfite and the oxidative bisulfite data should strictly be of the form `region_BS.txt` and `region_oxBS.txt`, respectively, where region is the name of the examined locus. The file containing the data of the conversion errors should have the ending `_errors.txt` but it does not need to begin with the name of the specific locus in case. All the entries in the data and error files should be comma-separated without empty spaces. Clicking on the question marks of the GUI next to each loading button the user can see sample input files. For the tool to run only the file with the BS-seq data is mandatory and the other two are optional. In case oxBS data is not provided the tool only predicts methylation levels and the corresponding efficiencies ignoring hydroxylation. Note that in this case the estimated efficiencies might wrongly compensate for the lack of hydroxylation information. E.g. the maintenance might appear to decrease, even though this does not



☑ CpG2
☑ CpG3
☑ CpG4
☑ CpG6
☑ Select All

```
day,CpG,TT,TC,CT,CC
0,2,7,26,26,232
0,3,14,22,27,213
0,4,7,16,32,223
0,6,7,20,30,227
1,2,11,27,20,157
1,3,1,18,27,152
1,4,3,21,27,158
1,6,2,23,24,162
3,2,22,24,21,132
3,3,22,20,30,114
3,4,11,28,27,126
3,6,11,16,28,145
6,2,79,54,40,153
6,3,85,15,48,159
6,4,68,34,39,167
6,6,39,39,46,211
```

```
day,CpG,TT,TC,CT,CC
0,2,10,21,23,150
0,3,13,12,21,143
0,4,8,20,18,151
0,6,4,17,15,166
1,2,16,40,29,242
1,3,17,33,23,230
1,4,14,29,32,236
1,6,12,26,36,247
3,2,22,37,24,107
3,3,24,21,29,106
3,4,19,30,27,112
3,6,13,21,23,124
6,2,66,17,17,76
6,3,66,9,14,87
6,4,44,23,15,88
6,6,23,17,30,110
```

```
day,label,CErr_BS,mCErr_BS,hmCErr_BS,CErr_oxBS,mCErr_oxBS,hmCErr_oxBS,process
0,day0,0.003,0.0709,0.0774,0.002,0.0935,0.0701,-
1,day1,0.002,0.0685,0.0411,0.002,0.0813,0.0939,rep
3,day3,0.001,0.0628,0.0721,0.001,0.09,0.0905,rep
6,day6,0.003,0.0829,0.0455,0.002,0.0737,0.0942,rep
```

(a) (b)

Figure 5.3: (a): The input panel of H(O)TA. (b): Example input files of BS and oxBS data and of the conversion errors' file.

happen in the real system. In case the error file is not provided, default error values based on averages of historical values are used.

Input data files

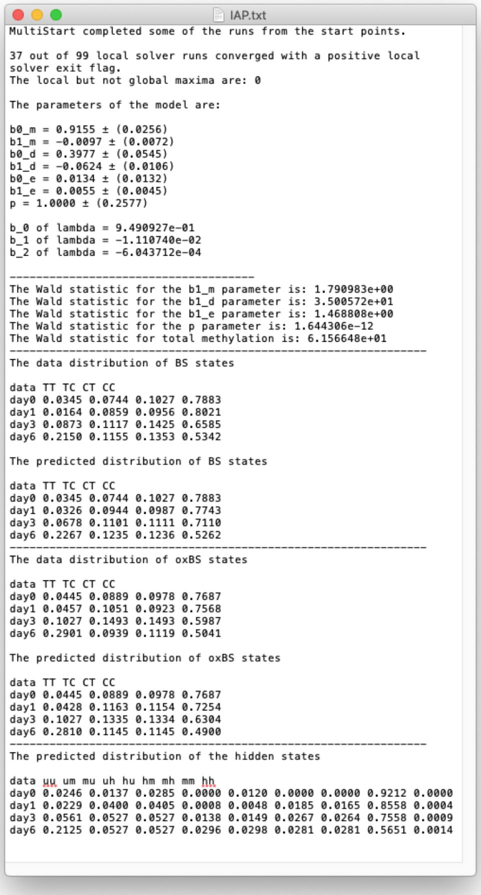
The format of the BS and oxBS data .txt files (Figure 5.3b, upper row) is extremely simple. Every row of the file corresponds to the measurements that have been taken for a particular time point and a particular CpG (first and second column, respectively). After the first row with six column headers "day, CpG, TT, TC, CT, CC", the data is listed. The first column is the day of the measurement and the number of the CpG follows. The next four columns contain the absolute number of times the states TT, TC, CT and CC have been measured.

Conversion errors file

Each row of the error file (Figure 5.3b, lower row) corresponds to a time point (first column) of measurements. The second column is the label that appears for the corresponding time point in the output plots. Columns 3-5 contain the conversion errors of the bisulfite setup and columns 6-8 the conversion errors of the oxidative bisulfite setup. Each red dashed arrow of Figure 5.1 corresponds to one error listed in the error file from left to right. The last column of the error file is the characterization of the process that happens between two observation time points. This entry can either be rep or no-rep. In the first case we assume that a number of cell replications equal to the difference between the two time points has happened, while in the second case no cell replication has happened.

5.2.3 Output

Once the input files have been loaded a panel with a check-box for each CpG of the locus appears. The user has to check the CpGs for which he would like to have estimations of the (hydroxy-)methylation levels and the activity of the enzymes. Regardless of the choice for the single CpGs the prediction of the behavior of the whole imported region (aggregated data of all CpGs) will always run and the aggregated results will appear in the right part of the main output window. In the left part of the window the prediction for each chosen CpGs is shown. The user can zoom in on the results of an individual CpG choosing it from the pop-up window in the upper left part of the screen. The output for a single CpG or the aggregated data can be extracted in both .pdf (export to .pdf button) and .txt (export to .txt button). In Figure 5.4 we see the .txt output file containing information about the optimization run, the estimated parameter values along with test statistics for the null hypothesis of a constant efficiency, as well as the predicted distribution for the hidden and the observable states. At last, the software gives also the user the option to export all



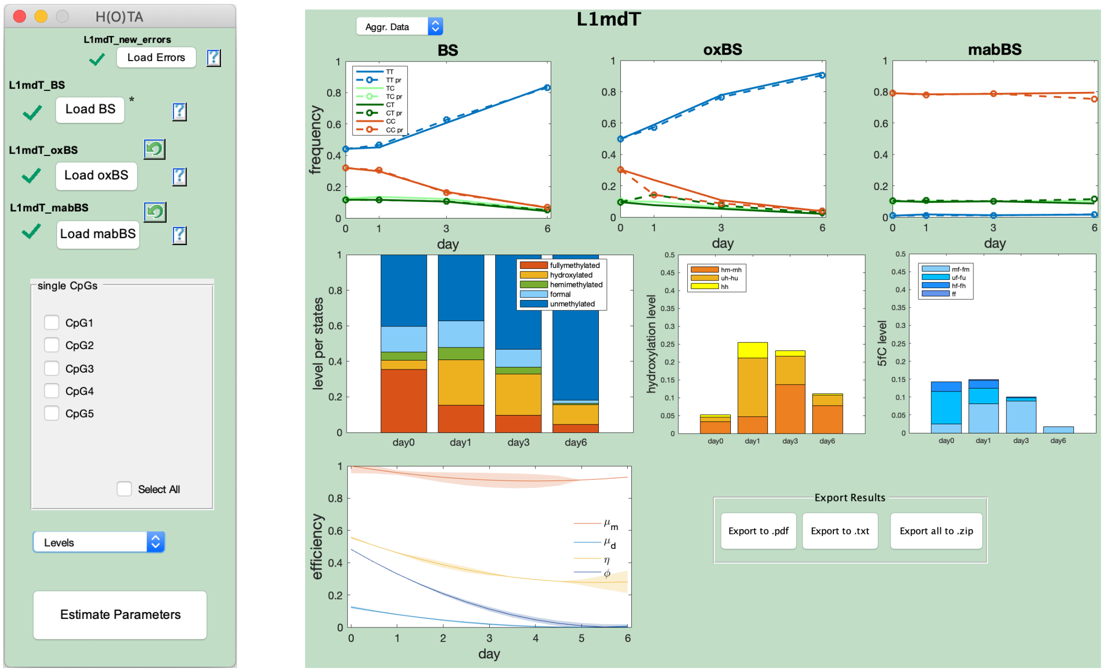
```
MultiStart completed some of the runs from the start points.

37 out of 99 local solver runs converged with a positive local
solver exit flag.
The local but not global maxima are: 0

The parameters of the model are:

b0_m = 0.9155 ± (0.0256)
b1_m = -0.0097 ± (0.0072)
b0_d = 0.3977 ± (0.0545)
b1_d = -0.0624 ± (0.0106)
b0_e = 0.0134 ± (0.0132)
b1_e = 0.0055 ± (0.0045)
p = 1.0000 ± (0.2577)

b_0 of lambda = 9.490927e-01
b_1 of lambda = -1.110740e-02
b_2 of lambda = -6.043712e-04

-------------------------------------
The Wald statistic for the b1_m parameter is: 1.790983e+00
The Wald statistic for the b1_d parameter is: 3.500572e+01
The Wald statistic for the b1_e parameter is: 1.468808e+00
The Wald statistic for the p parameter is: 1.644306e-12
The Wald statistic for total methylation is: 6.156648e+01
---------------------------------------------------------------
The data distribution of BS states

data TT TC CT CC
day0 0.0345 0.0744 0.1027 0.7883
day1 0.0164 0.0859 0.0956 0.8021
day3 0.0873 0.1117 0.1425 0.6585
day6 0.2150 0.1155 0.1353 0.5342

The predicted distribution of BS states

data TT TC CT CC
day0 0.0345 0.0744 0.1027 0.7883
day1 0.0326 0.0944 0.0987 0.7743
day3 0.0678 0.1101 0.1111 0.7110
day6 0.2267 0.1235 0.1236 0.5262
---------------------------------------------------------------
The data distribution of oxBS states

data TT TC CT CC
day0 0.0445 0.0889 0.0978 0.7687
day1 0.0457 0.1051 0.0923 0.7568
day3 0.1027 0.1493 0.1493 0.5987
day6 0.2901 0.0939 0.1119 0.5041

The predicted distribution of oxBS states

data TT TC CT CC
day0 0.0445 0.0889 0.0978 0.7687
day1 0.0428 0.1163 0.1154 0.7254
day3 0.1027 0.1335 0.1334 0.6304
day6 0.2810 0.1145 0.1145 0.4900
---------------------------------------------------------------
The predicted distribution of the hidden states

data uu um mu uh hu hm mh mm hh
day0 0.0246 0.0137 0.0285 0.0000 0.0120 0.0000 0.0000 0.9212 0.0000
day1 0.0229 0.0400 0.0405 0.0008 0.0048 0.0185 0.0165 0.8558 0.0004
day3 0.0561 0.0527 0.0527 0.0138 0.0149 0.0267 0.0264 0.7558 0.0009
day6 0.2125 0.0527 0.0527 0.0296 0.0298 0.0281 0.0281 0.5651 0.0014
```

Figure 5.4: H(O)TA's output file in .txt format

the results, i.e., all CpGs plus the aggregated data, of a locus in a .zip file (export all to .zip file).

5.3 H(O)TA v.2 - Hybrid model, further extensions

An upgraded version of the program, H(O)TA v.2 (beta), is built based on the hybrid extension of our model being described in Section 3.3 and incorporates the following extended features.



(a) (b)

Figure 5.5: (a): The input panel of H(O)TA v.2. (b): The main window of H(O)TA v.2 with two additional outputs. A plot for the fit of the mabBS data (first row, right) and a plot for the detailed prediction of formyl-cytosine modification levels (second row, right)

1. In addition to BS and oxBS data, it gets data from mabBS experiment (see Section 3.3.1 for details). The main window of H(O)TA v.2 is extended to include a plot for the fit of the mabBS data (Figure 5.5b, right in first row) and a plot for the detailed predicted distribution of formyl-cytosine modifications (Figure 5.5b, right in second row).

2. The user has an option to exclusively estimate the hidden states' levels, avoid running the optimization for inferring the enzymatic efficiencies (see Figure 5.5a). This option might be convenient to users who quickly want to determine hydroxy-, formal- and methyl-cytosine levels, and are not interested in the enzymes' behavior. We do this using MLE similarly to [74] but with the difference of additionally considering the conversion errors of each experiment. Hence, we are able to predict corrected and far more accurate 5hmC levels

compared with performing a simple subtraction. The advantage of this approach compared to conventional subtraction of BS and oxBS data is evident in Figure 5.6. One clearly sees, that the MLE method gives always more accurate estimations in comparison to subtraction. Simple subtraction method seems to work quite well when the 5hmC levels are very small[1], but it gets really off (almost 10% off from the real value) as the 5hmC levels increase. On the other hand, the absolute error of the MLE method remains constant for all 5hmC levels.

3. The last and most important extension of H(O)TA v.2 is that provides the user the possibility to define himself the underlying stochastic model he wants to run. In a simple a .txt model file the user can describe the transitions of each of the involved subprocesses. Then the tool automatically constructs the hidden state space, produce the stochastic matrix of each of the described subprocesses and finally estimate the values of the parameters of the given model. This feature can be extremely useful for testing different biological assumptions than the ones that have been made for the development of this thesis core-model on a single CpG dyad, or for constructing spatial epigenetic models that consider more than one CpG dyad to study the spatial characteristics (e.g. association and processivity) of certain enzymes, or/and to test the potential need of more complex functions than linear for the efficiencies' behavior, as it has been described in Section 3.3. In Figure 5.7 we see the creation of a single CpG model of quadratic efficiency functions that consists of 5 subprocesses: maintenance, de novo, hydroxylation, formylation and active demethylation.

[1] This a result of the high depth sequencing. For a small coverage subtraction can even give negative 5hmC levels.

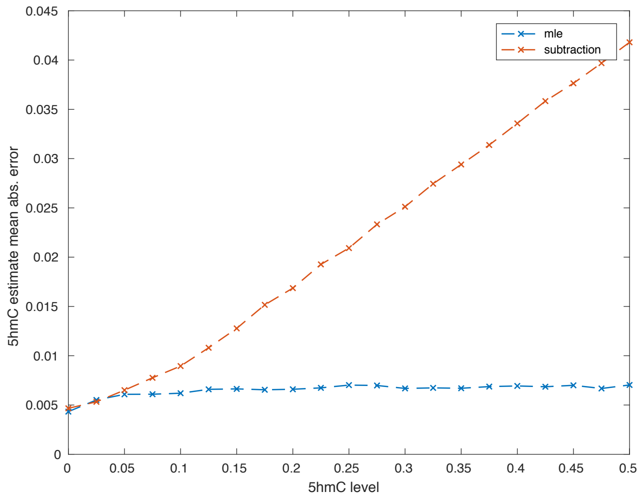
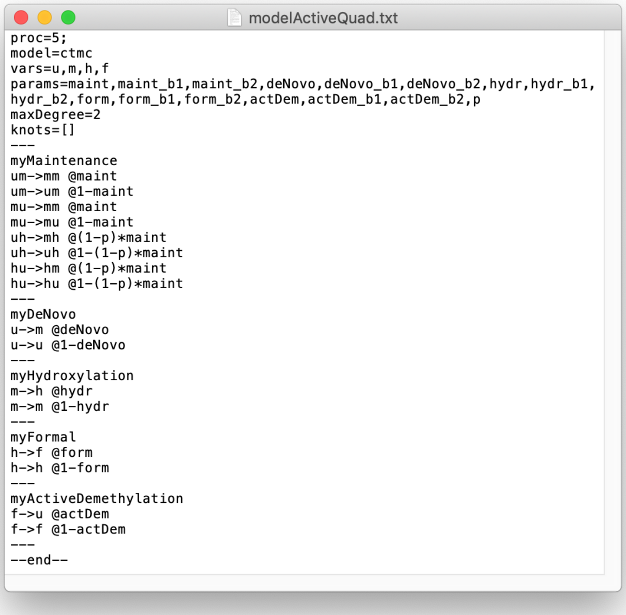


Figure 5.6: MLE vs subtraction estimators for predicting various hydroxylation levels of an artificially generated data set of average available depth sequencing ($5 \cdot 10^3$ x). The plot shows the average absolute errors among 10^3 repetitions for a certain total 5hmC level.



```
proc=5;
model=ctmc
vars=u,m,h,f
params=maint,maint_b1,maint_b2,deNovo,deNovo_b1,deNovo_b2,hydr,hydr_b1,
hydr_b2,form,form_b1,form_b2,actDem,actDem_b1,actDem_b2,p
maxDegree=2
knots=[]
---
myMaintenance
um->mm @maint
um->um @1-maint
mu->mm @maint
mu->mu @1-maint
uh->mh @(1-p)*maint
uh->uh @1-(1-p)*maint
hu->hm @(1-p)*maint
hu->hu @1-(1-p)*maint
---
myDeNovo
u->m @deNovo
u->u @1-deNovo
---
myHydroxylation
m->h @hydr
m->m @1-hydr
---
myFormal
h->f @form
h->h @1-form
---
myActiveDemethylation
f->u @actDem
f->f @1-actDem
---
--end--
```

Figure 5.7: An example of a model input file for H(O)TA v.2. The hybrid model describes a single CpG dyad which undergoes the subprocesses of maintenance and *de novo* methylation, as well as hydroxylation, formylation and active demethylation. The enzymatic efficiencies are quadratic functions of time.

Chapter 6

Whole Genome Analysis

Previous genome wide analyses on mESCs have shown a high or moderate decrease of DNA methylation when the cells during their transition from serum/LIF into 2i containing medium [18, 29, 94]. Furthermore, in some of these analyses it was shown that the oxidation of 5mC to 5hmC is likely to contribute to this DNA demethylation [18]. The aim of this chapter is to model the changes of DNA methylation in mESCs using precise strand specific information, obtained by genome-wide hairpin bisulfite sequencing under conventional and oxidative bisulfite (oxBS) conditions. We present, therefore, a genome-wide analysis based on genome-wide hairpin BS and oxBS sequencing data sampled from protocol RRHPoxBS, described in Section 2.2.2. Following our approach we reach in total around 3 million CpGs across the mouse genome in both WT and Tet TKO cells with a sequencing depth sufficient for comparative modeling.

We sequence six hairpin libraries of WT ESCs at three different time points: Serum/LIF (day0), 72h 2i (day3) and 144h 2i (day6), and four Tet triple KO cells: Serum/LIF, 96h 2i (day4), 168h 2i (day7). For WT we sequence one BS and one oxBS library for each sample, respectively. Using our previously described HMM, 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 oxidized cytosine derivatives.

The core of our computations remains the discrete time stochastic model presented in Chapter 3. The difference relies on the parameter estimation method we use for estimating the unknown model parameters. In Section 6.1 we describe in detail the developed computational methods. The frequentist approach of the maximum likelihood method we used before seems now to provide consistently biased results due to the much smaller available sequencing depth. To overcome the problems of the ill-suited MLE, we design a Bayesian inference method for estimating the unknown parameters, i.e., the enzymatic efficiency functions, of the HMM (Section 6.1.1). Besides, to identify genomic regions of common enzymatic activity, we develop a sophisticated clustering approach that takes into account the uncertainty around the Bayesian inference estimators (Section 6.1.2). At last, in order to investigate possible interactions among the different enzymes we compute their spatial cross-correlations over the entire genome (Section 6.1.3). In Section 6.2 we show the results from the application of these methods and the comparison with the more straightforward approaches. We present both the spatial and the temporal analysis results of our model output for the whole genome in Section 6.3. Finally, in Section 6.4 we discuss the biological findings of particular importance that arise from the comparative analysis of the model's results. For a detailed description of our parallel implementation to run the genome-wide data sets we refer the reader to Section B.2.

6.1 Computational Methods

6.1.1 Bayesian inference

Due to the low depth sequencing per time point and experiment (on average 40 for BS, 29 for oxBS in WT, and 14 for BS in Tet TKO, see Figure B.1b) we expect that the asymptotic properties of the MLE around the true parameter values do not hold [12, 57], especially in cases where the true parameter values are close to the problem's boundary constraints [80]. For this reason, we use a Bayesian Inference (BI) approach to get the posterior distribution of the model parameters, i.e, the efficiencies over time, for each CpG of the genome-wide dataset.

Metropolis-Hastings

To compute the model's unknown parameters we use BI by sampling from the multi-dimensional posterior

$$P(\mathbf{v} \mid \text{data}) = \frac{\mathcal{L}_2(\text{data} \mid \mathbf{v})P(\mathbf{v})}{\int_{\mathbf{v}} P(\text{data}, \mathbf{v})},$$

where $\mathcal{L}_2(\text{data} \mid \mathbf{v})$ is the likelihood defined in Section 3.2.2 and $P(\mathbf{v})$ is the prior distribution of the parameter vector $\mathbf{v}$. To avoid approximating the normalizing factor $\int_{\mathbf{v}} P(\text{data}, \mathbf{v})$ we apply a Metropolis-Hastings MCMC approach [33].

In general, Metropolis-Hastings algorithm starts from an initial sample point $\mathbf{v}_0$ of the parameter vector, i.e., $\mathbf{x} = \mathbf{v}_0$, generates for each current state $\mathbf{x}$ a new sample $\mathbf{y}$ based on a proposal distribution $g(\mathbf{y} \mid \mathbf{x})$, and it accepts the new state (sample) with acceptance probability

$$A(\mathbf{y} \mid \mathbf{x}) = \min\left(1, \frac{f(\mathbf{y})}{f(\mathbf{x})} \frac{g(\mathbf{x} \mid \mathbf{y})}{g(\mathbf{y} \mid \mathbf{x})}\right), \tag{6.1}$$

where f is a function proportional to the target distribution[1], i.e., here $f(\mathbf{v}) = \mathcal{L}_2(\text{data} \mid \mathbf{v})P(\mathbf{v})$. Intuitively, the ratio in Eq. (6.1) is defined to balance between a sampler that on the one hand tends to visit high probability density regions, i.e., $f(\mathbf{y})/f(\mathbf{x})$), while at the same time tries to avoid getting stuck at one region and explore a broad landscape of the parameter space. The last is expressed by the likelihood term $g(\mathbf{x} \mid \mathbf{y})/g(\mathbf{y} \mid \mathbf{x})$. Fulfilling the conditions of ergodicity [91] and irreducibility [67] for the above described Markov chain, the algorithm guarantees that after simulating for a sufficiently large number of steps the stationary distribution of the chain is the same as the target distribution we want to sample from [20].

Here to apply Metropolis-Hastings we make use of the following prior and proposal distributions.

Prior Distribution As prior distribution we choose for all CpGs the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where the mean $\boldsymbol{\mu}$ is the average of the estimated efficiencies for the individual loci of Section 4.1.1. Similarly, Σ is the average of the corresponding estimated covariance matrices.

Proposal Distribution To fulfill the constraints of the optimization problem regarding the parameter vector $\mathbf{v}$ we use an asymmetric, truncated proposal distribution. The bounds of the truncated proposal are determined s.t. the constraints for the efficiencies constantly hold for the time span of the observations, i.e., the efficiencies are in $[0, 1]$ for all $t \in [0, t_{\max}]$ (see Section 3.2.2). More concretely, let $v = |\mathbf{v}|$ denote the number of model parameters. Here it is $v = 7$ because we make the same linear assumption about the efficiency functions as in Section 3.2.2. In every state $\mathbf{x} \in \mathbb{R}^v$ of the MCMC we generate the next sample $\mathbf{y} \in \mathbb{R}^v$ from a product of truncated univariate normals

$$g(\mathbf{y} \mid \mathbf{x}) = \prod_i \mathcal{N}_{\mathrm{t}}(\mathbf{y}_i \mid \mathbf{x}_i, \sigma_i^2/c, a_i, b_i),$$

[1] The target distribution is the posterior of the parameters, $P(\mathbf{v} \mid \text{data})$.

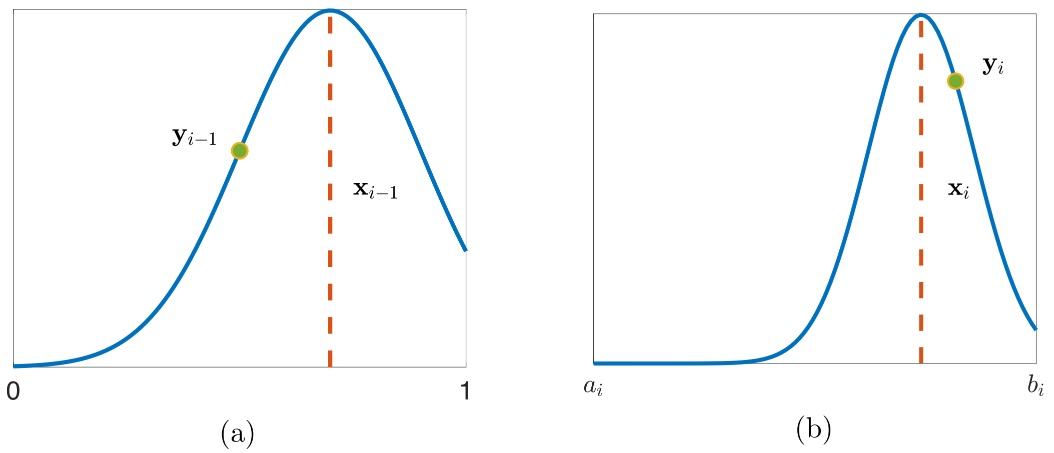


Figure 6.1: Metropolis-Hastings' update step: We sample a new efficiency vector using two truncated normal distributions in two steps: (a) Step 1: We sample the intercept $\mathbf{y}_{i-1}$ from the truncated normal with mean $\mathbf{x}_{i-1}$ and bounds [0, 1]. (b) Step 2: We sample the gradient $\mathbf{y}_i$ 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.

around the current MCMC point $\mathbf{x}$, where $\mathbf{x}_i$ refers to the i-th entry of the parameter vector for $i = 1, \ldots, 7$, 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 of the vector $\mathbf{y}$ where $\mathbf{y}_i$ refers to the gradient of an efficiency function and $\mathbf{y}_{i-1}$ to the corresponding intercept for $i = 1, \ldots, 7$. As illustrated in Figure 6.1, 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 $a_i = -\mathbf{y}_{i-1}/t_{\max}, b_i = (1 - \mathbf{y}_{i-1})/t_{\max}$. These bounds guarantee that each efficiency function is $\in [0, 1]$ for all observation time points. The bounds of the non-recognition probability p (see again Section 3.2.2) are set as those of an intercept, i.e., $[a_i, b_i] = [0, 1]$. Hence, the use of the above proposal distribution satisfies all the constraints for the parameter vector.

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 deviation of the proposal distribution[2] s.t. the average MCMC acceptance ratio is around 25% of the total number of generated samples [79]. In our runs we see that the above MCMC converges almost surely after 10^4 simulation steps. 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 matrix.

Hypervolume of the β confidence hyper-ellipse

To quantify the improvement of BI against the MLE regarding the decrease in the uncertainty of the parameter estimators we computed the average hypervolume corresponding to the covariance matrices of the estimators of each method. 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

$$V = \frac{2\pi^{v/2}}{v\Gamma(v/2)}(\chi^2_{crit})^{v/2}|\Sigma_{\mathbf{v}^*}|^{1/2},$$

where $|\Sigma_{\mathbf{v}^*}|$ is the determinant of the estimators' covariance matrix, $\Gamma(x)$ is the gamma function, and χ^2_{crit} is the critical value for chi-squared distribution $\chi^2(v)$ for a given confidence level β. In Figure 6.2 we illustrate the ellipse of the bivariate normal distribution and show the relation of it to the covariance matrix of the two-dimensional normal.

6.1.2 Clustering of enzymatic efficiencies

To split the genome in regions of similar enzymatic activity we extended and applied to the output produced by our model a sophisticated clustering approach, named

[2] 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.

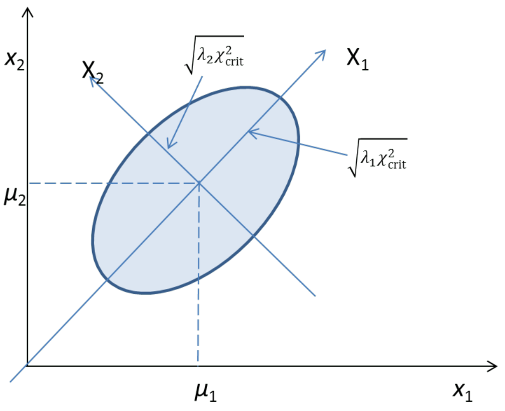

x_2
X_2
$\sqrt{\lambda_2\chi^2_{\text{crit}}}$
X_1
$\sqrt{\lambda_1\chi^2_{\text{crit}}}$
μ_2
μ_1
x_1

Figure 6.2: The ellipse of the bivariate normal has axes pointing in the directions of the eigenvectors X_1, X_2 of the covariance matrix Σ. The long axis of the ellipse points in the direction of the first eigenvector X_1 and the short 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 $l_i = \sqrt{\lambda_i \chi^2_{crit}}$.

k-error, that takes into account not only the estimated parameter vectors but also their standard deviations [49]. In general, it is well known that in case of measurements with uncertainty or estimation errors, incorporating a quantified notion of this uncertainty around the data points in the clustering process can produce different and closer to the "ground" truth clusters.

k-error clustering

The k-error clustering algorithm is a modification of the k-means algorithm that takes into account the uncertainties of each data point, i.e, the covariance matrix $\Sigma_{\mathbf{v}}$ of the parameter vector of the efficiencies $\mathbf{v}$ [49]. The difference between the two methods is nicely illustrated in Figure 6.3.

Let $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^v$ be the estimated parameter column vectors and $\Sigma_1, \ldots, \Sigma_N \in \mathbb{R}^{v \times v}$ the associated covariance matrices for N input CpGs. Let us assume that the estimated parameter vectors are independent and each arises from a v-variate normal distribution with one of k possible means $\theta_1, \ldots, \theta_k \in \mathbb{R}^v$, that is $\mathbf{v}_i \sim \mathcal{N}_v(\mu_i, \Sigma_i)$, where $\mu_i \in \{\theta_1, \ldots, \theta_k\}, k \leq N$, for $i = 1, \ldots, N$. Then, we seek to find the clusters

$C_1, \ldots, C_k$ such that each cluster C_j consists of all parameter vectors that have the same mean θ_j for $j = 1 \ldots k$.

More formally, if $S_j = \{i \mid \mathbf{v}_i \in C_j\}$ contains all CpGs i whose parameter vectors belong to cluster C_j, then $\forall i \in S_j$ it holds $\mu_i = \theta_j$, for $j = 1 \ldots k$. Hence, given all parameter vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$, and their error matrices $\Sigma_1, \ldots, \Sigma_N$, we search for a partition $S = (S_1, \ldots, S_k)$ and $\theta = (\theta_1, \ldots, \theta_k)$ that maximizes the likelihood:

$$\mathcal{L}_c(\mathbf{v}) = \prod_{j=1}^{k} \prod_{i \in S_j} \frac{1}{2\pi}^{p/2} |\Sigma_i|^{-1/2} e^{-1/2(\mathbf{v}_i - \theta_j)\Sigma_i^{-1}(\mathbf{v}_i - \theta_j)^\intercal}, \tag{6.2}$$

where $|\Sigma_i|$ is the determinant of matrix Σ_i for $i = 1, \ldots, N$.

As it is shown in [49], maximizing the likelihood of Eq. 6.2 is equivalent to minimizing the total squared Mahalanobis distance of the points that belong to a cluster from the cluster centroid. Hence, it is enough to solve

$$\min_S \sum_{j=1}^{k} \sum_{i \in S_j} (\mathbf{v}_i - \hat{\theta}_j)\Sigma_i^{-1}(\mathbf{v}_i - \hat{\theta}_j),$$

where $\hat{\theta}_j$ is the ML estimate of θ_j given by

$$\hat{\theta}_j = \Big(\sum_{i \in S_j} \Sigma_i^{-1}\Big)^{-1} \Big(\sum_{i \in S_j} \mathbf{v}_i \Sigma_i^{-1}\Big) \tag{6.3}$$

for $j = 1, \ldots, k$. Notice that the estimated centroid $\hat{\theta}_j$ of Eq.(6.3) is a weighted mean of the points in cluster C_j. We refer to it as the *Mahalanobis mean* of C_j. By using simple matrix algebra we can additionally compute the covariance matrix

$$\Psi_j = \mathrm{Cov}(\hat{\theta}_j) = \mathrm{Cov}\left(\Big(\sum_{i \in S_j} \Sigma_i^{-1}\Big)^{-1} \Big(\sum_{i \in S_j} \mathbf{v}_i \Sigma_i^{-1}\Big)\right) = \Big(\sum_{i \in S_j} \Sigma_i^{-1}\Big)^{-1}$$

associated with the estimated centroid $\hat{\theta}_j$.

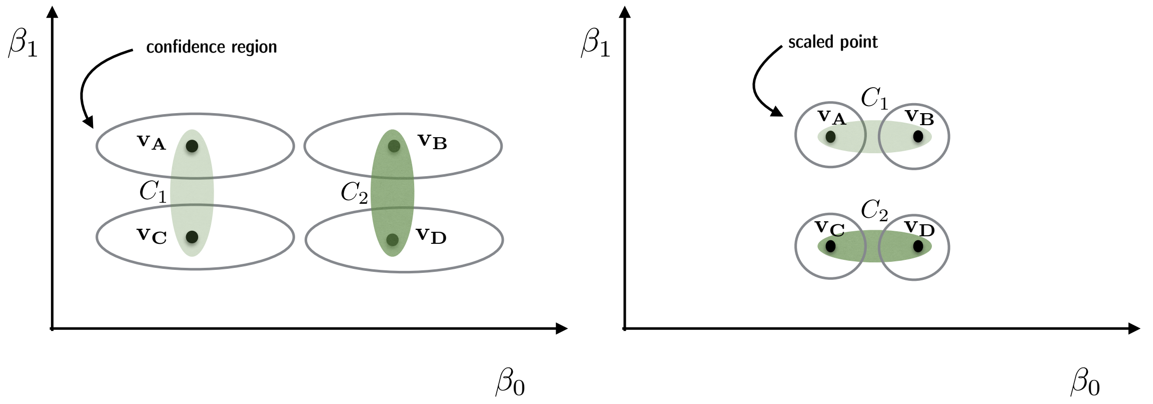


Figure 6.3: Illustration of the clustering of the estimated enzymatic efficiency with intercept β_0 and gradient β_1 for CpGs A, B, C, D. Left: k-means clustering considers only the euclidean distance between two parameter vectors without taking into account the uncertainty around them. Right: k-error clustering "scales" the distance between two parameter vectors based on their estimated covariance matrices.

After we randomly choose an initial set of k centroids[3], the k-error clustering algorithm follows from the above as an iteration over the next two steps until the chosen partitioning does not change.

1. For each CpG $i = 1, \dots, N$ assign its parameter vector $\mathbf{v}_i$ to the cluster C_{j^*} whose centroid is the closest using the squared Mahalanobis distance, i.e,

$$i \in S_{j^*}, \mathbf{v}_i \in C_{j^*} \iff j^* = \arg\min_j (\mathbf{v}_i - \hat{\theta}_j)\Sigma_i^{-1}(\mathbf{v}_i - \theta_j)^\intercal. \tag{6.4}$$

2. For clusters $C_1, \dots, C_k$ compute the new cluster centroids $\hat{\theta}_1, \dots, \hat{\theta}_k$ as the Mahalanobis means of the clusters (Eq. 6.3).

Note that the distance function used in Eq. 6.4 is chosen so that it guarantees the decrease of the objective function in each iteration of k-error [49], and as a consequence k-error algorithm will always converge after a finite number of steps.

[3] This initialization method is called Forgy method.

Metrics for deciding the number of clusters

In order to identify the "optimal" number of clusters we evaluate three different metrics for a range of number of clusters $k = 1, \dots, K$. We use Davies-Bouldin and Calinski-Harabasz criteria, as well as the widely used elbow method. The first two metrics evaluate the overall within-to-between cluster variability, each in a slightly different fashion. The third metric considers the sum of squared errors (SSE) of a certain clustering. The goal of it 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

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

For a given distance function d, S_i is a measure of within cluster i variance, i.e.,

$$S_i = \frac{1}{|C_i|} \sum_{\mathbf{v} \in C_i} \mathrm{d}(\mathbf{v}, m_i)$$

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 as the one that provides the smallest DB.

Calinski-Harabasz Criterion The Calinski-Harabasz criterion, alternatively 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

$$SS_W = \sum_{i=1}^{k} \sum_{\mathbf{v} \in C_i} \mathrm{d}(\mathbf{v}, m_i).$$

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): $\sum_{i=1}^{k} \sum_{\mathbf{v} \in C_i} \mathrm{d}(\mathbf{v}, m_i)$ for $k = \{1 \ldots K\}$. We choose the optimal k to be the point where the graph starts to flatten significantly. E.g., in the curve of Figure 6.4 the optimal number of clusters is three.

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,

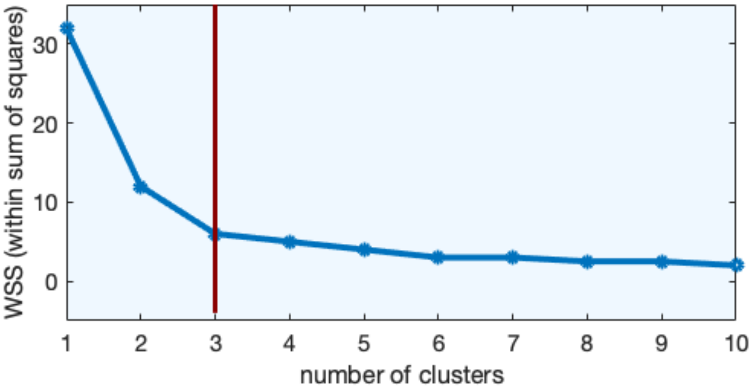


Figure 6.4: Elbow method: The "optimal" number of clusters is the point where the graph starts to smooth out, i.e., the "elbow" of the graph.

in case of k-means we use the squared 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.

6.1.3 Spatial and non-spatial cross-correlations

In this Section we formally define the measures we used to quantify the similarities between two different enzymatic activities (spatial cross-correlation), or the similarity of each enzymatic activity with itself (spatial autocorrelation), when these are seen as functions of space. We, also, shortly recap Pearson correlation.

Spatial cross-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

$$R_X(\tau) = \frac{\mathbb{E}[(X_s - \mu_{X_s})(X_{s+\tau} - \mu_{X_{s+\tau}})]}{\sigma_{X_s}\sigma_{X_{s+\tau}}}.$$

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

$$\rho_{X,Y}(\tau) = \frac{\mathbb{E}[(X_s - \mu_{X_s})(Y_{s+\tau} - \mu_{Y_{s+\tau}})]}{\sigma_{X_s}\sigma_{Y_{s+\tau}}}.$$

Here, 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, for a fixed τ, genomic position $s \in S(\tau)$ when both CpGs of genomic 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

$$\hat{\rho}(\tau) = \frac{1}{|S(\tau) - 1|\hat{\sigma}_{X_s}\hat{\sigma}_{Y_{s+\tau}}} \sum_{s \in S(\tau)} (X_s - \bar{X}_s)(Y_{s+\tau} - \bar{Y}_{s+\tau})].$$

Pearson correlation

The Pearson correlation coefficient between two random variables X, Y is defined as

$$\rho_{\text{Pearson}} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$$

The corresponding sample Pearson correlation estimate is

$$r = \frac{1}{(\mid G \mid -1)\hat{\sigma}_X\hat{\sigma}_Y} \sum_{s \in G} (X_s - \bar{X})(Y_s - \bar{Y}),$$

where G is the set of all CpG positions available in our data. Note that the Pearson correlation coefficient is equivalent to spatial cross-correlation for $\tau = 0$.

6.2 Results

Our input is double strand, 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 knock out (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 3.2.1 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 B.1a, B.1c) we assume a linear behavior of the efficiencies over time and we analyze the HMM as described in Section 6.1.1 for estimating both the values of (hydroxy-)methylation efficiencies and levels over time. Note that our input data is, due to the sequencing, relatively, equally distributed among the different chromosomes (Figure B.3, B.4). 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.

6.2.1 BI vs MLE

Fit of whole-genome data

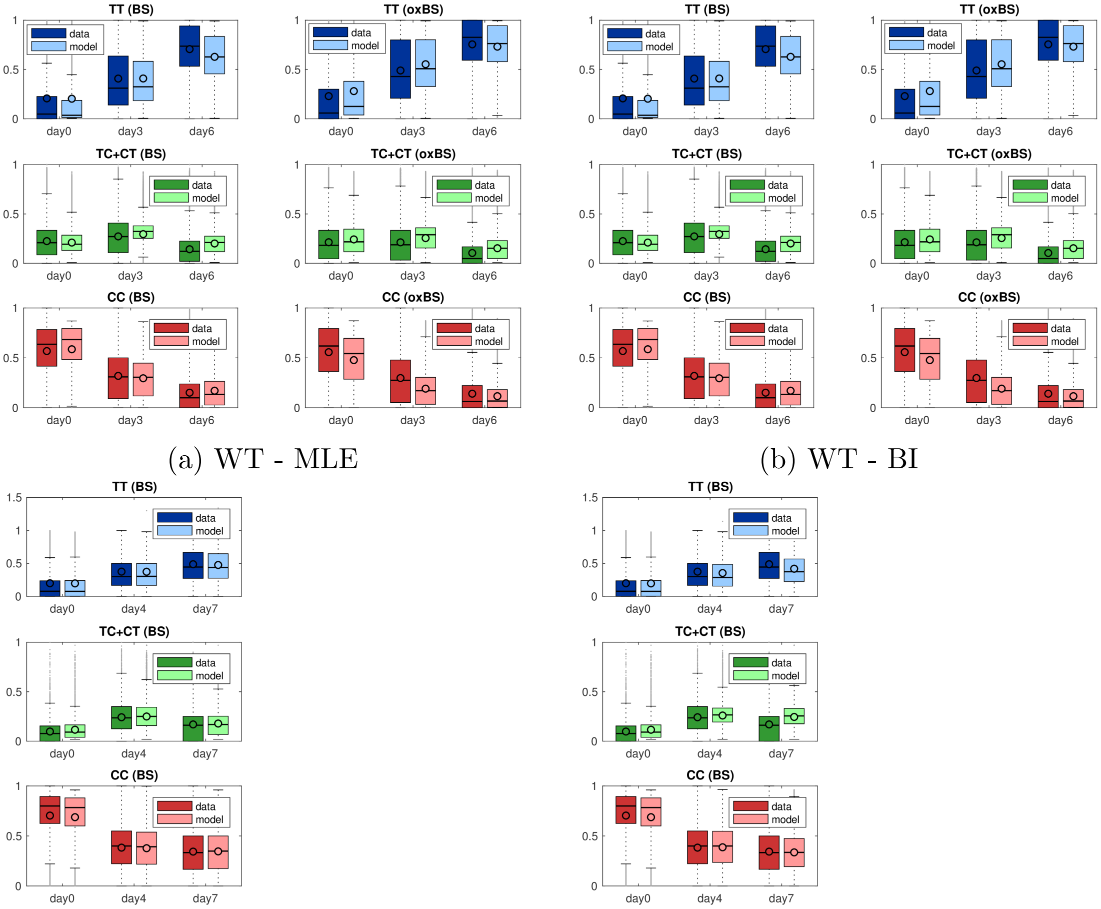
Using box plots, we compare the levels of CC, TT and CT-TC CpG dyads for the whole genome present in the data of BS and oxBS in WT (Figure 6.5a, 6.5b) and of BS in Tet TKO (Figure 6.5c, 6.5d) and the probabilities of the observable states predicted by the two HMMs using MLE or BI for estimating the model parameters.

The circles inside the plots correspond to 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 6.1 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 not surprising 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 6.6 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 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 predicts it to increase. 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.

Enzymatic activities' prediction

By making a comparison between the MLE and the BI methods' prediction of the enzymatic activities we have several reasons to trust the output of the BI method



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

Figure 6.5: 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.

more than this of the ML. First in WT data we observe that only the BI estimates are in line with the genome wide or the vast majority of individual examined loci behavior being described in Chapter 4 and in the literature [18, 94]. Second, 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 (see Section 6.3.1 and Figure 6.13 for details).

Table 6.1: 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.

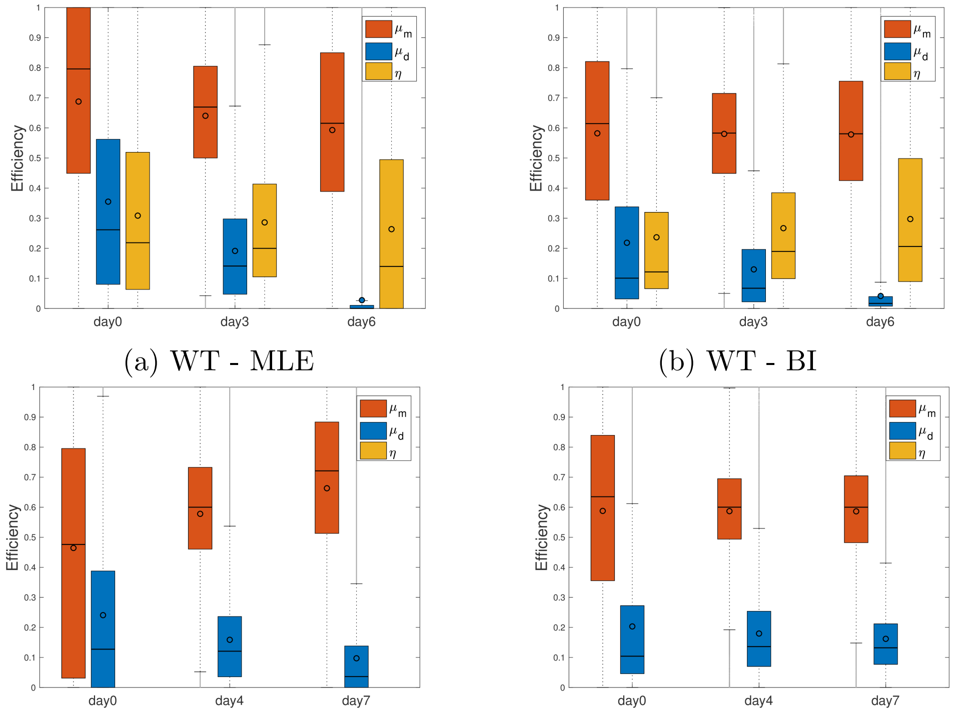
experiment - method	$\hat{D}_{KL}(P_{bs}\|\|\pi_{bs})$	$\hat{D}_{KL}(P_{ox}\|\|\pi_{ox})$
WT - MLE	0.1802	0.2369
WT - BI	0.2904	0.3941
Tet TKO - MLE	0.154	-
Tet TKO - BI	0.277	-

Observing the box plots of Figure 6.6 we note that the dispersion of the efficiencies values is evidently smaller for the BI estimates, which probably is a consequence of the higher precision of the BI estimates compared with the MLE. In particular, in WT the average volume of the 95% hyper-ellipse over all CpGs in case of MLE is 0.0024, while the average 95% 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 B.2 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). The small differences on the prediction of the hidden states despite the evident difference in the enzymes' efficiency estimators in particular for the Tet TKO case, indicate how critical an ML estimation bias can be for an accurate model prediction.

Overall, from the above we confirm that, indeed, 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 [5, 64]. Therefore, we use the BI estimates in all the output analysis we conduct in Section 6.3 and as the input to the clustering algorithm we describe in the sequel.

6.2.2 k-error vs k-mean

To identify CpGs with similar enzymatic behavior we applied the clustering method presented in Section 6.1.2 that takes into account the uncertainty around the esti-



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

Figure 6.6: Bar plots for maintenance (μ_m), *de novo* (μ_d) and hydroxylation (η) efficiencies over time taken by MLE (a), (c) and BI (b), (d) methods.

mated parameters, i.e., in our case the covariance matrix of the parameters' posterior distribution of the estimated efficiency linear functions.

For each total number of clusters $k = \{1, \ldots, 10\}$ we produce 100 initializations in order to avoid that the algorithm converges in a local optimum. Given the three different criteria described in Section 6.1.2 we observe that a typical k-means clustering algorithm would return four as the optimal number of clusters, while our approach always decides for two clusters (Figure 6.7).

In Figure 6.8 we see the optimal k-means clustering and in Figure 6.9 the optimal k-error clustering. We observe that in clusters 2, 3, and 4 of k-means algorithm the average maintenance methylation efficiency slightly increases over time which contradicts the decreasing concentrations of the H3K9me2 and Uhrf1[4] in 2i, especially assuming it happens on a genome wide level [94]. Based on this, we conclude that

[4] Uhrf1 is the main cofactor of Dnmt1.

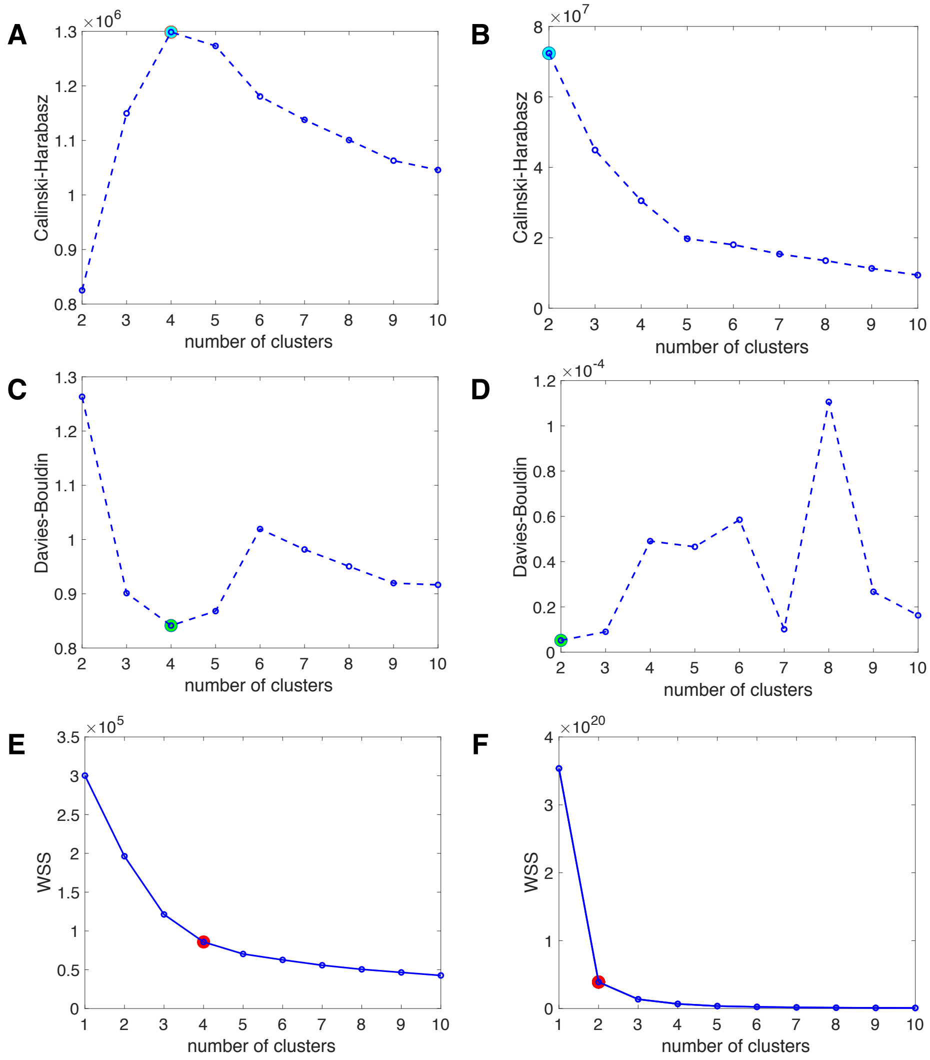


Figure 6.7: Optimal number of clusters for k-means and k-error algorithms according to three clustering validity metrics: Davies-Bouldin criterion for k-means **(A)** and k-error **(B)**. Calinski-Harabasz criterion for k-means **(C)** and k-error **(D)**. Elbow method for k-means **(F)** and k-error **(E)**.

the two additional clusters returned from k-mean are probably the result of noise on the estimation due to the insufficient depth sequencing.

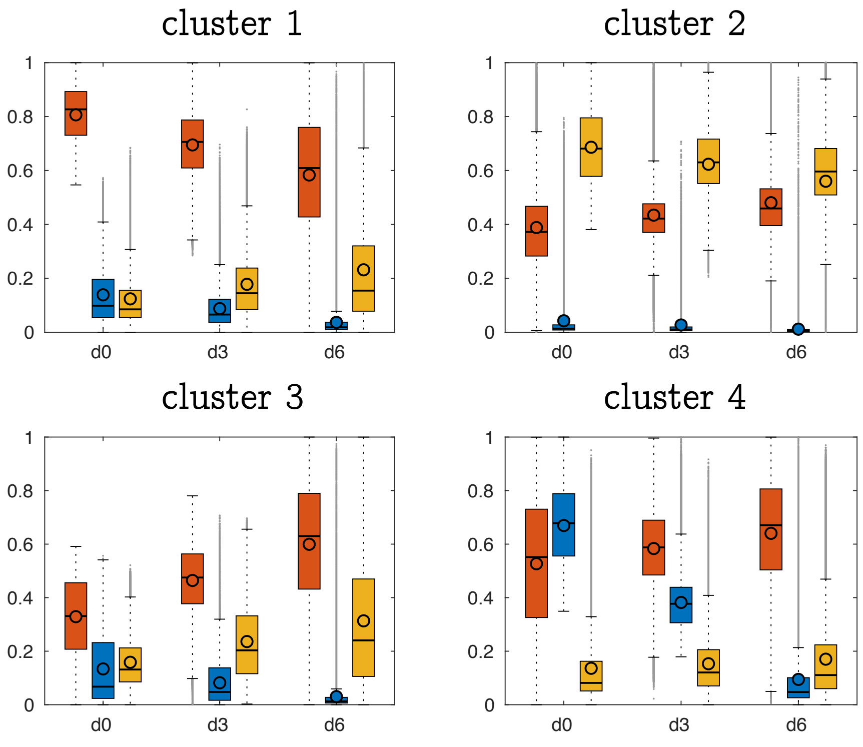


Figure 6.8: The optimal clustering of the enzymatic efficiencies over time based on the k-means algorithm and the squared euclidean distance. Red = maintenance methylation (μ_m), blue = *de novo* methylation (μ_d), yellow = hydroxylation (η).

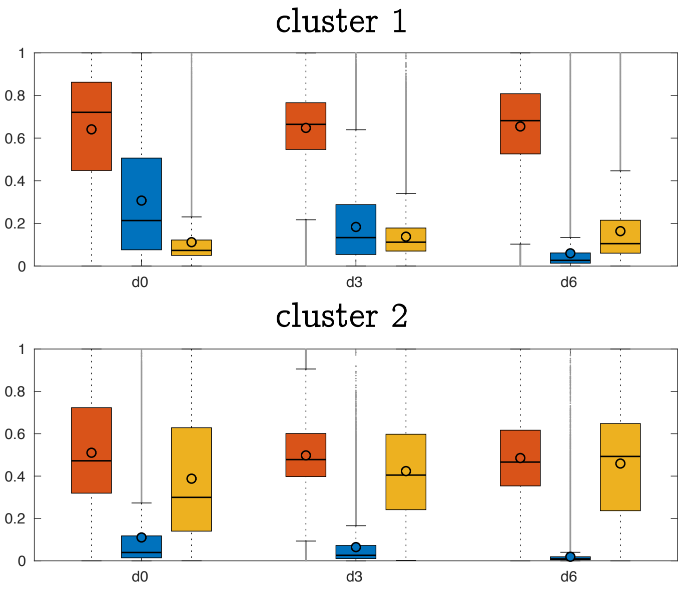


Figure 6.9: The optimal clustering of the enzymatic efficiencies over time based on the k-error algorithm and the squared Mahalanobis distance. Cluster 1 contains 855201 CpGs, while cluster 2 contains 702901 CpGs.

6.2.3 Enzymatic activities' correlations

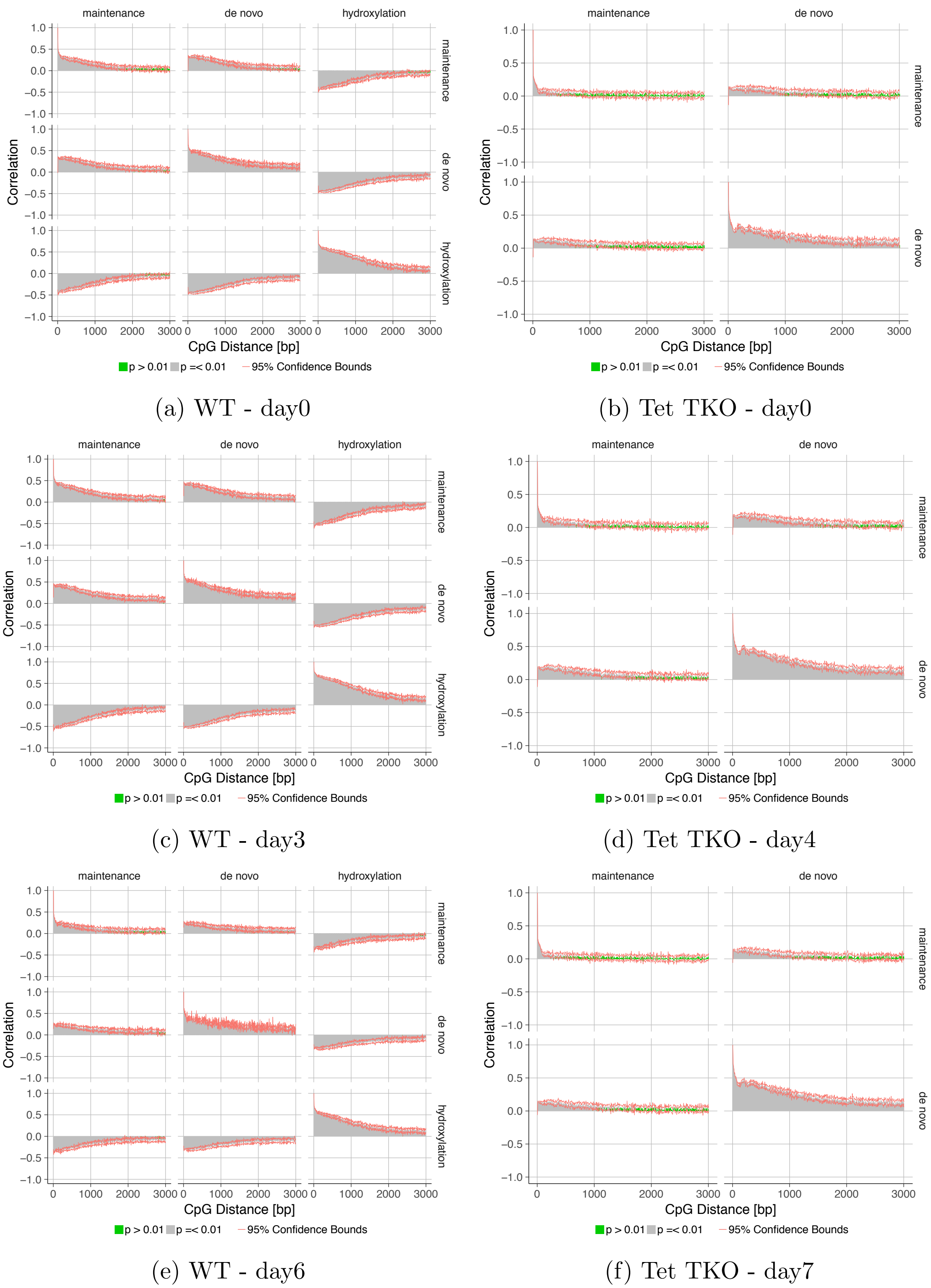
Spatial cross-correlations

To investigate the enzymatic antagonism between methylation and hydroxylation activities we calculated the spatial auto- and cross-correlations of methylation and hydroxylation efficiencies (Section 6.1.3). Choosing $\tau = 0, 5, \dots, 3000$ we plot in Figure 6.10 the sample autocorrelations and sample cross-correlations between all efficiencies at all time points in WT (Figure 6.10a, 6.10c, 6.10e) and Tet TKO (Figure 6.10b, 6.10d, 6.10f) 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.

In line with Figure 6.15, 6.17, we consistently[5] see a positive correlation between maintenance and *de novo* efficiency and a negative correlation between hydroxylation and both methylation efficiencies. With increasing distance of CpGs, all correlations get closer to zero. Maintenance autocorrelation drops rather quickly and becomes almost zero around 1500bp. After this point the autocorrelation is no longer significant (p-value > 0.01). In contrast, the autocorrelations of *de novo* and hydroxylation efficiency show initially higher values but also seem to smoothen out after around 2000bp on average.

Interestingly, in Tet TKO cells, the autocorrelation of maintenance is initially strongly reduced (≈ 0.25) and seems to flatten out earlier than in WT, around 500bp, 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 *de novo* autocorrelation seems to remain unaffected compared to WT cells. Overall the spatial autocorrelations do not indicate any change of the significant window size of dependent enzymatic activity over time for either cell type.

[5] This is true besides the case where $\tau = 0$, because obviously one CpG dyad can not be accessed by two enzymes at the same time.



(a) WT - day0

(b) Tet TKO - day0

(c) WT - day3

(d) Tet TKO - day4

(e) WT - day6

(f) Tet TKO - day7

Figure 6.10: 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

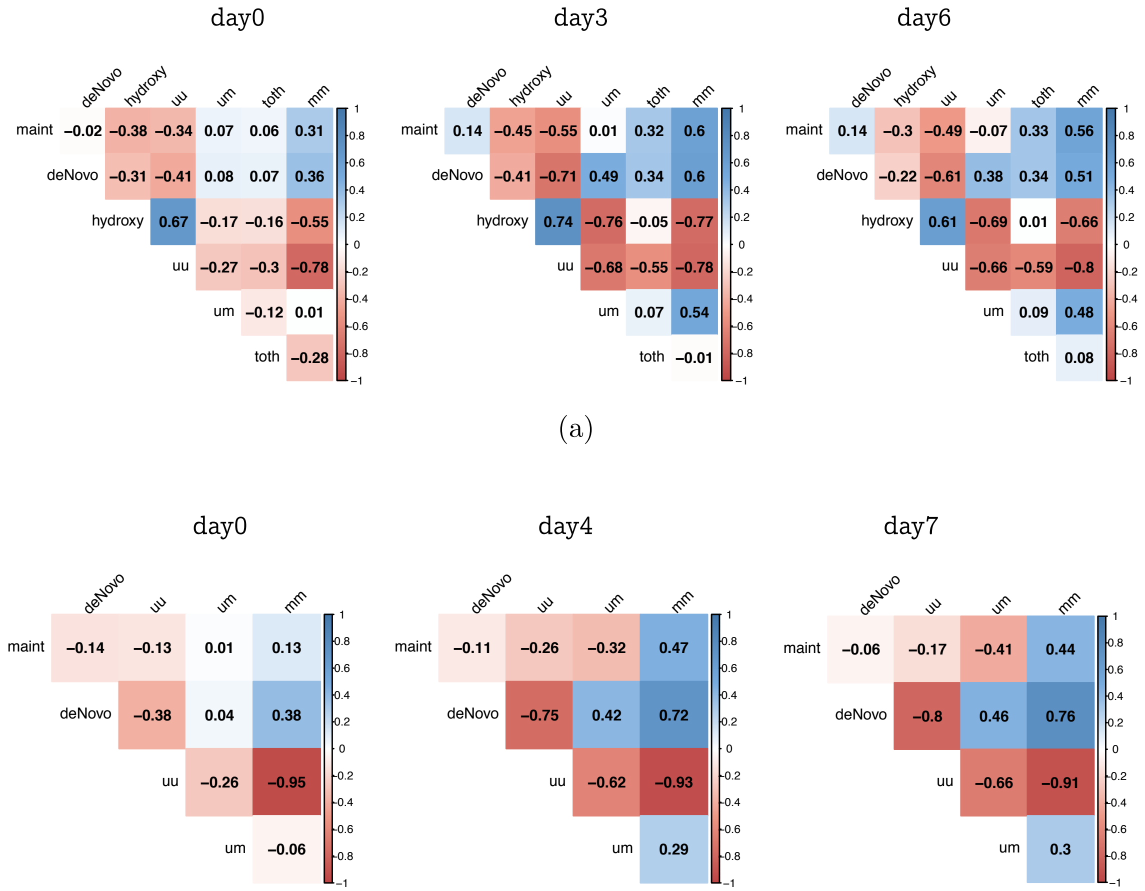
In addition to the spatial correlation, we calculated a Pearson correlation coefficient between the enzymes' efficiencies and the predicted modification levels. In WT ESCs, we observe a positive correlation of fully methylated CpGs with maintenance and *de novo* efficiency at day0 which increases for later time points. In addition, after day0 there is a positive correlation between *de novo* methylation efficiency and hemimethylated CpGs for day3 and day6.

Verifying the gene plots, we observe no correlation between hydroxylated CpGs and hydroxylation efficiency. Instead, hydroxylation activity strongly correlates with unmethylated CpGs. Consequently, we conclude that high hydroxylation activity is not sufficient to generate stable 5hmC but will always result in CpGs free of methylation. At last, there is also a strong anti-correlation between fully methylated and fully unmethylated dyads, probably indicating the tendency of distinct regions either to fully maintain or to completely lose their methylation pattern.

In Tet TKO, the correlation between maintenance methylation efficiency and fully methylated CpG dyads is reduced and in case of day0 is 0.13. We observe, instead, a stronger correlation of fully methylated CpGs with *de novo* methylation efficiency which points again towards a misregulated methylation activity in the absence of Tet enzymes.

6.3 Output Analysis

We split the biological analysis of the genome-wide results for both WT and Tet TKO cells into two sections. In Section 6.3.1 we present a spatial analysis, while in Section 6.3.2 we focus on a temporal analysis of the results arising from our approach.



(a)

(b)

Figure 6.11: Pearson correlation between enzymes' efficiencies and modification levels in WT ES cells (a) for day0, day3 and day6 and in Tet TKO ESCs (b) for day0, day4 and day7. 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.

6.3.1 Spatial analysis

CpG methylation in WT and Tet TKO ESCs

We report the approximate level and distribution of 5mC over the two DNA strands within the obtained RRHPoxBS data. In line with previous reports [18, 29, 94] 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 6.13A). On the contrary, Tet TKO cells present a higher methylation level under both, primed/Serum (75%) and naive/2i (40%) conditions. The majority of methylated CpGs is present in a symmetric methylation state under both cultivation conditions, but we detect as well a significant amount of hemimethylated CpGs.

Hemimethylation The number of hemimethylated CpGs in Tet TKO cells at day0 is strongly reduced compared to WT cells but it does not differ much for the rest days. In WT, 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.

Since hemimethylated CpGs are the result of either *de novo* methylation events and/or active and passive demethylation, a potential strand specific gene regulation mechanism should show a preference to a particular DNA strand. Analyzing, though, the strand specific methylation of genes transcribed from plus- and minus strand, we did not observe any methylation differences between genes expressed from upper or lower DNA strands. Hemimethylated CpGs are always equally distributed among both DNA strands and this is true for low/not expressed genes as well (Figure 6.12).

NonCpG methylation in WT and Tet TKO ESCs Frequently, DNA methylation occurs outside of a CpG context [77, 100]. It is known that methylation in a nonCpG context is, due to its asymmetric nature, exclusively the result of *de novo*

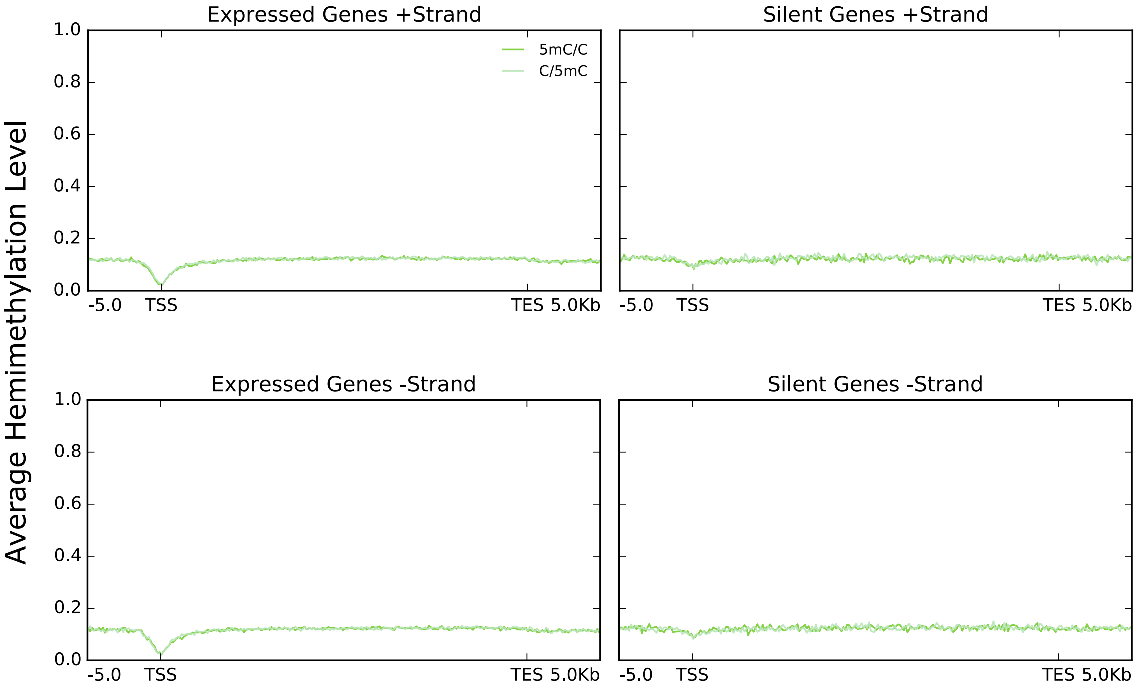


Figure 6.12: Hemimethylated CpGs detected by RRHPoxBS across expressed and not/low expressed genes. Dark green = 5mC/C, light green = C/5mC

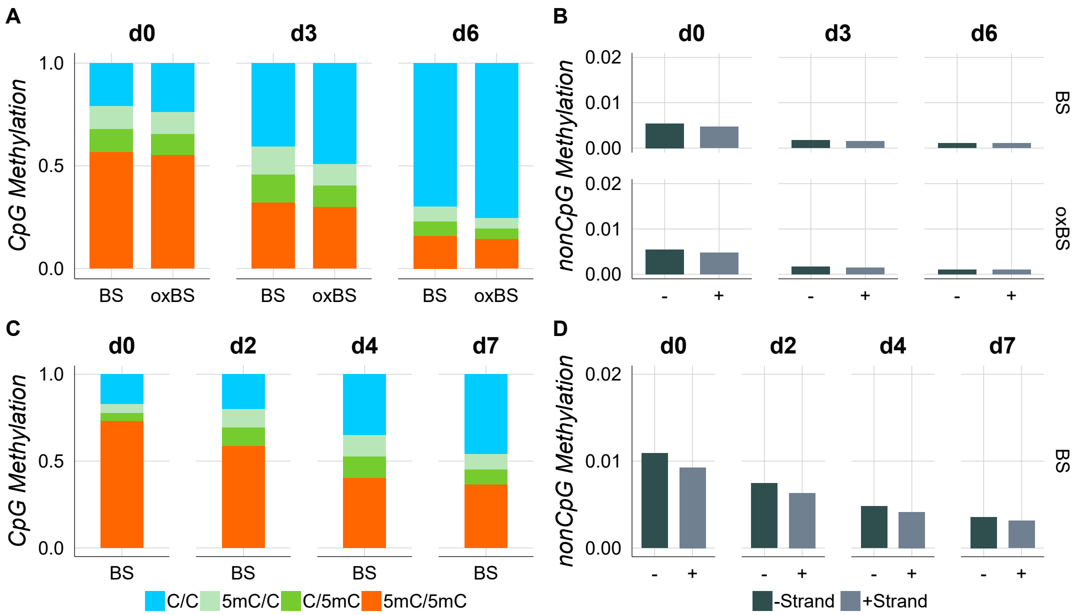


Figure 6.13: CpG and nonCpG methylation. Genome-wide CpG methylation levels of WT ES cells cultivated under Serum/LIF conditions (d0), and their shift to 2i after 72h (d3), 144h (d6) **(A)**. NonCpG methylation levels of WT cells **(B)**. CpG methylation levels of Tet TKO ES cells **(C)**. NonCpG methylation levels of Tet TKO ES cells **(D)**.

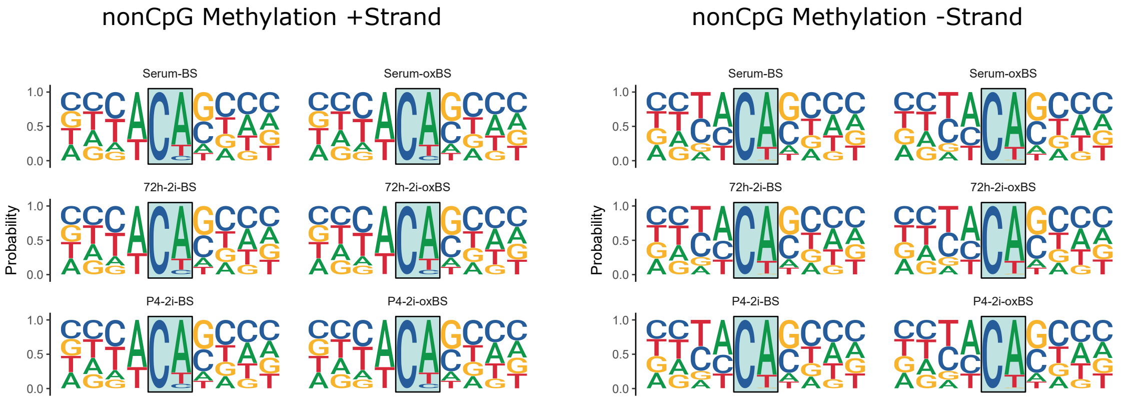


Figure 6.14: Occurrences of nonCpG methylation in Serum and 2i cultivated WT ES cells. Size of bases indicate the probability at a given position. nonCpG with 4 bases up- and downstream are shown.

activity [25, 1]. We determine the sequence occurrence of nonCpG methylation in our WT samples by considering only nonCpG positions which are (i) methylated above the conversion error, (ii) show at least three methylated reads and (iii) a coverage of $\geq 10x$. Considering both strands and in accordance with the literature, we find that the most common methylated sequence after CpG on both DNA strands is present in a CpA context, mostly located in regions with high CpG methylation levels (Figure 6.14). Restricting, however, the analysis only at the plus strand, the most common nonCpG methylation can be found in a ApCpA sequence context (Figure 6.14).

WT ESCs show approximate 1% nonCpG methylation at day0 which quickly declines in 2i (Figure 6.13B). In contrast, Tet TKO cells exhibit twice as much non-CpG methylation under Serum/LIF conditions and furthermore, nonCpG methylation seems to be more stable during 2i cultivation under naive conditions. Even at day7, we still detect considerable levels of methylated CpGs in a nonCpG context Figure 6.13D).

Enzymatic efficiency profile across genes

Using Bayesian Inference (BI) to identify the unknown HMM parameters (Section 6.1.1), we predicted the efficiencies of maintenance methylation, *de novo* methylation and hydroxylation activity based on BS and oxBS data for WT and Tet KO ESCs. Based on the model's output we, first, investigated the efficiency profiles of Dnmts and Tets across genes. Both, maintenance and *de novo* methylation activity show initially a strong activity at the gene body (≥ 0.6 for maintenance, and ≥ 0.1 for *de novo*), whereas at the transcription start site (TSS), efficiencies are strongly reduced (Figure 6.15A). Note, that in case of *de novo* methylation, the activity at the TSS drops to zero. Hydroxylation presents, on the other hand, the inverse behavior; it shows reduced activity at the gene body and high efficiency at the TSS. Over time we see an increase of Tet activity at the gene body, whereas *de novo* activity shows a strong reduction. The temporal profile of maintenance activity suggests no further changes at the gene body or the TSS.

5mC and 5hmC follow in their profile the activity of *de novo* and maintenance. Both modifications are enriched at the gene body and reduced at the TSS. The continuously observed simultaneous occurrence of reduced levels of 5hmC along with strong Tet activity across the TSS (Figure 6.15B) seems maybe puzzling at first sight, but it is probably the consequence of missing 5mC substrates, due to either passive (see Section 4.2) or/and active transform of 5hmC to C [94]. Indeed, our model predicts that 5hmC is very probably not recognized by Dnmt1 after replication (on average with probability around 75%) and by that enhances passive demethylation.

In the absence of Tet enzymes there are distinct differences in maintenance and *de novo* methylation activity compared to WT data. For instance, maintenance methylation activity in Tet TKO cells is elevated at the TSS compared to WT ESCs (Figure 6.16). This increase of efficiency becomes even more pronounced over time. Initially, *de novo* methylation shows no visible increase either at the TSS or the gene

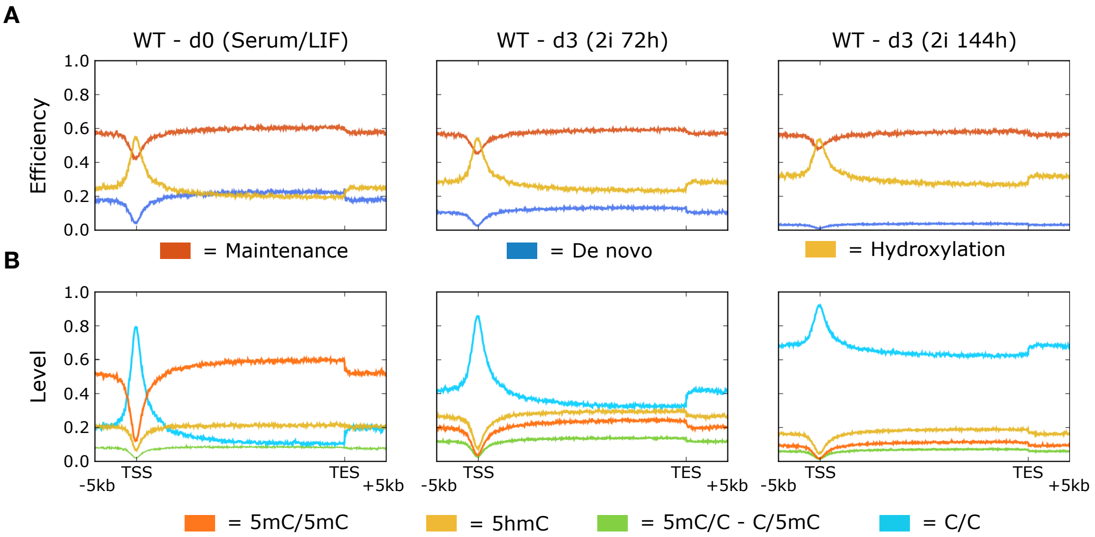


Figure 6.15: Average enzymatic efficiencies and CpG methylation pattern of WT ESCs. Average maintenance (red), *de novo* (blue) and hydroxylation (yellow) activity across genes during Serum-to-2i transition **(A)**. Average CpG (hydroxy-)methylation pattern across genes during Serum-to-2i shift **(B)**. Symmetric methylated CpG dyads (5mC/5mC, red); asymmetric CpG methylation (5mC/C or C/5mC, green), all combinations of hydroxylated dyads (5hmC/* or */5hmC, yellow), unmethylated CpG dyads (C/C, blue).

body. It is particularly compelling, though, that it seems to decrease much slower (if not remaining stable) over time than it does in WT data. At day6 of the WT *de novo* is zero, whereas in the Tet TKO cells there is a very significant amount of *de novo* activity at day7.

Note that this striking prediction of our model for remaining *de novo* activity in Tet TKO is independently[6] verified by the previously reported nonCpG methylation in the same cells (see Figure 6.13D).

Distinct profiles at highly and low expressed genes

Methylation at promoters and TSS is known to correlate with gene expression [31, 7, 44]. We investigated whether the enzyme efficiencies show similar relations. For our analysis, we used a previous published transcriptome of mESCs under Serum/LIF

[6] NonCpG data is not a part of the data given as input to the model.

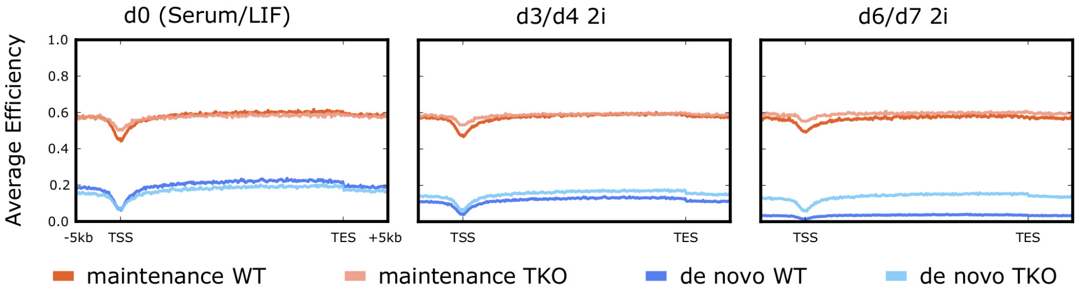


Figure 6.16: Efficiencies in WT and TKO ESCs. Comparison of maintenance and *de novo* methylation efficiencies in WT and TKO ESCs. red = maintenance WT, light red = maintenance TKO, blue = *de novo* WT, light blue = *de novo* TKO.

conditions [18]. Calculating the median of transcripts per million (TPM), we considered genes with a TPM above or equal to 0.065 as highly expressed and genes with a TPM below 0.065 as not/low expressed.

Profiles of highly expressed genes in Figure 6.17A,B match nicely those of the combined analysis (Figure 6.15) which is easily explained by the fact that the majority of the genes are expressed. However, not/low transcribed genes show a diverse pattern, particularly at the TSS. Low gene expression comes together with a higher maintenance and *de novo* activity, and an evidently reduced efficiency of the Tet enzymes. In addition, we capture for these genes a more pronounced *de novo* efficiency, also, across the gene body.

Performing the same split in Tet TKO cells (Figure 6.17C) we observe, similarly to the gene plots, an increase of maintenance efficiency at the TSS for expressed, but only a mild change for not/low expressed genes. However, not expressed genes show a reduced *de novo* activity across the gene body, when comparing Tet TKO with WT cells.

Protein binding sites and histone marks

The opposing behavior of methylation and hydroxylation activity is observed, as well, at protein binding sites and selected histone marks (Figure 6.18). In case of

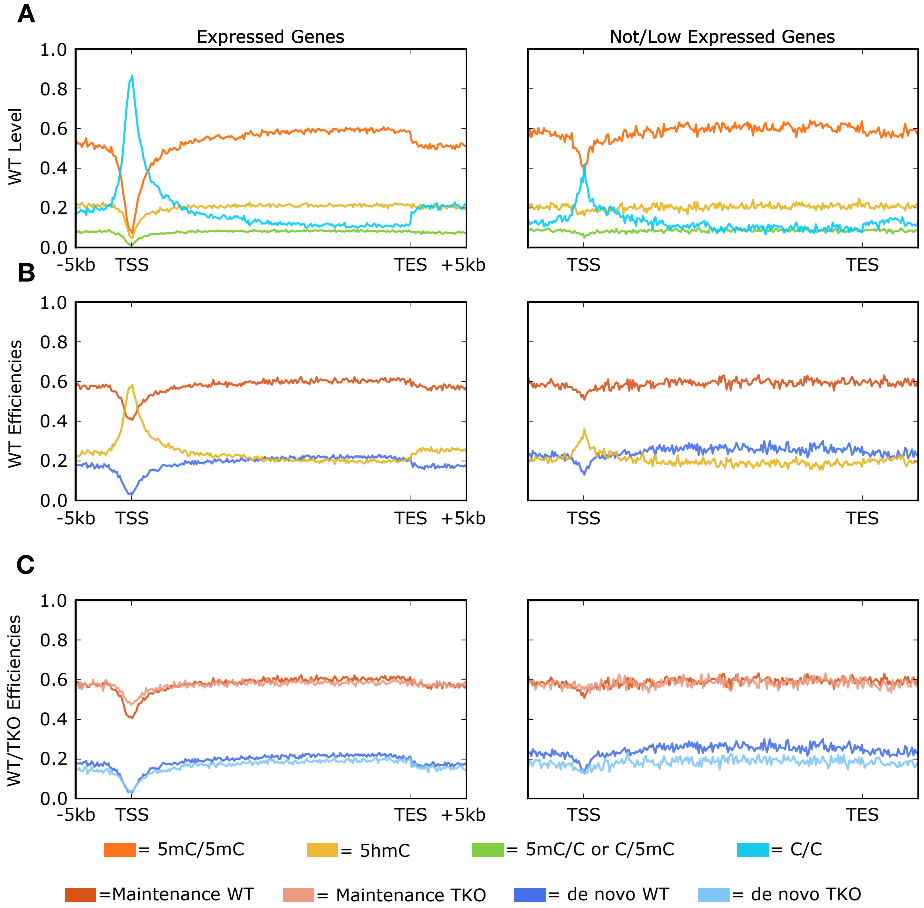


Figure 6.17: Modification levels across expressed and not/low expressed genes under Serum/LIF conditions; red = 5mC/5mC, yellow = 5hmC, green = 5mC/C or C/5hmC, blue = C/C **(A)**. Average efficiency profile across expressed and non/low expressed genes **(B)**. Comparison of WT and TKO cell efficiency across expressed and non/low expressed genes **(C)**. Red = maintenance WT, light red = maintenance TKO, blue = *de novo* WT, light blue = *de novo* TKO, yellow = hydroxylation.

H3K9me2, known to recruit Dnmt1 [94], we see a higher maintenance and *de novo* methylation efficiency together with a strongly reduced Tet activity. Similar profiles are also observed for H3K9me3 and H3K36me3.

In case of the open chromatin mark H3K4me3 the model reveals a high hydroxylation and reduced methylation activity (Figure 6.18). Likewise, we observe, as expected, a high hydroxylation accompanied by a reduction of maintenance and *de novo* methylation for the binding sites of Tet1, Sox2, Nanog and Oct4. Again, as in the TSS of expressed genes, *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.

Regarding the differences in case that the Tet enzymes are knocked out we note that across histone modifications and protein binding sites the methylation efficiencies show similar tendencies as across genes at day0. We see again that *de novo* activity remains, with the exception of H3K9me3, in most of the cases unaffected, while maintenance usually increases at the TSS. The most pronounced increase of maintenance methylation efficiency can be observed at the binding sites of Tet1 and the pluripotency factors Nanog, Sox2 and Oct4, which, among the selected loci, are the ones with the highest Tet activity in WT. In addition, we also notice a clear but smaller maintenance increase in regions which display H3K4me1 and H3K4me3 in WT ESCs (Figure 6.19).

Individual chromosomes and repetitive elements

Plotting the efficiencies and the modification levels for each of the 21 main chromosomes of the mESCs (Figure B.5a, B.6a) indicates only minor differences between them. The only exception is chromosome Y with a more pronounced maintenance and *de novo* activity and also an almost zero hydroxylation activity, which can prob-

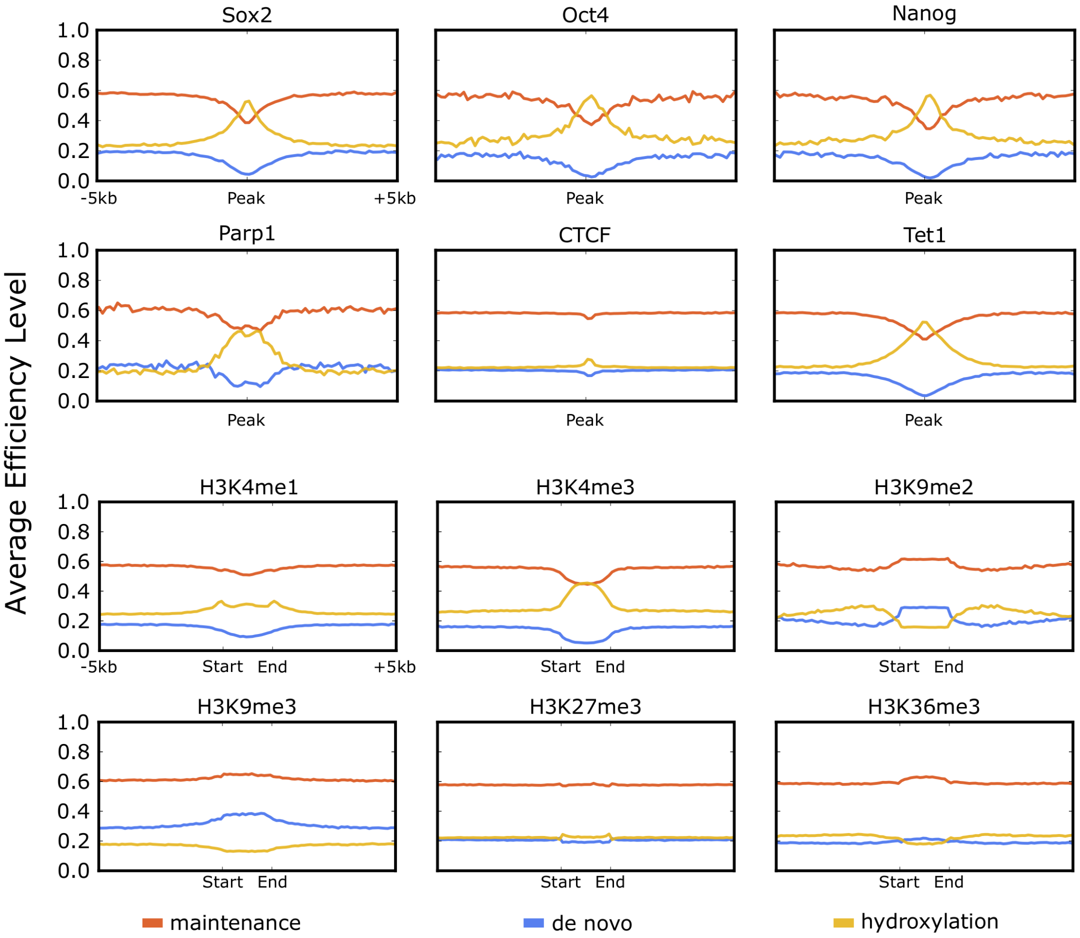


Figure 6.18: Enzymatic activity at TFBS and across histone marks. Average efficiency profiles for Dnmts and Tets at TFBS and across HMods in WT ESCs. red = maintenance methylation, blue = *de novo* methylation, yellow = hydroxylation.

ably be attributed to the smaller number of genes, and consequently promoters, that this chromosome contains. Nevertheless, in general, the average profiles of each chromosome seem to well coincide among the whole genome.

Since the majority of the mammalian genome is composed of repetitive elements (REs), we also examined whether a subset of REs would reflect the genome's average behavior. Figures B.8 to B.10 show, respectively, the efficiencies and the modification levels 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 efficiency and the level profiles of individual chromosomes, as well as the average gene

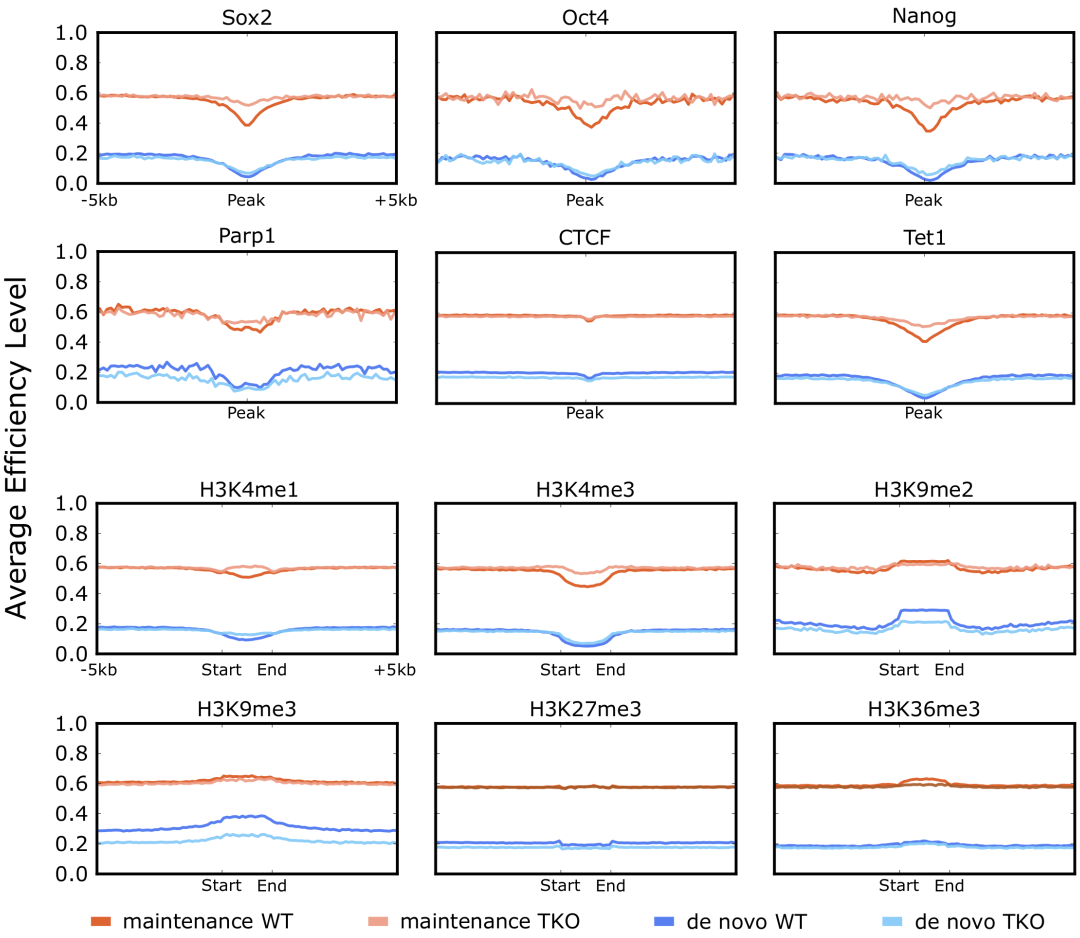


Figure 6.19: Enzymatic activity at TFBS and across histone marks. Comparison of maintenance and *de novo* methylation efficiencies in WT and Tet TKO ESCs. Red = maintenance WT, light red = maintenance Tet TKO, blue = *de novo* WT, light blue = *de novo* Tet TKO.

profiles (Figure 6.15, 6.16). Exceptions among REs are Intracisternal A particle and major satellites which exhibit a considerably 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. They seem, henceforth, to 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 coincides. In addition, *de novo* methylation appears again reduced for day0, but remains present even after continuous incubation in 2i media (Figure B.10).

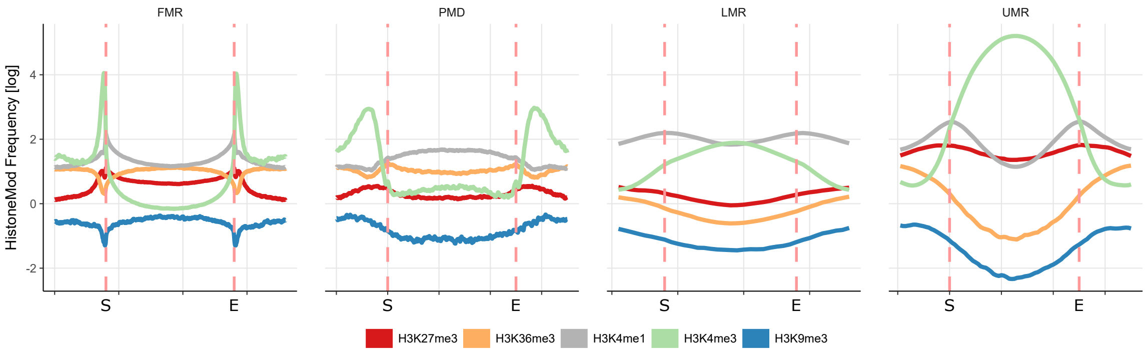


Figure 6.20: Profile of ENCODE histone modifications across HMRs, PMDs, LMRs and UMRs. Histone modifications are color coded; red dashed lines indicate start (S) and end (E) of a given segment.

Enzymatic 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 [56, 86]. We used MethylSeekR, a computational method published by Burger et al. 2013[13], to subset the genome into four distinct segment classes: Highly methylated regions (HMRs, alternatively FMRs), partially methylated domains (PMDs), low methylated regions (LMRs), as well as unmethylated regions (UMRs). For an optimal segmentation, we used a whole genome bisulfite sequencing data set of primed mESCs, published by Ficz et al. in 2013 [18] and subsequently compared the segmentation to RRHPoxBS data set. Plotting the available histone marks from ENCODE across the individual segment types reveals similar pattern as described before by [13], which gives us confidence that the segmentation was performed accurately.

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 ESC kept under Serum/LIF exhibit a hypermethylated phenotype (HMRs: 85.5%, PMDs: 12.6%, LMRs: 0.4%, UMRs: 1.5%) (Figure 6.22A, 6.22B). Assigning, next, the (hydroxy-)methylation level and the enzymatic efficiencies to each segment type, we observe the following. Consistently with the genome-wide

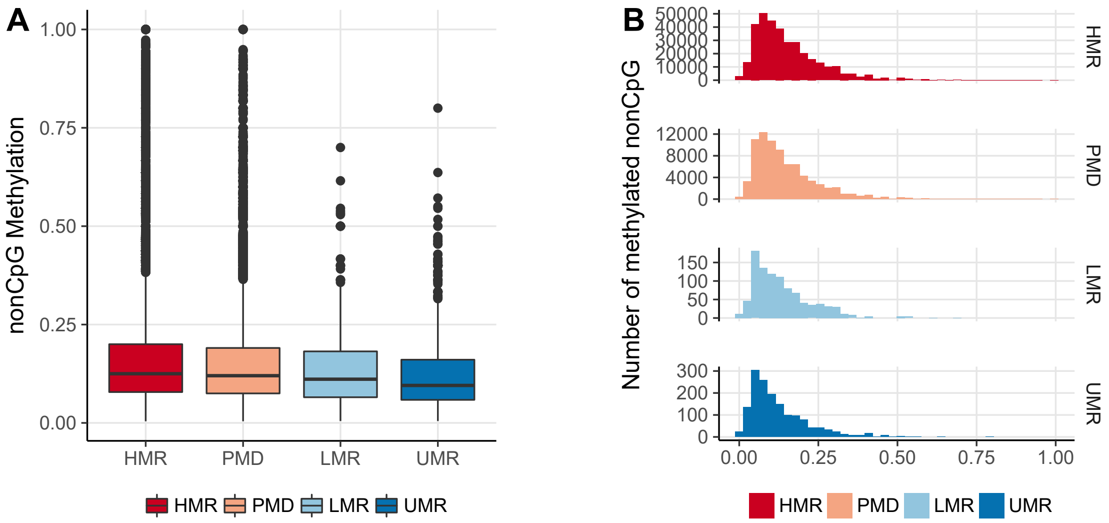


Figure 6.21: Methylation level and distribution of methylated nonCpGs in HMRs, PMDs LMRs and UMRs. Methylation level (A). Distribution of nonCpG methylation in HMRs, PMDs, LMRs and UMRs (B)

.

methylation data, we see high levels of 5mC at HMRs and PMDs, whereas LMRs and UMRs exhibit low methylation levels (Figure 6.22E). 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 6.22E, 6.22G).

Regarding nonCpG methylation, we observe that the majority of all nonCpG in our data set corresponds to HMRs and PMDs, whereas only a small fraction can be found in LMRs and UMRs (Figure 6.21). Again, this observation nicely matches our model prediction according to which HMRs and PMDs exhibit higher *de novo* methylation activity mainly caused by Dnmt3a/b.

Finally, 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. Since the truthful inheritance of DNA methylation pattern can only be ensured by correct maintenance activity, the timing of DNA replication might influence the efficiency of maintenance methylation. Therefore, we

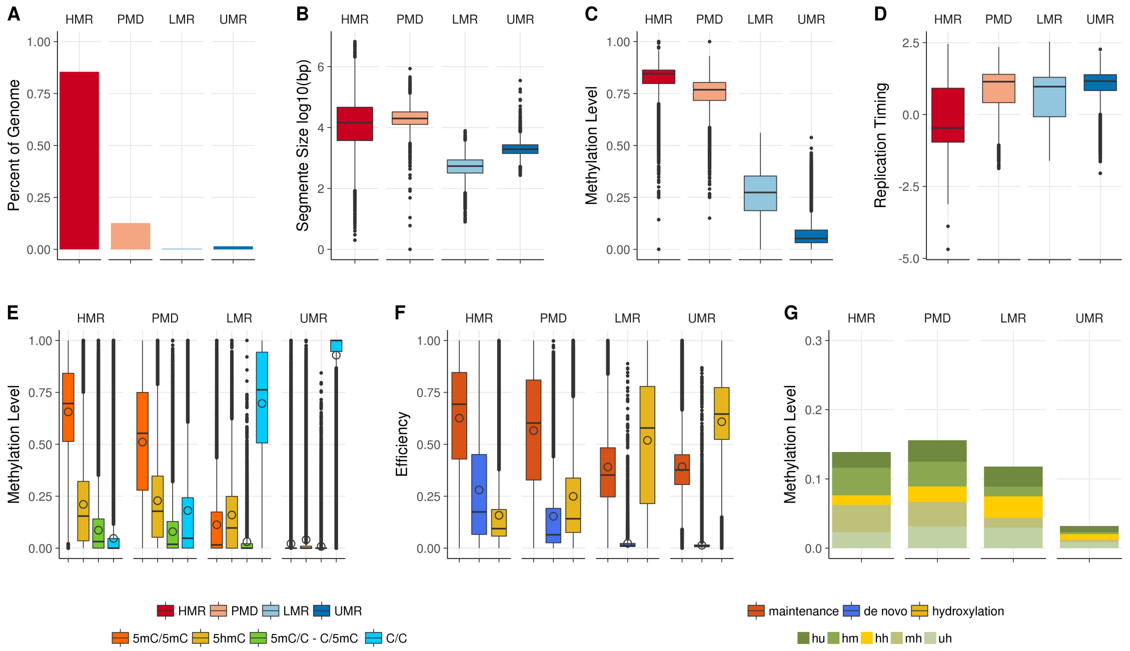


Figure 6.22: Outcome of the segmentation using MethylSeekR on mESCs under Serum/LIF conditions. Number of HMRs, PMDs, LMRs and UMRs after segmentation. **(A)**. Size distribution of the individual segments **(B)**. Methylation level of segments according to Ficz et al 2013. **(C)**. Replication timing based on the data from Hiratani et al. 2008 **(D)**. Methylation distribution based on RRHPoxBS **(E)**. maintenance (red), *de novo* (blue) and hydroxylation (yellow) efficiency **(F)**. 5hmC distribution in HyperD and HypoDs **(G)**.

compared the replication timing of the distinct segments using the replication information of three ESC lines published by Hiratani et al. [36]. In this context we observe, that HMRs tend to replicate later than PMDs, LMRs and UMRs (Figure 6.22D), which leads to the conclusion that later replication accompanies higher methylation.

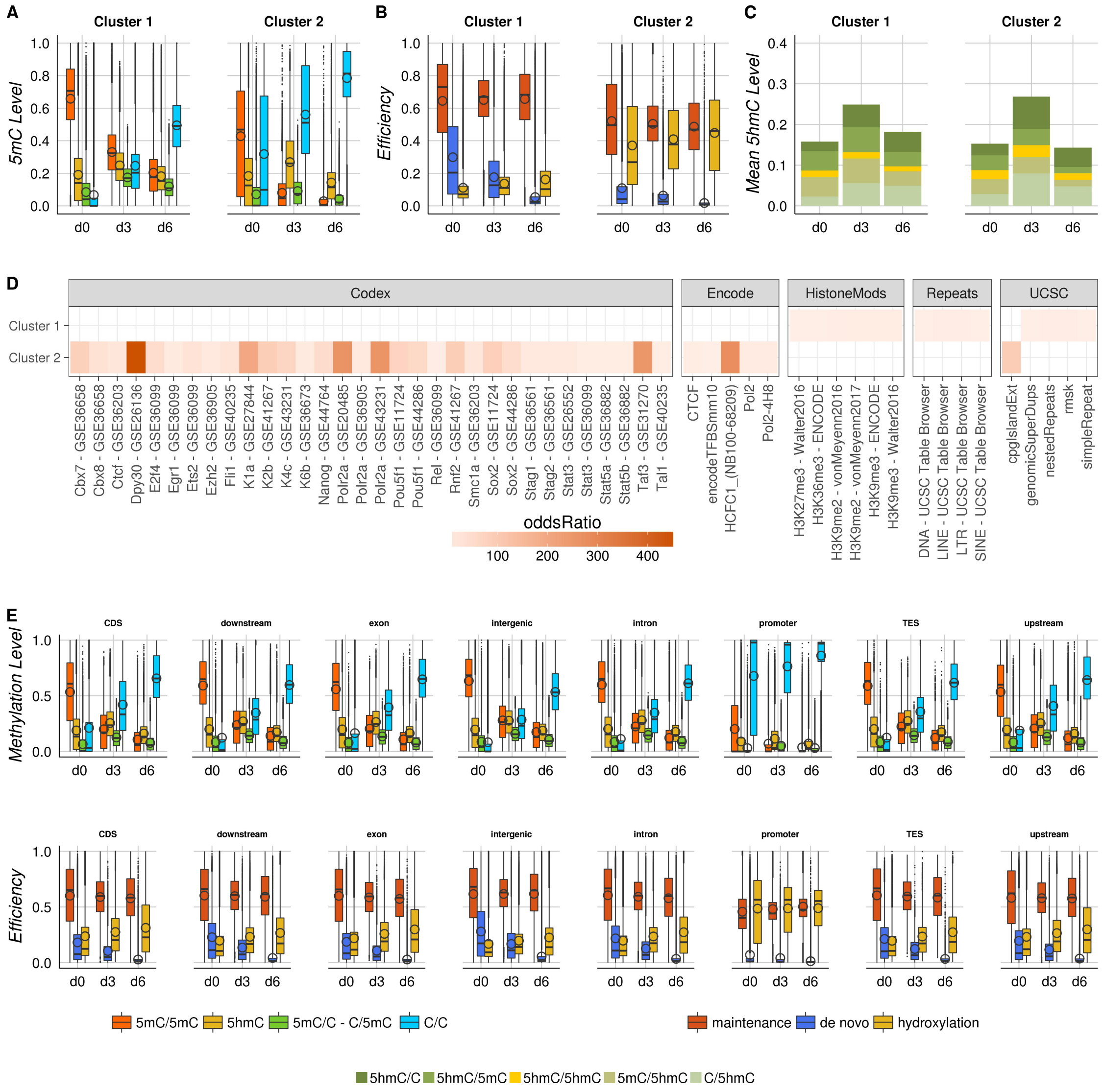


Figure 6.23: **M**ethylation profile of identified efficiency clusters **(A)**. Efficiency Profile of identified clusters **(B)**. Mean 5hmC level and distribution **(C)**. LOLA enrichment analysis of clustered CpGs **(D)**. Methylation and efficiency profile of annotated genomic features **(E)**.

6.3.2 Temporal analysis

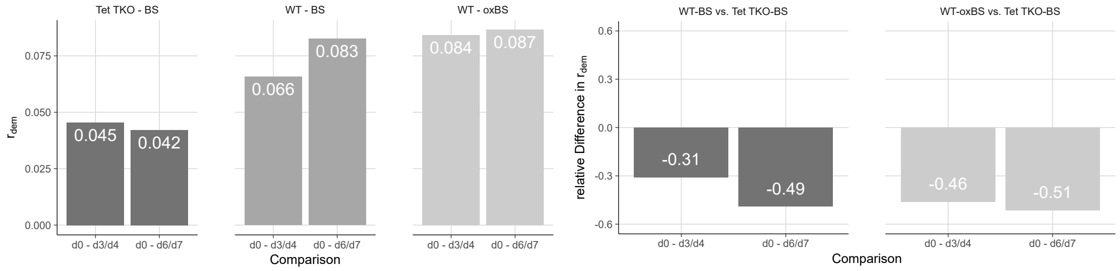
Demethylation kinetics

Previous studies indicate some disagreement about the importance of the role of Tet enzymes in the rate of the demethylation process, and whether the formation of hydroxylation is one of the major mechanisms that contribute to demethylation or not. In [97], for instance, the authors observe up a global to 10% increase of 5mC under the depletion of only Tet1 enzyme. In contrast, in [94] the authors claim is that Tet TKO cells exhibit almost the same demethylation kinetics as WT ES cells during their transition from Serum to 2i. The measurements of 5mC presence in the second study though have been done with the use of mass spectrometry and they do not come from sequencing.

In order to determine, here, the difference in the demethylation rate in the absence of Tets we calculated the increase of unmethylated cytosines for time points $t = \{3, 6\}$ in WT and for $t = \{4, 7\}$ in Tet TKO cells using the equation:

$$r_{\mathrm{dem}}(t) = \frac{\mathrm{TT}(t) - \mathrm{TT}(0)}{t}.$$

WT ES cells show an increase of unmethylated CpGs of more than 8.3% (day6 of BS experiment), whereas Tet TKO cells exhibit demethylation rates of 4.2% as shown in Figure 6.24a. In other words, we observe a vast decrease of the demethylation rate in the absence of Tet enzymes by around 50% (Fig. 6.24b). In fact, this decrease is even more pronounced, around 52%, if one considers, additionally, the hydroxylated cytosines as a non-methylated modification (day6 of oxBS experiment). Hence, the above data clearly demonstrates that there is a considerable contribution of Tet enzymes to the rate of DNA demethylation kinetics.



(a) (b)

Figure 6.24: (a) Demethylation rate in WT and Tet TKO cells (b) Relative difference in demethylation rate between WT and Tet TKO cells.

Two main profiles of enzymatic temporal activity

The profiles of expressed and non-expressed genes suggest that CpGs display distinct patterns of enzymatic efficiencies depending on their genomic location. To investigate this further, we clustered all individual CpGs of the genome based on their efficiencies and their temporal changes in order to identify distinct enzymatic kinetics during the Serum/2i transition. As we already saw, our approach always decides for two clusters (Figure 6.7).

In more detail, cluster 1 contains 855201 CpGs and it is characterized by high maintenance ($\approx 60\%$) and *de novo* activity ($\approx 30\%$) at day0, whereas the activity of Tet enzymes ($\approx 10\%$) is rather low (Figure 6.23B). At the same time we observe high methylation levels at day0 (Figure 6.23A). Over time, we observe a strong decrease in *de novo* methylation together with a nearly stable maintenance and an increasing hydroxylation efficiency. 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 6.23A).

Cluster 2 contains 702901 CpGs and it is mainly characterized by a high hydroxylation activity, which further increases over time (Figure 6.23B). Maintenance efficiency ($\approx 50\%$) is considerably lower compared to cluster 1 and appears to slightly decrease during the transition to 2i. Additionally, we observe in this cluster a very low *de novo* activity of Dnmts, which almost vanishes over time. 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 enough, despite the difference in the absolute hydroxylation efficiency in two clusters, their demethylation rates appear to be very similar (Figure 6.23A). In addition to the shared temporal increase of 5hmC from day0 to day3, we observe comparable 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 strand. Nevertheless, the distribution of 5hmC is distinct for each cluster. Whereas most CpGs in cluster 1 exhibit a 5hmC/5mC or 5mC/5hmC state, the majority of 5hmC in cluster 2 is present paired with unmethylated cytosine on the opposite strand (5hmC/C or C/5hmC).

Conduct of LOLA enrichment analysis on both clusters reveals an enrichment of transcription factor binding sites, euchromatic histone modifications and CpG islands for cluster 2 [82]. The list of transcription factor binding sites includes typical stem cell markers as Oct4, Nanog and Sox2. Cluster 1, on the other hand, discloses an enrichment in heterochromatic histone marks and repetitive elements (Figure 6.23D).

In Tet TKO cells both clusters show higher methylation levels than WT, with this being particularly pronounced in cluster 2, and retain a notably amount of 5mC even at day7 in 2i containing media (Figure 6.25A). Maintenance methylation efficiency seems to be unchanged for cluster 1 Tet TKO cells and it exhibits a noticeable increase for cluster 2. *De novo* methylation, on the other hand, stays rather stable over time and persists even at the latest time point for both clusters.

Separate genomic regions Finally, to map certain genomic regions with either cluster of enzymatic activity we grouped the CpGs based on their genomic context.

The revealed conserved methylation and efficiency patterns show that all examined genetic features, clearly, belong to cluster 1, apart from promoter regions which belong to cluster 2. Hence, for all features, but promoters, maintenance methylation appears to be again stable over time and *de novo* and hydroxylation efficiency exhibit the same tendency, with minor deviations (Figure 6.23E). On the contrary, high hydroxylation efficiency, moderate maintenance and only marginal *de novo* methylation is what we observe in promoters. Since promoter regions are usually located around the TSS, this is in agreement with the profile plots across genes, which unveiled similar dynamics at the TSS (Figure 6.15).

At last, in Tet TKO cells, we make two main observations: First that *de novo* activity is more uniformly distributed among all distinct genomic features, exhibiting, similarly to the gene plots, an almost stable activity over time, and second that a clear increase of maintenance methylation is present only in the promoters (Figure 6.25D).

6.4 Discussion

In our study, we investigated how Dnmts and Tets contribute to a stable methylome with unmethylated and methylated domains and furthermore examine how changes in the activity of individual enzymes shape new methylation patterns.

To address those questions, we get data from RRHPoxBS (see Section 2.2.2), a unique 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. Using our developed HMM analysis we are able to estimate the detailed distribution of (hydroxy-)methylation states and the activities of Dnmts and Tets for each individual CpG in the genome.

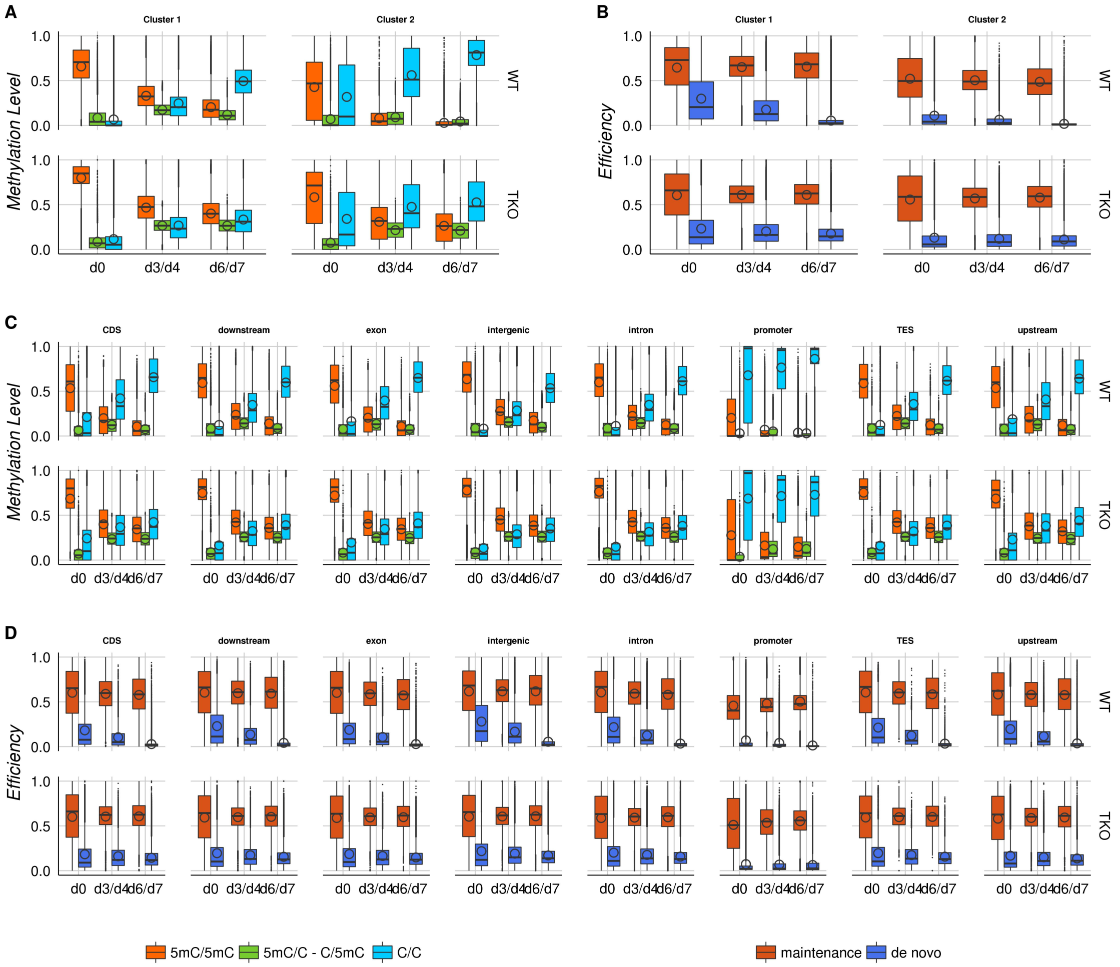


Figure 6.25: Comparison of clustered CpGs between WT and TKO ESCs. Methylation profile of clustered CpGs (**A**), Efficiency profile of clustered CpGs (**B**). Methylation profile of annotated genomic features **C**. Efficiency profile of annotated genomic features (**D**).

The measured CpG methylation levels of RRHPoxBS for Serum/LIF (65%) and 2i (20%) conditions are in line with previous described methylation levels from RRBS and WGBS [18, 94]. The observed reduction of nonCpG methylation after the incubation in 2i, is in agreement with the previous observed loss of Dnmt3a and 3b under naive conditions [18]. At last, 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 (hydroxy-)methylation levels.

6.4.1 Biological findings

In the next paragraphs we summarize the most important biological findings of our analysis.

Uniformly distributed hemi-methylation among the two strands

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, the evaluation of the double strand information obtained from RRHPoxBS does not reveal any strand specific distribution of hemimethylated CpGs in relation to transcription. Instead, we see that hemimethylation is equally distributed between both strands and follows the behavior of symmetric CpG methylation. This suggests that hemimethylation is more likely the side-effect of *de novo* methylation or active and passive demethylation events, rather than the result of a selective strand specific mechanism.

Two profiles of enzymatic temporal activity

The combination of Dnmt and Tet activity defines the methylation status of each CpG. The clustering of the enzymatic efficiencies of our model reveals that every CpG of the genome belongs to one of two main clusters regarding its enzymatic activity. In line with the results of individual loci analysis (Chapter 4) and previous findings of others [94], both clusters predict a genome-wide reduced (w.r.t. serum) but stable in 2i maintenance (≈ 0.6), a declining *de novo*, and a slightly increasing hydroxylation efficiency. The reduction of maintenance efficiency was recently related to the rearrangement of H3K9me2 under 2i conditions [94] and the later stable behavior can be explained if the reorganization of H3K9me2 is an early event and completed within the first 24h upon the transition to 2i. The progressive decrease of *de novo* methylation activity fits to a gradual degradation and transcriptional halt of Dnmt3a/b as described in [18] and finally the predicted increase of hydroxylation activity is congruent with the observed up-regulation of Tet enzymes in 2i medium [18, 29].

In cluster 1, containing the majority of the CpGs, we observe a pronounced maintenance and *de novo* activity and a relatively low hydroxylation, while for cluster 2 we see a significantly reduced maintenance, almost zero *de novo* and particularly high hydroxylation. As a result, CpGs of cluster 2 belong mainly to strongly unmethylated regions. An enrichment analysis clearly reveals the functional role of Tet enzymes. While cluster 1 corresponds to inactive epigenetic marks, cluster 2 is associated with euchromatic histone marks, transcription factor binding sites and CpG islands.

Opposed Tet and Dnmt activity

Spatial cross-correlation computation shows that a low methylation efficiency is usually accompanied by a high hydroxylation efficiency and vice versa, defining domains of low and high methylation levels, respectively. From the profiles of enzymatic activity across genes, in transcription binding factors, and in the histone modifications, as well as from our clustering, we verify that our model undoubtedly suggests that for a given CpG methylation and hydroxylation efficiencies are not exclusive, but they show an antagonistic behavior. As already mentioned, we observe high maintenance and compelling *de novo* efficiency at the majority of the genome. The activity of Tet enzymes, on the other hand, is highest at UMRs and LMRs such as promoters, TFBS (Sox2, Pou5f1) and especially TSS. Very recent studies based on chromatin immuno precipitation support our findings revealing binding of Dnmt3a/b at the gene body and HMRs, whereas Tet1 binding was observed across methylation valleys (LMRs and UMRs) [4, 27].

Local control of Tets - creation of stable 5hmC

We observe in general that 5hmC represents a fraction of 5mC for all genomic regions. In LMRs, however, it seems that a specific combination of methylation and hydroxylation efficiencies is sufficient to maintain a constantly high amount of 5hmC which exceeds that of 5mC. Since LMRs represent mostly enhancers [86], this finding is in accordance with previous observations which link enhancers' function to the presence of 5hmC [96, 87, 42]. Overall, the high hydroxylation efficiencies observed in our study, suggest a tight regulation of Tet enzymes. In this context, several mechanisms such as histone modifications, the expression of Tet isoforms, but also post-translational modifications, or the interaction with cofactors are possible.

Tets - guardians against methylation spreading

In the absence of Tets we observe a clear misregulation in both maintenance and *de novo* methylation efficiency. In particular, with the exception of day0, we see, compared to WT, 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 ESCs. A misregulation of Dnmt1 is further supported by the spatial autocorrelation of maintenance in Tet TKO cells.

The almost stable estimated *de novo* efficiency under 2i conditions in Tet TKO is surprising, considering the down regulation of Dnmt3a/3b in WT ES cells. However, the apparent presence of Dnmt3a/3b under 2i conditions in Tet TKO cells is strongly supported by the persistent nonCpG methylation in these cells.

Taken together, we hypothesize that Tet enzymes work against methylation and they enhance gene expression in three ways: (i) They guarantee an instant conversion of 5mC and act against its establishment during a cell replication mainly via passive but possibly also via active demethylation. (ii) They further inhibit the effectiveness of the maintenance machinery over regions such as enhancers and promoters that should remain unmethylated. (iii) They ensure an efficient down-regulation of the *de novo* enzymes, which can not be observed in their absence. As a result of all the above factors, we observe under the presence of Tet enzymes a vastly increased by around 50% demethylation rate, which demonstrates their great role in a proper demethylation process.

Chapter 7

Conclusions & Future Work

7.1 Conclusions

In this thesis we tried to unveil the driving forces behind the DNA demethylation in ESCs. Based on newly generated hairpin (double strand) BS and oxBS data at single CpG resolution, we built a novel hidden Markov model that is able to accurately estimate the 5hmC levels of a CpG dyad as well as the activities of the involved enzymes over time.

Our model strongly suggests a passive demethylation mechanism as the main driving force of losing methylation in mESCs. Our results show that the main cause of demethylation in ESCs is a combination of maintenance methylation impairment and the presence of 5hmC modification which, as the model reveals, is not recognized by the maintenance enzymes. This mechanism seems to be, consistently, further enhanced by a constantly decreasing *de novo* methylation and an increasing hydroxylation activity over time. A formally developed hybrid generalization of the core-model can, in case of available data from additional sequencing experiments, estimate also the levels of further oxidative cytosine forms as 5fC and 5CaC and, in this way, is able to also identify active demethylation.

To facilitate the usefulness of our method we implemented the above stochastic models in an easy-to-use software tool named H(O)TA that gives biologists the option to upload their own data and get estimations for (hydroxy-)methylation levels and more importantly quantitative information about the enzymatic activity, which otherwise can not be quantified from in vivo experiments. The tool's extension H(O)TA v.2, which is currently available as a beta version, provides the user the additional possibility to run self-defined epigenetic models.

Developing the appropriate computational methods to cope with the small coverage, we have been able to apply our model on a genome-wide scale for both WT and Tet KO ESCs data. Implementing a sophisticated clustering approach, we showed that the whole genome can be separated into only two large clusters of different enzymatic behavior. Protein enrichment analysis revealed that the two different clusters correspond impressively well to distinct genomic regions. The first cluster includes mainly inactive epigenetic marks as repetitive elements, whereas the second one is strongly associated with active marks such as transcription factors, enhancers and promoters connected with the ESCs phenotype. Based on a comparison of the model genome-wide output in WT and Tet TKO cells and several layers of analysis, we conclude that the seemingly critical contribution of Tets in demethylation and accordingly in gene expression occurs in three distinct ways: First by acting against methylation retainment via non-recognition by the Dnmts, second, by the further inhibition of the maintenance machinery activity in regions that should remain unmethylated, and third, by ensuring an efficient down-regulation of the *de novo* enzymes. As a result of these contributions, the total rate of the demethylation process in mESCs is twice as fast in case of the presence of Tet enzymes compared to their absence.

7.2 Future Work

Possible research directions that can be based on the models, the subsequent analysis, and the corresponding software tools established in this thesis, are the following:

The construction of models that consider more than one CpG dyad will offer the chance to investigate more closely the spatial correlations and the interactions of certain enzymes. In this direction it would be of interest to extend works as [8, 60] in order to investigate the neighborhood interdependencies, the in vivo processivity, and the association-dissociation properties of Dnmts and Tets. This could be easily done by setting up multi-CpG models in H(O)TA and using also the same locus-specific, or genome-wide data that has been used here. In addition, the combination of the above models with data from knockout experiments would further help us towards unveiling the individual role of each of the Tets, or clarifying the differences between Dnmt3a and Dnmt3b.

The upgrade of H(O)TA into a complete tool of general use, might be of help to the epigenetics community also in other directions. For instance, incorporating into the tool the option to get data from alternative sequencing experiments such as TAB-seq, fCAB-seq, CAB-seq, or redBS-seq, would offer the possibility to distinguish between all oxidative cytosine modifications. An application of the hybrid extension of the core-model in cell types of pure active demethylation like monocyte-macrophages, or brain cells would also be possible, given the availability of such data. Meaningful research questions in these cell types would be to infer the period of the active demethylation cycle, and to investigate whether some modifications are more stable than others. In this fashion it would also be worth exploring whether distinct loci of the same cell type can have an active demethylation cycle of different length.

As the holy grail of the epigenetics field is to find the relation between DNA methylation and gene expression, a very relevant research problem is the identification of the role of 5hmC and the further oxidative forms 5fC, 5CaC in gene expression. In this context, similarly to [85], would be essential clarifying how certain regions, seemingly as LMRs in ESCs, can retain stable low-methylation patterns in spite of repeated cell replications. At last, it would also be very interesting to extend the methods developed in [44] to get as input either the genome-wide (hydroxy-)methylation levels, or the enzymatic activities estimated by our model. Such a method combined ideally with histone modifications and transcription factor binding sites temporal data could potentially improve to a very great extent our understanding on the synergy between the various molecular players and gene expression.

Appendix A

Individual Genomic Loci Analysis

A.1 HPoxBS Protocol Details

500 ng of mESC DNA was cleaved with 10 units of restriction enzymes for 5h in a 30 µl 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.

A.1.1 Primer- and reference sequences

Table A.1 shows the sequence of the nine different hairpin linkers used to covalent link both DNA strands. We included unmodified cytosine C, 5mC (M) and 5hmC (H) into the hairpin linker to follow the conversion of these modifications during BS and oxBS treatment. By mapping the sequencing information to these reference sequences we determined the position of each (modified or not) cytosine and by this we were able to calculate its possible conversion rates (and therefore the conversion errors) for each time point and each genomic region as follows. 5hmC, for instance, should be converted after oxBS treatment to 5fU and will be seen as T after sequencing. Hence, to get the right conversion rate of 5hmC during oxBS treatment, we check for each sequenced hairpin molecule the state of the 5hmC position (either C or T) and we divide then the number of T's we sampled at this position by the total number of measurements (coverage). The conversion rates for C and 5mC were calculated in the same fashion. We followed the above procedure for all analyzed loci of Chapter 4 besides Snrpn[1]. In Table A.2 we give the primer sequences and in Table A.3 the corresponding reference sequence for each sequenced region.

[1] For Snprn we had to use a hairpin linker without 5mC or 5hmC and therefore could not calculate the sample specific conversion error. Instead we applied the mean errors of all other analyzed loci.

Table A.1: Sequences of the hairpin linker for the analyzed loci; 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$M$CCATDDDDDDDDDATGGG$H$CC
L1mdA-HP	*Pho*-TNAGGG$M$CCATDDDDDDDDDATGGG$H$CC
L1mdT-HP	*Pho*-CCGGAGGG$M$CCATDDDDDDDDDATGGG$H$CCT
mSat-HP	*Pho*-GNCGGG$M$CCATDDDDDDDDDATGGG$H$CC
Afp-HP	*Pho*-CGGGG$M$CCATDDDDDDDDDATGGG$H$CC
Ttc25-HP	*Pho*-CGGGG$M$CCATDDDDDDDDDATGGG$H$CC
Zim3-HP	*Pho*-CGGGG$M$CCATDDDDDDDDDATGGG$H$CC
Snrpn-HP	*Pho*-GGGCCTADDDDDDDDDTAGGCCCCATG

Table A.2: Primer for amplification of the analyzed loci 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
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

A.2 Input Data

In Table A.4, A.5 we show the data for the DNA loci L1mdA, L1mdT, IAP, mSat, 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. 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

Table A.3: Reference Sequences used for 5mC and 5hmC analysis; M = 5mC, H = 5hmC

IAP	TGTCACTCCCTGATTGGCTGCAGCCCATCGGCCGAGTTGACGTCACGGGGAAGGCAGAGCACATGGAGTAGAGAACCACCCTC GGCATATGCGCAGATTATTTGTTTACCAC**TNAGGGMCCATDDDDDDDDATGGGHCC**TAAGTGGTAAACAAATAATCTGCGCAT ATGCCGAGGGTGGTTCTCTACTCCATGTGCTCTGCCTTCCCCGTGACGTCAACTCGGCCGATGGGCTGCAGCCAATCAGGGAG TGACA
L1mdA	TCCAATCGCGCGGAACTTGAGACTGCGGTACATAGGGAAGCAGGCTACCCGGGCCTGATCTGGGGCACAAGTCCCTTCCGCTC GACTCGAGACTCGAGCCCCGGGCTACCTTGCCAGCAGAGTCTTGCCCAACACCCGCAAGGGCCCACACGGGACTCCCCACGGG ACCC**TNAGGGMCCATDDDDDDDDATGGGHCC**TNAGGGTCCCGTGGGGAGTCCCGTGTGGGCCCTTGCGGGTGTTGGGCAAGAC TCTGCTGGCAAGGTAGCCCGGGGCTCGAGTCTCGAGTCGAGCGGAAGGGACTTGTGCCCCAGATCAGGCCCGGGTAGCCTGCT TCCCTATGTACCGCAGTCTCAAGTTCCGCGCGATTGGATTGGGGCAGGCACTGTGATCCACTC
L1mdT	CCCGGGACCAAGATGGCGACCGCTGCTGCTGTGGCTTAGGCCGCCTCCCCAGCCGGGTGGGCACCTGT CCT**CCGGAGGGMCCATDDDDDDDDATGGGHCCT**CCGGAGGACAGGTGCCCACCCGGCTGGGGAGGCGG CCTAAGCCACAGCAGCAGCGGTCGCCATCTTGGTCCCGGG
mSat	CGCCCGAGACAAGGTGATTCTAGTTATTATAATGGACAGCGTAGACAAAAGAATGTTTATAATAACAT ACCCAGTAATGGTCAGCACAGGAGAGGTGAAATTTATAATGGCATGACTCGGTTG**GWCGGGMCCATDD** **DDDDDDATGGGHCC**GWTTCAACCGAGTCATGCCATTATAAATTTCACCTCTCCTGTGCTGACCATTAC TGGGTATGTTATTATAAACATTCTTTTGTCTACGCTGTCCATTATAATAACTAGAATCACCTTGTCTC GGGCG
Afp	TTTTGTTATAGGAAAATAGTTTTTAAGTTACAAAGCATCTTACCTATCCCAAACTCATTTTCGTGCAA TGCTTTGGACGCAGCGAAATGTAGCAGGAGGATGAGGGAAGCGGGTGTGATCCACTTCATGGCTGCTG GTTCCTTCACCGCAGGCAGTGCTGGAAGTGGGATGTTT**CGGGGMCCATDDDDDDDDATGGGHCC**CGAA ACATCCCACTTCCAGCACTGCCTGCGGTGAAGGAACCAGCAGCCATGAAGTGGATCACACCCGCTTCC CTCATCCTCCTGCTACATTTCGCTGCGTCCAAAGCATTGCACGAAAATGAGTTTGGGATAGGTAAGAT GtTTTGTGATTT
Ttc25	CCAGTAGATCCTCAGCTGGGGGCAGGGATCTATTCCATCACTCCCCTTCCGTGTCGGGATTTCGTGCA GCTCAGACGGGTCCAAGTCTTACACAAGCTGTCCTAACTGCTGTGCGTTTATATAACAACTACCCGGT TGTGTTTAGAAAACACTGTTTT**CGGGGMCCATDDDDDDDDATGGGHCC**CGAAAACAGTGTTTTCTAAA CACAACCGGGTAGTTGTTATATAAACGCACAGCAGTTAGGACAGCTTGTGTAAGACTTGGACCCGTCT GAGCTGCACGAAATCCCGACACGGAAGGGGAGTGATGGAATAGATCCCTGCCCC
Zim3	CCCGGCCACCATAGTCGGATTATCCGTGGGCGGGGTGAGATGGACGGAGCGCCTTGCAGACCTCAGGA AAACCTCCCCACGCCTGTCCGGCCTTGGCTTGGTGACAGGGAAACTGGCTGGACT**CGGGGMCCATDDD** **DDDDDATGGGHCC**CGAGTCCAGCCAGTTTCCCTGTCACCAAGCCAAGGCCGGACAGGCGTGGGGAGGT TTTCCTGAGGTCTGCAAGGCGCTCCGTCCATCTCACCCCGCCCACGGATAATCCGACTATGGTGGCCG GGCAAGGACCACAC
Snrpn	AGAATTTACAAGTTTAGTTGATTTTTTTCGCTCCATTGCGTTGCAAATCACTCCTCAGAACCAAGCGT CTGGCATCTCCGGCTCCCTCTCCTCTCTGCGCTAGTCTTGCCGCAATGGCTCAGGTTTGTCGCGCGGC TCCCTACGCATG**GGGCCTADDDDDDDDTAGGCCCCATG**CGTAGGGAGCCGCGCGACAAACCTGAGCCA TTGCGGCAAGACTAGCGCAGAGAGGAGAGGGAGCCGGAGATGCCAGACGCTTGGTTCTGAGGAGTGAT TTGCAACGCAATGGAGCGAGGAAGGTCAGCTGGGCTTGTGGATTCTAGTAGTGAAAGTGTATTTTATT TGATTA

Table A.4: Input BS and oxBS data and conversion errors (repetitive elements)

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

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

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

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

of the corresponding locus that have been observed in each of the four observable states (TT, TC, CT and CC) for every day of measurerement.

A.3 Single CpGs' Results

In Figure A.1, A.2 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

Table A.5: Input BS and oxBS data and conversion errors (single copy genes)

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

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

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

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

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.

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 and in IAP 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.

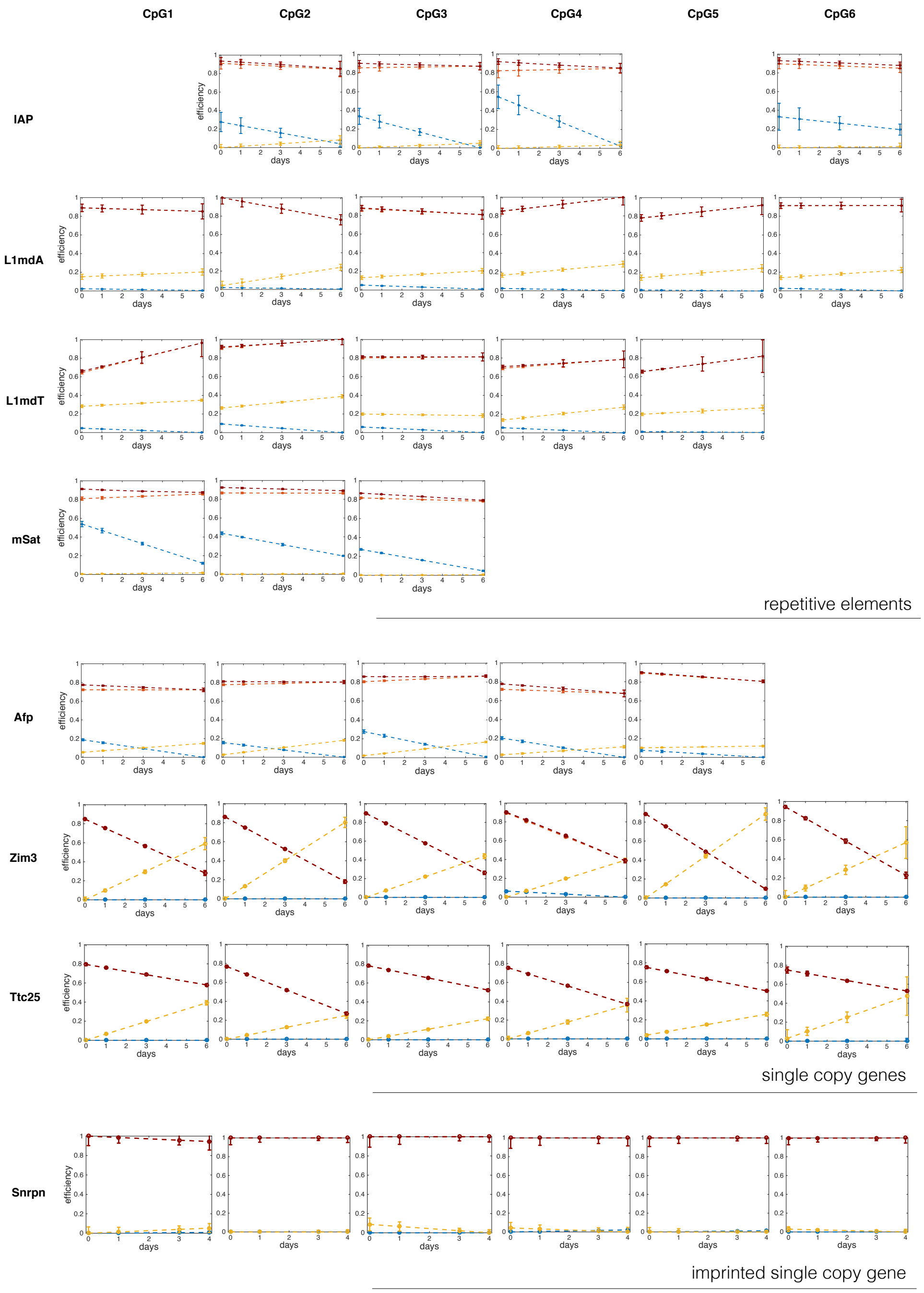


Figure A.1: Estimated efficiencies and standard deviations for each single CpG dyad of loci IAP, L1mdA, L1mdT, mSat and the single copy genes Afp, Ttc25, Zim3 and Snrpn over time. Maintenance (red), *de novo* (blue), hydroxylation (yellow) and total efficiency on a hemimethylated CpG (dark red). In the 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, (8) CpGs, analyzed in L1mdA, (Zim3).

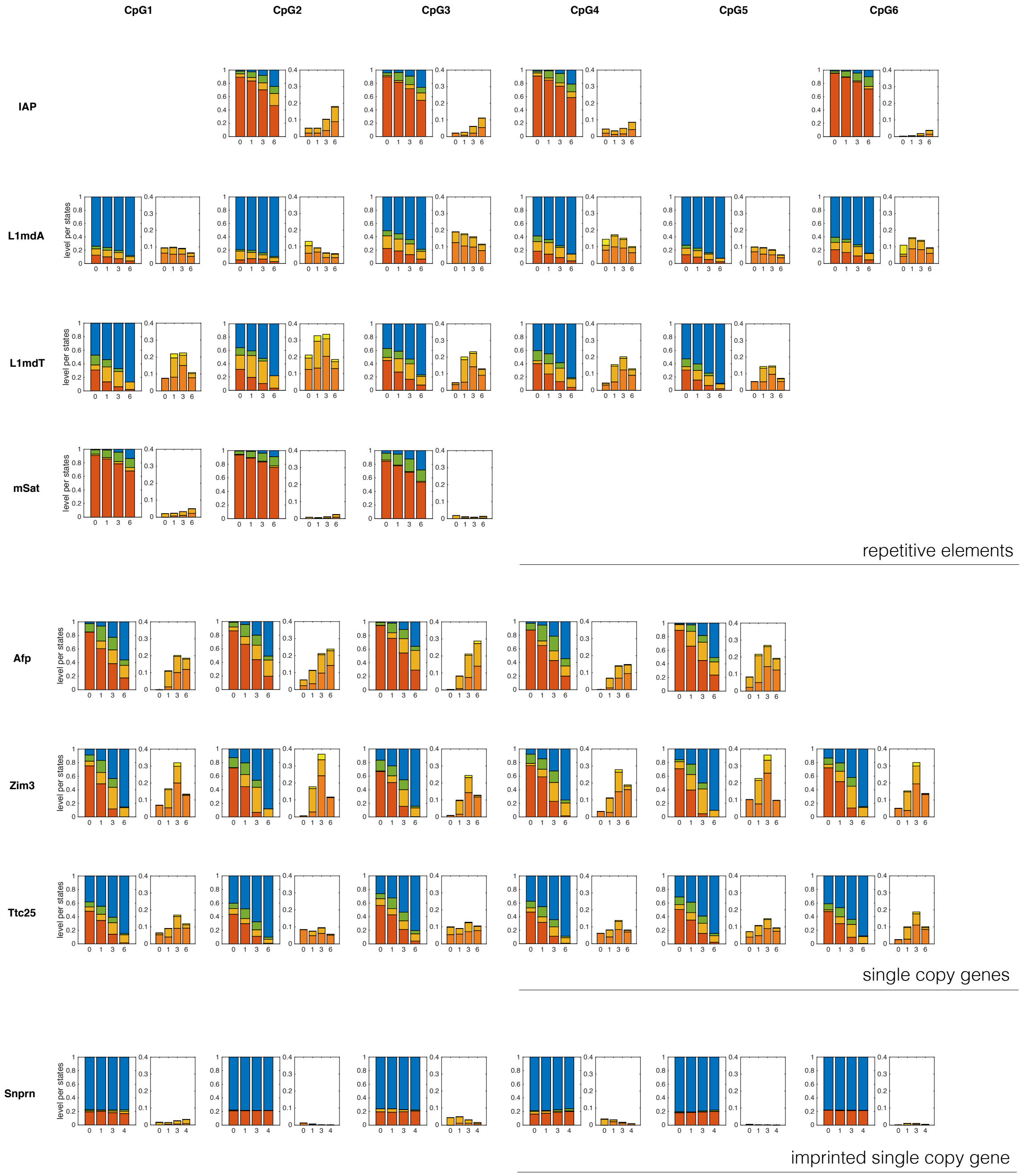


Figure A.2: Prediction of the (hydroxy-)methylation levels for each single CpG dyad of loci IAP, L1mdA, L1mdT, mSat and the single copy genes Afp, Ttc25, Zim3 and Snrpn over time. The left diagram depicts the amount of fully methylated (mm) sites in red color, hemimethylated (um and mu) sites in green color, and unmethylated (uu) sites in blue color. 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.

Appendix B

Whole Genome Analysis

B.1 RRHPoxBS Protocol Details

1.2μg DNA is divided equally into three 0.5ml reaction tubes and digested in a 20μl reaction using 20U of HaeIII (NEB), AluI (NEB) and HpyCH4V (NEB), respectively. The reactions are incubated overnight for a minimum of 12h at 37°C. Restriction enzymes are inactivated at 80°C for 30min. The reactions are pooled and subjected to a ligation step. During this process, hairpin linker and sequencing adapter are introduced to the opposed 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°C. 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.

Table B.1: Conversion rates of cytosine variants included in the TrueMethyl Spike in 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

B.1.1 Read mapping and methylation calling

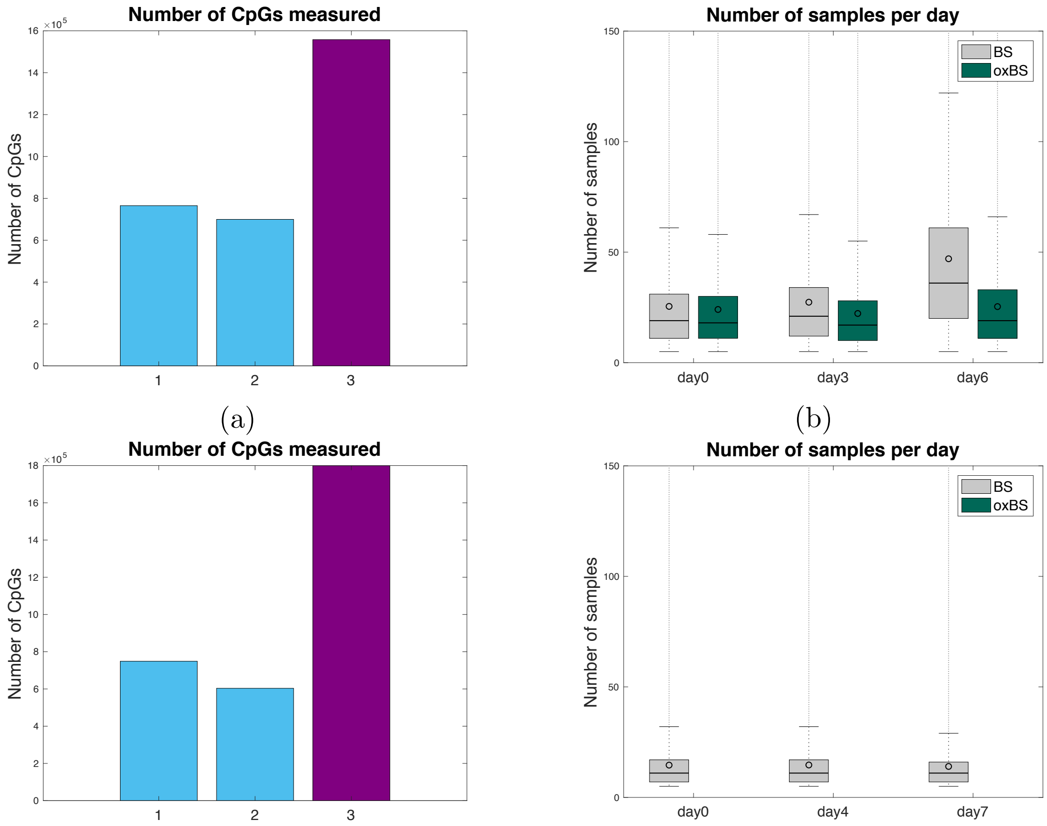
The sequences were aligned as suggested by Porter et al. [72]. In detail; reads were trimmed for adapter, hairpin-linker and 3 quality (Q≥20) with TrimGalore! [2] and cutadapt [63]. Trimmed read pairs were aligned with the Smith-Waterman algorithm allowing for bisulfite induced mismatches. The two bisulfite converted strands were used to deduce the original genomic sequence. Mismatches other than G-to-A and T-to-C were replaced with N. The resulting sequences were aligned to the mouse genome (mm10) with GEM-mapper (beta build 1.376) [62], after which the methylation information was reintroduced with a custom pileup function based on HTSJDK [54] and ratios for the four methylation states were calculated for each cytosine. The pipeline was implemented with the Ruffus pipeline framework [24].

B.1.2 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.

Table B.2: Conversion rates of cytosine variants included in the TrueMethyl Spike in 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



(a)

(b)

(c)

(d)

Figure B.1: 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.

B.2 Implementation Details

The code of our implementation is available in `https://github.com/kyriakopou/hydroxyGit`.

All functions that form the computational core for both the individual loci (packed into H(O)TA) and the genome-wide implementation are located in the `hydroxyGit` directory, while the functions related to processing of the input genomic data, the Bayesian inference method, and the computational approaches used to analyze the model's output such as the clustering and the spatial correlations are located in the subdirectory `hydroxyGit/genomeWide`.

To run the whole genome data we installed MATLAB Parallel Computing Toolbox in the DEEP cluster located in Max Planck Institute for Informatics (MPII) consisting of 32 machines each having 16 double threaded CPUs. In order to avoid memory leakage we handled the 3,022,903 CpGs of the WT cells and the 3,151,985 of the Tet TKO cells (Figure B.1) as follows. We assigned each of the chromosomes to a certain machine, by splitting its total CpGs in jobs of 500 consecutive CpGs. We run in parallel the 500 CpGs of a job in the 32 (16 x 2) threads of the machine and sequentially one job after the other.

An interested user that has access to the DEEP cluster can submit to the cluster queue the list of chromosomes he would like to run by typing in the command line

```
$./submitChromosomes.sh .
```

Listing B.1: submitChromosomes.sh

```
#!/bin/bash
#give the list of chromosomes to run
chrList=( $(seq 1 10))
for i in in ${chrList[*]}
do
```

```
qsub -v chr=$i -N chr_$i runChromosome.csh
done
exit 0
```

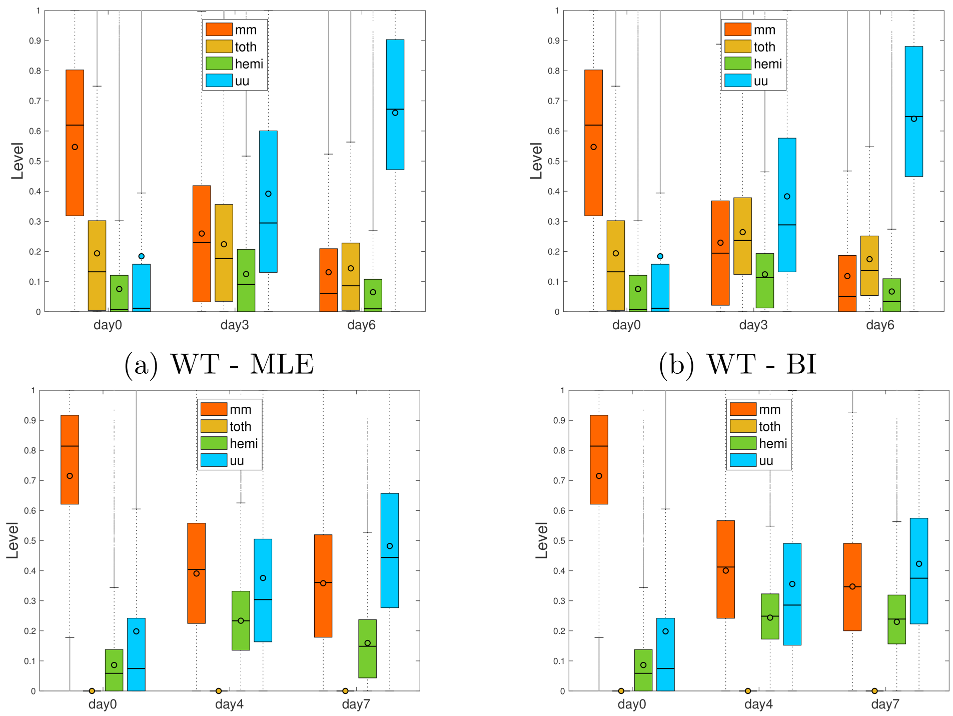
Listing B.2: runChromosome.csh

```
#!/bin/csh
#$ -cwd -V
#$ -l mem_free=50G,h_vmem=80G
#$ -m e -M <Email Address>
matlab -nodesktop -nosplash < initOther.m > matOutFiles/
    matChr_$chr.log
```

Listing B.3: initChr.m

```
chrToRun=getenv('chr');
chrToRun=str2num(chrToRun)
cd /code/MATLAB/HydroxyMethylation/;
addpath(genpath(pwd));
dataPath='/code/MATLAB/HydroxyMethylation/genomeWide/RRHPBS.data';
cd genomeWide;
runDataGW(dataPath, chrToRun);
exit;
```

Then each of the chromosomes, e.g., here 1 . . . 10, included in the array ‘chrList’ of the bash script submitChromosomes.sh (Listing B.1) is sent to a different machine of the cluster. Optionally, the user can choose to receive a notification mail once the analysis of a chromosome is done by inserting his email address in the corresponding field of runChromosome.csh (Listing B.2). The analysis is started by the call to the Matlab script initChr.m (Listing B.3), where the chosen chromosome is given as input to Matlab function runDataGW().



(a) WT - MLE

(b) WT - BI

(c) Tet TKO - MLE

(d) Tet TKO - BI

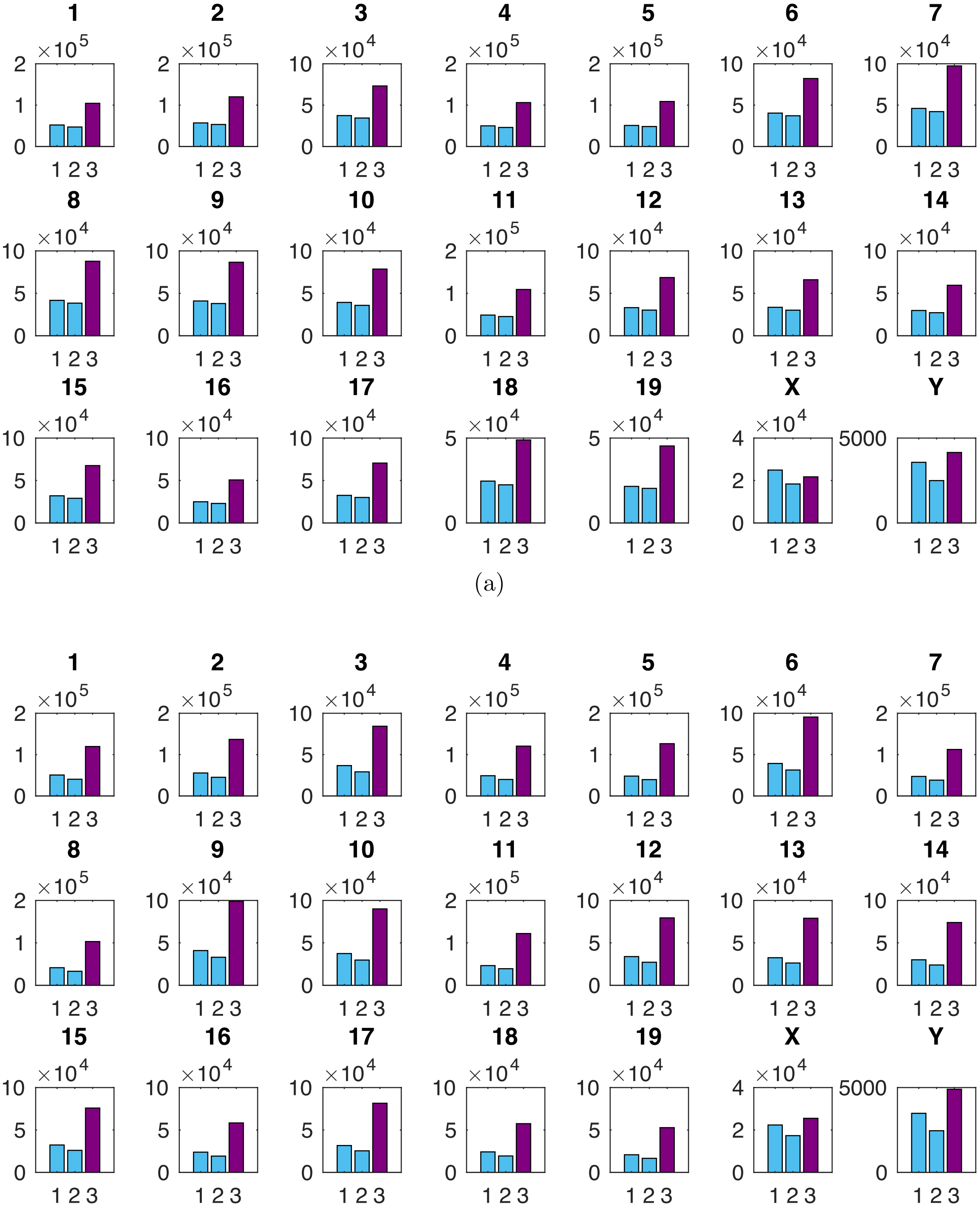
Figure B.2: Bar plots for the hidden states levels for all CpGs in the genome estimating the parameters with MLE (a), (c) and BI (b), (d). Red = symmetric methylated CpG (mm - 5mC/5mC), yellow = 5hmC in all possible combinations (toth - 5hmC/C, C/5hmC, 5hmC/5mC, 5mC/5hmC, 5hmC/5hmC), green = hemi methylated CpGs (hemi - 5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

B.3 BI vs MLE

In Figure B.2 we plot the different prediction of hidden states' probabilities between the ML and BI methods for both WT and TET TKO cells.

B.4 ESCs Chromosomes' Results

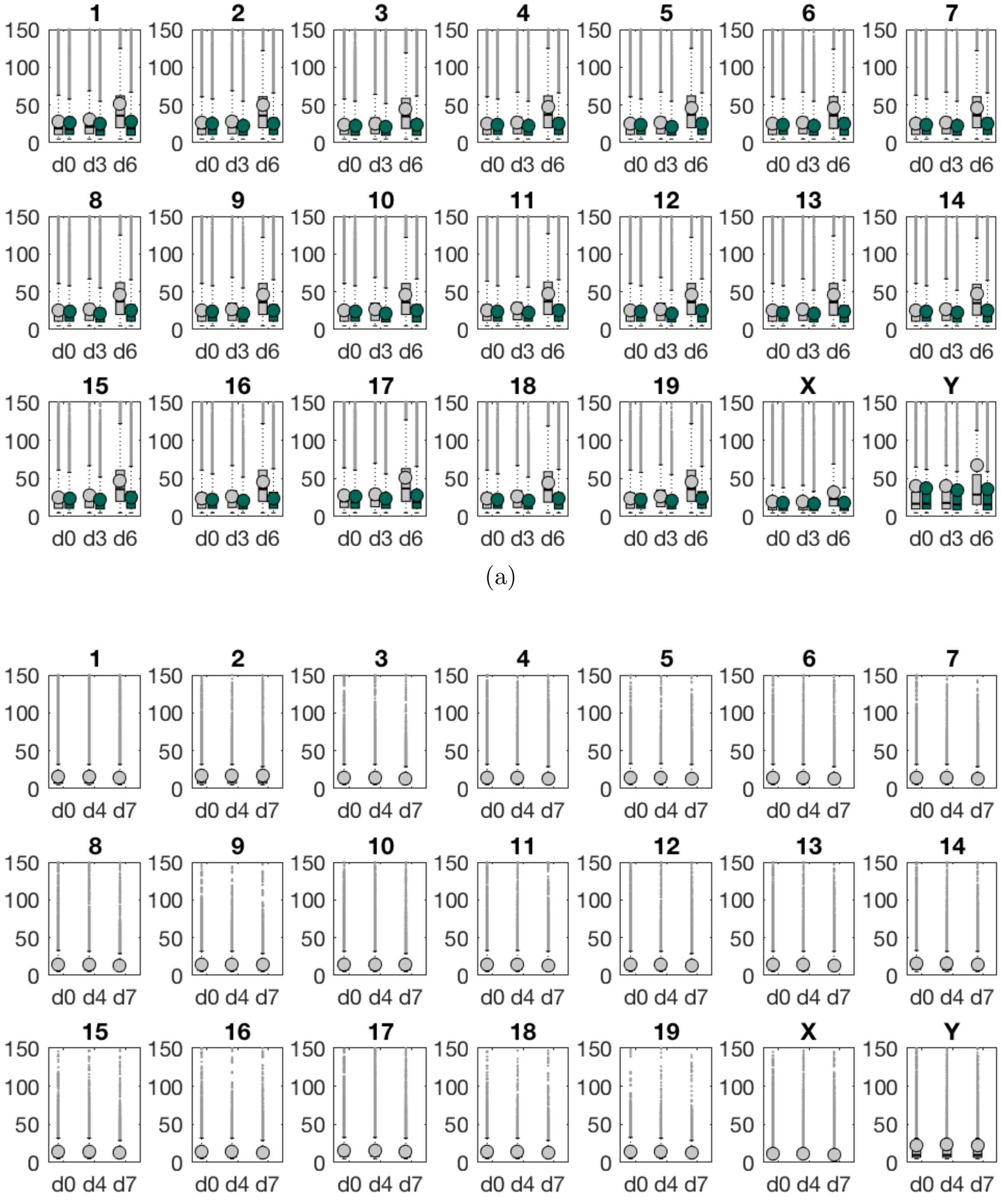
In this Section we provide the input data information plots as well as the output of our model for each of the 21 main chromosomes of the ESCs. In Figure B.3a, B.3b we plot the number of CpGs for each chromosome with one, two or three observation days in WT and Tet TKO cells, respectively. We plot the average number of samples (depth sequencing) for each chromosome in WT (Figure B.4a) and Tet TKO (Figure B.4b). In Figure B.5a, B.5b we show the efficiencies over time computed by BI and in Figure B.6a, B.6b we report the prediction of the model for the hidden states probabilities in each chromosome in WT and Tet TKO cells.



(a)

(b)

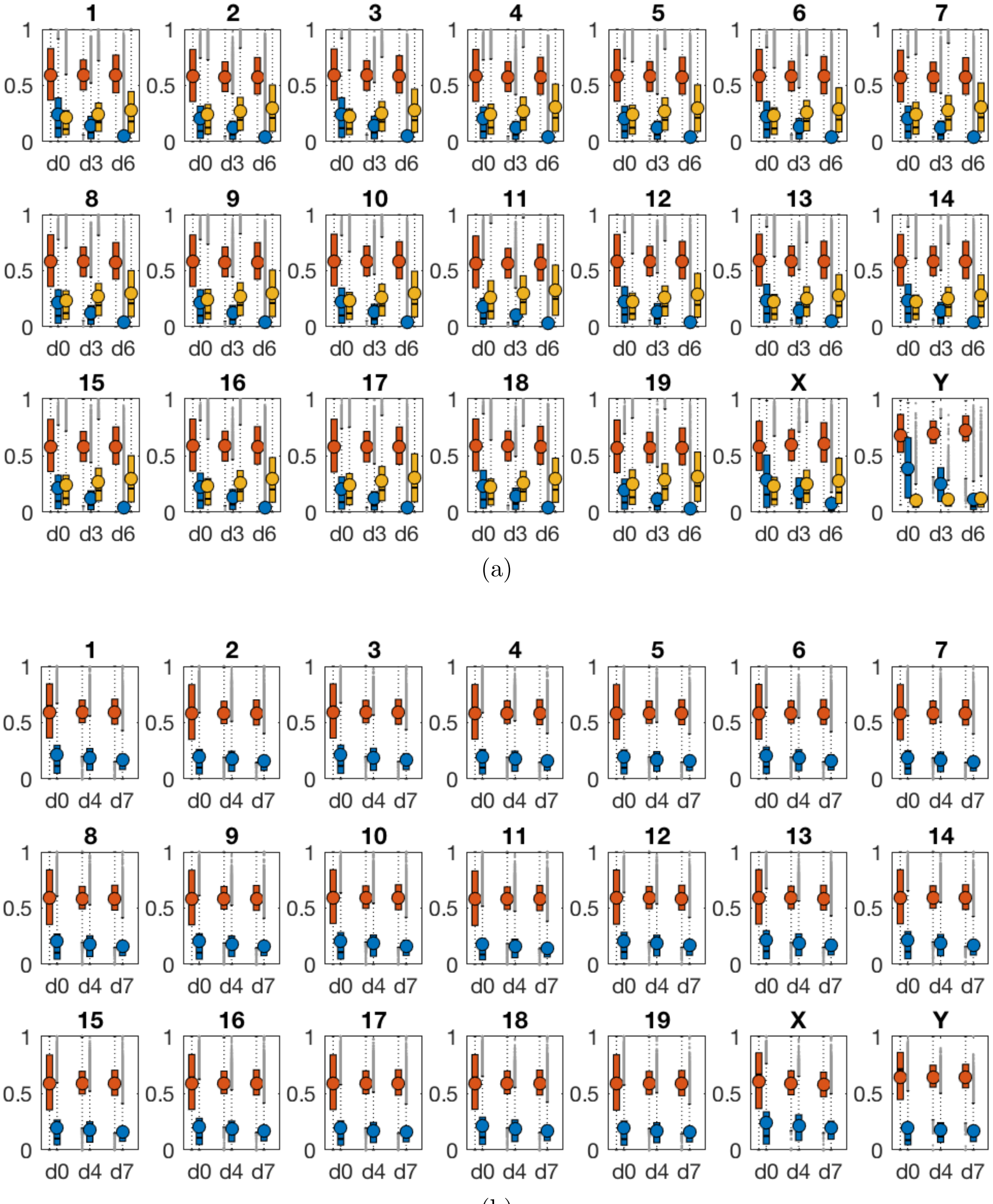
Figure B.3: Number of CpGs (y-axis) with one, two or three observation days (x-axis) for each chromosome in (a) WT and (b) Tet TKO data.



(a)

(b)

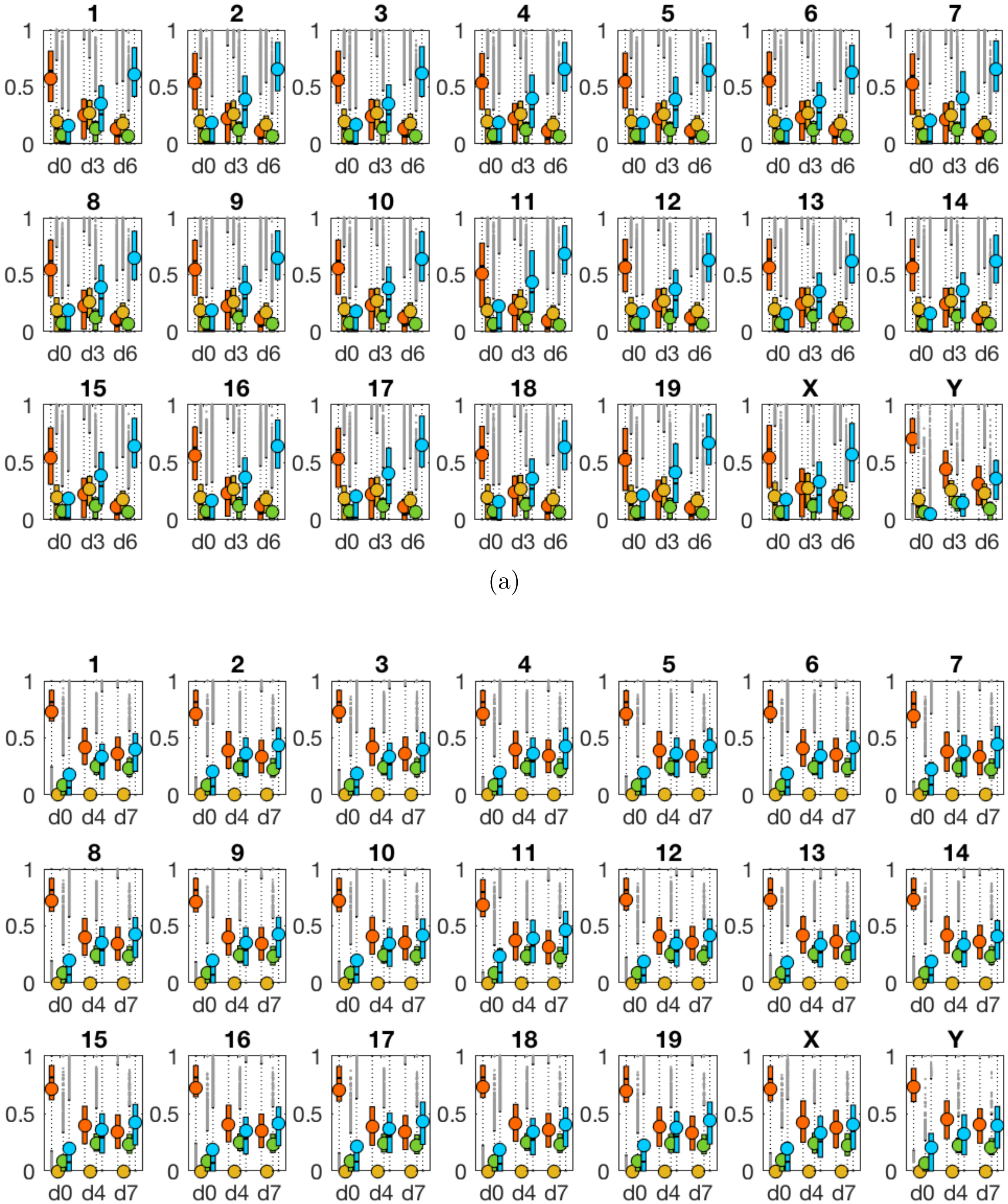
Figure B.4: Average number of single CpG independent samples, i.e, depth sequencing, (y-axis) per day (x-axis) for each chromosome in WT and Tet TKO data.



(a)

(b)

Figure B.5: Bar plots for the maintenance (red), *de novo* (blue) and hydroxylation (yellow) efficiencies over time taken by BI method for each individual chromosome in (a) WT and (b) Tet KO data.



(a)

(b)

Figure B.6: Bar plots for the hidden states' levels over time of each individual chromosome in (a) WT and (b) Tet TKO data. 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)

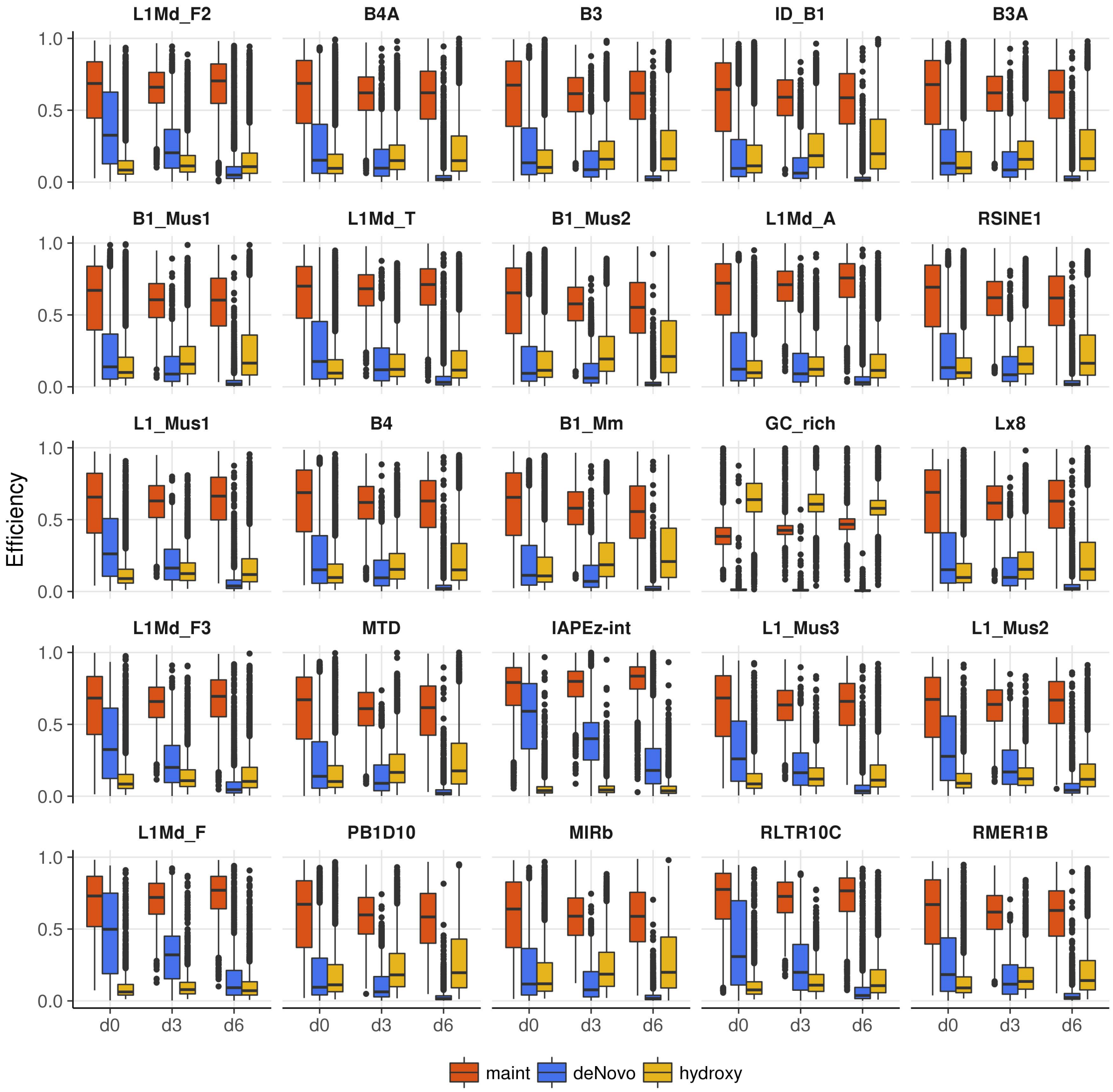


Figure B.7: Efficiency profiles of the 25 most frequent repetitive elements in WT ES cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. Red = maintenance, blue = *de novo*, yellow = hydroxylation.

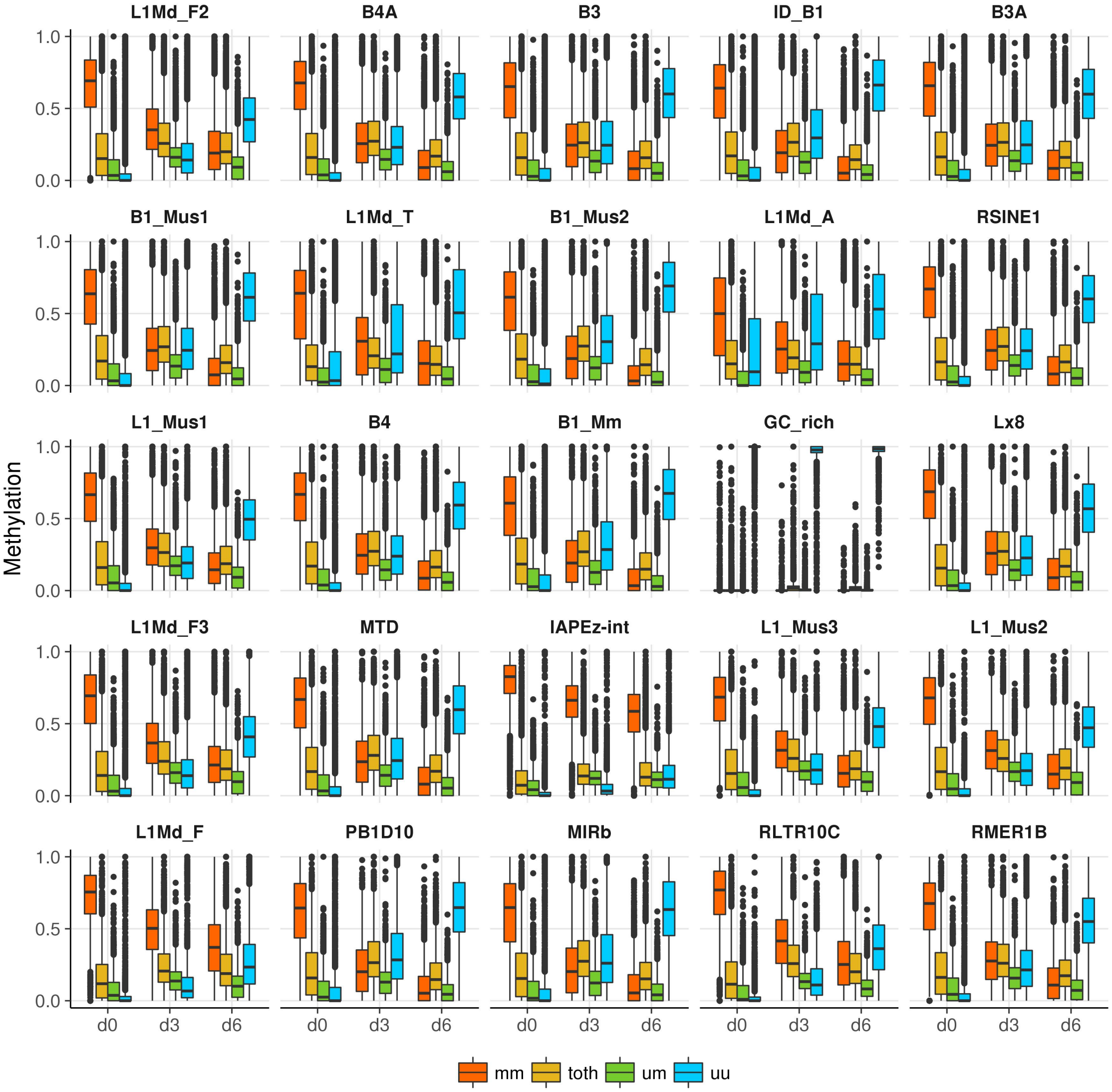


Figure B.8: Methylation level at the 25 most frequent repetitive elements in our analysis for WT cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. Red = fully methylated CpGs (5mC/5mC), green = hemimethylated CpGs (5mC/C or C/5mC), yellow = 5hmC, blue = unmethylated CpGs (C/C).

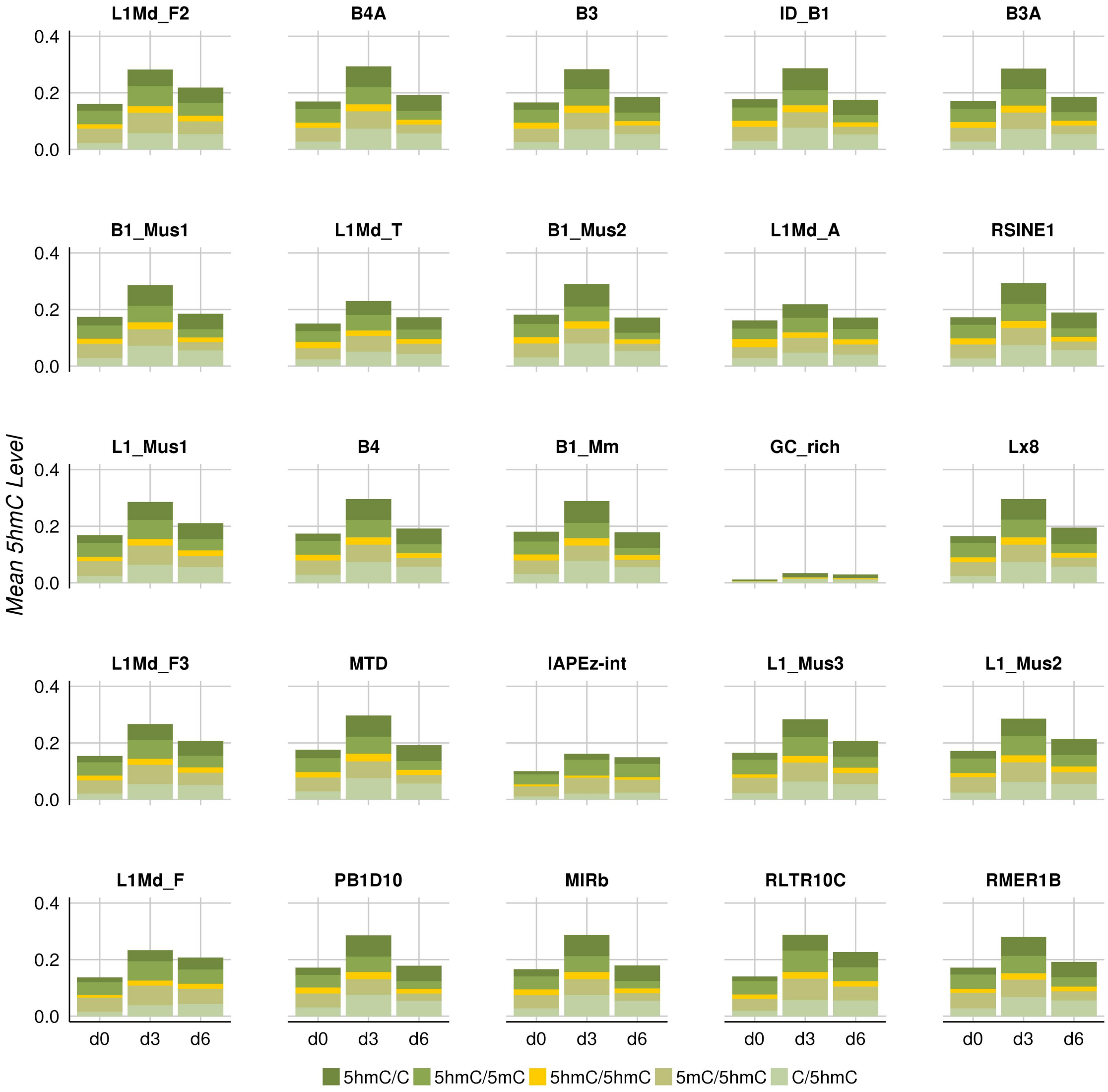


Figure B.9: Level and distribution of 5hmC across the 25 across frequent repetitive elements in WT cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC.

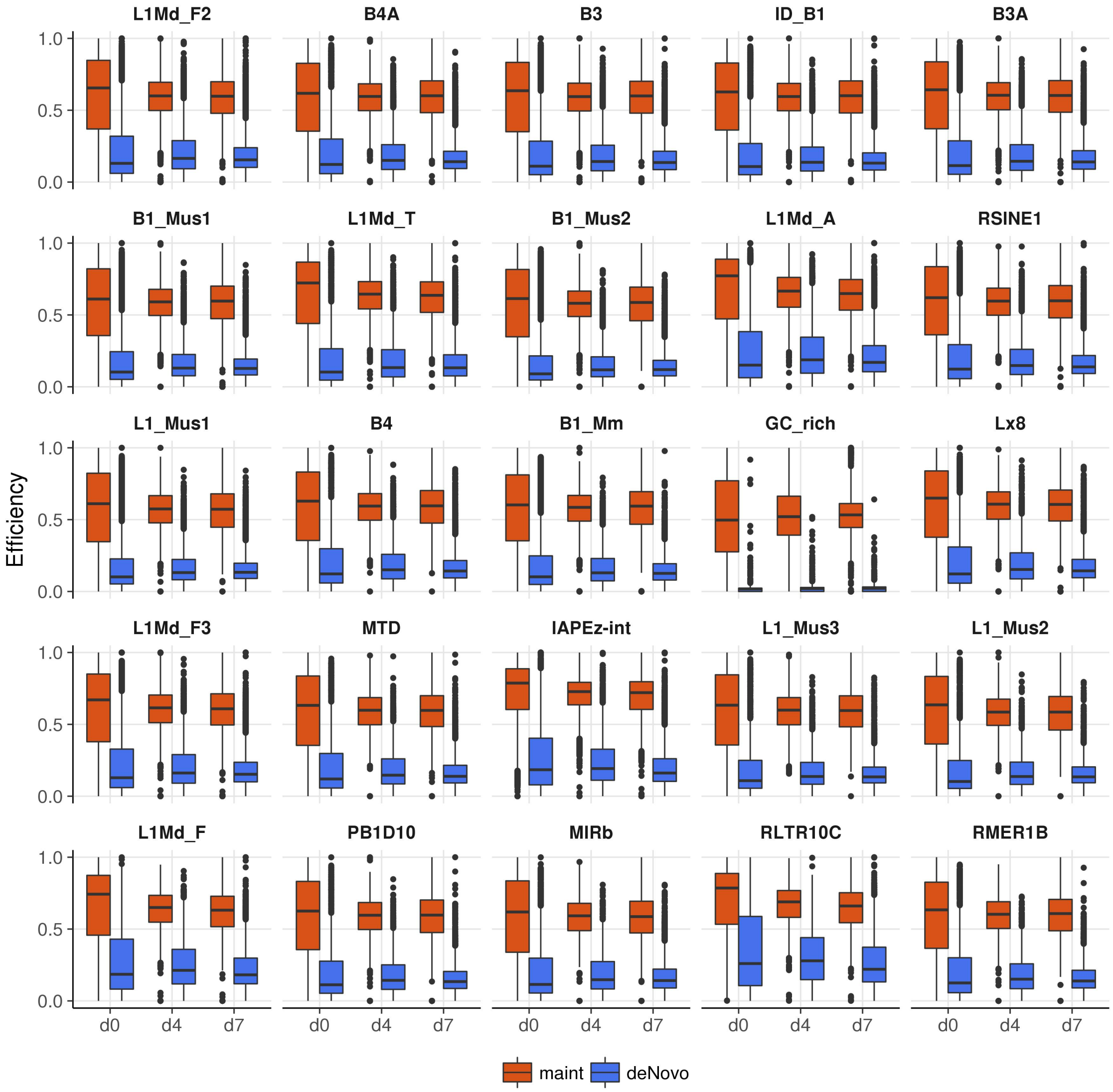


Figure B.10: Efficiency profiles of the 25 most frequent repetitive elements in our analysis for Tet TKO cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. Red = maintenance, blue = *de novo*.

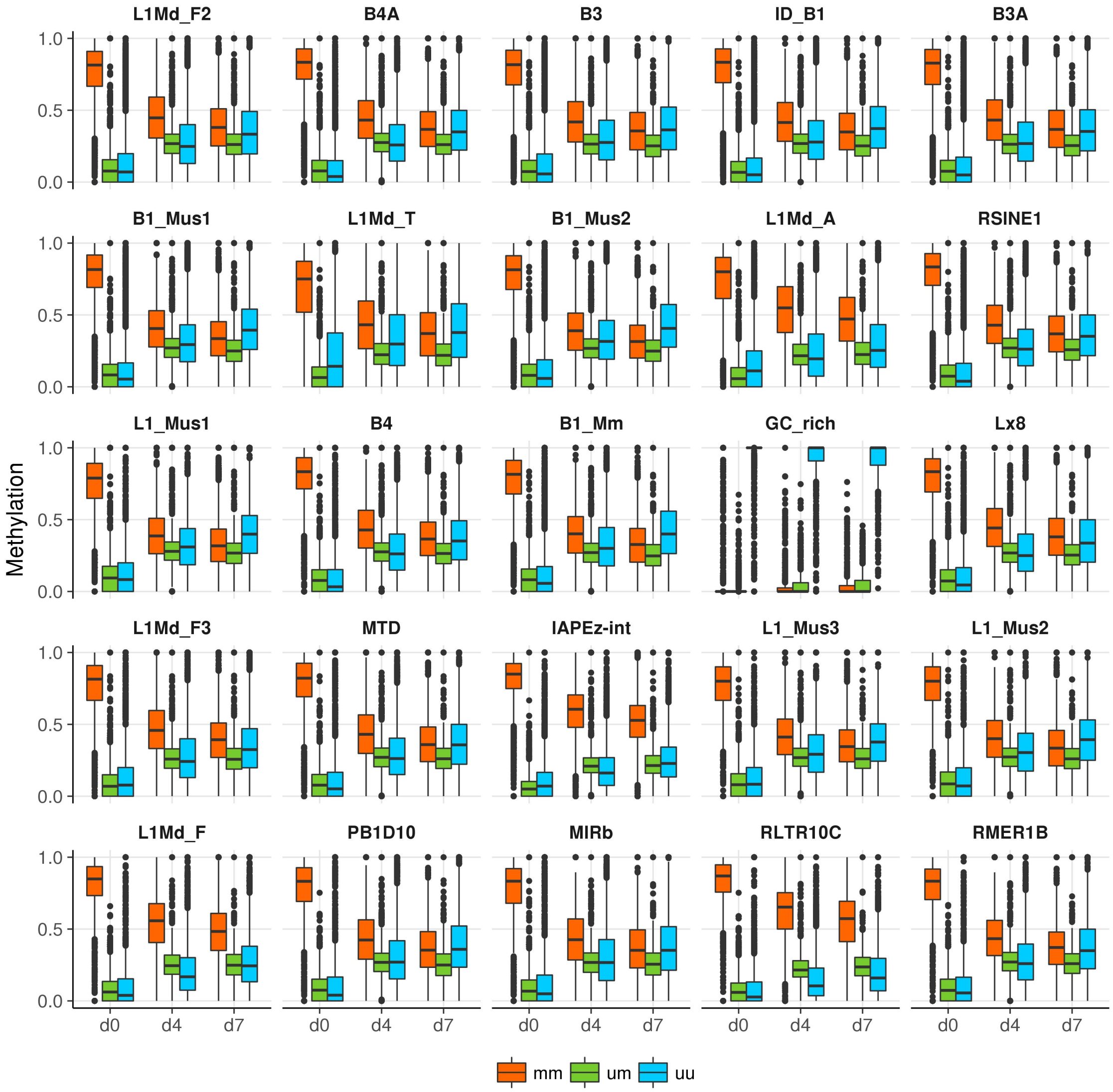


Figure B.11: Modification levels of the 25 most frequent repetitive elements in our analysis for Tet TKO cells. Elements are presented in decreasing order, most frequent left top, least frequent right bottom. Annotation according to UCSC. Red = fully methylated CpGs (5mC/5mC), green = hemimethylated CpGs (5mC/C or C/5mC), blue = unmethylated CpGs (C/C).

Bibliography

[1] 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.

[2] Babraham Bioinformatics - Trim Galore! `http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/`.

[3] 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.

[4] 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.

[5] Peter Beerli. Comparison of bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*, 22(3):341–345, 2005.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483, 1978.

[17] 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.

[18] 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.

[19] Audrey Q Fu, Diane P Genereux, Reinhard Stöger, Alice F Burden, Charles D Laird, and Matthew Stephens. Statistical inference of in vivo properties of human dna methyltransferases from double-stranded methylation patterns. *PLoS One*, 7(3):e32225, 2012.

[20] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[21] 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.

[22] 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.

[23] 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.

[24] Leo Goodstadt. Ruffus: a lightweight python library for computational pipelines. *Bioinformatics*, 26(21):2778–2779, 2010.

[25] Humaira Gowher and Albert Jeltsch. Enzymatic properties of recombinant dnmt3a dna methyltransferase from mouse: the enzyme modifies dna in a non-processive manner and also methylates non-cpa sites. *Journal of molecular biology*, 309(5):1201–1208, 2001.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[34] 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.

[35] 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.

[36] 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.

[37] Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.

[38] 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.

[39] 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.

[40] Albert Jeltsch. On the enzymatic properties of dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme, 2006.

[41] 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.

[42] 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.

[43] Sara Kangaspeska, Brenda Stride, Raphaël Métivier, Maria Polycarpou-Schwarz, David Ibberson, Richard Paul Carmouche, Vladimir Benes, Frank Gannon, and George Reid. Transient cyclical methylation of promoter dna. *Nature*, 452(7183):112, 2008.

[44] Chantriolnt-Andreas Kapourani and Guido Sanguinetti. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412, 2016.

[45] 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.

[46] Gun-Do Kim, Jingwei Ni, Nicole Kelesoglu, Richard J Roberts, and Sriharsa Pradhan. Co-operation and communication between the human maintenance and de novo dna (cytosine-5) methyltransferases. *The EMBO journal*, 21(15):4183–4195, 2002.

[47] Andrei Kolmogoroff. Über die analytischen methoden in der wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931.

[48] 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.

[49] Mahesh Kumar and Nitin R Patel. Clustering data with measurement errors. *Computational Statistics & Data Analysis*, 51(12):6084–6101, 2007.

[50] 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.

[51] 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.

[52] Konstantin Lepikhov, Mark Wossidlo, Julia Arand, and Jörn Walter. DNA methylation reprograming and DNA repair in the mouse zygote. *International Journal of Developmental Biology*, 54(11):1565, 2010.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] Scott J Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[58] Matthew C Lorincz, Dirk Schübeler, Shauna R Hutchinson, David R Dickerson, and Mark Groudine. Dna methylation density influences the stability of an epigenetic imprint and dnmt3a/b-independent de novo methylation. *Molecular and cellular biology*, 22(21):7572–7580, 2002.

[59] 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.

[60] Alexander Lück, Pascal Giehr, Jörn Walter, and Verena Wolf. A stochastic model for the formation of spatial methylation patterns. In *International Conference on Computational Methods in Systems Biology*, pages 160–178. Springer, 2017.

[61] 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.

[62] 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.

[63] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011.

[64] Daniel McNeish. On using bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5):750–773, 2016.

[65] 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.

[66] Raphaël Métivier, Rozenn Gallais, Christophe Tiffoche, Christine Le Péron, Renata Z Jurkowska, Richard P Carmouche, David Ibberson, Peter Barath, Florence Demay, George Reid, et al. Cyclical dna methylation of a transcriptionally active promoter. *Nature*, 452(7183):45, 2008.

[67] Esa Nummelin. *General irreducible Markov chains and non-negative operators*, volume 83. Cambridge University Press, 2004.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] Sriharsa Pradhan, Albino Bacolla, Robert D Wells, and Richard J Roberts. Recombinant human dna (cytosine-5) methyltransferase i. expression, purification, and comparison of de novo and maintenance methylation. *Journal of Biological Chemistry*, 274(46):33002–33010, 1999.

[74] 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.

[75] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[76] 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.

[77] 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.

[78] Arthur D Riggs. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.

[79] 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.

[80] Ronald Schoenberg. Constrained maximum likelihood. *Computational Economics*, 10(3):251–266, 1997.

[81] 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.

[82] 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.

[83] David Shen, Zaizai Lu, et al. Computation of correlation coefficient and its confidence interval in sas. *SUGI: Paper*, pages 170–31, 2006.

[84] 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.

[85] 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.

[86] 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.

[87] 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.

[88] 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.

[89] Marco Taboga. *Lectures on probability theory and mathematical statistics*. CreateSpace Independent Pub., 2012.

[90] 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.

[91] Luke Tierney. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

[92] 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.

[93] Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.