Spatial statistical modelling of epigenomic variability

Chantriolnt-Andreas Kapourani

Doctor of Philosophy
Institute of Adaptive and Neural Computation
School of Informatics
University of Edinburgh
2019

Dedicated to Erisa
my brother Lazaros
and
to my parents
for encouraging me
to pursue my dreams

Declaration

I declare that this thesis has been composed by myself and that this work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work which has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly indicated at the beginning of each chapter. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Edinburgh, 2019

Chantriolnt-Andreas Kapourani
2019

Acknowledgements

I would like to sincerely thank my enthusiastic supervisor Guido Sanguinetti for his kindness, guidance and support throughout the duration of this thesis. I owe him a huge debt of gratitude for teaching me most of what I know now, and giving me the freedom to delve into new subjects in computational biology, machine learning and statistics. Similarly, I express my gratitude to my co-supervisor Duncan Sproul for his enthusiasm and insightful comments, especially on the biological aspects of my research. Without them this work would have been a terrible failure.

I would also like to thank Matthias Hennig and Iain Murray for their constructive comments and assisting me with my annual reviews. In addition, I want to extend my appreciation to my collaborators Ricard Argelaguet, Stephen Clark, Oliver Stegle, John Marioni and Wolf Reik; it was refreshing and challenging to work on exciting biological questions. A special thanks goes to the staff and fellow students of the CDT, especially Rafael, for creating such a warm atmosphere, it was a pleasure to be part of this community.

I am grateful to have been able to work with the wonderful group of people that Guido has assembled. I will miss dearly the warm and friendly atmosphere in the group, our daily coffee mornings, and our outdoor summer socials. I am indebted to my friends and colleagues Alina, Anastasis, Botond, Christos, Daniel, David, Dimitris, Edward, Emily, Gabriele, Giulio, Michael, Michalis, Tom, Van-Anh, Veronica, Ylenia, and Yuanhua, for making these years so enjoyable. I particularly want to thank Michalis, Dimitris and Yuanhua for the fruitful research discussions we had together and their invaluable help and support.

Undoubtedly, I would like to express my gratitude to my family for their unwavering support and making me who I am today; I hope this piece will make them proud. Finally, I would like to thank Erisa for her love, encouragement and understanding through this process, and her endless patience with me. I would not be here without her.

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

Abstract

Each cell in our body carries the same genetic information encoded in the DNA, yet the human organism contains hundreds of cell types which differ substantially in physiology and functionality. This variability stems from the existence of regulatory mechanisms that control gene expression, and hence phenotype. The field of epigenetics studies how changes in biochemical factors, other than the DNA sequence itself, might affect gene regulation. The advent of high throughput sequencing platforms has enabled the profiling of different epigenetic marks on a genome-wide scale; however, bespoke computational methods are required to interpret these high-dimensional data and investigate the coupling between the epigenome and transcriptome.

This thesis contributes to the development of statistical models to capture spatial correlations of epigenetic marks, with the main focus being DNA methylation. To this end, we developed BPRMeth (Bayesian Probit Regression for Methylation), a probabilistic model for extracting higher order methylation features that precisely quantify the spatial variability of bulk DNA methylation patterns. Using such features, we constructed an accurate machine learning predictor of gene expression from DNA methylation and identified prototypical methylation profiles that explain most of the variability across promoter regions. The BPRMeth model, and its algorithmic implementation, were subsequently substantially extended both to accommodate different data types, and to improve the scalability of the algorithm.

Bulk experiments have paved the way for mapping the epigenetic landscape, nonetheless, they fall short of explaining the epigenetic heterogeneity and quantifying its dynamics, which inherently occur at the single cell level. Single cell bisulfite sequencing protocols have been recently developed, however, due to intrinsic limitations of the technology they result in extremely sparse coverage of CpG sites, effectively limiting the analysis repertoire to a semi-quantitative level. To overcome these difficulties we developed Melissa (MEthyLation Inference for Single cell Analysis), a Bayesian hierarchical model that leverages local correlations between neighbouring CpGs and similarity between individual cells to jointly impute missing methylation states, and cluster cells based on their genome-wide methylation profiles.

A recent experimental innovation enables the parallel profiling of DNA methylation, transcription and chromatin accessibility (scNMT-seq), making it possible to link transcriptional and epigenetic heterogeneity at the single cell resolution. For the scNMT-seq study, we applied the extended BPRMeth model to quantify cell-to-cell chromatin accessibility heterogeneity around promoter regions and subsequently link it to transcript abundance. This revealed that genes with conserved accessibility profiles are associated with higher average expression levels.

In summary, this thesis proposes statistical methods to model and interpret epigenomic data generated from high throughput sequencing experiments. Due to their statistical power and flexibility we anticipate that these methods will be applicable to future sequencing technologies and become widespread tools in the high throughput bioinformatics workbench for performing biomedical data analysis.

Lay Summary

How can a single cell — the fundamental unit of life — accurately orchestrate the development of complex life forms? How can the same DNA sequence give rise to diverse cell types with different physiology and functionality? The answer is that although each cell has access to the same book, the whole DNA sequence, they interpret it in a cell-type specific manner by regulating the expression of certain genes. The field of epigenetics studies how changes in biochemical factors, other than the DNA sequence itself, might affect gene regulation.

This thesis contributes to the development of statistical models to capture spatial correlations of epigenetic marks, with the main focus being DNA methylation. To this end, we developed a computational analysis pipeline for explicitly modelling and quantifying the variability of DNA methylation, when measured as an average across millions of cells. We demonstrated that by capturing higher order information of epigenetic marks we can accurately predict the expression of nearby genes.

Single cells have inherently diverse epigenetic patterns which can be masked when assaying a bulk population. In theory, single cell technologies will enable us to characterise in an unbiased way the single cell epigenetic heterogeneity. However, due to technology limitations the data are extremely sparse, resulting in missing information for the majority of epigenetic marks across the genome. To overcome these difficulties we developed a Bayesian method that captures both local and cell-to-cell variability, and subsequently jointly imputes the missing epigenetic states and identifies cell sub-populations from the data. In addition, recent experimental methods have enabled the profiling of multiple molecular modalities from the same single cell, allowing us to investigate the dynamic coupling between different biological layers. To leverage the richness of the data, we developed a statistical model to quantify cell-to-cell epigenetic heterogeneity and subsequently link this heterogeneity to gene expression levels.

In summary, this thesis proposes statistical methods to model and interpret epigenomic data generated from high throughput sequencing experiments. Due to their statistical power and flexibility we anticipate that these methods will be applicable to future sequencing technologies and become widespread tools in the high throughput bioinformatics workbench for performing biomedical data analysis.

Contents

List of Figures

List of Tables

Nomenclature

Mathematical Notation

x, X	Scalars (lowercase or uppercase letters)
$\mathbf{x}$	Vectors (lowercase bold letters)
$\mathbf{X}$	Matrices (uppercase bold letters)
$\mathcal{D}$	Dataset
$\mathbf{x}^\top, \mathbf{X}^\top$	Transpose of vector $\mathbf{x}$ and matrix $\mathbf{X}$
$\mathbf{x} = (x_1, \ldots, x_N)^\top$	Column vector with N elements
$\mathbf{I}_N$	The $N \times N$ identity matrix
$\hat{\theta}$	Point estimate for parameter θ
$p(x, y)$	Probability of x and y
$p(x \mid y)$	Probability of x conditioned on y
$x \perp\!\!\!\perp y \mid z$	Variable x is independent of y conditioned on z
$\langle \cdot \rangle_{p(x)}$	Expectation with respect to distribution $p(x)$
$\mathrm{KL}[p \,\|\, q]$	The KL divergence from distribution p to distribution q
$\mathrm{H}[p]$	The entropy of the probability distribution p
$\Phi(\cdot)$	The CDF of the standard normal distribution
$\nabla_\theta f$	Gradient of function f with respect to θ
$\mathbb{I}(\cdot)$	Indicator function
$\{\ldots\}$	A set, depending on context
$\log$	The natural logarithm

Acronyms / Abbreviations

e.g.	For example
i.e.	That is to say
w.r.t.	With respect to
5mC	5-methylcytosine
ARI	Adjusted rand index
AUC	Area under the receiver operating characteristic curve
BIC	Bayesian information criterion
BPRMeth	Bayesian probit regression for methylation
BS-seq	Bisulfite sequencing
CAVI	Coordinate ascent variational inference
CDF	Cumulative distribution function
CGI	CpG island
ChIP-seq	Chromatin immunoprecipitation sequencing
DAG	Directed acyclic graph
DGM	Directed graphical model
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
ELBO	Evidence lower bound
EM	Expectation maximisation
ENCODE	Encyclopaedia of DNA elements
ESC	Embryonic stem cell
ESS	Effective sample size
GEO	Gene expression omnibus

GLM	Generalised linear model
GO	Gene ontology
KL	Kullback-Leibler
MAP	Maximum a posteriori
MB	Markov blanket
MCMC	Markov chain Monte Carlo
Melissa	Methylation inference for single cell analysis
MFVI	Mean field variational inference
MLE	Maximum likelihood estimation
mRNA	Messenger RNA
NGS	Next generation sequencing
NOMe-seq	Nucleosome occupancy and methylation sequencing
PCR	Polymerase chain reaction
PDF	Probability density function
RBF	Radial basis function
RF	Random forest
RMSE	Root mean square error
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RRBS	Reduced representation bisulfite sequencing
scNMT-seq	Single cell nucleosome, methylome and transcriptome sequencing
SVM	Support vector machine
TF	Transcription factor
TSS	Transcription start site
WGBS	Whole genome bisulfite sequencing

Chapter 1

Introduction

How can a single cell — the fundamental unit of life — accurately orchestrate the development of complex life forms? How can the same DNA sequence give rise to diverse cell types with different physiology and functionality? Clearly, these mysteries of life sciences have fascinated scientists, and humanity as a whole, over the past years. The fundamental goal of molecular biology is to understand the structure and functional organisation of DNA and how variations lead to different phenotypes and eventually disease; however, this task is extremely difficult due to the inherent complexity and stochasticity of biological systems. Thanks to advances in sequencing technology and vast reductions in cost researchers can assay multiple molecular layers at unprecedented spatial and temporal resolution, enabling quantitative modelling of these complex biological processes.

Providing a historical perspective, the Human Genome Project (Lander et al., 2001) was completed at the turn of the 21^{st} century with an estimated cost of roughly three billion dollars. Now, thanks to the parallelisation of sequencing and novel technologies we can assay multiple biological layers, such as the ENCODE project (Dunham et al., 2012), and hundreds of thousands of individuals, see UK Biobank cohort study (Sudlow et al., 2015); while at the same time single cell genomics are promising to create comprehensive reference maps of all human cells (Regev et al., 2017), paving the path towards precision medicine (Ashley, 2016). In addition, global effort studies such as the Earth BioGenome Project (Lewin et al., 2018), plan to sequence all 1.5 million known species of animal, plant, protozoa, and fungi on Earth, revolutionising our understanding of biology and evolution.

This new era with exponential growth of biomedical data[1] poses computational challenges in handling and processing the raw data. Nevertheless, even beyond technical aspects, bespoke computational methodologies must be developed to transform these high-dimensional data into scientific knowledge and new hypotheses. Therefore, computational biology is a highly interdisciplinary field between biology, computer science and statistics, that requires deep understanding of biological processes while at the same time developing efficient yet realistic

[1] The amount of health-related data is estimated to double every 73 days by 2020 (Nature Editorial, 2016).

statistical models to analyse and distil meaningful patterns from these data. It is our belief that progress in biological research and computational modelling can be symbiotic. The high dimensionality, volume, noise, and heterogeneity of biological data pose formidable challenges to traditional statistical and machine learning models; on the other hand, computational models might potentially reveal new biological insights leading to novel research directions.

Due to the inherent stochasticity and noisiness of biological experiments, this thesis focuses on developing probabilistic models that take into account the uncertainty associated with the data. The Bayesian paradigm provides an excellent basis for quantifying uncertainty and incorporating biological prior knowledge in a principled way, resembling the data driven scientific process itself: we start off with some initial models that encode our prior knowledge, then we revise our models accordingly in the light of new evidence, and finally use the updated knowledge to explain the (potentially) noisy phenomena and make optimal decisions. To compactly represent the modelling assumptions of high-dimensional probability distributions and exploit their independence structure for developing efficient inference algorithms, we use a formalism called probabilistic graphical models (Koller and Friedman, 2009; Lauritzen, 1996). Due to its graph-theoretic representation, this formalism has the added benefit of effortlessly interpreting and communicating the modelling assumptions which are essential for cross-disciplinary fields, such as computational biology.

This thesis concerns the development of statistical models for analysing and deciphering epigenomic data, with the main focus being DNA methylation. Epigenetics can be loosely described as the field that studies how changes in biochemical factors, other than the DNA sequence itself, affect gene regulation. In particular, the thesis presents methods for capturing spatial correlations of epigenetic marks both on bulk and single cell studies, and algorithms for performing robust and efficient inference on large-scale probabilistic models.

1.1 Contributions

The aim of this thesis is to use probabilistic machine learning to interpret epigenomic data generated from high throughput sequencing experiments. The contributions of the thesis can be summarised as follows:

1. The development of a computational analysis pipeline for modelling bulk bisulfite sequencing experiments, which extracts higher order methylation features that quantify the spatial variability of DNA methylation, accurately predicts gene expression levels, and identifies prototypical methylation profiles that explain most of the variability across promoter regions (Kapourani and Sanguinetti, 2016).

2. The substantial extension of this model both to accommodate different sequencing technologies and to improve the scalability of the algorithm (Kapourani and Sanguinetti, 2018a); and its subsequent application on a novel multi-omics protocol scNMT-seq, to

quantify cell-to-cell chromatin accessibility heterogeneity around promoter regions and link this heterogeneity to transcript abundance (Clark et al., 2018).

3. The development of a Bayesian hierarchical method to model single cell bisulfite sequencing data, which exploits the local correlation between neighbouring genomic sites and the similarity across individual cells to jointly impute missing methylation states and identify cell sub-populations based on their methylomes (Kapourani and Sanguinetti, 2018b).

1.2 Thesis layout

It is anticipated that the thesis will be of interest to researchers across different disciplines. Hence, a substantial effort is dedicated to provide the necessary background and make the thesis accessible to a general audience in molecular biology, computational statistics and computer science. The thesis is organised as follows. Chapter 2 provides the necessary background to molecular biology, focusing on the epigenetic control of gene expression and the different sequencing technologies for assaying (epi)genomic data. Chapter 3 introduces probabilistic machine learning, how the Bayesian paradigm provides a natural approach to learn from data, and the computational challenges when performing inference for complex probabilistic models. Chapter 4 presents the BPRMeth method for modelling spatial correlations in bulk DNA methylation data and demonstrates the significance of developing bespoke computational approaches for analysing complex biological data. Chapter 5 concerns the extension of the BPRMeth model and its application on the scNMT-seq study to link chromatin accessibility heterogeneity with transcription abundance. Chapter 6 introduces the Melissa model that jointly imputes methylation states and clusters single cell based on their methylomes, and evaluates the robustness and accuracy of the method both on real and synthetic data in various genomic contexts. Finally, chapter 7 concludes the thesis with a summary of the main contributions and discusses possible avenues for future research, such as performing integrative modelling of high throughput epigenomic data.

Chapter 2

Molecular biology and epigenetics

This chapter introduces the relevant biology for the thesis. Section 2.1 briefly outlines the foundations of molecular biology and section 2.2 introduces the different epigenetic marks and their role in gene regulation. Section 2.3 concerns different sequencing technologies for assaying (epi)genetic data. Section 2.4 concludes this chapter by introducing single cell multi-omics: a powerful technology for assaying multiple biological layers in the same single cell.

2.1 Foundations of molecular biology

Deoxyribonucleic acid (DNA) is a macromolecule that carries the hereditary information used in the development and function of all known living organisms. DNA was first discovered by Swiss chemist Friedrich Miescher, in 1869, who was analysing the pus of discarded surgical bandages (Dahm, 2008; Miescher-Rüsch, 1871). Important work from Phoebus Levene[1] and Erwin Chargaff[2] on the primary chemical components of DNA, combined with X-ray crystallography (Franklin and Gosling, 1953), were vital for the discovery of the three-dimensional double-helical model for the structure of DNA (Watson and Crick, 1953). Each strand of the DNA is composed of a series of *nucleotides*, and each nucleotide consists of two parts: a sugar-phosphate group and one of four nitrogen-containing bases: adenine (A), cytosine (C), guanine (G), and thymine (T). The nitrogenous bases of the two separate strands are bound together, by strict rules of base complementarity (A with T and C with G), with hydrogen bonds forming the double-helix DNA (see figure 2.1A).

The sugar-phosphate backbone of the DNA is formed by alternating the sugar and phosphate groups of successive nucleotides. The sugar-phosphate links are resistant to cleavage compared to hydrogen bonds, which makes it possible to unwind the two strands without breaking the DNA backbone. This process is essential for DNA *replication*, where each separated strand

[1] Phoebus Levene discovered the order of the three major components of a single nucleotide (phosphate-sugar-base); and the carbohydrate components of DNA (deoxyribose) and RNA (ribose) (Levene, 1917).

[2] Chargaff's rule states that across all organisms the total number of purines (adenine and guanine) is equal to the total number of pyrimidines (thymine and cytosine) (Chargaff, 1950).

acts as a template for the formation of an entire new strand with the help of DNA polymerase enzyme (Alberts et al., 2015). This process ensures that after cell division the two daughter cells will inherit the hereditary information from the parent cell[3]. Within eukaryotic cells, the DNA is stored in the cell nucleus, and is typically organised in distinct structures, called *chromosomes*.

DNA can be thought of as a precious book, storing all genetic information in a permanent repository, which however is stuck in the library (i.e. the cell nucleus). To use this information in a flexible manner, the cell produces a dedicated molecule, called RNA (ribonucleic acid). RNA is more flexible since it serves as a temporary copy of stretches of DNA, and the cell can make multiple copies of the same information. In addition, RNA can transfer the genetic information outside the library of the nucleus, where DNA cannot reach (Richards and Hawley, 2011). RNA is structurally similar to DNA, with the following main differences: (1) the sugar in RNA is ribose instead of deoxyribose, (2) RNA mostly occurs in single-stranded form[4] and (3) the base thymine is replaced by uracil (U) which can base-pair with adenine, as shown in figure 2.1A. RNA contains a variety of forms — including messenger RNA (mRNA), transfer RNA (tRNA), micro RNA (miRNA), and ribosomal RNA (rRNA) — which are involved in different biological functions. This thesis will focus on mRNAs, which contain the coding information that the cell can use to produce *amino acids*, the building block of proteins.

2.1.1 Gene expression

During *transcription*, the enzyme RNA polymerase II (Pol II) is used to synthesize RNA from a specific region of DNA — in a similar fashion to DNA replication — where a complex of proteins are required to unwind the DNA and insert nucleotide bases into the growing strand of RNA. Genomic regions of DNA that encode discrete hereditary characteristics are known as *genes*[5] and the resulting mRNA transcribed from these regions can be used to build protein molecules, using a process called *translation* (Crick, 1958). The *genome* is the complete set of genetic information in an organism, including both genes and non-coding DNA. These two major processes — transcription, where DNA information is transferred to mRNA, and translation, where mRNA is 'read' according to the *genetic code*[6] to produce proteins — constitute gene expression, which is the workhorse for the correct functionality of the cell. This process is also referred to as the central dogma of molecular biology (Crick, 1970), since it explains the flow of genetic information in a biological system, as illustrated in figure 2.1B.

[3] The DNA replication machinery duplicates eukaryotic human DNA at a rate of 50 nucleotides per second, with spectacular accuracy of 1 error every $10^7 - 10^8$ bases copied (Kunkel, 2004).

[4] RNA can take complex forms involving single-stranded and double-stranded regions, which is implicated in diverse biological processes. The interested reader should consult Selega (2018) for an in depth overview of RNA secondary structure.

[5] Gerstein et al. (2007) discuss how the definition of genes has changed over the years.

[6] The genetic code is a set of rules defining how the 4-letter DNA alphabet is translated into the 20-letter code of amino acids (Crick, 1968).

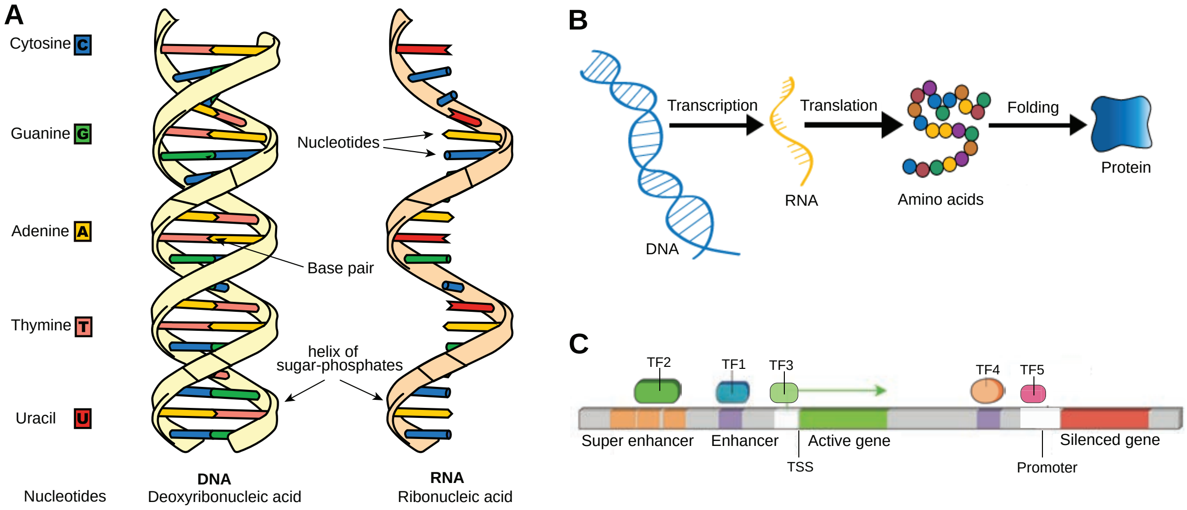


Figure 2.1 Main concepts of molecular biology. (**A**) Comparison of DNA (left) and RNA (right) structures; figure adapted from Sponk/Wikimedia Commons/CC BY-SA 3.0. (**B**) Schematic illustration of the central dogma of molecular biology; figure adapted from biosocialmethods.isr.umich.edu. (**C**) Simplistic model of transcription regulation in a representative genomic region. Transcription factors work in a coordinated fashion and bind to regulatory regions, such as promoters and enhancers, resulting in active (green) or silenced (red) genes.

This definition however does not provide the whole story of DNA complexity. It is estimated that protein-coding regions comprise of only 2% of the mammalian genomes (Elgar and Vavouri, 2008), leaving open the question of the importance of the remaining 98% of the non-coding DNA, which is often called 'junk DNA' (Palazzo and Gregory, 2014; Pennisi, 2012). The international Encyclopaedia of DNA Elements (ENCODE) project estimated that more than 80% of the human genome is comprised of regulatory elements (Dunham et al., 2012; Kellis et al., 2014), indicating the importance of non-coding DNA in regulating of gene expression. These findings were later criticised by researchers who claim that biochemical activities in DNA should not be conflated with biological function (Doolittle, 2013; Ponting and Hardison, 2011); based on comparative genomics methods the fraction of the human genome that is functional ranges between 8% - 10% (Rands et al., 2014).

2.1.2 Gene regulatory elements

Each cell in our body carries the same genetic information encoded in our DNA, yet the human organism contains more than 200 cell types (Alberts et al., 2015), which differ substantially in physiology and functionality. Even more astonishing is the fact that the development of complex life forms are a result of a preset programme encoded in a single cell right after fertilization, the *zygote*[7]. If all cells have almost identical DNA — and encode the same genes — how can

[7] The zygote (from Greek ζυγωτός meaning 'joined') is a one-cell totipotent embryo formed as a result of the union between the egg and the sperm at fertilisation (Eckersley-Maslin et al., 2018)

they ensure a robust and correct progression of development and differentiate to a variety of cell types that give rise to different phenotypes, for example neuronal or liver cells?

The answer is that the DNA sequence of an organism does not directly affect its phenotype; rather phenotype results from each cell expressing distinct cell-type specific genes in an orchestrated manner, which give rise to their unique form and properties (Richards and Hawley, 2011). For instance, retinal cells produce rhodopsin, a specific light sensitive protein that enables vision in low-light conditions. However, liver cells do not express rhodopsin, instead they express a set of genes required for specific functions of the liver. In addition, there is a subset of genes, known as *housekeeping genes*, that are involved in basic cell maintenance and are expressed at relatively constant rates across conditions (Eisenberg and Levanon, 2013).

The ability of the cell to coordinate and control the synthesis of proteins is termed *regulation of gene expression*; and is estimated that more than 8% of mammalian proteins are involved in this complex process (Alberts et al., 2015). Regulation occurs at different levels. *Transcriptional regulation* refers to the process of transcription of DNA to mRNA; and modulates which genes are transcribed, their order of transcription, and how frequently or rapidly these genes are transcribed. The produced RNA is additionally controlled — post-transcriptional regulation — by specific RNA binding proteins and microRNAs[8] that modulate alternative splicing[9] (Modrek and Lee, 2002), polyadenylation[10] (Colgan and Manley, 1997), and generally the stability of mRNA transcripts (Franks et al., 2017). Finally, the resulting proteins might also be subject to post-translational modifications (Mann and Jensen, 2003). This thesis concerns the regulation at the transcriptional level, and the terms gene expression regulation and transcription regulation will be used interchangeably throughout this document.

In mammalian organisms, the initiation of the transcriptional machinery often requires hundreds of proteins to bind in specific regions of the DNA and act in concert for regulating expression of each gene (see figure 2.1C). The region that these protein complexes bind is known as the *promoter*. The promoter is usually located at the beginning of the transcribed gene — called the *transcription start site* (TSS) — and can span a sequence of thousands of base pairs (bp). The promoter sequence contains regulatory elements that allow the recruitment of specific regulatory proteins, called *transcription factors* (TFs) (Latchman, 1997; Mitchell and Tjian, 1989). TFs contain specific DNA binding domains that recognize DNA motifs comprising of 6 to 12 nucleotides (Spitz and Furlong, 2012), and working in a coordinated fashion can facilitate (as activators) or inhibit (as repressors) the recruitment and stabilization of RNA Pol II to the genes that they regulate (Ptashne and Gann, 2002). The importance of TFs in cellular

[8] microRNAs (miRNAs) are estimated to target and repress more than 60% of the mammalian mRNAs (Friedman et al., 2009; Lewis et al., 2005).

[9] In a newly synthesized mRNA (also called pre-mRNA), *splicing* removes non-coding regions, the *introns*, while keeping the protein coding *exons*, resulting in mature mRNA. Through the process of alternative splicing, different combinations of exons are included to form mature mRNA, allowing a single gene to encode multiple proteins.

[10] Polyadenilation or poly-A tail is a stretch of adenine bases added at the $3'$ end of RNA.

development was elucidated in 2006 by Yamanaka's lab, who demonstrated that the expression of only four reprogramming TFs (*Oct4*, *Sox2*, *cMyc*, and *Klf4*) are necessary to reprogram adult human fibroblasts into induced pluripotent stem cells (iPSCs) (Takahashi et al., 2007; Takahashi and Yamanaka, 2006); a discovery that fuelled the research in regenerative medicine. However, a variety of DNA binding proteins are not classified as TFs due to their lack of DNA sequence specificity. The binding affinity of TFs is also influenced by different epigenetic marks, such as DNA methylation and chromatin accessibility (Clark et al., 1997; Prendergast and Ziff, 1991) (see below); although a small group of transcription factors — called pioneer factors — can bind to inaccessible chromatin and facilitate the remodelling and de-compaction of chromatin.

Transcription regulation is not restricted only to promoter regions and may also require other regulatory regions, termed *enhancers*, which can be found tens of kilobases upstream or downstream of the genes they regulate (Blackwood and Kadonaga, 1998). The canonical understanding is that enhancers need to interact with the transcription initiation complex at the promoter region to facilitate initiation of transcription (Ong and Corces, 2011). For this to happen, a group of activator proteins needs to bind to the enhancer, causing the DNA to bend around and create loop structures, allowing the physical interaction of enhancers and promoter regions. Active enhancers are associated with accessible chromatin (Creyghton et al., 2010) and occupancy of *p300* enzymes, which are global co-activators that are involved in the regulation of TFs (Janknecht and Hunter, 1996; Ogryzko et al., 1996). In general, enhancers are not directional (as opposed to promoters), are quite small and span a few hundreds of base pairs, and can regulate multiple promoters in a context specific manner; making it difficult to obtain a complete map of interactions between enhancers and promoters (Krivega and Dean, 2012). Recently, two independent studies identified genomic regions comprising multiple enhancers that regulate cell-specific genes; and this cluster of enhancers is termed *super enhancers* (Parker et al., 2013) or stretch enhancers (Whyte et al., 2013).

Clearly, the dynamic interplay between these numerous regulatory elements is tightly regulated at different levels. Given that a large fraction of the genome is comprised of regulatory elements, it is not surprising that we still have limited understanding of the precise mechanisms involved in transcription regulation.

2.2 The regulatory role of the epigenome

So far we have focused on the regulatory role of transcription factors that recognize and bind to specific sequences of DNA. However, a transcription factor only model is not sufficient to define the stable and wide spectrum of gene expression patterns present across differentiated cells; implying that other stable biochemical factors might affect gene regulation, and hence phenotype, independent of changes in the DNA sequence (Bird, 2007). These marks have fallen under the broad (and sometimes confusing) term *epigenetics*.

2.2.1 A brief history of epigenetics

The term epigenetics[11] was first coined by geneticist and embryologist Conrad Waddington, as a result of the old debate between the schools of preformationism and epigenesis (Waddington, 1939). The theory of preformationism — whose roots date back to Greek scholar Pythagoras and were widely adopted by medieval Europeans — stated that the sperm or egg already contained a miniaturised minihuman, the homunculus, which simply expanded during development[12] (Speybroeck, 2002). An early voice against Pythagoras' theory was Aristotle. In his work "*On the generation of animals*", he offered an alternative theory: what was passed from male to female during intercourse was not matter, but a *message* or *movement* which would carry the instructions for the formation of the foetus (Mukherjee, 2017). With the invention of the microscope in the 17th century, it was confirmed that during development organs where formed from initially homogeneous material in the egg (Lappalainen and Greally, 2017); a process coined as *epigenesis* by William Harvey circa 1650 (Villota-Salazar et al., 2016).

In his attempt to bridge the gap between the disciplines of embryology and genetics, Waddington (1942) introduced the term *epigenetics*, by fusing the words epigenesis and genetics. To Waddington, epigenetics was the study of the epigenesis, that is, the interpretation of the genotype to give rise to diverse phenotypes during development (Jaenisch and Bird, 2003). Waddington also proposed the concept of an *epigenetic landscape* (Waddington, 1957), in which a cell — depicted by a ball — rolls down the hillside representing the various cell fate decisions made during development, as illustrated in figure 2.2. Importantly, the landscape surface is influenced by an interaction of gene networks that create a bifurcating delta of valleys, which symbolize commitment to specific cell lineages (Deichmann, 2016). Nanney (1958) re-interpreted epigenetics as the process responsible for cellular memory, and based on this interpretation, Riggs (1975) and Holliday and Pugh (1975) identified that DNA methylation could be a potential transcriptional regulator that could persist through cell division. Russo et al. (1996) re-defined epigenetics as: "the study of mitotically and / or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence". However, this definition tells us only what epigenetics is not and is constrained by requiring heritability. A unifying definition of epigenetic events was recently proposed by Adrian Bird as "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird, 2007); including the possibility that epigenetic events might act as buffers for allowing genetic variability.

It is apparent that the definition of epigenetics has evolved to reflect what could be studied at each period, where early definitions were limited to testing cellular developmental events, whereas recent definitions encompass our increased knowledge of the molecular mechanisms

[11] Literal translation from the Greek word επιγενετιϰή is *outside (or over) conventional genetics.*

[12] A charm of preformation — at least for medieval Christians — was that it was infinitely recursive; hence all future humans had tasted the fruits of the *ancestral sin* of Adam and Eve's rebellion in the Garden of Eden, since each of us were present inside Adam's body (Mukherjee, 2017).

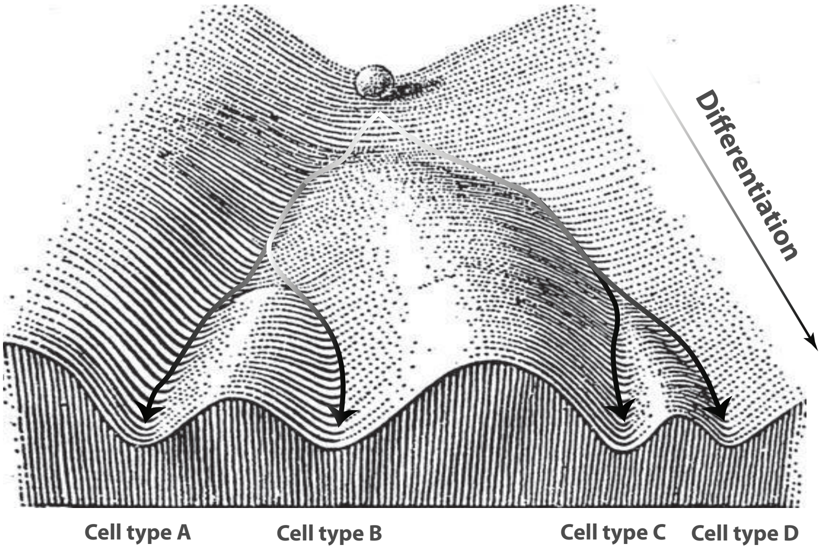


Figure 2.2 Waddington's epigenetic landscape. Differentiation pathways are indicated by dark arrows. Each cell, represented by a ball, rolls down the hillside and can take permitted trajectories indicating cell fate decisions, which will eventually result in differentiated cell types. Figure adapted from Waddington (1957).

underlying regulation of gene expression (Felsenfeld, 2014); and inevitably, this has led to debates on which molecular factors constitute epigenetic and which not (Madhani et al., 2008; Ptashne, 2013; Ptashne et al., 2010). The semantics of the term epigenetics are not a focus of this study, rather we are mostly interested in understanding the role of these mechanisms which are, in fact, important for a considerable part of the phenotype of complex life forms.

2.2.2 Chromatin organisation

In eukaryotic cells, the genomic DNA is arranged into *chromatin*, a complex structure consisting of DNA, proteins and RNA. This structure is essential, since the large linear DNA[13] needs to be packed by a factor of 10,000 to fit into a nucleus of roughly 10 μm diameter (Woodcock and Ghosh, 2010). To achieve this condensed form, DNA is organized in hierarchical chromatin structures (see figure 2.3). The fundamental layer of this hierarchy constitutes the *nucleosomes*, which comprise of 147 base pairs of DNA wrapped in approximately 1.75 superhelical turns around eight histone protein cores (two of each H2A, H2B, H3 and H4 histones) to form an octameric nucleosome (Luger et al., 1997). Two successive nucleosomes are connected together by short segments of linker DNA (∼10-80 bp long), and this specific structure is often referred to as 'beads on a string' (Bell et al., 2011). This poly-nucleosome string is further folded in a 30 nm compact fibre, which is stabilised by the binding of the H1 histone (Felsenfeld and Groudine, 2003). The genome is further organised in chromosomal loop structures formed by binding of the protein *CTCF* and the protein complex *cohesin* (Hnisz et al., 2016); which

[13] The 3.2 Gb human genome, when expanded, corresponds to approximately 2 meters of linear DNA.

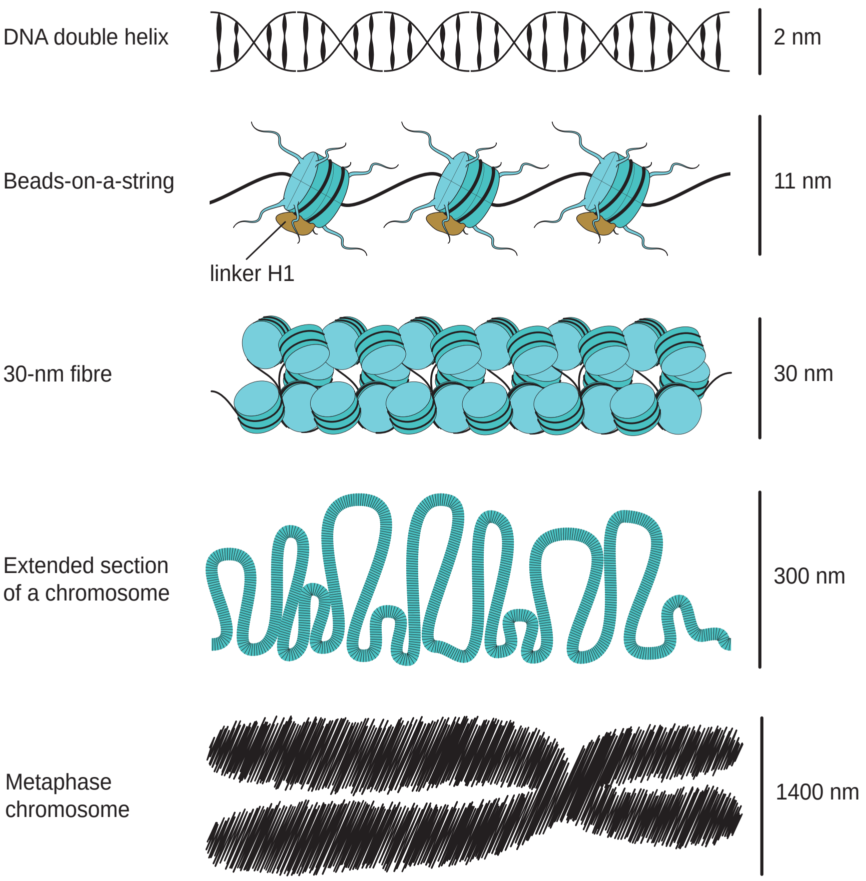


Figure 2.3 Levels of chromatin organisation in eukaryotic cells. Figure adapted from Müller (2016).

form a number of configurations including topologically associated domains (TADs) that divide the genome into structurally separate segments, although their role and structure is still not completely understood (Dixon et al., 2012; Sexton and Cavalli, 2015). Chromatin loop structure is crucial for gene-enhancer targeting — hence gene regulation — by limiting enhancer's ability to interact with genes outside the loop (Bell et al., 1999; Kellum and Schedl, 1991).

The highly compact conformation of chromatin, which is associated with gene poor and transcriptionally inactive regions, is known as *heterochromatin*; whereas the accessible chromatin with generally gene rich regions is termed *euchromatin*. Active genes are also characterized by regions with low nucleosome occupancy, termed as nucleosome depleted regions (NDRs) (Bell et al., 2011). Locally the chromatin is quite dynamic and undergoes a substantial reorganisation during maternal to zygote transition and generally during early embryo development (Eckersley-Maslin et al., 2018). Histone modifications, non-coding RNA and DNA methylation are associated with chromatin organisation in a highly interrelated manner (Lee, 2011); however, it is still not clear how these configurations are established and stably inherited through cell division (Cedar and Bergman, 2009).

2.2.3 Histone modifications

Histones undergo reversible post-translational modifications which play a fundamental role in gene regulation by allowing or blocking access to transcription factors (Allfrey et al., 1964; Kornberg and Thomas, 1974). These covalent modifications occur at the amino termini (tail domains) of the histones and are important for the structural stability of the nucleosomes (Biswas et al., 2011). Different modifications are associated with condensed or accessible chromatin, and the most well studied ones are acetylation and methylation of the lysine (amino acid K) residue[14] (see figure 2.4). However, these modifications normally do not persist across cell generations and need continual maintenance by transcription factors or other mechanisms; hence it is hotly debated whether or not they constitute an epigenetic mark (Ptashne, 2013).

Acetylation of the histone tail, via attachment of histone acetyltransferases (HATs), is normally associated with accessible chromatin and high transcriptional activity[15] (Brownell and Allis, 1995), whereas histone deacetylases (HDACs) detach the acetyl groups resulting in condensed chromatin, which blocks the binding of transcription factors (Haberland et al., 2009). During histone tail methylation, we can have mono-, di-, and tri-methylation of the lysine residue, and the degree of methylation and residue's position have diverse functional properties. The canonical understanding is that tri-methylation of histone H3 on lysine 4 (H3K4me3) is tightly associated with promoters and active transcription (Ruthenburg et al., 2007). On the other hand, tri-methylation of histone H3 on lysine 27 (H3K27me3) is prominent mark of gene silencing and repression of developmental genes (Aldiri and Vetter, 2012; Boyer et al., 2006). Notably, the active H3K4me3 and repressive H3K27me3 might occupy the same promoter regions. These *bivalent domains* are important during embryo development (Gerstein et al., 2007), and is believed that their role is to keep the chromatin in a *poised* state, enabling it to be rapidly activated (Voigt et al., 2013).

The variety and interactive properties of the different histone marks in a context dependent manner are postulated to constitute a *histone code* (Jenuwein and Allis, 2001). According to this hypothesis, combinations of histone modifications might dictate distinct functional events (Turner, 2000). Recently, these signatures of co-occurring histone marks have been summarised on a genome wide scale, mainly using hidden Markov models (HMMs) (Ernst and Kellis, 2010). Also, two independent studies from Karlic et al. (2010) and Dong et al. (2012) illustrated that histone modifications levels are predictive of gene expression, confirming the functional importance of these marks. In addition, Benveniste et al. (2014) showed that histone modifications could be accurately predicted only from transcription factor binding patterns; which might indicate that associations between histone modifications and gene expression might be indirect effects explained by transcription factors.

[14] The interested reader should consult Kouzarides (2007) for a thorough review on all — known at the time — covalent modifications of the core histones.

[15] One example being the acetylation of histone H3 on lysine 27 (termed as H3K27ac), which is associated with active enhancer regions.

2.2.4 DNA methylation

The most well studied epigenetic mark, and the main focus of this thesis, is *DNA methylation*. DNA methylation was first discovered in calf thymus cells (Hotchkiss, 1948), however, it was not until mid-1970s that researchers identified a transcriptionally repressive role of this epigenetic mark (Holliday and Pugh, 1975; Riggs, 1975). DNA methylation occurs when a methyl group is attached to a DNA nucleotide — mediated through the action of DNA methyltransferase (DNMT) enzymes — and in many eukaryotes this covalent modification is observed almost exclusively on carbon 5 of cytosine residues (which we refer to as 5-methylcytosine or 5mC) (Jaenisch and Bird, 2003). In mammals, 5mC predominantly occurs in the context of CpG dinucleotides, that is, a cytosine followed by a guanine, where ‘p’ stands for the phosphate group linking C and G. Also, 5mC is typically symmetric across the two complementary strands. This symmetry allows the methylation to be inherited and maintained through mitotic cell divisions by specific enzymes that recognise hemi-methylated CpG palindromes[16]. In keeping with the model, the DNMT1 enzyme is linked to the preservation of a methylated state by identifying and ‘completing’ hemi-methylated CpG sites; hence is essential for propagating epigenetic information across cell divisions (*maintenance* methyltransferase) (Bird, 2002; Pradhan et al., 1999). DNMT3A and DNMT3B are also important for maintenance (Chen et al., 2003), however, their main role is to perform *de novo* methylation of unmethylated CpG sites and they are essential for setting up DNA methylation patterns in early development (Okano et al., 1999, 1998).

Methylcytosines are prone to mutations resulting to under representation of CpG dinucleotides in the mammalian genome (Bird, 1980; Scarano et al., 1967). These genome-wide CpG poor regions, are punctuated by *CpG islands* (CGIs) (Bird et al., 1985): around 29,000 short (from 200 bp up to a few kilobases) CpG-rich regions with elevated G+C base composition (Cross et al., 1994; Gardiner-Garden and Frommer, 1987). The mammalian genomes show a *bimodal* methylation pattern: almost all CpGs are methylated, except those located in CGIs which to a large extent are associated with absence of DNA methylation (Deaton and Bird, 2011; Illingworth and Bird, 2009). Approximately 70% of promoter regions are associated with a CGI — including virtually all housekeeping genes — so that promoters are usually classified as CpG poor and CpG rich (Saxonov et al., 2006). Hyper-methylation of CGIs near promoter regions is generally associated with transcriptional repression (Schübeler, 2015); however, outside of this well documented case, the association between DNA methylation across promoter-proximal regions and transcript abundance is considerably weaker and poorly understood (Jones, 2012; Varley et al., 2013). Figure 2.4 shows a cartoon with the role of epigenetic marks in regulating gene expression.

DNA methylation is highly dynamic and is thought to reset cell fates during early mammalian development. Right after fertilization, we observe a demethylation process, allowing

[16] In contrast to DNA replication, the error rates of DNA methylation are much higher: ∼1 error for every 25 methylated sites copied (Laird et al., 2004). This results in cell populations that have diverse methylation patterns (Silva et al., 1993).

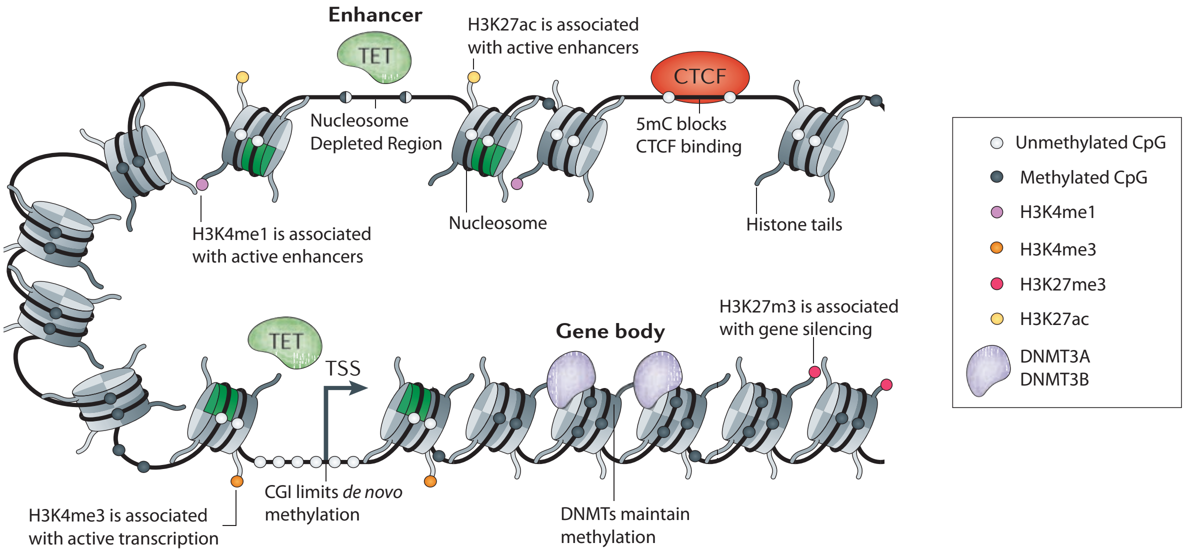

Enhancer
TET
H3K27ac is associated with active enhancers
CTCF
Nucleosome Depleted Region
5mC blocks CTCF binding
H3K4me1 is associated with active enhancers
Nucleosome
Histone tails
○ Unmethylated CpG
● Methylated CpG
H3K4me1
H3K4me3
H3K27me3
H3K27ac
DNMT3A DNMT3B
H3K27m3 is associated with gene silencing
Gene body
TET
TSS
H3K4me3 is associated with active transcription
CGI limits *de novo* methylation
DNMTs maintain methylation

Figure 2.4 Epigenetic marks and their role in gene expression. Figure adapted from Jones (2012).

the zygote to become totipotent (Eckersley-Maslin et al., 2018). Specifically, the maternal genome undergoes passive demethylation during cleavage divisions (Li, 2002), however, the paternal genome becomes globally unmethylated within hours of fertilization through active demethylation (Mayer et al., 2000; Oswald et al., 2000). This active demethylation process is predominantly mediated by the ten-eleven translocation (TET) family of proteins (Li, 2002), which catalyse the successive oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Kohli and Zhang, 2013). The exact role of these modifications is not clearly understood — mainly due to their low abundance — although significant enrichment of 5hmC is observed in *embryonic stem cells* (ESCs) and in the brain (Pastor et al., 2011).

CpG methylation is implicated in diverse biological processes of direct clinical relevance. The most well studied example is monoallelic X-chromosome inactivation in the female mammalian embryo (Avner and Heard, 2001; Mohandas et al., 1981). In this process, DNA methylation acts as cellular memory, since it does not intervene to silence active genes, but propagates the silent state leading to long-term inactivation of the associated gene (Bird, 2002; Lock et al., 1987). Also, it is associated with genomic imprinting[17] (Li et al., 1993; Reik and Walter, 2001), silencing of transposable elements[18] (Waterland and Jirtle, 2003) and more recently with carcinogenesis (Baylin and Jones, 2011; Thienpont et al., 2016). DNA methylation could regulate expression through several ways: (1) exclusion of proteins that affect transcription and chromatin structure, the clearest example is CTCF binding (Hark et al., 2000), (2) attraction

[17] Phenomenon in which certain genes are selectively inactivated from one of the two parental alleles.

[18] Sequences of DNA scattered throughout the genome that move from one genomic location to another; were first discovered by Nobel laureate Barbara McClintock in the 1940s.

of methyl-CpG binding proteins, such as MeCP2[19] (Lewis et al., 1992) , and (3) interaction with histone marks (Mutskov et al., 2002; Tamaru and Selker, 2001) which indirectly change the chromatin state. Also, recent genome-wide statistical analyses have shown that histone modifications and DNA methylation are highly correlated with DNA sequence motifs (Whitaker et al., 2015); again demonstrating the high interrelation between the different molecular layers[20].

2.3 Sequencing technologies

DNA sequencing refers to the process of determining the precise order of nucleotides of an organism's DNA. Until early-2000s the most widely used protocol for DNA sequencing was the *chain termination* method developed by Sanger et al. (1977), which provided the basis for the Human Genome Project (Lander et al., 2001; Venter et al., 2001); an international research project for determining the complete sequence of nucleotides that make up the human genome. Over the years, the Sanger biochemistry technique was substantially improved and semi-automated largely because of efforts to sequence the human genome. However, DNA sequencing using this technology — often termed as first-generation sequencing — had inherent limitations in scalability, costs and parallelisation (Mardis, 2008).

The advent of massively parallel DNA sequencing platforms in mid-2000s — coined as *next generation* sequencing (NGS) or *high throughput* sequencing (HTS) — allowed great advancements in biological and biomedical research, mainly due to improvements in cost, speed and sequencing parallelisation (Metzker, 2010). By reducing the sequencing costs more than two orders of magnitude compared to traditional Sanger biochemistry, DNA sequencing platforms have become a versatile tool on the hands of individual investigators, who can now routinely perform comprehensive analyses of genomes (Shendure and Ji, 2008). Due to the variety of NGS features, multiple platforms exist in the marketplace[21], each having clear advantages for particular applications over others. However, the sequencing workflow of the different platforms follows the same basic steps. First, the DNA is sonicated so it breaks up in smaller DNA fragments. Sequencing *libraries* are then prepared by ligating specific adaptor oligos to both ends of the DNA fragments. Importantly, little amount of input DNA is required to generate a library. Subsequently, the libraries are subjected to next generation sequencing where millions to billions of short *reads* (typically between 25 - 250 bp) are produced in a single run (Mardis,

[19] Genetic mutation of MeCP2 is associated to Rett syndrome (Amir et al., 1999).

[20] This rather long (and dense in biological jargon) detour of epigenetics does not intend to discourage the reader. On the contrary, the aim is to appreciate the complex interplay of these biological processes, which are highly coordinated in space and time. To better understand biological processes and eventually disease progression, these biological layers should be studied simultaneously. Due to limitations in both technological advancements and computational modelling, these layers are often studied independently — which is mostly the case for this thesis as well — hence results should be interpreted with caution, due to indirect effects from the various unobserved factors.

[21] Including Illumina, Roche/454, SOLiD and Complete Genomics. For excellent reviews comparing the different platforms see Metzker (2010) and Goodwin et al. (2016).

2008). Finally, depending on the application we map the sequenced reads on a *reference* genome or we perform *de novo* assembly for organisms whose genomes are unavailable (Zerbino and Birney, 2008). To obtain more reliable quantification of genetic information, one should perform deep sequencing which will result in greater *coverage* of the genome (Mardis, 2008).

In addition to genome analysis, modifications in the protocol design have allowed rapid development of diverse epigenome mapping assays — see figure 2.5 for a selection of sequencing protocols — at a scale and depth that was previously impossible. The basic idea is that prior to library preparation, the signal of the epigenetic mark is mapped to the DNA sequence by enriching for specific sequence fragments or chemically modifying the DNA sequence. The Encyclopaedia of DNA Elements (ENCODE) project (Dunham et al., 2012) is an international consortium, initiated by the National Human Genome Research Institute (NHGRI), for identifying and mapping the functional and regulatory DNA elements of the human genome. High quality epigenomic and transcriptomic data are generated using NGS technology from bulk populations of different cell types, including pluripotent, cancer and adult cell lines. Due to the variety of functional elements, an important challenge is to perform integrative modelling of the different molecular layers. In chapter 4, we will use specific cell lines from the ENCODE project for evaluating the proposed model for capturing spatial variability of methylation patterns (Kapourani and Sanguinetti, 2016).

The majority of NGS datasets is generated from bulk populations of millions of input cells, and the analysis of these populations provides us with average behaviours over cell ensembles (Shapiro et al., 2013). However, these ensemble-based approaches are insufficient to capture the diversity and heterogeneity of cells across different conditions, since even seemingly homogeneous cell sub-populations display transcriptional, epigenetic and phenotypic variability (Schwartzman and Tanay, 2015; Stegle et al., 2015). Recent advancements in sequencing technology have enabled the measurement of different biological layers at the single cell resolution, providing us with unprecedented information for studying cell identity and function in normal development and disease (Gawad et al., 2016).

2.3.1 Quantifying gene expression

RNA-seq is the method of choice for performing digital transcriptome profiling and provides far more precise measurements of gene expression compared to other methods (Marioni et al., 2008; Mortazavi et al., 2008). Initially, RNA is isolated from the tissue and to analyse the signal of interest the population of RNA might be further filtered, e.g. if we are interested for capturing mRNA, the protocol would involve poly(A) tail selection. Subsequently, the RNA is fragmented and reverse transcribed — by a reverse transcriptase enzyme — for the synthesis of complementary DNA (cDNA). Millions of short reads are then generated by sequencing the cDNA libraries using next generation sequencing (Pepke et al., 2009; Wang et al., 2009). The resulting reads are then aligned and mapped to a reference genome using efficient methods,

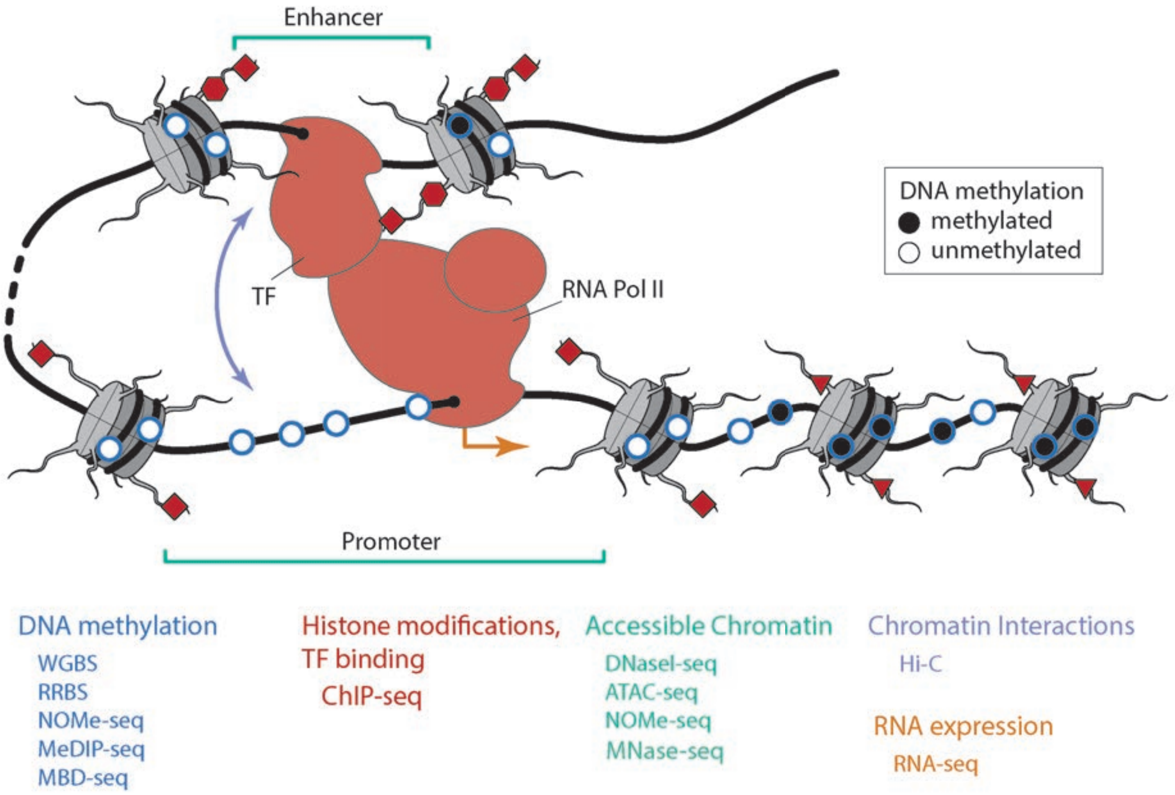


Figure 2.5 Sequencing technologies for quantifying epigenetic marks. Colours indicate different epigenetic marks with a selection of common sequencing assays used to quantify them. Figure adapted from Müller (2016).

such as *Bowtie* (Langmead et al., 2009) and *HISAT* (Kim et al., 2015) or gene expression can be quantified directly with alignment-free methods such as *Kallisto* (Bray et al., 2016) and *Salmon* (Patro et al., 2017). The total number of reads mapped to each gene acts as a proxy for measuring (relative) gene expression levels and statistical models use a normalised version of the count reads for downstream analyses (Aleksic et al., 2014; Garber et al., 2011; Robinson and Oshlack, 2010), including differential expression analysis between groups of samples (Anders and Huber, 2010; Robinson et al., 2010) or clustering similarly expressed genes (Si et al., 2013).

Recent advancements in the field have enabled profiling of single cells using RNA-seq (scRNA-seq) (Tang et al., 2009), which has already led to profound discoveries in biology research, ranging from identifying sub-populations of rare cell types and cellular states (Buettner et al., 2015; Zeisel et al., 2015), to reconstructing linage hierarchies (Treutlein et al., 2014), and to characterising cellular heterogeneity during embryo development and cancer (Patel et al., 2014). The experimental design is similar to bulk RNA-seq, however, the initial step requires physical isolation of individual cells. A widely used protocol is fluorescence-activated cell sorting (FACS) that isolates cells into microwell plates (Islam et al., 2011), whereas recent high-throughput[22]

[22] Here the term high-throughput refers to sequencing in parallel a large number of single cells.

approaches use microfludics and droplet systems (Macosko et al., 2015; Mazutis et al., 2013) to profile the transcriptome of thousands of single cells. Although we have seen an exponential scaling in profiling of single cells, this inevitably leads to obtaining low sequencing depths (due to prohibitive sequencing costs), where only a handful of gene transcripts are captured and quantified. Statistical approaches are being developed to fully exploit and interpret scRNA-seq data (Stegle et al., 2015), by appropriately normalising expression counts (Vallejos et al., 2017), disentangling technical from biological variability (Brennecke et al., 2013; Vallejos et al., 2015), accounting for amplification bias and batch effects[23] (Hicks et al., 2015; Islam et al., 2014), and modelling zero inflation due to capture inefficiencies and low sequencing depth (Pierson and Yau, 2015; Risso et al., 2018).

2.3.2 Analysing protein interactions with DNA

During the past few years there has been a remarkable progress in determining transcription factor binding sites and characterising histone modifications on a genome-wide scale. The main driving force is the chromatin immunoprecipitation followed by sequencing (ChIP-seq) protocol (Barski et al., 2007; Johnson et al., 2007). ChIP-seq is a general purpose assay enabling a genome-wide view of DNA-protein interactions *in vivo* (Park, 2009). During the ChIP process, the protein of interest is cross-linked to the DNA where it is bound. Subsequently, the chromatin is sonicated into small DNA fragments and the protein of interest is selectively enriched by immunoprecipitation using factor-specific antibodies. The cross-linking is subsequently reversed and oligonucleotide adaptors are attached to the resulting purified DNA to enable standard sequencing library construction (Park, 2009). A major challenge of ChIP-seq protocols is antibody specificity and sensitivity (Teytelman et al., 2013), however, well standardised protocols with high quality control have been established, especially for the immunoprecipitation of histone modifications (Landt et al., 2012).

The resulting sequenced reads are then mapped to the reference genome using standardised tools, and depending on the biological question different downstream analyses can be performed. The most common question is the localisation of the specific DNA-protein interactions, which is achieved using peak-calling algorithms (Zhang et al., 2008) that identify signal enrichment (peaks) relative to the background signal when omitting the immunoprecipitation step (Rozowsky et al., 2009). Another essential question is identifying differential binding patterns between two different conditions (e.g. before and after specific gene knock-out). To this end, most methods use the number of reads mapping to a specific region of interest (Ross-Innes et al., 2012); however, they ignore the shape of ChIP-seq peaks and recent powerful methods have been developed to account for the spatially distributed patterns of ChIP-seq data (Schweikert et al., 2013).

[23] Systematic biases that frequently arise from undesirable (and often unrecognised) differences in sample processing.

2.3.3 Charting the DNA methylome

The gold-standard method for the detection of cytosine DNA methylation on a genome-wide scale at the single nucleotide resolution is bisulfite treatment of DNA followed by high-throughput sequencing (BS-seq)[24] (Frommer et al., 1992). Bisulfite treatment efficiently converts unmethylated cytosines to uracils, while methylated cytosines remain intact. Uracils are subsequently read as thymines by DNA polymerase, hence, after polymerase chain reaction (PCR) amplification, the unmethylated cytosines appear as thymines (Krueger et al., 2012; Laird, 2010). To obtain the methylation information of each cytosine, reads are mapped back to a reference genome allowing changes of cytosines to thymines during the mapping procedure (Krueger and Andrews, 2011). Reads containing a thymine where the reference in that molecule contains a cytosine indicate that the cytosine was unmethylated, whereas reads containing a cytosine indicate that the cytosine in the reference genome was methylated (Schultz et al., 2012).

Different protocols employing the same principle of bisulfite treatment have been proposed. Of these, the most widely adopted is whole genome bisulfite sequencing (WGBS) which in theory can assay the whole methylome landscape of around 28 million CpGs in the human genome (Lister et al., 2009; Ziller et al., 2013). To additionally obtain an accurate estimate of methylation level in each CpG site, an average of more than 20 reads per CpG is often required, making the WGBS technology rather expensive for individual research groups. A variant of BS-seq technology, termed reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2005, 2008), uses methylation-sensitive restriction enzymes, such as *MspI* that recognises CCGG motifs, to cleave the DNA at genomic regions with high CpG content prior to size selection and bisulfite treatment. This results in measuring in greater coverage and at lower cost the methylation level of around 10% of total CpGs, which however predominantly reside near promoter regions and CGIs. Figure 2.6 provides a schematic comparison between WGBS and RRBS protocols.

Bisulfite treatment has damaging effects to the DNA, hence large amounts of input DNA from a population of cells is often required for performing BS-seq experiments (Bock, 2012). Although each cytosine from a single cell can either be methylated or unmethylated, when surveying the bulk population we will obtain heterogeneous reads for each CpG site. The amount of methylation at each CpG is referred to as the *methylation level* or *methylation rate*, and expresses the fraction of methylated cytosines out of the total reads covering the specific site (Schultz et al., 2012). Based on methylation levels at each CpG site, analysis tasks may include data visualisation to identify global changes in distribution of DNA methylation, e.g.

[24] A plethora of additional protocols have been proposed for measuring DNA methylation, including methylated DNA immunoprecipitation (MeDIP-seq) and methyl binding protein enrichment (MBD-seq and MethylCap-seq). Comparative analyses show high agreement between the competing protocols, however, they have substantial differences in CpG coverage and experimental costs (Bock, 2012; Harris et al., 2010). This thesis focuses on modelling bisulfite sequencing data and excellent reviews on the additional protocols can be found in Laird (2010) and Li et al. (2010).

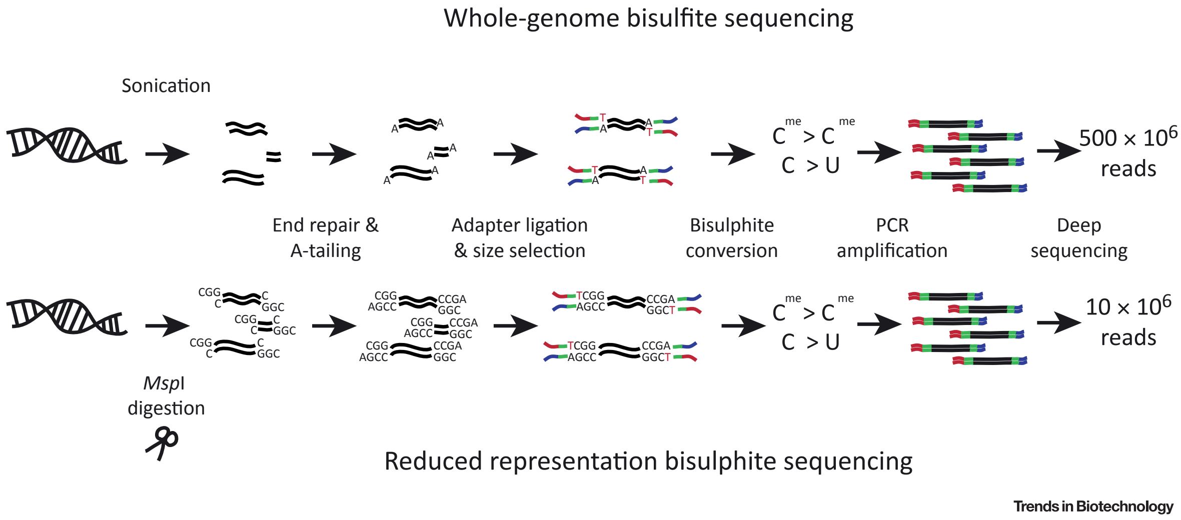


Figure 2.6 Schematic comparison of WGBS (top) and RRBS (bottom) protocols. Figure adapted from Karemaker and Vermeulen (2018).

aberrant hyper-methylation in cancer cells (Smiraglia et al., 2001; Sproul et al., 2011), and identify differences between groups of samples by differential analysis (Bock, 2012). There are a number of statistical methods for calling differentially methylated loci (DML), see Robinson et al. (2014), and most of them rely on modelling the count methylation data using the beta-binomial distribution that accounts for *overdispersion*[25] (Dolzhenko and Smith, 2014; Feng et al., 2014).

Although single CpG sites may be biologically relevant (Xu et al., 2007), researchers are often interested in computing the methylation level over a genomic region. The most common way to summarise the methylation signal over a genomic region is to take the arithmetic mean of methylation levels at sites within the region, which is termed as *mean methylation level* (Schultz et al., 2012). Using segmentation algorithms and thresholding (Burger et al., 2013), researchers have identified low-methylated regions (LMRs) (Stadler et al., 2011) and partially methylated domains (PMDs) (Gaidatzis et al., 2014), whereas a plethora of statistical methods have been proposed to detect differentially methylated regions (DMRs) (Hansen et al., 2012; Mayo et al., 2015; Rackham et al., 2017), which are believed to be associated with cell-type specific transcriptional activity of the associated genes (Bock, 2012). Most studies use DMR detection as a pre-filtering step, and then correlate mean methylation levels across each region with gene expression (Bock et al., 2012; Hansen et al., 2011). However, using this simplistic encoding of DNA methylation as a simple average, (1) we cannot capture the complex patterns of methylation in a given region, and (2) we do not fully exploit the richness of BS-seq data that provide us with single nucleotide information. In chapter 4 we propose *BPRMeth*, a probabilistic model for capturing and quantifying spatial variation in DNA methylation

[25] Overdispersion is common in biological experiments and occurs due to multiple sources of variation: technical variation due to error measurements coming from the experiment design and biological variation between the subjects of interest.

patterns, and we show that promoter-proximal *methylation profiles* are highly correlated with gene expression levels (Kapourani and Sanguinetti, 2016).

Single cell bisulfite sequencing

Although bulk BS-seq experiments have opened the way for mapping the epigenetic landscape, they fall short of explaining the epigenetic variability and quantifying their dynamics, which inherently occur at the single cell level (Schwartzman and Tanay, 2015). In a recent study, Landau et al. (2014) exploited the single molecule nature of BS-seq experiments to study the intratumour methylation variation in chronic lymphocytic leukaemia and their findings suggest a higher intrasample variability of DNA methylation patterns across the genome of malignant cells. Using RRBS experiments they observed two methylation patterns: (1) concordant reads with distinct methylation states across cell populations and (2) discordant reads with locally disordered methylation within cells. This study demonstrated that bulk approaches are insufficient to capture the actual epigenetic heterogeneity of cells across different conditions.

This shortcoming has been addressed within the last five years through the development of protocols that allow profiling of DNA methylation at the single cell resolution. Due to the bisulfite-induced DNA degradation, BS-seq protocols were initially prohibitive for small amounts of input DNA (Karemaker and Vermeulen, 2018), however, modifications of this technology have improved the recovery rate allowing for genome-wide coverage (Schwartzman and Tanay, 2015). The first single cell method, called scRRBS, was based on enrichment of CpG dense regions (Guo et al., 2013). Later, whole genome approaches, including scBS-seq (Clark et al., 2017; Smallwood et al., 2014) and scWGBS (Farlik et al., 2015), were developed based on post-bisulfite adapter-tagging (PBAT) (Miura et al., 2012), in which bisulfite conversion precedes library preparation so that DNA degradation does not destroy the adaptor-tagged fragments (Clark et al., 2016). Recent protocols allow for high-throughput measurement of single cells, snmC-seq (Luo et al., 2017) and sci-MET (Mulqueen et al., 2018), which have demonstrated the ability to identify neuronal subtypes from thousands of single cell methylomes[26]. Single cell methylome analysis will allow epigenomic studies of small population of cells or niches, including early mammalian development (Guo et al., 2014; Zhu et al., 2018) and patient derived samples; an attractive target of DNA methylation since it provides cell specific information that is more stable than gene expression profiles (Clark et al., 2016).

Due to sequencing costs and small amounts of input DNA per cell, these protocols usually result in really sparse genome-wide CpG coverage, normally ranging from 5% to 20%. The sparsity of the data represents a major hurdle for effectively using them to inform our understanding of epigenetic control of transcriptome variability, or to distinguish individual cells based on their epigenomic state. In chapter 6 we introduce *Melissa*, a Bayesian model for jointly

[26] The interested reader should consult Karemaker and Vermeulen (2018) for a comprehensive review on the existing single cell methylome profiling protocols.

imputing missing methylation states and clustering single cells based on their methylation profiles (Kapourani and Sanguinetti, 2018b).

Methylation microarrays

Bisulfite treatment combined with highly specialised microarrays are also widely used for assaying DNA methylation levels. This array-based technology quantifies methylation levels of pre-specified sequences containing CpGs, which are represented by dedicated probes on the microarray. The most widely adopted microarray technology is the Illumina Infinium 450K assay (Bibikova et al., 2011), which covers more than 450,000 CpGs (mainly near CGIs), whereas the most recent release of the platform, Infinium EPIC (Moran et al., 2016), covers more than 850,000 CpG methylation sites. The additional sites include regulatory regions identified by the ENCODE project. The main limitations of these technologies are the limited CpG coverage and relatively high technical bias (Lee, 2011; Teschendorff et al., 2011); however, the low experimental costs makes them an indispensable tool for measuring DNA methylation in large sample cohorts, e.g. when conducting epigenome wide association studies (EWAS) (Lappalainen and Greally, 2017). Signal intensities are measured for each probe and are summarised by β-values (or M-values), which are the average signal of methylated and unmethylated CpGs (Bock, 2012). In chapter 5, the BPRMeth model is extended to support data measured by methylation array platforms (Kapourani and Sanguinetti, 2018a).

2.3.4 Assessing accessible chromatin

Identification of accessible regions in a genome-wide scale is crucial for understanding gene regulation (Bai et al., 2010; Taberlay et al., 2014) and indirectly inferring TF binding sites (Pique-Regi et al., 2011). Most chromatin accessibility protocols rely on separating the genome by chemical means to isolate either the accessible or occupied regions, and subsequently quantify the isolated DNA using next generation sequencing (Tsompana and Buck, 2014). The most common protocols are DNaseI-seq (Thurman et al., 2012) and MNase-seq (Ponts et al., 2010), which involve digestion of DNA with the respective nuclease — deoxyribonuclease I and micrococcal nuclease, respectively — allowing for high resolution profiling of nucleosome free regions (DNaseI-seq) or TF and nucleosome occupancy regions (MNase-seq). In its quest to identify functional elements of the genome, the ENCODE project used extensively the DNaseI-seq protocol to identify DNaseI hypersensitive sites (DHS) (Thurman et al., 2012). A more recent protocol, termed assay for transposase accessible chromatin (ATAC-seq) (Buenrostro et al., 2015a), uses the hyperactive Tn5 transposase to fragment the DNA and attach adapter sequences in nucleosome free regions.

An alternative approach that simultaneously measures chromatin accessibility and endogenous DNA methylation, is nucleosome occupancy and methylation sequencing (NOMe-seq) (Kelly et al., 2012), where the M.CviPI methylase is used to methylate exposed GpC dinucleotides[27] while nucleosome protected DNA remains unmodified. Standard BS-seq protocols can then be used to identify nucleosome occupancy at the single nucleotide resolution. An attractive feature of NOMe-seq is that we can recover endogenous DNA methylation in parallel[28], and in contrast to count-based assays, such as ATAC-seq and DNaseI-seq, we can directly discriminate inaccessible chromatin from missing data (Clark et al., 2018). Due to its simple and fast protocol and its high sensitivity to low cell numbers, scATAC-seq (Buenrostro et al., 2015b; Cusanovich et al., 2015) is widely used for high throughput single cell profiling; and recently scNOMe-seq (Pott, 2017) has been applied to provide high resolution mapping of chromatin accessibility and DNA methylation at the single cell level. In addition to assessing accessible chromatin, it is also possible to capture large-scale chromosomal conformation using HiC-based methods in both bulk (Lieberman-Aiden et al., 2009) and single cells (Nagano et al., 2013).

2.4 Single cell multi-omics

Single cell sequencing technology provides information that is not confounded by the phenotypic heterogeneity of bulk technologies (Macaulay et al., 2017). However, to better understand the dynamic cellular behaviour in health and disease and link transcriptional and (epi)genetic heterogeneity, we need to simultaneously take measurements of multiple molecular types and develop efficient integrative methods for combining these modalities (Argelaguet et al., 2018). Single cell multi-omics platforms have recently emerged as a powerful approach to simultaneously investigate the dynamic coupling between different biological layers — including the genome, transcriptome and epigenome — at the single cell resolution.

One of the earliest multi-omics protocols is single cell genome and transcriptome sequencing (G&T-seq) (Macaulay et al., 2015). In G&T-seq, following full cell lysis, poly-A mRNA molecules are physically separated from the DNA and subsequently each feature is amplified and sequenced by conventional single cell methods. Extensions to the G&T-seq protocol allow for single cell methylome and transcriptome sequencing, scM&T (Angermueller et al., 2016) and scMT-seq (Hu et al., 2016), by treating with bisulfite the isolated DNA prior to amplification, while the captured mRNA is sequenced as before. The scTrio-seq protocol (Hou et al., 2016) takes into account copy number variations (CNVs)[29], in addition to transcriptome and methylome profiling. Also, scCOOL-seq (chromatin overall omic-scale landscape sequencing) (Guo et al., 2017) relies on the scNOMe-seq protocol to assay nucleosome positioning, DNA methylation, ploidy and CNVs from the same individual cell. Finally, the scNMT-seq (nucleosome, methylome

[27] Note that DNA methylation mostly occurs in the CpG context, hence this protocol allows us to later distinguish between the two epigenetic marks.

[28] The only issue being that during read alignment, G-C-G and C-C-G positions should be discarded due to inability to distinguish endogenous methylation from *in vitro* methylation and off-target effects of the enzyme, respectively.

[29] Variation in the genome in which sections of the DNA are repeated, with the number of repeats varying between individuals.

and transcriptome sequencing) protocol (Clark et al., 2018) enables parallel profiling of DNA methylation, transcription and chromatin accessibility and can reveal dynamic coupling between epigenomic layers in differentiating mouse embryonic stem cells. Briefly, single cells are lysed and accessible DNA is labelled using GpC methyltransferase. Subsequently, DNA and RNA are physically separated from a fully lysed cell, and conventional scRNA-seq protocols are used for transcriptome profiling, while the DNA is assayed using the scNOMe-seq protocol to measure DNA methylation and chromatin accessibility. Details of the scNMT-seq study are provided in chapter 5, together with extensive analysis on linking epigenetic heterogeneity with transcription activity using the *BPRMeth* model. Although these methods are promising to revolutionise our understanding of complex biological systems, computational and statistical models incorporating domain knowledge for data integration are much needed, since the dynamics of these pathways operate on different biological time scales (Kelsey et al., 2017).

Chapter 3

Probabilistic machine learning

This chapter provides a high level introduction to machine learning, with emphasis on the probabilistic approach to machine learning (section 3.1) which is closely related to the field of computational statistics. Section 3.2 explains how the Bayesian paradigm provides a natural approach for learning from data and section 3.3 introduces probabilistic graphical models, a powerful framework for compactly representing complex distributions. Bayesian approaches are appealing, however, they are often intractable and approximate algorithms are often sought to perform inference (section 3.4). The reader is assumed to have some familiarity with basic concepts in probability theory; if not, please consult Feller (1968) for a refresher.

3.1 Statistical modelling

The vast amounts of data generated across different fields, including biological systems, combined with the increased computational resources has led to a surge of interest in methods that can automatically uncover patterns in data, which ideally could be related to some true underlying mechanisms. The field of machine learning is concerned with developing algorithms that can learn from data and subsequently perform different kinds of decision tasks, such as making predictions of as yet unobserved quantities given our current observations (Ghahramani, 2015). Machine learning is closely related to different fields — including computer science, signal processing, physics, and statistics — and despite the current hype most of the core ideas existed a long time ago[1] (Friedman et al., 2001).

Despite the wealth of 'omics' data generated from next generation sequencing platforms, biological systems are far too complex for the modeller to design an accurate representation of the data generation process; hence, computational biology systems can be still thought of as relatively data poor (Lawrence et al., 2010). In practice we need to abstract the real system

[1] The statistician Rob Tibshirani has created a glossary comparing machine learning and statistics terminology, available at http://statweb.stanford.edu/~tibs/stat315a/glossary.pdf. However, it should be noted that machine learning focuses predominantly on predictions rather than understanding.

in a mathematical model, which should both be flexible enough to capture the regularities of the data and make accurate predictions, and constrained to allow us to answer hypotheses about the system. This restriction on the class of models used to explain the system leads to *model compromise* (Lawrence et al., 2010). One should not forget the words of George Box, "*all models are wrong, but some are useful*" (Box et al., 1978), and that modelling in science partly remains an art[2] and requires domain knowledge (McCullagh and Nelder, 1989). In addition to having complex systems, the observations obtained from biological experiments are often noisy, either intrinsic or extrinsic, and incomplete (Elowitz et al., 2002; Kærn et al., 2005). This introduces *uncertainty* at the input level which should be propagated in a consistent way during model predictions (Oakley and O'Hagan, 2004). A principled statistical way for formulating the problem scientifically and quantifying our knowledge about uncertainty is the Bayesian paradigm, that uses nothing but the rules of probability (see below). An additional compromise that we often make during statistical modelling is an *inference compromise* (Lawrence et al., 2010), which is prevalent in Bayesian statistics when combined with relatively complex mathematical models leading to intractable computations. This computational burden is one of the main factors preventing the adoption of Bayesian methods in large-scale applications (Bishop, 2006).

Machine learning models fall broadly into two main categories. In the *supervised* or *predictive* learning approach, given a training dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ comprising of N observations, the goal is to learn a mapping from the input variables $\mathbf{x}$ to the output variables y. We can formalise this problem as function approximation, where we assume that $y = f(\mathbf{x} \,|\, \boldsymbol{\theta})$ for some unknown (and often) parametrised[3] function f, and during inference we use the training data $\mathcal{D}$ to estimate the precise form of f, that is encoded in the set of parameters $\boldsymbol{\theta}$. Using the trained model, we can then make predictions for y^* for any new value of $\mathbf{x}^*$ present in a test set. From a probabilistic perspective we aim to model the distribution $p(y \,|\, \mathbf{x}, \boldsymbol{\theta})$[4] which takes into account the uncertainty about the value of y given the input $\mathbf{x}$. The input variables $\mathbf{x}$, also known as *covariates, features*, or *independent* variables, are often represented as D-dimensional vectors of numbers $\mathbf{x} = (x_1, \ldots, x_D)^\top$, although in general they can represent any complex object, including images, molecular shape, graphs, and time series. The form of the output variable y, also known as *response* or *dependent* variable, can be equally

[2] Scientists, as well as artists, should restrain themselves from falling in love with one model and excluding alternatives — also known as *no free lunch* theorem (Wolpert, 1996).

[3] Probabilistic models that have a fixed number of parameters are called *parametric*, whereas when the number of parameters grows with the data we assume *non-parametric* models (Ferguson, 1973), with the most notable example being Gaussian Processes (Rasmussen and Williams, 2006). This thesis focuses on parametric models, including generalised linear models and finite mixture models (see below).

[4] Notes on notation. We will regularly use the shorthand $p(x)$ to denote $P(X = x)$, the probability that a random variable X takes on the value x, except when the distinction is necessary. Also, we will not distinguish between discrete probabilities and probability densities. This notation might lead to ambiguities, however, we will avoid a cumbersome notation throughout the thesis. The *expectation* or average of a function $f(x)$ under the distribution $p(x)$ is denoted by $\langle f(x) \rangle_{p(x)} \overset{\text{def}}{=} \int f(x)\, p(x)\, dx$. For discrete x, we simply need to replace the integral over x with a sum. Integrals (or summations) without a range denote operations 'over all values' of x. Similarly the *variance* is denoted by $\text{var}\,[f(x)]_{p(x)} \overset{\text{def}}{=} \left\langle \left(f(x) - \langle f(x) \rangle_{p(x)}\right)^2 \right\rangle_{p(x)}$.

complex; however, in most approaches we assume that y is a categorical variable — in which case the model is used for *classification* — or a continuous variable — in which case we perform *regression*.

A classification example is to predict whether histone marks are modified or not, $y \in \{0, 1\}$, using as input features $\mathbf{x} \in \mathbb{R}^D$ the binding of transcription factors, DNA methylation, and possibly DNA sequence around the genomic regions of interest, e.g. see Benveniste et al. (2014). In a regression setting one might be interested in predicting gene expression levels, $y \in \mathbb{R}$, using DNA methylation patterns or histone marks as input features, e.g. see Kapourani and Sanguinetti (2016) and Dong et al. (2012). In many applications, we might need to transform the input features $\mathbf{x}$ to some new space where we expect the statistical model to capture richer information present in the data. This process is often called *feature extraction* and is used by the BPRMeth model in chapter 4 to extract higher order methylation features that quantitate the shape of DNA methylation patterns.

The second class of machine learning models is the *descriptive* or *unsupervised* learning approach. In this setting, the training dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ consists only of input vectors $\mathbf{x}$ without any corresponding response variables, and the goal is to discover interesting structure in the data beyond what would be considered unstructured noise (Ghahramani, 2004); that is, we need to build models of the form $p(\mathbf{x} \,|\, \boldsymbol{\theta})$. Note that in this setting we have a more difficult task, since we are not told what patterns to look for and generally we need to provide a multivariate probabilistic model of the input vector $\mathbf{x}$. The goal in unsupervised learning may be to determine the distribution of the input data, a task known as *density estimation*, or to identify groups of observations that have similar patterns using *clustering*. A notable example of clustering in computational biology is to identify cell sub-populations based on gene expression levels (Zeisel et al., 2015) or epigenetic marks (Kapourani and Sanguinetti, 2018b), as we propose in chapter 6. Another application of unsupervised learning is *dimensionality reduction*, where one projects the data from a high dimensional space to a lower dimensional subspace while still preserving important features of the data. This approach is prevalent in single cell biology, where for each cell we might have input features $\mathbf{x} \in \mathbb{R}^{20000}$, denoting the transcriptome state of each cell, e.g. see Pierson and Yau (2015).

3.1.1 Generalised linear models

Generalised linear models (GLMs) — originally proposed by Nelder and Wedderburn (1972) — form one of the cornerstones of probabilistic modelling and provide an elegant framework for unifying diverse statistical methods for regression and classification. GLMs extend the classical linear model for regression developed by Adrien-Marie Legendre in 1805 and later by Carl Friedrich Gauss (Gauss, 1809), who derived the Normal or Gaussian distribution and the method of least squares (Stigler, 1981).

The standard linear regression model assumes a linear mapping from D-dimensional input features $\mathbf{x}_n$ to response variables y_n using a set of unknown parameters $\boldsymbol{\theta}$ — also called regression coefficients. We also assume that the observations y_n are a realisation of an unobserved random variable that corrupts the linear relationship between y_n and $\mathbf{x}_n$. The linear regression model then takes the form

$$\begin{aligned} y_n &= \eta_n + \epsilon, \\ \eta_n &= \boldsymbol{\theta}^\top \mathbf{x}_n, \\ \mu_n &= \langle y_n \rangle_{p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})} = \eta_n, \\ \epsilon &\sim \mathcal{N}(\epsilon \,|\, 0, \sigma^2), \end{aligned} \tag{3.1}$$

where η_n denotes the *systematic* component (or linear predictor), ϵ denotes the *random* component (or error term) that follows a Normal distribution, $\boldsymbol{\theta}^\top \mathbf{x}_n$ is the inner or dot product between the vectors $\boldsymbol{\theta}$ and $\mathbf{x}_n$, and $\langle y \rangle_{p(x)}$ denotes the expectation of y with respect to the distribution $p(x)$. A common assumption of linear regression — as well as GLMs — is that the error terms ϵ are independent and identically distributed (i.i.d.) and have constant variance across response variables y_n (McCullagh and Nelder, 1989). In the linear model, the systematic component is identical to the expected value of the response variable, thus we can write $p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta}) = \mathcal{N}(y_n \,|\, \eta_n, \sigma^2)$.

In linear models we implicitly assumed that the dependent variable y_n — or to be precise, its expected value μ_n — can take any value in the real line, i.e. $\mu_n \in \mathbb{R}$. However, when the dependent variables are instantiations of count or proportion data, this assumption does not hold. Generalised linear models relax the *normality assumption* by letting $p(y_n \,|\, \mathbf{x}_n, \boldsymbol{\theta})$ to be any distribution in the *exponential family*[5] (Andersen, 1970). The formulation of linear models using (3.1) allows us to simplify the transition to GLMs; first the random component term may

[5] The exponential family has some appealing mathematical properties making them broadly applicable: (1) they can be summarised by a finite number of sufficient statistics, (2) they are the only family for which conjugate priors exist, which are useful for simplifying computations in Bayesian inference, and (3) the most common probability distributions — including the Normal, binomial, Bernoulli, and Poisson — are part of the exponential family (Murphy, 2012). The exponential family of distributions over x given parameters θ is of the form

$$p(x \,|\, \theta) = \frac{1}{Z(\eta(\theta))} h(x) \exp\big(\eta(\theta)\phi(x)\big).$$

Here θ are called the natural or canonical parameters, $\phi(x)$ is called a vector of sufficient statistics, $h(x)$ is a scaling constant (often set to 1), and $Z(\eta(\theta))$ is the partition function ensuring that the distribution is normalised. If $\eta(\theta) = \theta$ then the distribution is said to be in the canonical form. For a concise example, the Bernoulli distribution for $x \in \{0, 1\}$ can be written as

$$\mathcal{B}\text{ern}(x \,|\, \mu) = \mu^x (1-\mu)^{1-x} = (1-\mu) \exp\left(x \log \left| \frac{\mu}{1-\mu} \right|\right),$$

where $\theta = \log \left| \frac{\mu}{1-\mu} \right|$ is called the *log-odds ratio*, $\phi(x) = x$, and $Z = \frac{1}{1-\mu}$. We can then obtain the mean parameter from the canonical parameter using

$$\mu = \sigma(\theta) = \frac{1}{1 + \exp(-\theta)},$$

where $\sigma(\cdot)$ denotes the *sigmoid* or *logistic* function. Under this framework, the exponential family allows us to develop general purpose algorithms, e.g. see Hoffman et al. (2013).

Response	Mean	Link name	Link function	Mean function	Regression
Real	$\mu \in \mathbb{R}$	identity	μ	η	linear
Binary	$\mu \in [0, 1]$	logit	$\log \left\| \frac{\mu}{1-\mu} \right\|$	$\sigma(\eta)$	logistic
Binary	$\mu \in [0, 1]$	probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	probit
Count	$\mu \in \mathbb{R}_+$	log	$\log \|\mu\|$	$\exp(\eta)$	Poisson

Table 3.1 Link functions for common generalised linear models. The binary response includes the Bernoulli, binomial, and beta observation models, and $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution.

come from an exponential family, and secondly we make the mean of the distribution to be some invertible monotonic differentiable function of the systematic component, so

$$\mu_n = g^{-1}(\eta_n),$$

where $g(\cdot)$ is called the *link function* that allows us to move from the systematic components η_n to mean parameters μ_n. The inverse of the link function is often called the *mean function.* We are free to choose any link function so long as it is invertible and g^{-1} has the appropriate range. Table 3.1 shows common GLMs with some of the corresponding link functions.

3.2 Learning from data — the Bayesian paradigm

There are two broad interpretations of the concept of probability, the *frequentist* or *classical* and *Bayesian* or *evidential* probabilities. The frequentist interpretation — as the name implies — views probabilities in terms of the frequencies of random, repeatable events (Bishop, 2006). On the other hand, in the Bayesian view the probability is interpreted as a reasonable expectation representing our current knowledge (Cox, 1946) or as a quantification of somebody's coherent beliefs[6] (De Finetti, 1974). The Bayesian paradigm is intuitive and compelling when we want to learn from data: we explicitly quantify our uncertainty through some prior knowledge, then we revise this uncertainty in the light of new evidence, and finally we use the updated knowledge to make optimal decisions. As it will become apparent, the whole process — which is known as *Bayesian inference* — depends on using nothing but the rules of probability[7].

[6] The term Bayesian refers to the clergyman and mathematician Thomas Bayes, who first derived Bayes' theorem (using uniform priors), however, it was Pierre-Simon Laplace who introduced the more general form of the theorem and demonstrated the wide applicability of what is now called Bayesian probability (Stigler, 1986). There are different foundational views of Bayesian probabilities as well, mainly divided in *objectivist* and *subjectivist*, with the main practical difference being the construction of prior probabilities. The philosophically inclined reader should consult Dawid (2004) and Savage (1971). In this thesis we take a subjectivist view of Bayesian probability in the sense that it is a feature of our description of the world (Nau, 2001).

[7] For two random variables x and y, the *sum rule* states that $p(x) = \int p(x, y)\, dy$, and the *product rule* states that $p(x, y) = p(x \,|\, y)p(y)$. The quantity $p(x, y)$ is called the 'joint probability of x and y', $p(x \,|\, y)$ is the 'conditional probability of x given y', and $p(x)$ is called the marginal probability or simply the probability of x.

3.2.1 Bayesian inference

To be more precise, the Bayesian paradigm involves specifying our uncertainty about some variables or model parameters θ before observing any data, in the form of a *prior* distribution $p(\theta)$. Given observed data $\mathcal{D}$, we express the probabilistic relationship between the parameters and the data through the *likelihood* function $p(\mathcal{D} \,|\, \theta)$, also called the *sampling model*[8], which expresses the plausibility of the different values of parameters θ given the observed data. Using Bayes' theorem we combine these quantities to obtain the *posterior* distribution $p(\theta \,|\, \mathcal{D})$, that provides our updated uncertainty in θ after incorporating the information from the data

$$p(\theta \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)\, p(\theta)}{p(\mathcal{D})}. \tag{3.2}$$

Here the denominator $p(\mathcal{D})$ is a normalisation constant ensuring that $p(\theta \,|\, \mathcal{D})$ is a valid probability distribution and integrates to one. This quantity is often called the *evidence* or *marginal likelihood*, since we marginalise out, or integrate over, all the parameters of the model

$$p(\mathcal{D}) = \int p(\mathcal{D} \,|\, \theta)\, p(\theta)\, d\theta = \langle p(\mathcal{D} \,|\, \theta) \rangle_{p(\theta)} \,. \tag{3.3}$$

In words, we can state Bayes' theorem as: $\textit{posterior} \propto \textit{likelihood} \times \textit{prior}$, which in a sense resembles the scientific process itself, we start off with some initial models to explain observed phenomena and then we update our model accordingly after making observations. This concept of updating our models and beliefs arises rather naturally in the Bayesian paradigm, since our posterior estimates from one task can be encoded in the prior when the model is updated with more data (Bishop, 2006). Hence, the posterior summarises what we have learnt so far from the data $\mathcal{D}$, and we can use it to take optimal decisions, such as making predictions for new points $\mathbf{x}^*$ using the *posterior predictive* distribution

$$p(\mathbf{x}^* \,|\, \mathcal{D}) = \int p(\mathbf{x}^* \,|\, \theta)\, p(\theta \,|\, \mathcal{D})\, d\theta = \langle p(\mathbf{x}^* \,|\, \theta) \rangle_{p(\theta \,|\, \mathcal{D})} \,, \tag{3.4}$$

that gives us a probability distribution over $\mathbf{x}^*$. Note that by marginalising out the parameters θ, instead of substituting a point estimate, e.g. posterior mean, allows us to propagate our uncertainty in the model parameters when making predictions. This process of consecutive integrations over model parameters — as opposed to parameter optimisation — allows Bayesian approaches to avoid *overfitting* and generalise well on unseen data[9].

In some sense it seems that our work is done since all our modelling assumptions are encoded through the likelihood and the prior. Unfortunately, there is one key practical issue: in most cases computing the evidence, and consequently the posterior, and the predictive

[8] If we think of the process as a *generative model*, then the name makes sense since given a set of parameters θ we can sample, or generate, observations $\mathbf{x}$ from $p(\mathcal{D} \,|\, \theta)$.

[9] Note the similarity between equations (3.3) and (3.4), which leads to often referring to (3.3) as the *prior predictive*, since it allows us to make predictions without seeing any data!

distribution is intractable — even for tractable likelihood and prior distributions — since it involves integrations in high dimensional space. Conjugate priors should be used, if possible, to achieve tractable inference, however for most interesting models it is difficult to obtain conjugacy[10]. Dealing with this computational burden is the main challenge of Bayesian inference, and section 3.4 introduces approximate inference algorithms for tackling this intractability.

Bayesian modelling is widely applied in computational biology, since it allows us to incorporate biological prior knowledge either from earlier experiments or based on information in the literature (Beaumont and Rannala, 2004). For instance, Angus et al. (2010) proposed a general Bayesian framework for reconstructing gene regulatory networks from microarray time series gene expression data. To model gene-gene interactions across time and infer the gene regulatory network they used a subclass of dynamic Bayesian networks (see later). Subsequently, they introduced constraints in the network structure that reflect biological prior knowledge either from the literature or from different experimental data such as ChIP-on-chip — the precursor of ChIP-seq for microarray experiments — to capture transcription factor and gene interactions. Similar approaches of incorporating prior information for modelling transcriptional regulation were proposed by Sanguinetti et al. (2006), Sabatti and James (2005) and others. In the single cell paradigm, Huang and Sanguinetti (2017) performed transcriptome-wide splicing quantification from scRNA-seq data by learning an informative prior distribution from sequence features, which was then incorporated in a Bayesian model. In addition to incorporating prior knowledge in the model, the Bayesian framework admits a probability distribution over all quantities of interest, representing our uncertainty about their value. For instance, Vallejos et al. (2015) developed an integrated Bayesian method for modelling scRNA-seq data, which accounts for the high experimental noise and propagates uncertainty when computing quantities of interest. This allows to confidently detect lowly and highly variable genes within a population of cells (Vallejos et al., 2015), as well as identify genes undergoing changes in cell-to-cell heterogeneity across populations (Vallejos et al., 2017).

3.2.2 Maximum likelihood estimation

The most common approach for parameter estimation is *maximum likelihood estimation* (MLE). Under this approach we seek the value of θ for which the probability of the observed data, encoded by the likelihood function $p(\mathcal{D} \,|\, \theta)$, is maximised

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\; p(\mathcal{D} \,|\, \theta). \tag{3.5}$$

The assumption that the likelihood function encodes all the relevant information for estimating the parameters θ is known as the likelihood principle (Berger and Wolpert, 1988; Birnbaum, 1962). In most problems this quantity is typically cheap to compute, since we are interested

[10] By conjugacy we mean that for a given form of the likelihood, we seek a prior which gives a posterior distribution that has the same functional form as the prior.

in obtaining point estimates for θ instead of performing integrations as in Bayesian inference. Also, the whole process can be cast as an optimisation problem, for which efficient methods already exist. In the machine learning literature the negative log-likelihood function is known as the error function.

Note that both in the Bayesian and frequentist approaches the likelihood function $p(\mathcal{D} \mid \theta)$ plays a central role. Indeed, in the limit of large observation $N \rightarrow \infty$, and under mild conditions, the Bernstein-Von Mises theorem states that the posterior distribution of the parameters is asymptotically independent of the prior and converges to a $\mathcal{N}(\theta^{*}, 1/N)$ (Freedman, 1963). When the likelihood model is correct, θ^{*} coincides with the MLE since the likelihood overwhelms the prior, however, in the scarce data regime the Bayesian approach of incorporating prior information is essential for effective modelling. Although both frequentist and Bayesian methods may yield similar results, they make fundamentally different assumptions. In the frequentist paradigm, θ is fixed, and estimates are obtained by considering the distribution of all possible datasets $\mathcal{D}$. In contrast, the Bayesian viewpoint assumes a fixed dataset $\mathcal{D}$, and we express our uncertainty in the model parameters through a probability distribution over θ (Bishop, 2006).

Due to the parameter optimisation process, the maximum likelihood approach is often prone to overfitting if we train complex models on datasets of limited size. For example, in GLMs we might decide to increase the number of input features, i.e. increase model complexity, to obtain better performance on our dataset. However, by doing so we might unknowingly extract random noise, which would lead to poor generalisation performance. A common technique for controlling overfitting involves the addition of a *regularisation* term in the likelihood function to encourage, or shrink, parameter values towards zero. In the statistics literature these methods are known as *shrinkage* methods, such as the lasso (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970). In chapter 4 we will use this regularisation strategy to infer the underlying methylation profiles from BS-seq data. The issue of determining the model complexity is known as *bias-variance* trade-off in frequentist statistics (Friedman et al., 2001).

An alternative approach for performing point estimates is the *maximum a posteriori* (MAP) method, which maximises the mode of the posterior distribution of the unknown quantities θ

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\; p(\theta \mid \mathcal{D}) \propto \underset{\theta}{\operatorname{argmax}}\; p(\mathcal{D} \mid \theta)\, p(\theta), \tag{3.6}$$

where we ignore the model evidence $p(\mathcal{D})$ since it does not depend on the parameters θ. In contrast to MLE, the MAP estimate is often referred to as Bayesian, since it considers the parameters θ as random variables and imposes a prior distribution over them. We can interpret the MAP estimation in frequentist terms, if we think of the prior as the regularisation term in the maximum likelihood scenario. We should emphasise that both the MLE and MAP return single point estimates for the parameters θ without providing any measure of uncertainty. This is often problematic, since without taking into account the uncertainty of the parameters we will often underestimate the variance of the predictive distribution (Murphy, 2012).

3.2.3 Hierarchical Bayesian models

In some scenarios defining a specific prior distribution might be difficult, e.g. we might not have domain knowledge on the problem at hand. Until now, for notation simplicity we defined the prior distribution as $p(\theta)$, however, the prior itself might have its own parameters, often called *hyper-parameters*, which we will denote as τ. For example, if we assume the prior follows a Gaussian distribution, then $\tau = (\mu, \sigma^2)$. Then, to express our uncertainty in the hyper-parameters themselves we introduce a prior distribution on our priors, that is, we consider the τ parameters as random variables as well. This modelling approach is called *hierarchical Bayes*, since there are multiple levels of unknown quantities (Murphy, 2012). Of course the number of levels might increase by introducing priors over the hyper-priors, and so on. Bayesian inference then proceeds as usual by computing the posterior distribution of all random variables

$$p(\theta, \tau \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \theta)\, p(\theta \,|\, \tau)\, p(\tau)}{\int p(\mathcal{D}, \theta, \tau)\, d\theta\, d\tau}. \tag{3.7}$$

It is often feasible to integrate over θ, however, the complete marginalisation over all the variables might be intractable. In this case, an approximate solution is to set the hyper-parameters to a specific value obtained by maximising the model evidence obtained by first marginalising out the parameters θ,

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\; p(\mathcal{D} \,|\, \tau) = \underset{\tau}{\operatorname{argmax}} \int p(\mathcal{D} \,|\, \theta)\, p(\theta \,|\, \tau)\, d\theta. \tag{3.8}$$

In the statistics literature this approach is known as *empirical Bayes* or *type II maximum likelihood* (Gelman et al., 1995) and in the machine learning literature is called the *evidence approximation* (Mackay, 1999). Hierarchical models are important for analysing genome-wide experimental studies, due to the 'large p, small n' problem, where p denotes the number of genes, or genomic regions, and n the number of samples (Ji and Liu, 2010). In this setting, the top-level distribution $p(\tau)$ is estimated by the thousands of genes available, which allows the transfer of information across genes at the lower levels for performing reliable inference, e.g. see Smyth (2004) and Vallejos et al. (2015).

3.2.4 Bayesian model selection

So far we have mostly focused on performing inference at the parameter level. In many cases though, one might be interested in evaluating competing models and automatically choosing the one that is most plausible for a given dataset. For example, if we are interested in identifying cell sub-populations from gene expression levels, what is total number of clusters that best explain the data? Or in the regression setting, how to consistently select the 'right' model complexity that explains the observed data and at the same time avoids overfitting? *Model selection* is an important task in understanding and representing the observed data in an automatic way and provides an alternative to the classical hypothesis testing techniques (Barber, 2012).

Suppose we have a set of M competing model hypotheses $\{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, each associated with parameters $\{\theta_1, \ldots, \theta_M\}$, respectively, and we want to compare the performance of the models in fitting the dataset $\mathcal{D}$. The Bayesian view of model selection then involves computing the posterior distribution of model $\mathcal{H}_m$ using nothing but the rules of probability

$$p(\mathcal{H}_m \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_m)\, p(\mathcal{H}_m)}{\sum_{j=1}^{M} p(\mathcal{D}, \mathcal{H}_j)}, \tag{3.9}$$

where the denominator sums over the potentially huge space of possible models. Our prior beliefs about certain models are expressed in the prior distribution $p(\mathcal{H}_m)$. Note that the data dependent term $p(\mathcal{D} \,|\, \mathcal{H}_m)$ is the model evidence and represents the likelihood of the data $\mathcal{D}$ given the model $\mathcal{H}_m$. This quantity is obtained by integrating over the parameter space

$$p(\mathcal{D} \,|\, \mathcal{H}_m) = \int p(\mathcal{D} \,|\, \theta_m, \mathcal{H}_m)\, p(\theta_m \,|\, \mathcal{H}_m)\, d\theta_m. \tag{3.10}$$

Here $p(\mathcal{D} \,|\, \theta_m, \mathcal{H}_m)$ is the likelihood function and $p(\theta_m \,|\, \mathcal{H}_m)$ is the prior distribution over the parameters. We should emphasise that the dimensionality of the parameters θ_m need not be the same across different models. It is interesting to note that the model evidence in (3.10) is exactly the normalisation constant appearing in Bayes' theorem — given in (3.3) — with the only difference that we explicitly condition on the model hypothesis $\mathcal{H}_m$.

One should be cautious when interpreting the posterior distributions $p(\mathcal{H}_m \,|\, \mathcal{D})$. These are not absolute probabilities of how well the model fits to the data, rather they refer to relative probabilities under the set of the M competing model hypotheses. In addition, computing this quantity is often intractable since it requires a summation over all possible models. A common and simpler task is to compare two competing model hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$. Assuming that we have no preference for two competing models, i.e. $p(\mathcal{H}_1) = p(\mathcal{H}_2)$, we observe that the model evidence $p(\mathcal{D} \,|\, \mathcal{H}_m)$ is a key quantity for selecting between competing models, since it transforms our prior beliefs to posterior beliefs through consideration of the data

$$\underbrace{\frac{p(\mathcal{H}_1 \,|\, \mathcal{D})}{p(\mathcal{H}_2 \,|\, \mathcal{D})}}_{\text{posterior odds}} = \underbrace{\frac{p(\mathcal{D} \,|\, \mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_2)}}_{\text{Bayes factor}} \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)}}_{\text{prior odds}}. \tag{3.11}$$

We observe that the transformation from prior to posterior beliefs is the multiplication by

$$BF_{12} = \frac{p(\mathcal{D} \,|\, \mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_2)} \tag{3.12}$$

which we refer to as the *Bayes factor* (Kass and Raftery, 1995). The subscript '12' denotes the evidence of model $\mathcal{H}_1$ against $\mathcal{H}_2$, so larger values increase evidence in favour of $\mathcal{H}_1$. However,

$2\log\lvert BF_{12}\rvert$	BF_{12}	Evidence against $\mathcal{H}_2$
0 to 2	1 to 3	Barely worth mentioning
2 to 6	3 to 20	Substantial
6 to 10	20 to 150	Strong
> 10	> 150	Decisive

Table 3.2 Scale of evidence for interpreting Bayes factors. Twice the natural logarithm of the Bayes factor corresponds to the same scale with the deviance and likelihood ratio tests of classical statistics. Negative values of the natural logarithm (and using the same scale) correspond to evidence in favour of model hypothesis $\mathcal{H}_2$.

Bayes factors lack a probabilistic interpretation, that is, they lack a metric or probability to measure the strength of the evidence. A scale of evidence for interpreting Bayes factors was proposed by Jeffreys (1961) and Kass and Raftery (1995) and is shown in table 3.2.

The use of Bayes factors for hypothesis testing is very similar to the likelihood ratio test in classical statistics (Dickey and Fuller, 1981), with the difference that instead of maximising the likelihood, we integrate over the parameters to obtain the model evidence. As illustrated in figure 3.1, by using the model evidence we are less prone to overfitting and we may favour models of intermediate complexity, since we integrate out the parameter space θ[11]. This cannot be achieved using $p(\mathcal{D}\,|\,\hat{\theta}_m, \mathcal{H}_m)$ — where $\hat{\theta}_m$ is the MLE of the parameters — since more complex models will fit better the data and achieve higher likelihood. However, care must be taken when specifying the prior $p(\theta_m\,|\,\mathcal{H}_m)$, since it significantly affects the model evidence. To obtain robust inferential results, one should perform sensitivity analysis on the Bayes factors under different settings of the prior. For a detailed discussion on the applications and difficulties of Bayes factors, see Kass and Raftery (1995) and Gelman et al. (1995).

In general, computing the model evidence in (3.10) is intractable, since it requires integrating out the model parameters. In section 3.4 we will discuss how to approximate this quantity, however, a popular approximation is the *Bayesian Information Criterion* (BIC) (Schwarz, 1978), which is based on the asymptotic behaviour of Bayes estimators for large sample size N, and has the following form

$$\mathrm{BIC} = \log p(\mathcal{D}\,|\,\hat{\theta}, \mathcal{H}) - \frac{\nu}{2}\log N \simeq p(\mathcal{D}\,|\,\mathcal{H}), \tag{3.13}$$

where $p(\mathcal{D}\,|\,\hat{\theta}, \mathcal{H})$ is the maximised value of the likelihood function, and ν is the total number of model parameters (Fraley and Raftery, 2007). The BIC is a popular method for cheaply performing model selection, where the term $\frac{\nu}{2}\log N$ penalises the complexity of the model. A related metric is the *Akaike information criterion* (AIC)(Akaike, 1974), which however tends to favour more complex models.

[11] This is known as *Occam's razor* principle: "*plurality should not be assumed without necessity*", that is, one should pick the simplest model that adequately explains the data (Murphy, 2012).

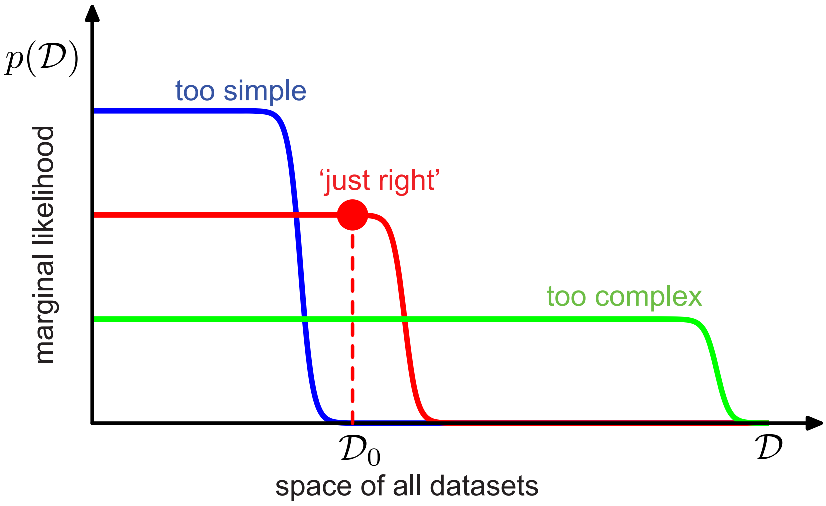


Figure 3.1 Schematic illustration of Bayesian model selection. The x-axis denotes the space of all possible datasets, and the y-axis is the marginal likelihood of three models of differing complexity. Note that the distributions are normalised, since the model evidence has to sum up to one when integrating over all possible datasets. Complex models (green) can model a wide range of datasets, however, to do so they must spread their probability mass. Simple models (blue) have high probability mass only on certain datasets, but are too simple to generate the dataset $\mathcal{D}_0$. The intermediate model (red), is just right and provides the largest evidence for the dataset $\mathcal{D}_0$. Figure adapted from Bishop (2006).

3.3 Probabilistic graphical models

In many real world phenomena, including biological systems, we need to define joint probability distributions over a large number of variables. However, these high dimensional models often have some logical structure, i.e. independence properties, that allows us to decompose the joint distribution into a product of factors, each defined over a subset of the variables. *Probabilistic graphical models* provide a general framework for quantifying uncertainty (using probabilities) and compactly representing and reasoning about high dimensional distributions by exploiting the independence properties (using graphs) present in the model (Koller and Friedman, 2009).

A probabilistic graphical model is a graph where each node represents a random variable (or a group of random variables) and edges represent statistical dependencies among the random variables. As we shall see shortly, what conveys important information is the lack of edges, since they inform us about various local *conditional independence* relationships encoded in the graph. Let x, y, and z be a set of random variables. Then, x and y are independent, denoted $x \perp\!\!\!\perp y$, if and only if $p(x, y) = p(x)\, p(y)$. Also, x and y are conditionally independent given z, denoted $x \perp\!\!\!\perp y \mid z$, if and only if

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z), \qquad \text{or alternatively,} \qquad p(x \mid y, z) = p(x \mid z),$$

such that $p(z) > 0$ for every possible value of the z variable (Dawid, 1979). Note that the joint distribution of x and y factorises into the product of the conditional marginal distributions.

This property allows us to both simplify the structure of the probabilistic model and make efficient computations when performing inference.

There are several classes of graphical models depending on whether the graph is directed, undirected, or a combination of both. *Undirected graphical models* (UGMs), also known as Markov random fields, imply no ordering on their factorisation and generally are suited for expressing soft constraints between random variables (Jordan, 2003). In this case, a variable is conditionally independent of all other variables given its direct neighbours in the graph. Common models include the Boltzmann machine (Ackley et al., 1985) and the Ising model in statistical physics (Glauber, 1963). The main focus of this thesis, though, is on *directed graphical models* (DGMs), also known as Bayesian[12] or belief networks (Pearl, 1988). A DGM is a directed acyclic graph (DAG)[13] of N random variables $\{x_1, \ldots, x_N\}$ that factorises the joint distribution

$$p(x_1, \ldots, x_N) = \prod_{n=1}^{N} p(x_n \,|\, x_{\mathrm{pa}(n)}), \tag{3.14}$$

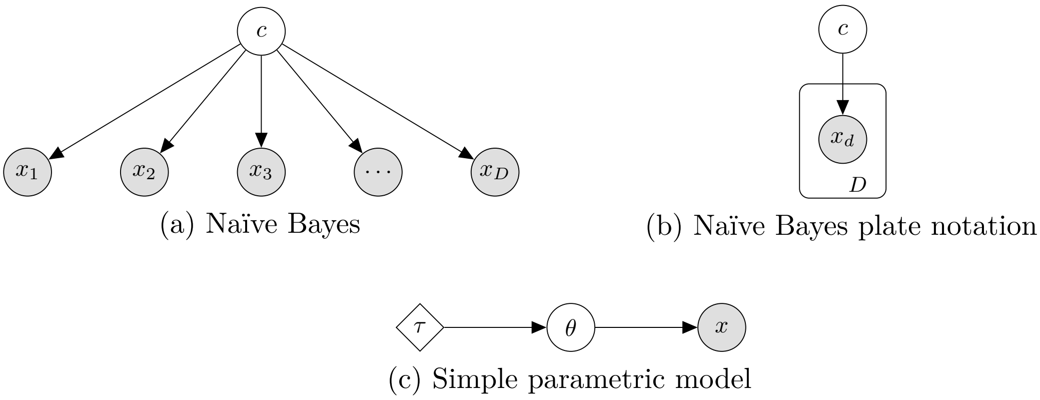
where $x_{\mathrm{pa}(n)}$ denotes the set of parents of node x_n, if node x_n has no parents we take $p(x_n \,|\, x_{\mathrm{pa}(n)}) = p(x_n)$. This states that the joint distribution is decomposed in factors, where each factor corresponds to the local conditional probability of the variable given the values of its parents. For the graphical model then to be completely defined we need to specify the graph structure and the form of the conditional probability distribution $p(x_n \,|\, x_{\mathrm{pa}(n)})$ at each node (Koller and Friedman, 2009). Hence, when designing a probabilistic model we should be careful to choose an ordering which allows us to factorise the joint in conditional distributions that we can evaluate. Once the basic modelling assumptions are formed, all questions are answered by performing inference on the quantities of interest from the graphical model.

To explicitly demonstrate this factorisation, figure 3.2a shows a concrete example of a DGM. Here, we distinguish between *observed* variables, which are denoted by shading the corresponding nodes, and unobserved variables. The unobserved variables are often called *latent* or *hidden* variables. The latent variables may represent hidden processes which cannot be directly measured, and as we shall shortly see they are essential elements of probabilistic modelling. The DGM in figure 3.2a corresponds to the joint distribution found in the *Naïve Bayes* model (Rish, 2001). Briefly, in Naïve Bayes we wish to classify an observed D-dimensional feature vector $\mathbf{x}$ to some output class variable c,

$$p(c, \mathbf{x}) = p(c)\, p(\mathbf{x} \,|\, c) = p(c) \prod_{d=1}^{D} p(x_d \,|\, c). \tag{3.15}$$

[12] Note that there is nothing inherently Bayesian about these models, they are just a way of defining probability distributions. However, by making hierarchical Bayesian models easy to represent this formalism has become popular within the Bayesian paradigm.

[13] That is, it contains no directed cycles, which is equivalent to the statement that there exists a topological ordering of the nodes such that there is no edge from any node to its ancestors (Bishop, 2006).



(a) Naïve Bayes

(b) Naïve Bayes plate notation

(c) Simple parametric model

Figure 3.2 Selection of directed graphical models. (a) Directed graphical model for Naïve Bayes and (b) the corresponding representation in plate notation. (c) Simple probabilistic model where parameters are explicitly denoted in the graph.

Note that both equation (3.15) and figure 3.2a contain the same information, that is, the observed features x_d are conditionally independent given the class c. However, if we marginalise out c, the observed features x_d are in general no longer independent

$$\sum_c p(c, \mathbf{x}) = p(\mathbf{x}) \neq \prod_{d=1}^{D} p(x_d). \tag{3.16}$$

As more variables and complex models are introduced, it is convenient to compactly express multiple nodes using the *plate notation*, see figure 3.2b, which is a useful representation for capturing replication in graphical models.

Note that until now we have made no distinction between data and parameters in our formulation and indeed it is natural to include parameters among the nodes in the graph. Also, in a fully Bayesian setting the parameters themselves are random variables, hence, during probabilistic inference there is no distinction between parameters and variables in the graphical model. Figure 3.2c demonstrates a parametric model factorised as $p(x, \theta \mid \tau) = p(x \mid \theta)\, p(\theta \mid \tau)$, where the observed data x are generated from a distribution parametrised by θ, which in turn admits a distribution parametrised by the hyper-parameter τ. When optimising with respect to τ, this graphical model corresponds to the empirical Bayes procedure. Note that factorising the joint in a different way, i.e. changing the direction of arrows, would imply a different interpretation of the data generation mechanism. Also, in the graphical representation, parameters that admit no probability distribution, such as τ, are denoted by a rhombus. This convention will make it clear if parameters are treated as fixed, e.g. for maximum likelihood estimation, or as random variables in the Bayesian formalism.

The conditional independence relationships embedded in the DGM can be efficiently deduced by the *d-separation* criterion, where 'd' stands for directed (Pearl, 1988). Consider a directed graph with three, non-intersecting, sets of nodes x, y, and z. We wish to deduce whether

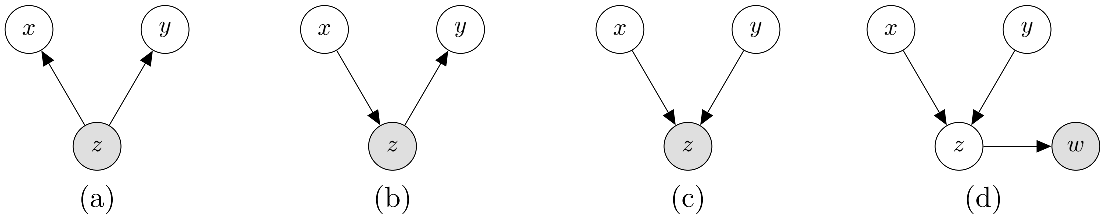


Figure 3.3 Illustration of the concept of d-separation. For both (a) and (b), the path from x to y is blocked by z, rule (i) of d-separation. (c) In this scenario, the path from x and y is unblocked, since it does not satisfy the rules of d-separation. (d) The path from x to y is unblocked, since even though z is unobserved we have conditioned on its descendant w. The rather unintuitive scenarios (c) and (d) can be understood by the *explaining away* principle in probabilistic reasoning (Jordan, 2003).

$x \perp\!\!\!\perp y \,|\, z$ is implied by the structure of the DGM. To do so, we query all possible paths from x to y. We say that a path is *blocked* if it includes a node such that either the arrows meet

(i) head-to-tail or tail-to-tail at the node, and the node is in z, or

(ii) head-to-head at the node, and neither the node, nor any of its descendants, are in z.

If all paths are blocked, then x is d-separated from y by z, and the conditional independence statement is satisfied[14] (Bishop, 2006). Figure 3.3 shows example DGMs that illustrate the concept of d-separation. An additional important concept is that of the *Markov blanket*. The Markov blanket of a node x, MB(x), comprises the set of parents, children and children's other parents. Then, every set of nodes $\mathcal{S}$ in the DGM is conditionally independent of x given its Markov blanket, i.e. $p(x \,|\, \mathrm{MB}(x), \mathcal{S}) = p(x \,|\, \mathrm{MB}(x))$. These concepts are useful for inference algorithms, such as Gibbs sampling and mean field variational inference (see below), that require evaluating full conditional distributions, i.e. $p(x \,|\, \mathcal{S})$ where $\mathcal{S}$ denotes all remaining nodes.

3.3.1 Data augmentation and the EM algorithm

As aforementioned, latent variables might arise in missing value problems, where we have incomplete data due to our inability to directly measure some underlying processes that describe the model. These models are known as *latent variable models* (LVMs). However, latent variables need not have any physical interpretation in the probabilistic model, but may be introduced to augment the observed data so as to make the model tractable and easy to analyse (Bishop, 2006). The main idea is that often the likelihood $p(x \,|\, \theta)$ of incomplete data x parametrised by θ might be complicated, hence optimising or inferring the posterior distribution of the parameters is infeasible. In such cases, the introduction of latent variables z might simplify the inference of the parameters θ by repeatedly solving complete data problems from $p(x, z \,|\, \theta)$; which we assume to be significantly easier, e.g. $p(x, z \,|\, \theta)$ might belong to the exponential family (Van Dyk and Meng, 2001).

[14] An efficient approach for computing d-separation is the Bayes ball algorithm (Shachter, 1998). More in-depth treatments of how to establish conditional independence, and general introduction to probabilistic graphical models, can be found in Bishop (2006), Koller and Friedman (2009), Barber (2012) and references therein.

An influential paper in 1977 from Dempster, Laird and Rubin introduced a powerful and general method for finding MLEs in latent variable models. In this work, Dempster et al. (1977) demonstrated that many seemingly unrelated problems in statistics, could be cast as LVMs and could be efficiently computed using an iterative process known as the *Expectation-Maximisation* (EM) algorithm. In the Bayesian statistics literature the idea of introducing latent variables to iteratively compute the posterior $p(\theta \mid x)$ was first demonstrated by Tanner and Wong (1987)[15]. In this class of *data augmentation* algorithms, in a similar fashion to EM, one alternates between generating z from the the predictive distribution $p(z \mid x)$ (imputation step) and sampling θ from the complete data posterior $p(\theta \mid x, z)$ (posterior step). In chapter 5 we will use the data augmentation strategy to obtain efficient algorithms for the Bayesian probit regression model.

The EM algorithm

Consider the probabilistic model of observed and latent variables $p(x, z \mid \theta)$ governed by a set of parameters θ. Our interest is to maximise the observed data log-likelihood $p(x \mid \theta)$ by marginalising over the latent variables

$$\log p(x \mid \theta) = \log \int p(x, z \mid \theta)\, dz, \tag{3.17}$$

direct maximisation of this quantity though may not be often possible in closed form. To simplify our problem we introduce a distribution $q(z)$ defined over the latent variables. Then, for any choice of $q(z)$ we obtain a *lower bound* on the log-likelihood using Jensen's inequality[16]

$$\begin{aligned}
\log p(x \mid \theta) &= \log \int q(z)\, \frac{p(x, z \mid \theta)}{q(z)}\, dz \\
&\geq \int q(z) \log \frac{p(x, z \mid \theta)}{q(z)}\, dz \\
&= \int q(z) \log p(x, z \mid \theta)\, dz - \int q(z) \log q(z)\, dz \\
&= \langle \log p(x, z \mid \theta) \rangle_{q(z)} + \mathrm{H}\left[q(z)\right] \\
&\overset{\mathsf{def}}{=} \mathcal{L}\left(q(z), \theta\right).
\end{aligned} \tag{3.18}$$

Here $\mathcal{L}\left(q(z), \theta\right)$ is a functional[17] of the distribution $q(z)$ and a function of the model parameters θ, and $\mathrm{H}[q(z)]$ denotes the entropy of the probability distribution $q(z)$. The basic idea of the EM algorithm is then to alternate between optimising the lower bound with respect to the

[15] Interestingly, at about the same time Swendsen and Wang (1987) introduced latent variables — known as *auxiliary variables* in the physics literature — to improve the speed of iterative simulation for Ising and Potts models in statistical physics. The historical development of these methods and the relationship between data augmentation and the EM algorithm are discussed in Tanner and Wong (2010).

[16] Jensen's inequality states that if f is a concave function, then $f(\langle x \rangle_{p(x)}) \geq \langle f(x) \rangle_{p(x)}$ (Jensen, 1906).

[17] A functional is a mapping that takes a function as input and returns the value of the functional as output. An example of a functional is the entropy which takes as input a distribution $p(x)$ and returns $\mathrm{H}\left[p(x)\right] = -\int p(x) \log p(x)\, dx$ as output.

distribution $q(z)$ and the parameters θ (Ghahramani, 2004). After initialising the parameters, the t^{th} iteration of the EM algorithm consists of the following two steps:

E-step: optimise $\mathcal{L}$ with respect to $q(z)$ while holding θ_{t-1} fixed

$$q_t(z) = \underset{q(z)}{\operatorname{argmax}}\, \mathcal{L}\left(q(z), \theta_{t-1}\right). \tag{3.19}$$

However, what would be the optimal distribution $q(z)$? To obtain a better intuition, we can rewrite the lower bound

$$\begin{aligned}\mathcal{L}(q(z), \theta) &= \int q(z) \log \frac{p(z \,|\, x, \theta)\, p(x \,|\, \theta)}{q(z)}\, dz \\ &= \int q(z) \log p(x \,|\, \theta)\, dz + \int q(z) \log \frac{p(z \,|\, x, \theta)}{q(z)}\, dz \\ &= \log p(x \,|\, \theta) - \mathrm{KL}\left[q(z) \,||\, p(z \,|\, x, \theta)\right],\end{aligned} \tag{3.20}$$

where the second term is known as the Kullback-Leibler divergence (Kullback and Leibler, 1951), or *relative entropy*, between $q(z)$ and the posterior distribution $p(z \,|\, x, \theta)$[18]. The KL divergence satisfies $\mathrm{KL}[q \,||\, p] \geq 0$, with equality if, and only if $q = p$; also, it should be pointed out that the KL divergence is not a symmetrical quantity, that is $\mathrm{KL}[q \,||\, p] \neq \mathrm{KL}[p \,||\, q]$. This means that for fixed θ, the lower bound $\mathcal{L}(q(z), \theta)$ is bounded above by the log-likelihood of the observed data $\log p(x \,|\, \theta)$, and achieves that bound only when the KL divergence vanishes. Hence, the E-step naturally sets $q(z)$ to the posterior distribution of the latent variables

$$q_t(z) = p(z \,|\, x, \theta_{t-1}). \tag{3.21}$$

M-step: optimise $\mathcal{L}$ with respect to θ while holding $q_t(z)$ fixed

$$\begin{aligned}\theta_t &= \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}\left(q_t(z), \theta\right) \\ &= \underset{\theta}{\operatorname{argmax}} \left| \langle \log p(x, z \,|\, \theta) \rangle_{q_t(z)} + \mathrm{H}\left[q_t(z)\right] \right| \\ &= \underset{\theta}{\operatorname{argmax}}\, \langle \log p(x, z \,|\, \theta) \rangle_{q_t(z)},\end{aligned} \tag{3.22}$$

where the entropy term is constant with respect to θ. The M-step will cause the lower bound $\mathcal{L}$ to increase, unless it is already at a maximum. In the next iteration of EM, we need to update $q(z)$, since it was determined using θ_{t-1} and is held fixed in the M-step, hence it will not match the new conditional $p(z \,|\, x, \theta_t)$. This iterative process is guaranteed to monotonically increase the log-likelihood of observed data until it reaches a (local) optimum (Dempster et al., 1977).

[18] Here we interpret the KL divergence as a dissimilarity measure between two distributions. In information theory, the KL divergence measures the average additional amount of information — typically in bits — required to encode data coming from a source p when we used model q to construct our coding scheme. David Mackay has written an excellent book relating statistics, machine learning and information theory (MacKay, 2003).

Using the EM algorithm we broke down the problem of directly optimising the log-likelihood into two stages which are often simpler to implement. However, in many interesting models it is often computationally intractable to perform the E-step or the M-step, or indeed both. To address these issues, different variants of the EM algorithm have been proposed; see McLachlan and Krishnan (2007) for additional information.

3.3.2 Finite mixture models

When performing probabilistic modelling often our goal is to determine the intrinsic structure of some observed data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. *Mixture models* provide a powerful and flexible framework for describing complex systems and approximating the distribution of the data to arbitrary accuracy (McLachlan and Peel, 2004). To do so, mixture models comprise of a finite[19] convex combination of probability distributions

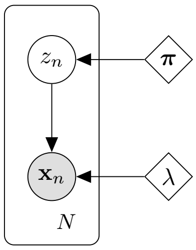
$$p(\mathbf{x}_n \mid \theta) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}_n \mid \lambda_k), \quad \text{such that} \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k \geq 0 \;\text{ for } k \in \{1, \ldots K\}, \qquad (3.23)$$

where $\theta = \{\lambda, \boldsymbol{\pi}\}$ is the set of all model parameters, $\lambda = \{\lambda_1, \ldots, \lambda_K\}$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. We refer to the parameters π_k as *mixing proportions*, or *mixing weights*, which are introduced to perform a weighted sum over the *mixture components* $p(\mathbf{x}_n \mid \lambda_k)$. Here we assume that the mixture components are coming from some parametric family, where each component has its own parameters λ_k. From (3.23) we observe that when $K = 1$ this reduces to just using a single distribution, and when K grows with the number of observations this model becomes non-parametric — since the θ parameters grow with N as well — and is similar to kernel density estimation (Marin et al., 2005).

A different interpretation of mixture models — and the one considered in this thesis — is that of performing model based clustering (McLachlan and Basford, 1988). In this setting, we assume that there is some sub-population structure in the observed data, which however is missing, and the goal of inference is to recover the assignment of observations to each sub-population. In this view, mixture models can be formulated as latent variable models, where the discrete latent states can be interpreted as allocating data points to specific mixture components (McLachlan and Peel, 2004). To do so, we introduce a latent categorical variable $z_n \in \{1, \ldots, K\}$, representing the component that is responsible for observation $\mathbf{x}_n$. Then, the prior probability that an observation will be assigned to cluster k is given by the the mixing proportions, that is $p(z_{nk}) = \pi_k$, where for notational convenience z_{nk} is used to denote $z_n = k$.

Here we perform MLE on model parameters and we defer the fully Bayesian treatment of mixture models (Diebolt and Robert, 1994; Richardson and Green, 1997) until chapter 5. The

[19] An infinite treatment of mixture models is also possible. In the Bayesian setting, one could introduce Dirichlet process (Ferguson, 1973) priors to obtain an infinite mixture model, which is widely used when we have no prior knowledge, or preference, about the number of mixture components, e.g. see Escobar and West (1995) and Neal (2000). In this thesis we restrict our discussion to finite mixture models.



$$\begin{aligned} z_n &\sim \mathcal{C}\mathrm{at}(z_n \,|\, \boldsymbol{\pi}), \\ \mathbf{x}_n \,|\, z_{nk} &\sim p(\mathbf{x}_n \,|\, \lambda_k). \end{aligned}$$

Figure 3.4 (Left) Probabilistic graphical representation of mixture models. (Right) Data generation process from a mixture model: we randomly sample one component with probabilities given by the mixing proportions, and then we generate an observation from the corresponding distribution. Here, $\mathcal{C}$at denotes the Categorical distribution, i.e. the Multinomial distribution over a single trial. In distribution $p(\mathbf{x}_n \,|\, \lambda_k)$, the conditioning on z_{nk} is implicitly denoted in the subscript of the λ parameter.

graphical representation of the mixture model is shown in figure 3.4. Given the dataset $\mathcal{D}$ and assuming i.i.d. observations, parameter estimation is achieved by maximising the log-likelihood

$$\log p(\mathcal{D}|\theta) = \sum_{n=1}^{N} \log \left| \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}_n \,|\, \lambda_k) \right| . \tag{3.24}$$

The summation inside the logarithm prevents the possibility of deriving an analytical solution. Hence, as the reader may have anticipated, we will use the EM algorithm to estimate the model parameters.

During the E-step, we optimise the lower bound $\mathcal{L}(q(\mathbf{z}), \theta)$ with respect to $q(\mathbf{z})$ while holding θ fixed, which results in setting $q(\mathbf{z})$ to the posterior distribution of the latent variables[20]. Applying Bayes' theorem and using the fact that the data are i.i.d., the posterior probability that observation $\mathbf{x}_n$ is generated from the k^{th} mixture component is

$$\gamma_n(k) \overset{\mathsf{def}}{=} q(z_{nk}) = p(z_{nk} \,|\, \mathbf{x}_n, \theta) = \frac{\pi_k \, p(\mathbf{x}_n \,|\, \lambda_k)}{\sum_{j=1}^{K} \pi_j \, p(\mathbf{x}_n \,|\, \lambda_j)} . \tag{3.25}$$

The quantity $\gamma_n(k)$ is also known as the *responsibility* that mixture component k takes for explaining observation $\mathbf{x}_n$ (Bishop, 2006). During the M-step we optimise $\mathcal{L}(q(\mathbf{z}), \theta)$ with respect to the model parameters θ while holding $q(\mathbf{z})$ fixed

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \, \langle \log p(\mathcal{D}, \mathbf{z} \,|\, \theta) \rangle_{q(\mathbf{z})} \\ &= \underset{\theta}{\operatorname{argmax}} \left| \langle \log p(\mathcal{D} \,|\, \mathbf{z}, \lambda) \rangle_{q(\mathbf{z})} + \langle \log p(\mathbf{z} \,|\, \boldsymbol{\pi}) \rangle_{q(\mathbf{z})} \right| \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \left| \langle \log p(\mathbf{x}_n \,|\, z_n, \lambda) \rangle_{q(z_n)} + \langle \log p(z_n \,|\, \boldsymbol{\pi}) \rangle_{q(z_n)} \right| \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \sum_{k=1}^{K} \left| \gamma_n(k) \log p(\mathbf{x}_n \,|\, \lambda_k) + \gamma_n(k) \log \pi_k \right| . \end{aligned} \tag{3.26}$$

[20] To keep the notation uncluttered, subscripts denoting the EM iterations are omitted.

Due to the above factorisation, we can optimise each set of parameters $\theta = \{\lambda, \boldsymbol{\pi}\}$ independently (McLachlan and Peel, 2004). Setting the partial derivatives of $\mathcal{L}(q(\mathbf{z}), \theta)$ with respect to π_k to zero and using Lagrange multipliers for the constraint $\sum_k \pi_k = 1$, the update for mixing proportions becomes

$$\hat{\pi}_k = \frac{1}{N}\sum_{n=1}^{N} \gamma_n(k). \tag{3.27}$$

Similarly, the partial derivatives of $\mathcal{L}(q(\mathbf{z}), \theta)$ with respect to model parameters λ_k are given by

$$\frac{\partial}{\partial \lambda_k}\mathcal{L} = \frac{\partial}{\partial \lambda_k}\left|\sum_{n=1}^{N} \gamma_n(k) \log p(\mathbf{x}_n \,|\, \lambda_k)\right| = \sum_{n=1}^{N} \frac{\gamma_n(k)}{p(\mathbf{x}_n \,|\, \lambda_k)} \frac{\partial\, p(\mathbf{x}_n \,|\, \lambda_k)}{\partial \lambda_k}. \tag{3.28}$$

Optimisation of (3.28) depends on the mathematical form of $p(\mathbf{x}_n \,|\, \lambda_k)$. For probability models belonging to the exponential family, maximisation of (3.28) results in closed form solutions; however, it is often intractable for complex models — including the *BPRMeth* model discussed in chapter 4 — and numerical optimisation strategies (Nocedal and Wright J., 1999) could be exploited.

3.4 Approximate Bayesian inference

As aforementioned, during probabilistic modelling we often need to make *inference compromise* due to intractable computations arising in complex mathematical models. This feature is prevalent in Bayesian statistics, where the main inferential operation is marginalisation, as opposed to optimisation in classical statistics (Lawrence et al., 2010). Let θ denote the set of random variables, $p(\theta)$ be the prior, $p(\mathcal{D} \,|\, \theta)$ be the likelihood, and $p(\theta \,|\, \mathcal{D})$ denote the posterior distribution. The key challenges in Bayesian inference can be defined as

(i) calculating the marginal likelihood $p(\mathcal{D}) = \langle p(\mathcal{D} \,|\, \theta)\rangle_{p(\theta)} = \int p(\mathcal{D} \,|\, \theta)\, p(\theta)\, d\theta$. By having access to $p(\mathcal{D})$ we can both evaluate the posterior at any input point and perform Bayesian model selection.

(ii) characterising the posterior distribution $p(\theta \,|\, \mathcal{D})$. In many situations the posterior itself might not be of direct interest, however, the objective of our analysis might require the posterior for carrying out downstream tasks. These include: (1) predicting new observations $\mathbf{x}^*$ using the predictive distribution $p(\mathbf{x}^* \,|\, \mathcal{D}) = \langle p(\mathbf{x}^* \,|\, \theta)\rangle_{p(\theta \,|\, \mathcal{D})} = \int p(\mathbf{x}^* \,|\, \theta)\, p(\theta \,|\, \mathcal{D})\, d\theta$ and (2) obtaining summary statistics of the form $\langle f(\theta)\rangle_{p(\theta \,|\, \mathcal{D})} = \int f(\theta)\, p(\theta \,|\, \mathcal{D})\, d\theta$ for some function of interest f, such as the posterior mean, where $f(\theta) = \theta$, or posterior variance, where $f(\theta) = \left(\theta - \langle\theta\rangle_{p(\theta \,|\, \mathcal{D})}\right)^2$ (Andrieu et al., 2003).

For many models of practical interest, the above integrations — or equivalently the expectations with respect to the corresponding probability distributions — may not have closed-form analytical solutions. Similarly, for discrete θ, the summation over all possible configurations might be computationally infeasible due to the exponentially many hidden states of the latent variables. In such situations, we must appeal to approximate inference methods which fall broadly into two main classes.

Sampling techniques, also known as *Monte Carlo* methods, approximate the posterior distribution via random sampling. The idea behind the broad family of Monte Carlo methods is to generate a set of statistical samples and use them as proxy for computing quantities of interest. These methods are very flexible and have been applied to a wide range of models, enabling the widespread adoption of Bayesian methods across various domains (Liu, 2001; Robert and Casella, 1999). Interestingly, although Monte Carlo methods were widely used in statistical physics from as early as 1950s (Metropolis et al., 1953; Metropolis and Ulam, 1949), their adoption in the Bayesian community was met with resistance and was viewed as antithetical to the Bayesian philosophy[21] (Tanner and Wong, 2010). It was not until the very early 1990s, with the seminal work of Gelfand and Smith (1990), that advanced Monte Carlo approaches became a mainstay in the field of computational statistics. Despite their widespread adoption, sampling based approaches can be computationally intensive and often are limited to relatively small-scale problems.

A complementary alternative to sampling based methods, are deterministic approximation schemes, such as *variational inference* (Jordan et al., 1999; Parisi, 1988) and *expectation propagation* (Minka, 1999). These approaches are based on analytical approximations to the posterior distribution, by formulating the Bayesian inference problem as the solution to an optimisation problem (Hoffman et al., 2013). These methods are increasingly popular in the machine learning community, due to their efficiency and scalability to large applications (Blei et al., 2003). However, the strong structural assumptions made about the form of the posterior, might often lead to rather poor results; and evaluating the approximation performance of variational inference is an active area of research (Yao et al., 2018).

3.4.1 Variational inference

Variational inference is an optimisation-based approach for approximating an intractable posterior using a restricted (and tractable) family of distributions. The name originates from the field of mathematical analysis known as *calculus of variations*[22], where we seek a function that minimises, or maximises, a functional. There is nothing intrinsically approximate in variational theory, however, variational methods can be used to find approximate solutions by restricting the range of functions over which the optimisation is performed (Jordan et al., 1999). For example, by assuming that the function has specific parametric form, such as a Gaussian distribution, or it factorises in a particular way.

[21] For example, O'Hagan (1987) expressed his criticism for using a 'frequentist' procedure in Bayesian inference, by writing a paper titled "*Monte Carlo is fundamentally unsound*".

[22] Developed by Leonhard Euler and Joseph-Louis Lagrange in the 18^{th} century.

The difficulty of computing the posterior in Bayesian inference is the evidence $p(\mathcal{D})$. Hence, in a similar fashion to the EM algorithm, we can lower bound the evidence log-likelihood by introducing a distribution $q(\theta)$ and using Jensen's inequality

$$\begin{aligned}\log p(\mathcal{D}) &\geq \int q(\theta) \log \frac{p(\mathcal{D}, \theta)}{q(\theta)}\, d\theta \\ &= \langle \log p(\mathcal{D}, \theta) \rangle_{q(\theta)} + \mathrm{H}\left[q(\theta)\right] \\ &\overset{\mathsf{def}}{=} \mathcal{L}\left(q(\theta)\right).\end{aligned} \tag{3.29}$$

Note that this bound is the same that we used for deriving the EM algorithm in (3.18), with the only difference that in the Bayesian paradigm there is no distinction between model parameters and latent variables and all these quantities are collectively denoted by the variable θ. In the machine learning literature, $\mathcal{L}$ is often called the *evidence lower bound* (ELBO) (Blei et al., 2017). Then, we choose a restricted family of distributions $\mathcal{Q}$, such that the expectations in (3.29) can be efficiently computed, and search over the space of *candidate* distributions $q(\theta)$ to maximise the ELBO

$$q^{*}(\theta) = \underset{q(\theta) \in \mathcal{Q}}{\operatorname{argmax}}\, \mathcal{L}\left(q(\theta)\right). \tag{3.30}$$

Hence, inference amounts to obtaining the optimised variational distribution $q^{*}(\theta) \simeq p(\theta \,|\, \mathcal{D})$, which can be used as a proxy for the posterior, e.g. to compute the posterior mean or the posterior predictive distribution.

An equivalent derivation is to seek the member of the family $\mathcal{Q}$ that is closest in KL divergence to the posterior distribution,

$$\begin{aligned}\mathrm{KL}\left[q(\theta) \,||\, p(\theta \,|\, \mathcal{D})\right] &= -\int q(\theta) \log \frac{p(\theta \,|\, \mathcal{D})}{q(\theta)}\, d\theta \\ &= \langle \log q(\theta) \rangle_{q(\theta)} - \int q(\theta) \log \frac{p(\mathcal{D}, \theta)}{p(\mathcal{D})} d\theta \\ &= -\mathrm{H}\left[q(\theta)\right] - \langle \log p(\mathcal{D}, \theta) \rangle_{q(\theta)} + \langle \log p(\mathcal{D}) \rangle_{q(\theta)} \\ &= -\mathcal{L}\left(q(\theta)\right) + \log p(\mathcal{D}).\end{aligned} \tag{3.31}$$

Here the model evidence $\log p(\mathcal{D})$ is constant with respect to $q(\theta)$, so as a function of the variational distribution, minimising the KL divergence is equivalent to maximising the ELBO. Note that the quality of the approximation depends solely on the flexibility of the family of the approximating distributions $\mathcal{Q}$. If we allow full flexibility on $\mathcal{Q}$, then the maximum of ELBO is obtained when the KL is zero, that is when $q(\theta) = p(\theta \,|\, \mathcal{D})$. However, working with the posterior is intractable, otherwise we would not resort to variational methods. Therefore, there is a compromise between choosing a restricted family of distributions $\mathcal{Q}$ that are computationally tractable and at the same time are sufficiently flexible to provide a good approximation to the posterior distribution. To obtain a better intuition about the optimal variational distribution we can rewrite the ELBO as

$$\begin{aligned}\mathcal{L}\left(q(\theta)\right) &= \langle \log p(\mathcal{D} \,|\, \theta)\rangle_{q(\theta)} + \langle \log p(\theta)\rangle_{q(\theta)} - \langle \log q(\theta)\rangle_{q(\theta)} \\ &= \langle \log p(\mathcal{D} \,|\, \theta)\rangle_{q(\theta)} - \mathrm{KL}\left[q(\theta) \,||\, p(\theta)\right].\end{aligned} \tag{3.32}$$

The first term is the expected log-likelihood of the data, and rewards variational distributions that place high mass on settings where the variables θ can explain the observations. The second term is the negative KL between the variational approximation and the prior, and encourages distributions that are close to the prior.

Mean field variational family

A popular approach for approximating the posterior distribution is to assume that the variational distribution factorises over some partition of the random variables θ,

$$q(\theta) = q(\theta_1, ..., \theta_M) = \prod_{m=1}^{M} q_m(\theta_m). \tag{3.33}$$

Note that the factorisation need not be over all individual random variables, in which case $M = D$, where D is the dimensionality of the posterior distribution. Hence, the approximation relies on making additional independence assumptions for the structure of the model, which may not necessarily be present in the true posterior distribution. We should emphasise that we make *no* further assumptions about the functional form of the variational distributions, and each group of random variables θ_m is governed by a distinct variational factor $q_m(\cdot)$ with its own parameters. Rather, the functional form is determined by the type of variables θ and the form of the model, and will be estimated by the variational algorithm — *free-form optimisation.*

The most common algorithm for optimising the ELBO is *coordinate ascent variational inference* (CAVI), which iteratively updates each factor $q_m(\theta_m)$, while holding the remaining factors fixed (Bishop, 2006). Substituting (3.33) to (3.29) and focusing on updating $q_i(\theta_i) \overset{\text{def}}{=} q_i$, we obtain

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_m q_m \Big[\log p(\mathcal{D}, \theta) - \sum_m \log q_m\Big] d\theta \\ &= \int q_i \Big[\int \log p(\mathcal{D}, \theta) \prod_{m \neq i} q_m \, d\theta_m\Big] d\theta_i - \int q_i \log q_i \, d\theta_i - \sum_{m \neq i} \int q_m \log q_m \, d\theta_m \\ &= \int q_i \log \widetilde{p}(\mathcal{D}, \theta_i) \, d\theta_i - \int q_i \log q_i \, d\theta_i + \text{const},\end{aligned} \tag{3.34}$$

where

$$\log \widetilde{p}(\mathcal{D}, \theta_i) \overset{\text{def}}{=} \langle \log p(\mathcal{D}, \theta)\rangle_{q_{m \neq i}}. \tag{3.35}$$

Here, $\langle \cdot \rangle_{q_{m \neq i}}$ denotes an expectation with respect to the distributions $q_m(\theta_m)$ for all $m \neq i$. We observe that (3.34) is the KL divergence between $q_i(\theta_i)$ and $\widetilde{p}(\mathcal{D}, \theta_i)$ and therefore will be

Algorithm 3.1 Coordinate ascent variational inference (CAVI)

```
initialize variational factors $q_m(\theta_m)$
while ELBO has not converged do
    for $i = 1, \ldots, M$ do
        Set $q_i(\theta_i) \propto \exp\left[\langle \log p(\mathcal{D}, \theta) \rangle_{q_{m \neq i}}\right]$
    end for
    Compute $\mathcal{L}\left(q(\theta)\right) \leftarrow \langle \log p(\mathcal{D}, \theta) \rangle_{q(\theta)} + \langle \log q(\theta) \rangle_{q(\theta)}$
end while
```

maximised when these two quantities are equal, which results in setting

$$q_i^*(\theta_i) = \frac{\exp\left[\langle \log p(\mathcal{D}, \theta) \rangle_{q_{m \neq i}}\right]}{\int \exp\left[\langle \log p(\mathcal{D}, \theta) \rangle_{q_{m \neq i}}\right] d\theta_i}. \tag{3.36}$$

In practice, we will find it more convenient to work in the logarithm space and identify the normalisation constant by inspecting the form of

$$\log q_i^*(\theta_i) = \langle \log p(\mathcal{D}, \theta) \rangle_{q_{m \neq i}} + \text{const.} \tag{3.37}$$

This process is then repeated for all factors q_m. Note that the expectations are computed with respect to all other factors, hence, we resort to an iterative algorithm, by first initialising q_m and then cycling through the factors until the ELBO has converged, as shown in algorithm 3.1. Using this algorithm we are guaranteed that the ELBO will reach a (local) optimum since the bound is convex with respect to each factor q_m (Boyd and Vandenberghe, 2004).

Typically we do not need to compute expectations with respect to all remaining factors, due to conditional independence relationships encoded in the joint distribution. Therefore, when representing the joint as a directed graphical model, for each variable θ_m we only need to reason about the variables belonging in its Markov blanket, hence CAVI can be seen as a *message passing* algorithm (Winn and Bishop, 2005). From (3.37) we observe that each variable is informed by the mean value — expectation — of the neighbouring variables, hence the name *mean field* variational inference, by analogy to such methods in statistical physics (Parisi, 1988).

It is often the case that the expectations in (3.37) are still intractable, even after assuming that the variational distribution factorises over the variables. One can either replace the 'problematic' factors with distributions that are point-wise lower bound to these factors (Jaakkola and Jordan, 2000) or perform stochastic optimisation of the variational objective (Hoffman et al., 2013; Paisley et al., 2012), leading to 'black box' variational inference (Ranganath et al., 2014). In addition, for many models the structural assumptions encoded in the variational distributions may lead to poor approximations; in which case Monte Carlo methods are a complementary alternative since they can be applied to a broader class of models.

3.4.2 Simple Monte Carlo

Consider the problem of computing the expectation of some function $f(x)$ with respect to a probability distribution $p(x)$, where $p(x) = p(\theta \,|\, \mathcal{D})$ for the Bayesian inference case,

$$\langle f(x) \rangle_{p(x)} = \int f(x)\, p(x)\, dx. \tag{3.38}$$

Assuming that we can draw a set of samples $x^{(s)}$ independently form the distribution $p(x)$, then (3.38) can be approximated by a finite sum

$$\langle f(x) \rangle_{p(x)} \simeq \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}), \qquad x^{(s)} \sim p(x). \tag{3.39}$$

This process of generating a large number of independent samples to approximate a complex expectation is known as *Monte Carlo integration* (Metropolis and Ulam, 1949). We should point out that (3.39) is an *unbiased estimator* and by the strong law of large numbers, as $S \to \infty$ this sample estimate will almost surely converge to the true expected value $\langle f(x) \rangle_{p(x)}$. Also, from the central limit theorem, and assuming that the variance of $f(x)$ is bounded, the variance of the Monte Carlo estimate will scale like $\text{var}\left[f(x)\right]_{p(x)} / S$, for independent samples $x^{(s)}$. Although the estimator's variance shrinks linearly with computational effort, the accuracy of the estimator is not directly related to the dimensionality of x, in contrast to standard numerical integration approaches (Neal, 1993; Robert and Casella, 1999). Also, we should note that the samples $x^{(s)}$ are drawn from $p(x)$, irrespective of the function $f(x)$; hence, we can use the same bag of samples to evaluate expectations of different functions of interest.

In practice though, generating independent samples directly from an arbitrary distribution $p(x)$ is often infeasible. The key idea behind most Monte Carlo methods, is then to use a *proposal distribution* from which we can easily sample from, and subsequently make corrections to achieve approximate samples from the target distribution. Most of these extensions will require that we can evaluate the target distribution up to a normalising constant

$$p(x) = Z^{-1}\, \widetilde{p}(x),$$

where $Z = \int \widetilde{p}(x)\, dx$. Note that the Bayesian setting is a specific case, were we normally cannot compute the marginal likelihood $Z = p(\mathcal{D})$ and we can readily evaluate $\widetilde{p}(x) = p(\mathcal{D} \,|\, \theta)\, p(\theta)$. Two classical generalisations of Monte Carlo methods are *rejection sampling* and *importance sampling* (Robert and Casella, 1999). However, these approaches have severe limitations in spaces of high dimensionality — curse of dimensionality (Bellman, 1961) — since they aim at capturing the target distribution immediately and obtain independent samples (see figure 3.5). In the next section we discuss a powerful framework called Markov Chain Monte Carlo (MCMC), which constructs a progressive picture of $p(x)$ by local correlated exploration of the parameter space until uncovering all regions of interest (Chib and Greenberg, 1995).

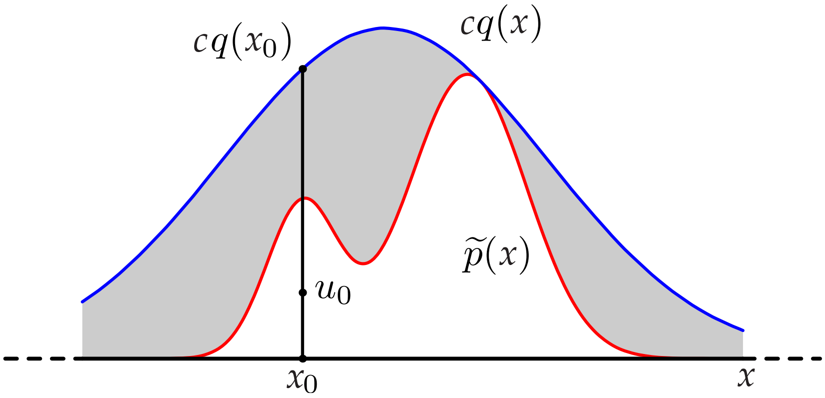


Figure 3.5 Illustration of the rejection sampling algorithm. In rejection sampling, we generate samples x_0 from an easy-to-sample proposal distribution $q(x)$, e.g. the Gaussian distribution. For each x_0, we also generate samples u_0 from a uniform distribution over $[0, cq(x_0)]$. Then, samples are rejected if they fall in the grey area between $\widetilde{p}(x)$ and the scaled version of $q(x)$. The scaling constant c is chosen such that $c\,q(x)$ is always above $\widetilde{p}(x)$. The proposal $q(x)$ should be as close as possible to $p(x)$ so the rejection rate is kept at minimum. However, this approach cannot scale to high dimensions, since the acceptance rate decreases exponentially with the dimensionality. Figure adapted from Bishop (2006).

3.4.3 Markov chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a powerful approach for circumventing the curse of dimensionality and sampling from a large class of distributions $p(x)$ without requiring the normalisation constant. This is accomplished by relaxing the requirement that the simulated draws should be independent. As the name implies, the key idea is to generate a correlated sequence of samples $x^{(s)}$ from a *Markov chain* that has the target distribution $p(x)$ as its *invariant* or *stationary* distribution (Gilks et al., 1995). In other words, in the limit of asymptotically large sample size $S \to \infty$, the parameters $x^{(s)}$ will be drawn from the correct distribution $p(x)$, and subsequently will be used to perform Monte Carlo integration using (3.39).

A *Markov chain* (Norris, 1998) is a sequence of random variables $x^{(1)}, \ldots, x^{(S)}$ such that the following conditional independence property holds

$$q\left(x^{(s+1)} \,|\, x^{(1)}, \ldots, x^{(s)}\right) = q\left(x^{(s+1)} \,|\, x^{(s)}\right). \tag{3.40}$$

This is called the Markov property, and specifies that each state is independent of all previous states given the current state. A Markov chain is then fully defined by the probability of its initial state $p(x^{(1)})$ and the conditional probabilities of the subsequent states in the form of *transition probabilities*

$$T\left(x^{(s+1)} \leftarrow x^{(s)}\right) \overset{\text{def}}{=} q\left(x^{(s+1)} \,|\, x^{(s)}\right), \tag{3.41}$$

where for discrete state space this is a conditional probability mass function and for continuous state space is a conditional probability density function. If the transition probabilities are fixed for all s, then the Markov chain is homogeneous.

A necessary condition for T is that is leaves the target distribution $p(x)$ *invariant*

$$p\left(x^{(s+1)}\right) = \int T\left(x^{(s+1)} \leftarrow x^{(s)}\right) p\left(x^{(s)}\right) dx^{(s)}. \tag{3.42}$$

In words, given a sample $x^{(s)}$ from $p(x)$, the marginal distribution over the next state $x^{(s+1)}$ is also the target distribution $p(x)$. In addition, the Markov chain should be both irreducible and aperiodic, which means that the chain should reach all states with non-zero probability in finite number of steps and that no states are only accessible at certain periods of time. If these two conditions are met, then the chain is said to be *ergodic*, that is, irrespective of the initial state $x^{(1)}$, the chain will converge to the required invariant distribution $p(x)$, which is called the equilibrium distribution. A sufficient, but not necessary, condition to ensure that $p(x)$ is the desired invariant distribution, is to choose a transition probability that satisfies reversibility or *detailed balance*

$$T\left(x^{(s+1)} \leftarrow x^{(s)}\right) p\left(x^{(s)}\right) = T\left(x^{(s)} \leftarrow x^{(s+1)}\right) p\left(x^{(s+1)}\right). \tag{3.43}$$

The MCMC algorithms that we describe below satisfy detailed balance, hence, the Markov chain will eventually generate samples that are approximately from the target distribution $p(x)$, e.g. the posterior. For more details on the above concepts see Tierney (1994) and Liu (2001).

Metropolis-Hastings

The *Metropolis-Hastings* algorithm, named after Metropolis et al. (1953) and Hastings (1970), is the workhorse of MCMC methods since most practical MCMC techniques can be considered as variants of this algorithm. In the Metropolis-Hastings algorithm the choice of the transition probability is defined in terms of a *proposal distribution* $q(x^* \mid x^{(s)})$ of moving from current state $x^{(s)}$ to a new state x^*. The Markov chain then moves towards x^* with acceptance ratio

$$\alpha = \min\left\{1, \frac{\widetilde{p}(x^*)\, q(x^{(s)} \mid x^*)}{\widetilde{p}(x^{(s)})\, q(x^* \mid x^{(s)})}\right\}. \tag{3.44}$$

If the proposal is accepted, then $x^{(s+1)} = x^*$, otherwise we maintain a record of the current state and set $x^{(s+1)} = x^{(s)}$. It is worth highlighting that we need to know the target distribution up to a constant of proportionality $p(x) = \widetilde{p}(x)/Z$, since Z cancels out in (3.44). The procedure for simulating a Markov chain using the Metropolis-Hastings algorithm is summarised in algorithm 3.2, where the accept / reject step corrects for the fact that the proposal distribution does not match the target distribution.

The algorithm is valid for any reasonable proposal distribution $q(x^* \mid x^{(s)})$, however, optimal proposals should be sought to perform effective exploration of the target distribution $p(x)$. A quantity that is important to monitor is the acceptance rate, i.e. the fraction of proposed draws that are accepted. When the acceptance rate is high, the Markov chain is probably

Algorithm 3.2 Metropolis-Hastings

```
initialize $x^{(1)}$ and number of iterations $S$
for $s = 1, \dots, S$ do
    Sample $u \sim \mathcal{U}[0, 1]$
    Propose $x^* \sim q\left(x^* \,|\, x^{(s)}\right)$
    Set $\alpha = \min\left\{1, \frac{\widetilde{p}(x^*)\, q(x^{(s)} \,|\, x^*)}{\widetilde{p}(x^{(s)})\, q(x^* \,|\, x^{(s)})}\right\}$    ▷ Acceptance ratio
    if $u \leq \alpha$ then
        Set $x^{(s+1)} \leftarrow x^*$    ▷ Accept proposal
    else
        Set $x^{(s+1)} \leftarrow x^{(s)}$    ▷ Reject proposal
    end if
end for
```

not exploring quickly the parameter space. On the other extreme of low acceptance rate the algorithm becomes inefficient, due to rejecting most of the proposed draws, because they probably lie in the parameter space where the probability of $p(x)$ is low (Robert and Casella, 1999). A common proposal distribution for continuous parameter space is the multivariate Gaussian distribution centred at the current value of parameters. When the proposal distribution is a *symmetric random walk* $q(x^* \,|\, x^{(s)}) = q(x^{(s)} \,|\, x^*)$ — such as the Gaussian distribution — the acceptance probability simplifies to the ratio of the (unnormalised) target distribution; this is the original Metropolis algorithm (Metropolis et al., 1953).

Gibbs sampling

Gibbs sampling is a widely applicable MCMC algorithm that circumvents the curse of dimensionality via conditioning, and is attributed to Geman and Geman (1984) who used this idea for analysing Gibbs distributions on lattices for image restoration. It was later popularised by Gelfand and Smith (1990), who demonstrated the wide applicability of the Gibbs sampler in Bayesian statistics. Consider a D-dimensional target distribution $p(x) = p(x_1, \dots, x_D)$ from which we wish to sample from. Instead of using a proposal distribution that will give a sample directly from the joint $p(x)$, the Gibbs sampler iteratively updates each variable x_d conditioned on the current values of the remaining variables, i.e. the full conditional $p(x_d \,|\, x_{\neg d}) \overset{\text{def}}{=} p(x_d \,|\, x_1, \dots, x_{d-1}, x_{d+1}, \dots, x_D)$. This idea is linked to the Hammersley-Clifford theorem, which states that, under mild conditions, the joint distribution is uniquely determined by the full set of conditionals (Besag, 1974). This way, the task of performing inference in a high dimensional space is simplified to sampling from (often) feasible one-dimensional conditional

Algorithm 3.3 Gibbs sampling

```
initialize $x^{(1)} = \left\{x_1^{(1)}, \dots, x_D^{(1)}\right\}$ and number of iterations $S$
for $s = 1, \dots, S$ do
    for $d = 1, \dots, D$ do
        Sample $x_d^{(s+1)} \sim p\left(x_d \,|\, x_1^{(s+1)}, \dots, x_{d-1}^{(s+1)}, x_{d+1}^{(s)}, \dots x_D^{(s)}\right)$
    end for
end for
```

distributions. A cycle through all the variables is usually considered a single iteration of the Gibbs sampler (see algorithm 3.3).

The sequence of samples from the full conditionals constitutes a valid Markov chain whose invariant distribution is the target joint distribution $p(x)$. In fact, the Gibbs sampling update is equivalent to using Metropolis-Hastings with proposal distribution

$$q\left(x^* \,|\, x^{(s)}\right) = p\left(x_d^* \,|\, x_{\neg d}^{(s)}\right) \mathbb{I}\left(x_{\neg d}^* = x_{\neg d}^{(s)}\right), \tag{3.45}$$

where $\mathbb{I}(\cdot)$ is the indicator function ensuring all components other than x_d remain unchanged. By using this proposal distribution we will actually accept every sample since the acceptance probability in (3.44) is identical to one

$$\frac{p(x^*)\, q(x^{(s)} \,|\, x^*)}{p(x^{(s)})\, q(x^* \,|\, x^{(s)})} = \frac{p(x_d^* \,|\, x_{\neg d}^*)\, p(x_{\neg d}^*)\, p(x_d^{(s)} \,|\, x_{\neg d}^*)}{p(x_d^{(s)} \,|\, x_{\neg d}^{(s)})\, p(x_{\neg d}^{(s)})\, p(x_d^* \,|\, x_{\neg d}^{(s)})} = 1, \tag{3.46}$$

where we used the fact that $x_{\neg d}^{(*)} = x_{\neg d}^{(s)}$. Since we accept all proposed samples and there are no free parameters to 'tune' the proposal distribution, the Gibbs sampler can be applied fairly automatically. Figure 3.6 illustrates the Gibbs sampler for a two-dimensional posterior distribution.

A major drawback of Gibbs sampling is that it is quite slow to move around the parameter space when the variables are highly correlated, since the updates are governed by the conditional distributions. Different strategies, such as 'blocking' and 'collapsing', have been proposed to make Gibbs sampling more efficient, e.g. see Liang et al. (2010). In cases where some of the full conditionals $p(x_d \,|\, x_{\neg d})$ are not tractable, Metropolis-Hastings updates can be embedded within the Gibbs algorithm to update only the intractable subset of variables (Andrieu et al., 2003). We should highlight that the data augmentation algorithm of Tanner and Wong (1987), amounts to a special case of Gibbs sampling, where we group the variables in two classes: the hidden variables and the parameters. Gibbs sampling is widely applied when representing joint distributions as directed graphical models, where the conditional independence assumptions are encoded in the graph. This way, the full conditional of each variable x_d is simplified to

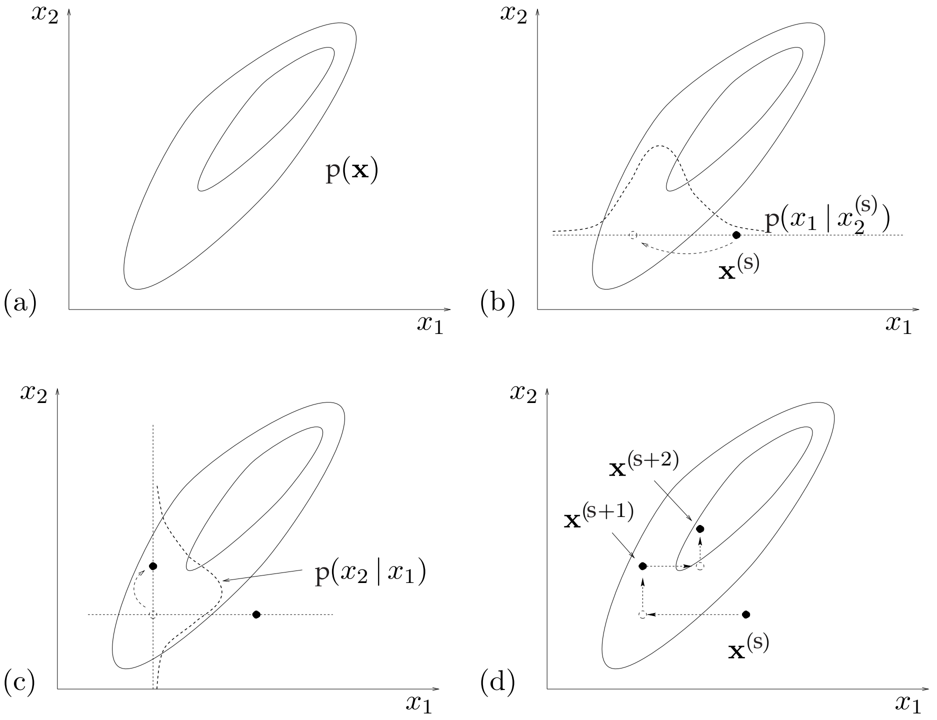


Figure 3.6 The Gibbs sampling algorithm for a two-dimensional posterior distribution. (a) Joint distribution $p(x)$ from which we wish to obtain samples from. (b) Being in state $x^{(s)}$, we draw x_1 from the full conditional distribution $p(x_1 \,|\, x_2^{(s)})$. (c) Then, we continue by sampling x_2 from the conditional $p(x_2 \,|\, x_1)$. (d) Example iterations of Gibbs sampling; note that the trajectories are parallel to each axis, since we update each variable by keeping the rest fixed. Figure adapted from MacKay (2003).

looking at its Markov blanket, that is $p(x_d \,|\, x_{\neg d}) = p(x_d \,|\, \mathrm{MB}(x_d))$, and forms the basis for general purpose software, such as *BUGS* (Bayesian Updating with Gibbs Sampling) (Lunn et al., 2009) which is widely used in computational statistics. Note the similarity between the Gibbs sampler and mean field variational inference. In Gibbs sampling we iteratively sample from each variable's full conditional, whereas in variational inference we take expectations over the current estimates of the neighbouring variables to obtain the variational distribution.

Remarks

To circumvent the curse of dimensionality, MCMC methods produce dependent samples using a Markov chain, which can be viewed as performing 'local exploration' of the parameter space, while predominantly ignoring regions of insignificant probability. However, MCMC is not a panacea and has weaknesses which should be addressed with caution. One issue is the local-trap problem, where the sampler gets trapped in a specific mode of the target distribution. Although in theory the local proposals will eventually visit all modes, they might spend disproportional time on each mode, providing biased estimates when summarising the samples from a finite MCMC chain.

In addition, since we create a Markov chain the early samples will depend on the initial state $x^{(1)}$, which presumably is not drawn from the target distribution. However, there is no guarantee that after S samples our chain will have reached the invariant distribution and adequately explored the target distribution. Convergence diagnostics have been proposed to resolve this issue, although they do not provide conclusive tests of convergence (Brooks and Gelman, 1998). On the contrary, the deterministic nature of the variational inference machinery makes it easier to assess convergence by looking at a single number, the difference in ELBO between successive iterations. Once the MCMC samples are considered to have converged to the target distribution, the samples prior to this point are often discarded as *burn-in* to remove dependence on the initial distribution (Gelman et al., 1995). Also, the correlation between successive samples has an important effect on the approximation we obtain by Monte Carlo integration using (3.39). In general, MCMC methods require a larger number of samples to achieve the same estimator variance as independent Monte Carlo estimates. A useful statistic is the effective sample size

$$ESS = \frac{S}{\tau} = \frac{S}{1 + 2\sum_{k=1}^{\infty} \rho_k(x)}, \tag{3.47}$$

where $\rho_k(x)$ is the autocorrelation at lag k for variable x, and τ is the autocorrelation time (Geyer, 1992). High τ indicates slow mixing of the chain, where the mixing is defined as the number of iterations required to obtain decorrelated samples in the chain (Robert and Casella, 1999). Hence, the ESS provides an estimate of the equivalent number of independent samples that the Markov chain represents, and can be used to compare competing MCMC methods if we standardise for computational run time.

Chapter 4

BPRMeth: Quantifying spatial correlations of DNA methylation

As outlined in chapter 2, DNA methylation is a heritable epigenetic mark that plays an important role in gene regulation and is associated with a broad range of biological processes of direct clinical relevance. The canonical understanding is that hyper-methylation of CpG islands (CGIs) near promoter regions is generally associated with transcriptional repression (Schübeler, 2015); however, outside of this well documented case, the association between DNA methylation across promoter-proximal regions and transcript abundance is considerably weaker and poorly understood (Jones, 2012; Varley et al., 2013).

The gold standard method to measure DNA methylation on a genome-wide scale is bisulfite treatment of DNA followed by sequencing, termed as BS-seq (see section 2.3) (Frommer et al., 1992). However, despite the widespread take up of bulk BS-seq technology, statistical modelling of such data is still challenging, yet it is crucial in order to uncover biological regulatory mechanisms. Analysis of BS-seq data has mainly focused on identifying differentially methylated regions (DMRs) across different conditions. Some notable DMR methods are BSmooth (Hansen et al., 2012), M^3D (Mayo et al., 2015) and ABBA (Rackham et al., 2017). While DMR detection methods are often crucial ingredients in exploratory data analysis pipelines, they do not provide a clear platform to quantitatively understand the relationship between DNA methylation and gene expression. Most studies use DMR detection as a pre-filtering step, and then simply correlate mean methylation levels across each region — often taken to be promoter-proximal — with gene expression. Adopting this simple approach, genome-wide studies have reported only modest correlation between mean methylation and gene expression, with Pearson's correlation coefficient $r \simeq$ -0.3 (Bock et al., 2012; Hansen et al., 2011).

In this chapter[1], we argue that part of the difficulty in quantitatively associating methylation levels with gene expression resides in the simplistic encoding of DNA methylation across a

[1] Most of the material in this chapter have appeared before in Kapourani and Sanguinetti (2016) and the manuscript was written by myself with Guido Sanguinetti providing feedback and editing.

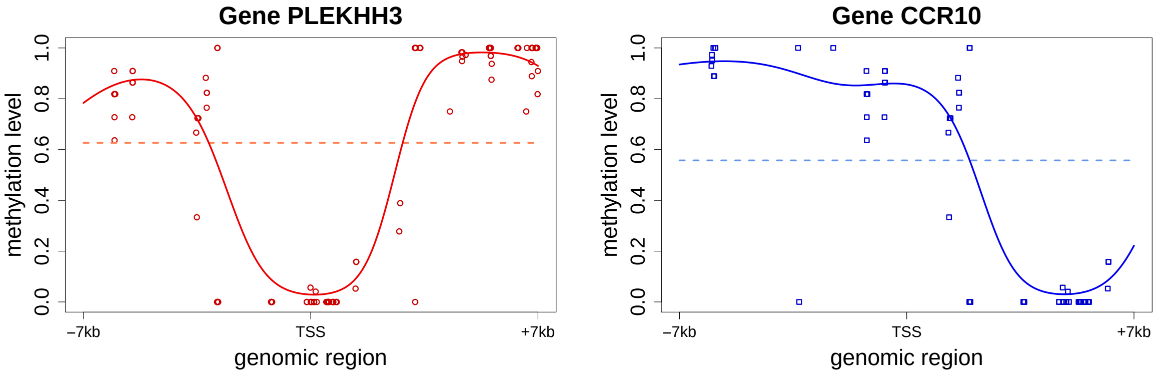


Figure 4.1 Promoter-proximal regions with characteristic methylation patterns. Methylation patterns for the PLEKHH3 and CCR10 genes from the K562 cell line over ±7 kb promoter region. Each point represents the relative CpG location w.r.t. TSS and the corresponding DNA methylation level. The dashed horizontal lines show mean methylation levels. The shapes of methylation profiles are very different, however, the mean methylation level cannot explain them. Also, note that there are no CpG measurements in the (-6 kb, -4 kb) region for the CCR10 gene, and the inferred methylation profiles can be thought of as imputing the missing values by taking into consideration the spatial correlation of nearby CpGs.

region as a simple average. DNA methylation often displays reproducible, spatially correlated patterns (*methylation profiles*); figure 4.1 shows two example promoter-proximal regions from an ENCODE dataset (Dunham et al., 2012), which clearly display such spatial correlations. This spatial reproducibility was exploited by Mayo et al. (2015) to provide more powerful tests for calling DMRs, and by Vanderkraats et al. (2013) to group genes with similar differential methylation patterns and corresponding expression changes. These results suggest that a precise quantification of the spatial variability in the DNA methylation mark may aid the quest to quantitatively understand the interplay between methylation and transcription. Here we propose a probabilistic model of methylation profiles which allows us to associate with each region of interest a set of features capturing precisely the methylation profile across the region. We then show that, using such features, we can construct an accurate machine learning predictor of gene expression from DNA methylation, achieving test correlations twice as large as previously reported.

4.1 Methods

Here we introduce BPRMeth (Binomial Probit Regression for Methylation)[2], a probabilistic machine learning methodology to quantify the profile of DNA methylation across genomic regions from bulk BS-seq data. A schematic illustration of the proposed method is shown in figure 4.2. Briefly, the method is based on a generalised linear model (GLM) of basis function

[2] In a similar fashion to the evolution of epigenetics in the past 50 years, the meaning of BPRMeth will evolve during this thesis, and in Chapter 5 we will re-introduce it as Bayesian Probit Regression for Methylation.

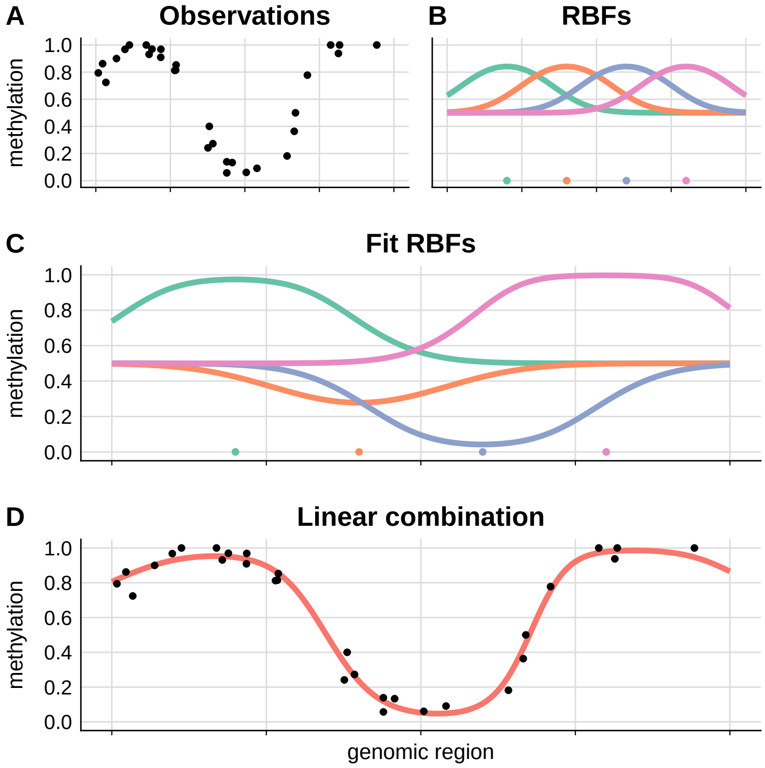


Figure 4.2 BPRMeth model overview. The inputs to the model are the CpG methylation levels (**A**) and number of basis functions (**B**); we consider radial basis functions (RBFs) by default (see below). The BPRMeth model estimates the optimal coefficients for each RBF (**C**) and infers the underlying methylation profile by linear combination of the fitted RBFs (**D**).

regression coupled with a binomial observation likelihood, which allows us to associate each region with a set of basis function coefficients that capture the methylation profile. We show how such higher-order features can then be used in downstream analysis to yield a significantly improved estimate of the correlation between methylation and gene expression, and to identify prototypical methylation profiles that explain most variability across promoter regions.

4.1.1 Modelling DNA methylation profiles

As in most NGS-based assays, the output of a BS-seq experiment is a set of reads aligned to the genome; the main difference is that the bisulfite treatment converts any unmethylated cytosine to thymine. Thus, the same base on the genome will appear as cytosine on some reads, and as thymine on others; the ratio of reads containing a cytosine readout to total reads gives a measurement of the sample methylation level. This measurement process at a single cytosine can be naturally modelled with a binomial distribution, where the number of successes represents the number of reads on which the cytosine actually appears as C, and the number of attempts is the total number of reads mapping to the specific site. Let ν be the total number of reads that are mapped to a specific CpG site, and let s of these reads contain methylated cytosines. Then, for each CpG site we assume that $s \sim \mathcal{B}\text{inom}(\nu, \rho)$, where ρ is the unknown methylation level.

In this chapter, and in many practical studies, we are interested in learning the methylation patterns of fixed-width genomic regions, e.g. promoters or enhancers. Each genomic region m, where $m = 1, \ldots, M$, can be represented as a vector of CpG locations $\boldsymbol{x}_m$, where each entry corresponds to the location of the CpG in the genomic region relative to a reference point, such as the transcription start site (TSS). It should be noted that the vector lengths I_m may vary between different genomic regions, since they depend on the number of actual CpG dinucleotides found in each region. For each region m, we also have a vector of observations $\mathbf{y}_m$, containing the methylation levels of the corresponding CpG sites; each entry consists of the tuple $y_{mi} = (s_{mi}, \nu_{mi})$, where, s_{mi} is the number of 5mC reads mapped to the i^{th} CpG site in region m, and ν_{mi} corresponds to the total number of reads. Collectively, the data for a given region are summarised as $\mathcal{D}_m = \{\boldsymbol{x}_m, \mathbf{y}_m\}$.

We are often interested in comparing epigenetic patterns, e.g. the task might be to cluster promoter regions with respect to methylation patterns. However, working directly with the observed data $\mathcal{D}_m$ might be complicated due to the variability in the vector lengths. To enable comparisons between genomic regions we formulate our problem as a regression problem, where the methylation profile associated with a genomic region m is defined as a (latent) function $f\colon m \rightarrow (0, 1)$, which takes as input the genomic coordinate along the region and returns the propensity for each locus to be methylated. More specifically, for a specific region m we assume that each observable follows a binomial distribution

$$s_{mi} \sim \mathcal{B}\text{inom}(\nu_{mi}, \rho_{mi}), \tag{4.1}$$

where the unknown 'true' methylation level ρ_{mi} has as covariates the CpG locations x_{mi}. Then, we define the binomial regression model as

$$\begin{aligned} \eta_{mi} &= \mathbf{w}_m^{\top} \mathbf{x}_{mi}, \\ f_m(x_{mi}) &= \rho_{mi} = g^{-1}(\eta_{mi}), \end{aligned} \tag{4.2}$$

where $\mathbf{w}$ are the regression coefficients, $\mathbf{x}_{mi} = (1, x_{mi})$ are the covariates, and $g(\cdot)$ is the link function that allows us to move from the systematic components η_{mi} to mean parameters ρ_{mi}. The *probit regression* model is obtained if we define $g^{-1}(\cdot) = \Phi(\cdot)$ — where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution — ensuring that f takes values in the $[0, 1]$ interval (see subsection 3.1.1).

Feature extraction

As shown in figure 4.1, we do not expect a linear relationship between the methylation levels y and the CpG locations x. In order to enforce spatial smoothness and obtain a compact representation for this function in terms of interpretable features, we represent the profile function as a linear combination of fixed non-linear basis functions[3] $h_d(\cdot)$ of the input space x,

$$\eta = w_0 + \sum_{d=1}^{D} w_d h_d(x) = \mathbf{w}^\top \mathbf{h}(x), \tag{4.3}$$

where $\mathbf{x} \overset{\text{def}}{=} \mathbf{h}(x)$ is a vector of all basis function values at input location x, and $\mathbf{w} \in \mathbb{R}^{D+1}$ represents the regression coefficients for each basis function. This 'basis function trick' allows us to expand our input space (i.e. feature expansion), by replacing one independent variable with multiple variables that are derived from it in a non-linear fashion (Murphy, 2012). We should emphasise that even though η is still linear with respect to the parameters $\mathbf{w}$, the latent function $f(x)$ — which is of interest — is non-linear due to the presence of the probit transformation.

In this chapter, we consider radial basis functions[4] (RBFs) since they are local functions of the input variable, so that changes in one locus of the input space do not affect all other loci. For a single input observation x, the RBF takes the following form

$$h_d(x) = \exp\left(-\gamma \,||\, x - \mu_d \,||^{\,2}\right). \tag{4.4}$$

Here μ_d denotes the location or centre of the d^{th} basis function in the input space and γ controls the spatial scale. To ensure that the extracted features, encoded by $\mathbf{w}_m$, can be directly compared and used for downstream analysis, the RBF parameters are held fixed and will be the same for all genomic regions. More specifically, the centres μ_d are assumed by default to be equally spaced across the genomic region, and the spatial scale is set empirically to $\gamma = D^2/(|l_{min}| + |l_{max}|)^2$, so that γ depends on the number of basis functions, where $|l_{min}|$ and $|l_{max}|$ denote the absolute values of the minimum and maximum locations in each genomic region, respectively. For example, in a ±2 kb promoter region we would have $-l_{min} = l_{max} = 2000$, however, for numerical stability we scale the genomic locations to the $[-1, 1]$ interval.

Parameter estimation

Given the latent function $f_m(x)$, the observations y_{mi} for each CpG site are independent and identically distributed binomial variables, so we can define the joint log-likelihood for region m in factorised form

$$\log p(\mathbf{y}_m \,|\, \mathbf{X}_m, \mathbf{w}_m) = \sum_{i=1}^{I_m} \log \left| \mathcal{B}\text{inom}\left(s_{mi} \,|\, \nu_{mi}, \Phi(\mathbf{w}_m^\top \mathbf{x}_{mi})\right)\right|, \tag{4.5}$$

[3] An alternative and powerful approach is to *adapt* the basis functions to the data during training, with the most successful example being artificial neural networks (Bishop et al., 1995), which are currently re-branded as *deep learning* (LeCun et al., 2015). The price to be paid for this type of models is that the likelihood function is no longer a convex function of the model parameters, and in general requires large amounts of training data.

[4] In chapter 5 we enhance the BPRMeth model to support additional basis functions.

where $\mathbf{X}_m = \{\mathbf{x}_{m1}, \ldots \mathbf{x}_{mI}\}$. Notice that the BPRMeth model explicitly accounts for the coverage variability across CpG sites through the use of the binomial observation model: as the variance of a binomial distribution decreases rapidly with the number of attempts, the model will be very strongly constrained by highly covered sites. Hence, it handles in a principled way the uncertainty present in low coverage reads during the analysis of BS-seq data.

Inferring the methylation profiles $f_m(x)$ for each genomic region is equivalent to optimising the model parameters $\mathbf{w}_m$. The parameters $\mathbf{w}_m$ can be considered as the extracted features which quantitate precisely notions of shape of a methylation profile. Optimising $\mathbf{w}_m$ involves maximising (4.5) for each genomic region; however, when increasing the number of basis functions, we also increase the resolution for the shape of the methylation profiles, which might lead to over-fitting. To ameliorate this issue, we maximise a penalised version of (4.5), by adding a regularisation term to the log-likelihood function encouraging the weights to decay to zero

$$\mathcal{J}(\mathbf{w}_m) = \log p(\mathbf{y}_m \,|\, \mathbf{X}_m, \mathbf{w}_m) - \lambda \mathbf{w}_m^\top \mathbf{w}_m, \tag{4.6}$$

where λ is the regularisation parameter controlling the amount of shrinkage on the regression coefficients. This approach is known as ridge regression or L_2 regularisation (Friedman et al., 2001). Direct maximisation of $\mathcal{J}(\mathbf{w}_m)$ is intractable due to the presence of the probit transformation, hence, we perform numerical optimisation using the conjugate gradients[5] method (Hestenes and Stiefel, 1952). The conjugate gradients approach is a first order numerical optimisation algorithm which requires deriving the gradient of (4.6) w.r.t. parameters $\mathbf{w}_m$

$$\nabla_{\mathbf{w}_m} \mathcal{J}(\mathbf{w}_m) = \nabla_{\mathbf{w}_m} \left[\sum_{i=1}^{I_m} \log \left| \mathcal{B}\text{inom}\left(s_{mi} \,|\, \nu_{mi}, \Phi(\mathbf{w}_m^\top \mathbf{x}_{mi}) \right) \right| - \lambda \mathbf{w}_m^\top \mathbf{w}_m \right]. \tag{4.7}$$

4.1.2 Predicting gene expression

To quantitatively predict expression at each promoter region, we construct another regression model whose input are the higher-order methylation features extracted from each promoter-proximal region. The performance of the regression model is evaluated by computing the root-mean squared error (RMSE, equation (D.2)) and the Pearson's correlation coefficient (r, equation (D.3)) between the predicted and the measured (log-transformed) gene expression levels. We compare the prediction performance of BPRMeth with the standard approach (Bock et al., 2012; Hansen et al., 2011), which uses the average methylation level across a region as input feature (this approach can be thought of as fitting a constant function across each genomic region). We have tested both a linear regression model and a variety of non-linear models, such as support vector machines (SVM) for regression (Schölkopf and Smola, 2002), random forests

[5] The conjugate gradients method can be replaced by any numerical optimisation approach, e.g. BFGS or gradient descent (Nocedal and Wright J., 1999).

(RF) (Breiman, 2001) and multivariate adaptive regression splines (MARS) (Friedman, 1991). For the rest of the analysis we use the SVM regression, since it consistently outperforms the competing methods (see table 4.1).

In addition to the methylation profile features, we consider two supplementary sources of information which could plausibly act as confounders[6] in the predictions. The first feature accounts for the goodness-of-fit of each methylation profile to the observed methylation data using the RMSE as error measure, intuitively quantitating the noisiness in the methylation profile. The second feature considers the number of CpG dinucleotides present in each promoter region. It is implicated that CpG density may play a functional role in regulating gene expression, with the main evidence being the existence of CpG islands (Deaton and Bird, 2011).

4.1.3 Clustering methylation profiles

In the BPRMeth model, the observed spatially correlated methylation patterns are treated as samples taken from underlying continuous smooth processes, which may encode specific biological function. The underlying idea for clustering methylation profiles assumes that genomic regions belonging to the same cluster might share similar functionality for gene regulation. A similar reasoning is extensively applied for clustering time-course expression data from microarray experiments, where genes with similar expression profiles belong to the same functional group (Luan and Li, 2003; Qin and Self, 2006; Song et al., 2007; Storey et al., 2005). Our approach differs in three main aspects: (1) we are modelling spatial correlations across the genome instead of time-course data, (2) the observation model is binomial, whereas in microarray expression data we often assume a Gaussian likelihood, and (3) we have varying observations (and often of different length) between genomic regions, while measurements in time-course data mostly occur at the same time points and are of equal length.

To cluster methylation profiles across genomic regions we consider a finite mixture modelling approach (see subsection 3.3.2) (McLachlan and Peel, 2004). We assume that the methylation profiles can be partitioned into at most K clusters, and each cluster k is modelled using the BPRMeth likelihood as the observation model; effectively regions belonging to the same cluster will share the same regression coefficients $\mathbf{w}_k$. The log-likelihood function for the mixture model is defined as

$$\log p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta}) = \sum_{m=1}^{M} \log \left| \sum_{k=1}^{K} \pi_k \, p(\mathbf{y}_m \,|\, \mathbf{X}_m, c_m = k, \mathbf{w}_k) \right|, \tag{4.8}$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_k, \mathbf{w}_1, \ldots, \mathbf{w}_k)$, $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}$, π_k are the mixing proportions (with $\pi_k \in (0, 1)\ \forall k$ and $\sum_k \pi_k = 1$), and c_m are latent variables indicating the component that is responsible for each genomic region. To estimate the model parameters $\boldsymbol{\theta}$ the Expectation Maximisation (EM) algorithm (Dempster et al., 1977) is considered (see subsection 3.3.1).

[6] In statistics confounding factors refer to (unaccounted) variables that affect both the dependent and independent variables leading to inaccurate associations.

Algorithm 4.1 EM algorithm for BPRMeth model

```
1:  initialize number of clusters $K$ and regularisation parameter $\lambda$
2:  for $m = 1, \ldots, M$ do
3:      Set $\mathbf{w}_m \leftarrow \underset{\mathbf{w}_m}{\text{argmax}} \left| \log p(\mathbf{y}_m \,|\, \mathbf{X}_m, \mathbf{w}_m) - \lambda \mathbf{w}_m^\top \mathbf{w}_m \right|$      ▷ Run BPRMeth
4:  end for
5:  initialize $\{\mathbf{w}_k, \pi_k\} \leftarrow$ k-means$(\mathbf{w}, K)$
6:  while EM has not converged do
7:      E-step
8:      Set $\gamma_m(k) \leftarrow \frac{\pi_k \, p(\mathbf{y}_m \,|\, \mathbf{X}_m, c_m = k, \mathbf{w}_k)}{\sum_{j=1}^{K} \pi_j \, p(\mathbf{y}_m \,|\, \mathbf{X}_m, c_m = j, \mathbf{w}_j)}$      ▷ Update responsibilities
9:      M-step
10:     Set $\pi_k \leftarrow \frac{1}{M} \sum_m \gamma_m(k)$      ▷ Update mixing proportions
11:     Set $\mathbf{w}_k \leftarrow \underset{\mathbf{w}_k}{\text{argmax}} \; \ell(\mathbf{w}_k)$      ▷ Update coefficients
12: end while
```

Briefly, the EM algorithm maximises a lower bound on the log-likelihood by alternating between inferring the latent variables given the parameters (E-step), and optimising the parameters given the posterior statistics of the latent variables (M-step). The procedure for clustering methylation profiles using EM is shown in algorithm 4.1. All quantities are relatively easy to compute, except step 11, which requires optimising the regression coefficients $\mathbf{w}_k$, that is,

$$\ell(\mathbf{w}_k) = \sum_m \gamma_m(k) \sum_i \log \left| \mathcal{B}\text{inom}(s_{mi} \,|\, \nu_{mi}, \Phi(\mathbf{w}_k^\top \mathbf{x}_{mi})) \right| . \tag{4.9}$$

Direct optimisation of $\ell(\mathbf{w}_k)$ with respect to parameters $\mathbf{w}_k$ has no closed-form analytical solution, hence, we resort again to numerical optimisation strategies, with the gradient derived as

$$\begin{aligned} \nabla_{\mathbf{w}_k} \ell(\mathbf{w}_k) = \sum_m \gamma_m(k) \sum_i \Big(& s_{mi} \; \Phi(\mathbf{w}_m^\top \mathbf{x}_{mi})^{-1} - \\ & (\nu_{mi} - s_{mi}) \big(1 - \Phi(\mathbf{w}_m^\top \mathbf{x}_{mi})\big)^{-1} \Big) \phi(\mathbf{w}_m^\top \mathbf{x}_{mi}) \mathbf{x}_{mi}, \end{aligned} \tag{4.10}$$

where $\phi(\cdot)$ is the probability density function (pdf) of the standard normal distribution. This variant of the EM algorithm is known as generalised EM, and it is proved to converge to the maximum likelihood estimate (Wu, 1983). It should be noted that the regularised version of the BPRMeth likelihood, given in (4.6), can be easily incorporated in the clustering approach for updating the regression coefficients.

4.2 Datasets

To evaluate the performance of BPRMeth we use the following three cell lines that are publicly available from the ENCODE project consortium:

1. K562 cell line, coming from a human female with chronic myelogenous leukaemia.
2. GM12878 lymphoblastoid cell line, produced from the blood of a female donor with northern and western European ancestry by EBV transformation.
3. H1-hESC embryonic stem cells, coming from a human male.

4.2.1 Data preprocessing

The RRBS data for all three cell lines are produced by the Myers Lab at HudsonAlpha Institute for Biotechnology (GEO: GSE27584). The data are already aligned to the *hg19* human reference genome, and can be downloaded from the web accessible database at UCSC (https://genome.ucsc.edu/ENCODE/). The methylation data are preprocessed as follows: (1) Download BED formatted files for each cell line, including the available replicate files. (2) Pool replicates to obtain higher read coverage on each CpG site. (3) Ignore strand information. (4) Discard CpGs with less than 4 read coverage. (5) Discard the sex chromosomes, as well as the M chromosome, i.e. mitochondrial DNA. (6) Group together CpGs to create promoter methylation regions.

To investigate the correlation between DNA methylation profiles and gene expression levels, we use the corresponding paired-end RNA-seq data produced by Caltech (GEO: GSE33480). The RNA-seq data are mapped to the *hg19* human reference genome using TopHat and transcription quantification, in FPKM (Fragments Per Kilobase transcript per Million mapped reads), is produced using Cufflinks (Trapnell et al., 2012). The expression data are then preprocessed as follows: (1) Download GTF formatted files for each cell line. (2) Convert to BED format using the BEDOPS tool (Neph et al., 2012). (3) Discard the sex chromosomes, as well as the M chromosome. (4) Each RNA-seq file contains metadata information for each transcript, such as TSS and transcript type (e.g. protein coding, pseudogene, etc.). (5) Keep only protein coding genes using the metadata information. (6) Then $\log_2$ transform the gene expression values measured in FPKM. (7) Using the TSS information create promoter regions by taking N base pairs upstream and downstream w.r.t TSS, resulting in promoter regions of length 2N base pairs.

4.3 Results

4.3.1 Methylation profiles are highly correlated with gene expression

Initially, we examine whether gene expression levels might be predictable from DNA methylation patterns alone. We therefore extract higher-order features from promoter regions of ±7 kb

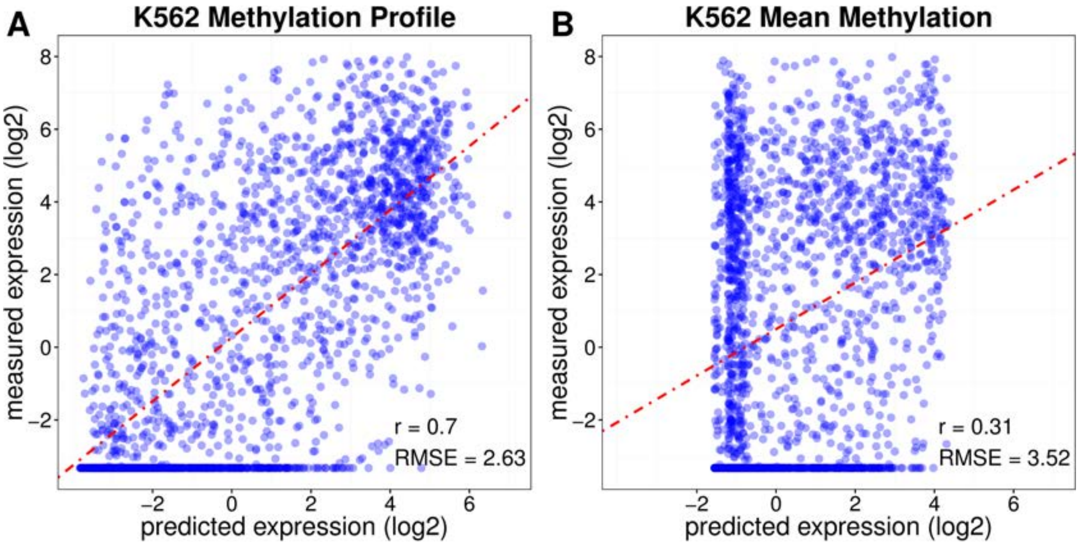


Figure 4.3 Quantitative relationship between DNA methylation and gene expression in the K562 cell line. (**A**) Scatter plot of predicted gene expression using the BPRMeth model on the x-axis versus the measured ($\log_2$-transformed) gene expression values for the K562 cell line on the y-axis. Each shaded blue dot represents a different gene and the darker the colour, the higher the density of points. The red dashed line indicates the linear fit between the predicted and measured expression values, which are highly correlated. (**B**) Scatter plot of predicted and measured gene expression values when using the average methylation level as input feature in the SVM model; correlation has decreased substantially.

around the TSS by learning the corresponding methylation profiles using the BPRMeth model. To ensure that the promoter-proximal regions will have enough data to learn reasonable methylation profiles, we discard regions with less than 15 CpGs and restrict our attention to regions which exhibit spatial variability in methylation levels. We applied the same preprocessing steps for the three ENCODE cell lines, which resulted in 7,093 promoters for K562, 6,022 for GM12878 and 5,753 for H1-hESC cell line. We model the methylation profiles using nine RBFs, which results in ten extracted features including the bias term. In addition to these features, we use the goodness-of-fit in RMSE and the CpG density across each region. We then train the SVM regression model on the resulting 12 features using a random subset of 70% of the promoter-proximal regions. We test the model's ability to quantitatively predict expression levels on the remaining 30% of the data. Our results show a striking improvement in prediction accuracy when compared to using the mean methylation level as input feature.

Figure 4.3A shows a scatter plot with predicted and measured gene expression values for the K562 cell line, with Pearson's correlation coefficient $r = 0.7$ and RMSE $= 2.63$, demonstrating that the shape of methylation patterns across promoter-proximal regions is well correlated with transcript abundance. Figure 4.3B shows the performance of the regression model when using the mean methylation level as input feature. It is evident that this approach cannot capture the diverse patterns present across promoter-proximal regions, leading to poor prediction accuracy ($r = 0.31$ and RMSE $= 3.52$). Notice that the mean methylation approach erroneously

	SVM	Random Forest	MARS	Linear Regression
K562	0.7	0.7	0.68	0.66
GM12878	0.61	0.61	0.58	0.52
H1-hESC	0.5	0.49	0.49	0.48

Table 4.1 Pearson's correlation coefficient r between predicted and measured gene expression levels using different regression models for all cell lines considered in this study.

predicts gene expression values only in the (-2, 4) interval, whereas the BPRMeth model captures more accurately the dynamic range of expression. Interestingly, the mean approach erroneously predicts the majority of genes to have expression value around -1, clearly indicating that summarising DNA methylation by a single average is insufficient to capture the complex relationship with expression. In addition, one should observe the horizontal stripe around -3 on both figures: these are genes whose lack of expression cannot be attributed to DNA methylation patterns, possibly implicating other regulatory mechanisms (e.g. histone marks, binding of transcription factors, and chromatin accessibility), or difficulties in the measurement process of RNA-seq experiments (e.g. due to mRNA capturing inefficiencies or genes having multiple promoters). Table 4.1 shows the prediction performance across all cell lines using different regression models.

We then consider the relative importance of the various features in predicting gene expression: in particular, we are interested in determining whether including goodness-of-fit or CpG density as covariates has any impact on predictive performance. For each cell line, we learn five SVM regression models, each having a different number of input features. The first four models consider as input the extracted higher-order methylation features with a combination of the two additional features we described in the previous section, whereas the last model takes the average methylation level as input feature. To statistically assess our results, we perform 20 random splits in training and test sets and evaluate the model performance on the corresponding test sets. Figure 4.4 shows boxplots of Pearson's r for the three ENCODE cell lines, where each boxplot indicates the performance of the prediction model on the 20 random splits of the data. The results demonstrate that by considering higher-order features we can build powerful predictive models of gene expression; and in the case of K562 and GM12878 we have more than two-fold increase in correlation.

Concentrating on the importance of the additional features for the prediction process, we observe that the addition of CpG density does not have a significant prediction improvement compared to using only the shape of methylation profiles as input features (paired Wilcoxon test p-value = 0.22, 0.18 and 0.02 for K562, GM12878 and H1-hESC, respectively). On the other hand, the goodness-of-fit of the methylation profile in terms of RMSE significantly increases the prediction performance (paired Wilcoxon test p-value = 4.8e-05, 4.8e-05 and 0.0001 for K562, GM12878 and H1-hESC, respectively). In addition, we explore the effect of considering

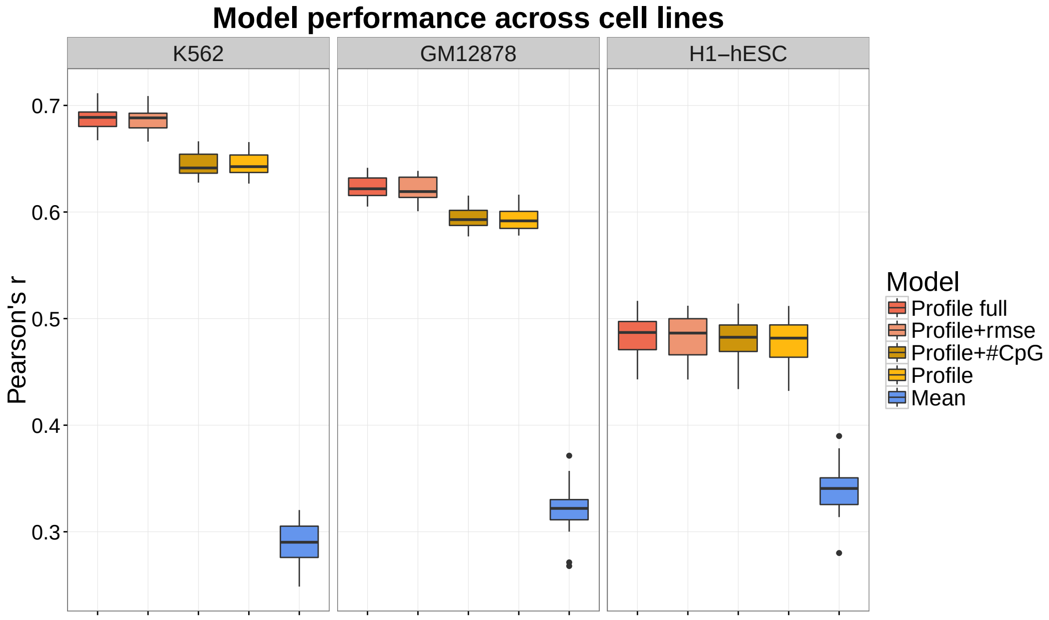


Figure 4.4 Boxplot of Pearson's r between predicted and measured expression for the three ENCODE cell lines with different input features for the SVM regression model. The 'Profile full' model corresponds to the extracted BPRMeth features plus the two additional features. Each boxplot indicates the performance using 20 random splits of the data in training and test sets. Paired Wilcoxon test shows that the high quantitative relationship between the shape of DNA methylation and expression exists in various cell lines, and is significantly better predictor than using the average methylation level (p-value = 8.4e-12). Regarding the two additional features, we observe that the goodness-of-fit measured in RMSE has a positive impact in correlation, whereas the CpG density does not improve the prediction performance. Paired Wilcoxon tests between K562 and other cell lines, show that K562 has significantly higher prediction accuracy (p-value = 4.8e-05 for both GM12878 and H1-hESC).

different promoter region windows. Table 4.2 shows Pearson's r when considering various length promoter-proximal regions around the TSS. In general, the BPRMeth model maintains its high predictive power across all cell lines for varying promoter region windows.

Cell Line	±2 kb	±3 kb	±4 kb	±5 kb	±6 kb	±7 kb	±8 kb	±9 kb
K562	0.63	0.69	0.69	0.67	0.67	**0.70**	0.67	0.67
GM12878	0.62	0.62	0.64	0.61	0.62	**0.61**	0.61	0.61
H1-hESC	0.46	0.49	0.48	0.43	0.49	**0.50**	0.47	0.49

Table 4.2 Pearson's r between predicted and measured expression for varying promoter region windows.

4.3.2 Methylation profiles are predictive of gene expression across cell lines

We demonstrated that gene expression levels are effectively predicted from the BPRMeth model by using higher-order methylation features among various cell lines. Next, we further explore if

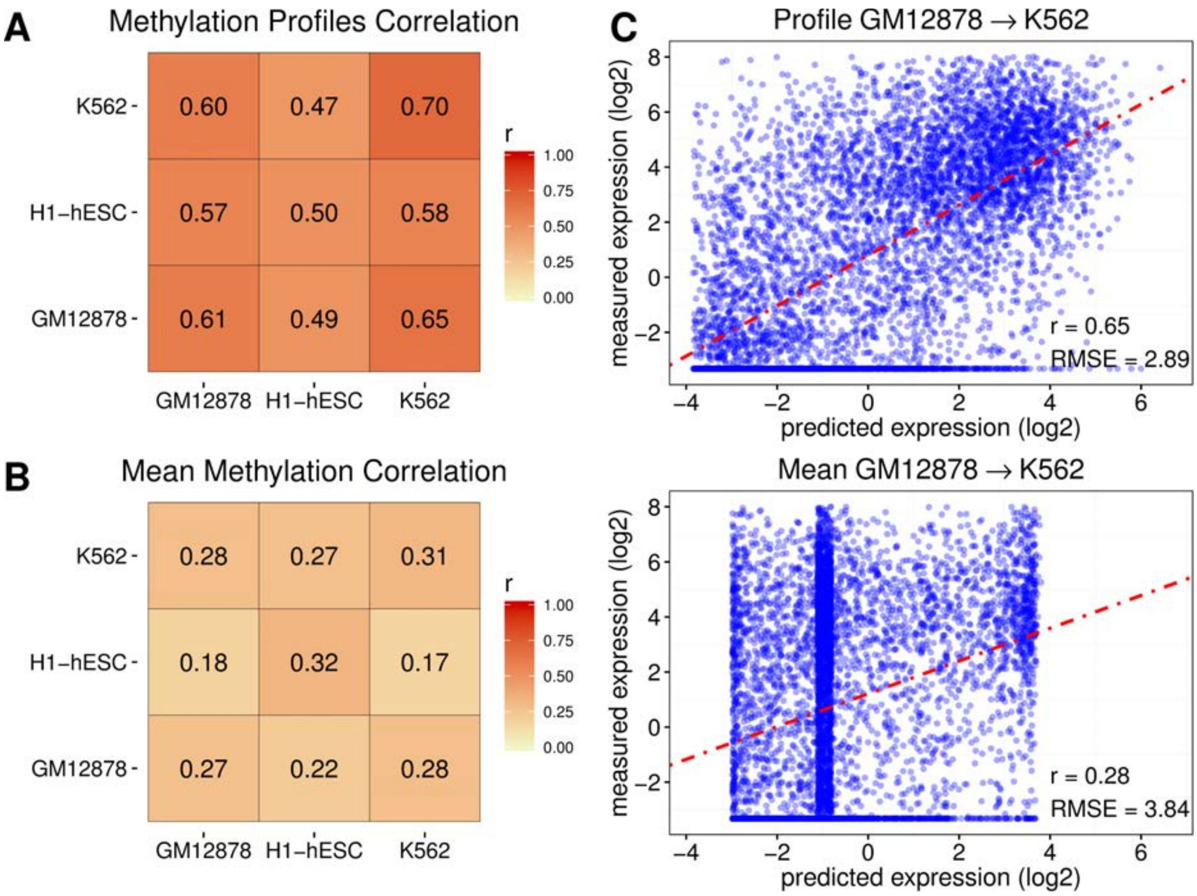


Figure 4.5 Prediction accuracy across cell lines. (**A**) Confusion matrix of Pearson's r across cell lines when using the BPRMeth model with nine RBFs as input features to the regression model. Each (i, j) entry of the confusion matrix, corresponds to training a regression model from the i^{th} cell line and predicting gene expression levels for the j^{th} cell line. The colour of the confusion matrix corresponds to Pearson's r value, the darker the colour the higher the correlation. (**B**) The corresponding correlation coefficients r when using the mean methylation level as input feature to the regression model. (**C**) Application of the model learned from GM12878 cell line to predict expression levels of the K562 cell line, using methylation profiles (top) and mean methylation levels (bottom) as input features.

the proposed model maintains predictive power across different cell lines. That is, we apply the regression model trained on one cell line to predict expression levels in another cell line, by using the inferred methylation profiles in those cell lines as input features to the regression model. Figures 4.5A-B show confusion matrices of Pearson's correlation coefficients for the cross-cell line prediction process, using BPRMeth and the mean methylation level approach, respectively. Figure 4.5C shows an example of applying the model learned from GM12878 methylation patterns to predict expression levels of the K562 cell line. The BPRMeth model effectively predicts gene expression (r = 0.65 and 0.49 for predicting K562 and H1-hESC, respectively), while, the mean methylation approach provides a poor correlation (r = 0.28 and 0.22 for predicting K562 and H1-hESC, respectively).

The results indicate that the quantitative relationship between methylation profiles and mRNA abundance is not cell line specific, and the model captures patterns of association between these two layers that hold across different cell lines. Although the proposed models have high prediction accuracy across all cell lines, the H1-hESC cell line shows consistently weaker correlations. This finding is in line with recent studies reporting weaker correlations of gene expression and chromatin features for the H1-hESC cell line (Dong et al., 2012), and with observations that mRNA-encoding genes in stem cells are transcriptionally paused during cell differentiation (Min et al., 2011).

4.3.3 Clustering methylation profiles across promoter-proximal regions

We next use the higher-order methylation features to cluster DNA methylation patterns across promoter-proximal regions and examine whether distinct methylation patterns across different cell lines are associated to gene expression levels. We apply the same preprocessing steps described in the previous section and we consider genomic regions of ±7 kb around the TSS. The total number of clusters was set to five, after applying the Bayesian Information Criterion (BIC) for model selection. We model the methylation profiles at a slightly lower spatial resolution, using four RBFs, as we are interested in capturing broader similarities between profiles, rather than fine details. Figure 4.6A shows the five distinct methylation profiles that were inferred from each cell line after applying the EM algorithm. To investigate the association of promoter methylation profiles and transcription, in figure 4.6B we show boxplots with the corresponding mRNA expression levels that are assigned to each cluster for each cell line. From the resulting methylation profile clusters, we seek to characterize the common features that are responsible for the corresponding mRNA abundance.

As expected, clusters corresponding to hyper-methylated regions (Cluster 4, green) are associated with repressed genes across all cell lines, confirming the known relationship of DNA methylation around TSS with gene repression. Also, two distinct patterns emerge: an S-shape profile (Cluster 5, yellow) with hypo-methylated CpGs upstream of TSS, which become gradually methylated at the gene body, and the reverse S-shape pattern (Cluster 3, orange). Genes associated with these profiles have intermediate expression levels for K562 and GM12878, and relatively high expression for H1-hESC. The most interesting pattern is the U-shape methylation profile (Cluster 2, blue), with a hypo-methylated region around the TSS surrounded by hyper-methylated domains. These profiles are associated with high transcriptional activity at their associated genes across all cell lines (t-test p-value $<$ 2.2e-16 for all paired cluster comparisons across cell lines). Surprisingly, uniformly low-methylated domains (Cluster 1, red) seem in general to be lowly expressed, except from the H1-hESC cell line, suggesting a different type of relationship between DNA methylation and expression in embryonic stem cells. The clustering analysis confirms, in a complementary way, that DNA

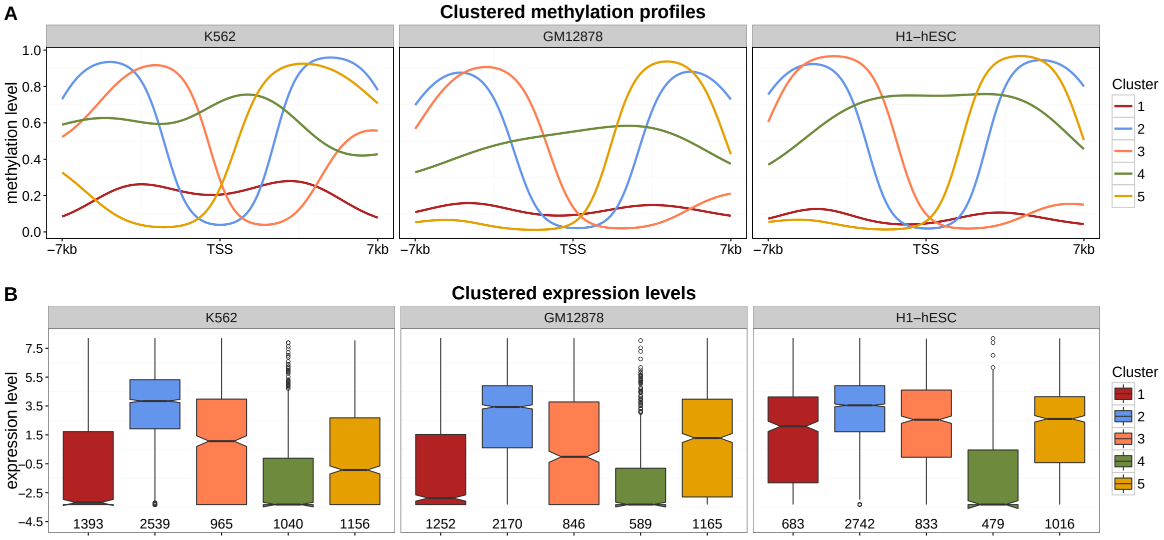


Figure 4.6 Clustering DNA methylation profiles across promoter-proximal regions. (**A**) Five clustered methylation profiles over ±7 kb promoter region w.r.t. TSS in the direction of transcription for the ENCODE cell lines. Each methylation profile is modelled using four RBFs. (**B**) Boxplots with the corresponding expression levels of the protein-coding genes assigned to each cluster for each of the three cell lines. The colours match with the clustered methylation profiles shown above. The numbers below each boxplot correspond to the total number of genes assigned to each cluster. T-test shows that the U-shape methylation profiles (Cluster 2, blue) correspond to significantly higher expression (p-value < 2.2e-16) compared to the expression of genes assigned to the remaining methylation profiles.

methylation profiles and transcript abundance are tightly connected to each other, and this relationship can be generalized across all cell lines considered in this study.

To provide a biological insight on the potential methylation mechanisms that regulate transcription, we consider the purity of the clustering across different cell lines, i.e. which fraction of genes assigned to a certain cluster in a certain cell line are assigned to the same cluster in the other cell lines (see figure 4.7). Surprisingly, around 68% of the genes assigned to the U-shape profile are present in all three cell lines, while the intersection of genes assigned to the other clusters ranges between 20% to 40%. Interestingly, the promoter-proximal regions clustered to the U-shape methylation profile are dominated by CGIs. Of all common promoters assigned to the U-shape profiles, 95.6% are CGI associated. Not surprisingly, hyper-methylated promoters are only 35.7% CGI associated, however, uniformly low-methylated promoters are 65.9% CGI associated. This suggests that promoters associated with totally unmethylated CGIs surrounded by hyper-methylated domains are transcriptionally active across cell lines. Indeed, we find that 35% of the U-shape profile genes are associated with a curated set of housekeeping genes (Eisenberg and Levanon, 2013). On the contrary, only a small fraction of genes assigned to hyper-methylated domains or uniformly low-methylated domains are housekeeping genes (1.4% and 17.7% respectively). Finally, around 22% of the genes assigned to the S-shape and reverse S-shape profiles are associated with housekeeping genes.

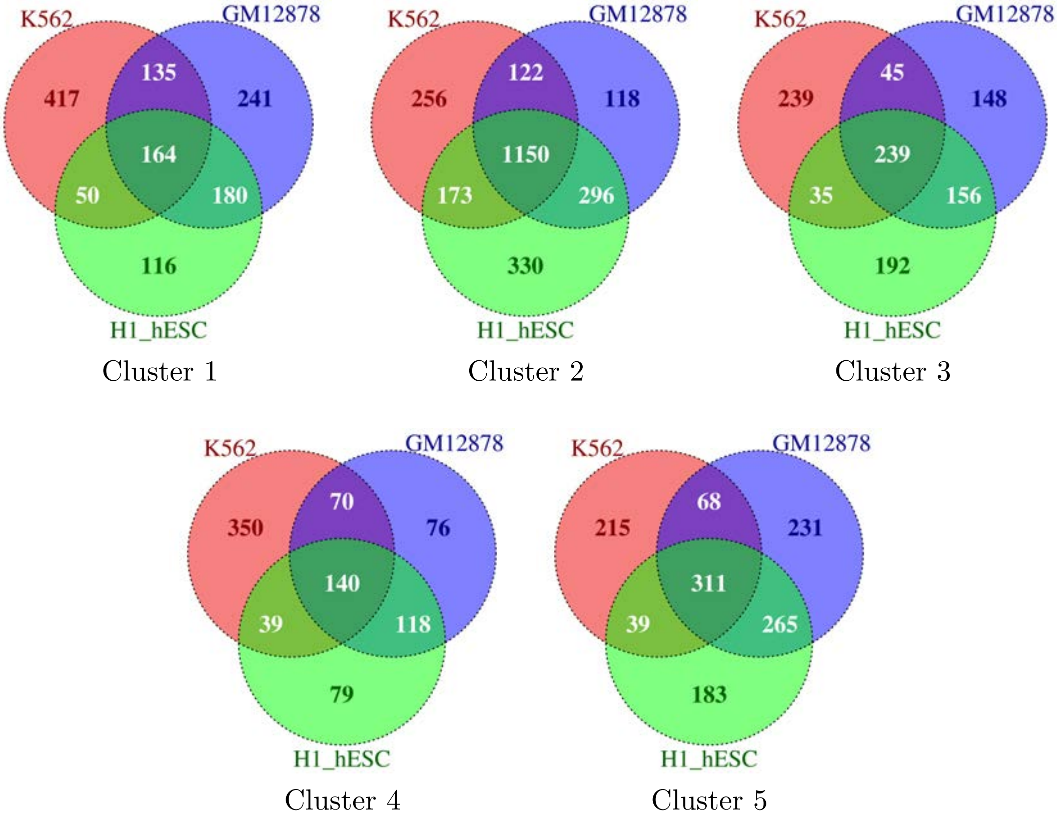


Figure 4.7 Venn diagrams showing the clustering purity across cell lines. Genes assigned to the U-shape profile (Cluster 2) have high intersection across the three cell lines, while the intersection of the genes assigned to the other clusters is considerably lower.

4.4 Discussion

Alterations of DNA methylation marks are associated with regulatory roles and are involved in many diseases, with the most notable example being cancer (Baylin and Jones, 2011). Therefore, unravelling the function of DNA methylation and its association with transcription abundance, is essential for understanding biological processes and developing biomarkers for disease diagnostics (Laird, 2003). Our results demonstrate that representing methylation patterns by their average level is insufficient to unravel the link between DNA methylation and gene expression, and one should consider the shape of the methylation profiles at the vicinity of the promoters. The contributions of this chapter are twofold. First, we introduced a generic modelling approach based on a GLM of basis function regression to quantitate spatially distributed methylation profiles via the BPRMeth model. These higher-order methylation features enabled us to build a powerful predictive model for gene expression in various cell lines, which more than doubled the predictive accuracy of current methods based on mean methylation levels.

Second, we demonstrated how the BPRMeth features can be used in downstream analyses by clustering spatially similar methylation profiles. We revealed five distinct groups of methylation patterns across promoter regions that are well correlated with gene expression and are well reproducible across different cell lines. Some of these patterns recapitulate existing biological knowledge. The U-shape methylation profile, consisting of hypo-methylated CGIs followed by hyper-methylated CGI shores (Irizarry et al., 2009), has been identified in different studies, and is termed as 'canyon' (Jeong et al., 2014) or 'ravine' (Edgar et al., 2014). Our findings are in line with Edgar et al. (2014), where ravines are in general positively correlated with mRNA abundance. Since the main difference of the U-shape methylation profile and the uniformly low-methylated profile is the CGI shore methylation, our results support the hypothesis that hyper-methylation on the edges of CGIs enhances transcriptional activity.

The existence of U-shape methylation profiles may help to explain observations that the methylation of gene body is sometimes positively correlated with transcript abundance (Lou et al., 2014; Varley et al., 2013). We hypothesize that these regions may correspond to U-shape methylation profiles, or a mixture of U-shape and S-shape methylation profiles. Another relevant study, showed that hyper-methylation of CGI shores on the mouse genome was associated with increased Dnmt3a activity, which resulted in positive correlation with transcriptional activity; indicating that methylation outside of CGIs may be used for maintaining active chromatin states for specific genes (Wu et al., 2010).

As an extension of this analysis, further work could include extending the BPRMeth model to relate differential methylation profiles with differential gene expression levels, and evaluate the importance of profile changes in regulation of gene expression across different cell types. More generally, it is increasingly clear that transcriptional activity is regulated by a complex and still incompletely understood interaction network of molecular players, including DNA methylation, histone modifications and transcription factor binding. Several recent computational studies have highlighted the dependencies between these players (Benveniste et al., 2014; Dong et al., 2012). The BPRMeth model provides an effective way of recapitulating DNA methylation patterns using higher-order features, and may therefore play an important role in building more effective integrative models of high-throughput data.

Chapter 5

BPRMeth extension and single cell multi-omics study

In chapter 4 we introduced BPRMeth, a platform to quantitatively understand the relationship between DNA methylation and transcript abundance from bulk BS-seq experiments. Although bulk BS-seq experiments have paved the way for mapping the methylome landscape, they fall short of explaining epigenetic heterogeneity and quantifying its dynamics, which inherently occur at the single cell level (Schwartzman and Tanay, 2015). Recent advancements in sequencing technology have enabled the development of protocols that allow profiling of DNA methylation at the single cell resolution, such as scBS-seq (Smallwood et al., 2014) and scRRBS (Guo et al., 2013). Now that single cell technologies are coming of age, rigorous analytical tools are required to uncover the role of epigenetic marks in major biological processes. BPRMeth is based on a generalised linear modelling approach, making it a flexible and versatile tool that can be easily enhanced with additional observation models to enable analysis of single cell and methylation array studies[1] (section 5.1).

Although the BPRMeth model was initially designed for quantifying DNA methylation profiles, it can be readily applied to analyse observations from different sequencing technologies, that make similar assumptions about spatial (or time) correlations of (epi)genomic data. Indeed, in section 5.4 we apply the enhanced BPRMeth model to analyse data generated from the scNMT-seq protocol (Clark et al., 2018), which enables parallel profiling of chromatin accessibility, DNA methylation, and transcription at the single cell level. In this study, BPRMeth is used to quantify cell-to-cell *chromatin accessibility heterogeneity* around promoter regions and subsequently link accessibility heterogeneity to gene expression levels[2].

[1] Material in sections 5.1 and 5.2 have appeared before in Kapourani and Sanguinetti (2018a) and the manuscript was written by myself with Guido Sanguinetti providing feedback and editing.

[2] This published work (Clark et al., 2018) was conducted in collaboration with Wolf Reik's, Oliver Stegle's and John Marioni's research groups. My contributions in this study were to perform statistical analysis on the data to obtain associations between the different molecular layers and link cell-to-cell chromatin heterogeneity with transcript abundance. These will form the main content of section 5.4.

5.1 BPRMeth model extension summary

We considerably extended the implementation of the BPRMeth model to provide a flexible environment for analysing and modelling spatial patterns of DNA methylation and similarly structured data from a variety of experimental platforms. The major features of the enhanced BPRMeth model are as follows:

1. Support for analysing single cell methylation data, by using a Bernoulli likelihood model.
2. Support for data measured by methylation array platforms which return a methylation level in (0,1); achieved by using a Beta likelihood model.
3. Support for Bayesian estimation via Gibbs sampling and mean field variational inference, enabling model selection and uncertainty quantification in all model quantities *a posteriori*.
4. Support for differential methylation analysis across conditions, such as cell populations or individuals, using Bayes factors.
5. Support for *Fourier basis functions*, as well as radial and polynomial basis functions, which may prove useful for analysing data with expected periodicity, e.g. for nucleosome positioning data generated from NOMe-seq (Kelly et al., 2012).

5.1.1 Software implementation

Given the continuing popularity of epigenetic assays and their rapid expansion in the clinical setting, the BPRMeth model is also provided as an R package available through Bioconductor[3]. A schematic workflow diagram of the BPRMeth package is shown in figure 5.1 (left). We should emphasise that although the focus until now has been on modelling promoter-proximal regions around the transcription start site (TSS), the BPRMeth package can be applied on arbitrary genomic features of interest, including enhancers, CTCF binding regions, Nanog regulatory regions or others. Hence, the BPRMeth model can become a widespread tool in the high throughput bioinformatics workbench.

The operational characteristics of the software are as follows: Methylation and annotation files are given as input to create genomic regions of pre-specified length. Next a basis object is required to transform the input methylation data, e.g. the `create_rbf_object` function will produce an RBF object. The `infer_profiles_`(`vb` or `mle`) functions are used to infer the latent methylation profiles (i.e. extract methylation features). Equivalently, the `cluster_profiles_`(`vb` or `mle`) functions are used to cluster genomic regions. The output of the algorithm can then be used for downstream analyses, such as predicting gene expression levels (e.g. using the `predict_expr` function) or quantifying levels of accessibility heterogeneity across single-cells (see section 5.4). To visualise the results, the objects produced from the model are given as input to `plot_infer_profiles` or `plot_cluster_profiles`. An example of the graphical output of the software is given in figure 5.1 (right).

[3] http://bioconductor.org/packages/BPRMeth, DOI: 10.5281/zenodo.2566628

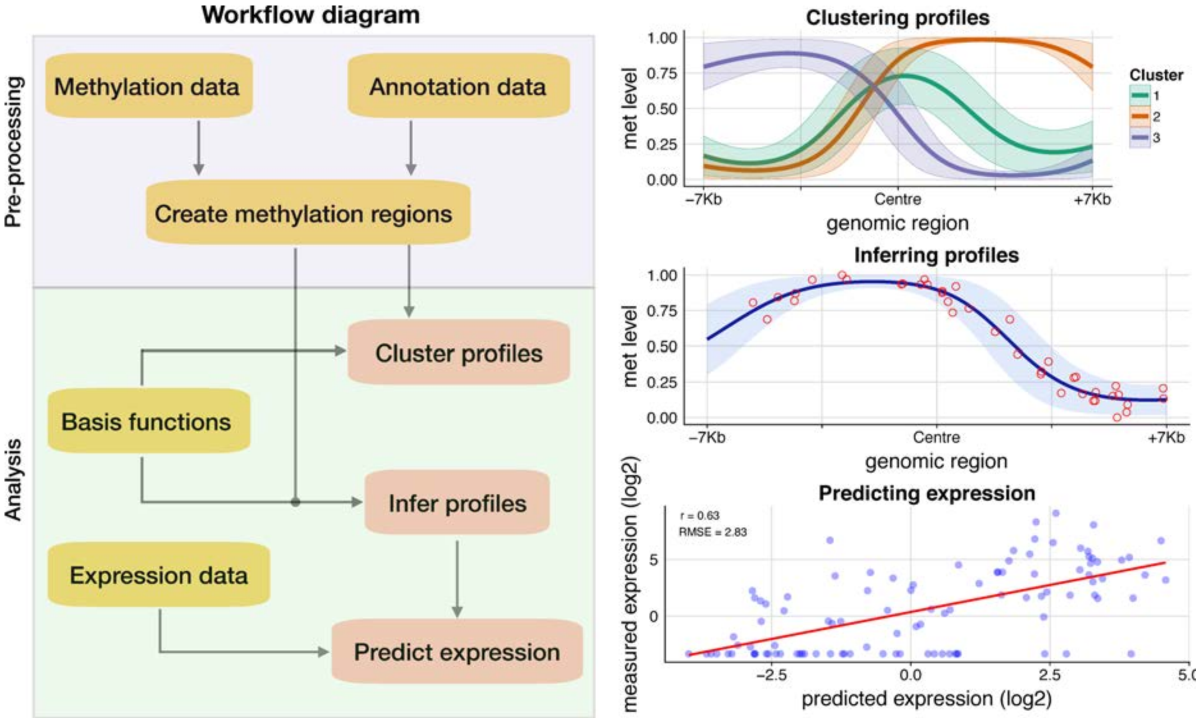


Figure 5.1 Schematic workflow of BPRMeth package (left) with example output graphs (right).

5.1.2 Modelling single cell methylation data

Single cell bisulfite sequencing protocols provide us with single base-pair resolution of CpG methylation states (Clark et al., 2016). Since we assay the DNA of a single cell, the methylation level for each CpG site is predominantly binary, either methylated or unmethylated. However, due to each chromosome having two copies, a small proportion of CpG sites have a non-binary nature (see figure 5.2). To avoid ambiguities, hemi-methylated sites — sites with 50% methylation level — are filtered prior to downstream analysis, and for the remaining sites binary methylation states are obtained from the ratio of methylated read counts to total read counts (Angermueller et al., 2016).

A natural way to model each binary CpG site is through a Bernoulli observation model. Assume that for a specific genomic region, we have observed I CpGs sites $\{(x_1, y_1), \ldots, (x_I, y_I)\}$, where x_i denote the CpG locations and $y_i \in \{0, 1\}$ the methylation state. Then, the log-likelihood of the Bernoulli probit regression model is given by

$$\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \log \left| \prod_{i=1}^{I} \mathcal{B}\text{ern}(y_i \mid \Phi(\mathbf{w}^\top \mathbf{h}(x_i))) \right| = \sum_{i=1}^{I} \log \left| \mathcal{B}\text{ern}(y_i \mid \Phi(\mathbf{w}^\top \mathbf{x}_i)) \right|. \quad (5.1)$$

Here $\mathbf{x}_i \overset{\text{def}}{=} \mathbf{h}(x_i)$ denotes the basis function transformed CpG locations x_i, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_I\}$, $\mathbf{w} \in \mathbb{R}^D$ represents the regression coefficients, and Φ is the inverse probit function ensuring that the underlying (latent) function takes values in the $[0, 1]$ interval. As in the case of binomial

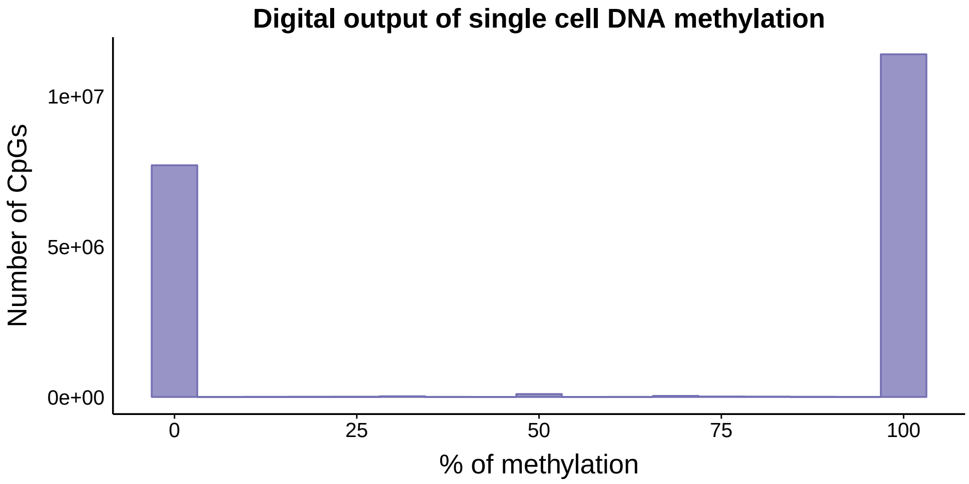


Figure 5.2 Digital output of single cell DNA methylation. Histogram of the distribution of CpG methylation values for 10 randomly sampled single cells from the Angermueller et al. (2016) study. As expected, the proportion of binary CpGs is very high (around 98.8%) and only around 0.5% of CpG sites are hemi-methylated.

probit regression, direct maximisation of this quantity is intractable due to the presence of the probit transformation, hence, we perform numerical optimisation which requires the gradient of (5.1) w.r.t. model parameters $\mathbf{w}$, which is given by

$$\nabla_{\mathbf{w}} \log p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w}) = \sum_{i=1}^{I} \left| \Big(y_i \Phi(\mathbf{w}^\top \mathbf{x}_i)^{-1} + (1 - y_i)\big(1 - \Phi(\mathbf{w}^\top \mathbf{x}_i)\big)^{-1} \Big) \phi(\mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i \right|, \qquad (5.2)$$

where $\phi(\cdot)$ is the probability density function for the standard normal distribution. Notice that we can easily incorporate a penalised version, similar to equation (4.6), and cluster single cell methylation profiles using EM, see algorithm 4.1, with the appropriate modifications.

5.1.3 Modelling methylation array data

To infer methylation profiles from DNA methylation array experiments —and generally any continuous observations that lie in (0, 1) interval — we use a beta observation model, which has been successfully applied for analysing DNA methylation arrays in several earlier publications (Siegmund, 2011). Briefly, the output of Illumina Infinium platforms (Moran et al., 2016) is a set of β-values which are computed as the ratio of intensities between methylated (m) and unmethylated (u) alleles, that is,

$$\beta = \frac{\max(m, 0)}{\max(m, 0) + \max(u, 0) + 100}.$$

The values of this statistic lie in the (0,1) interval, where a value of zero indicates a completely unmethylated CpG dinucleotide, and a value of one denotes a completely methylated

CpG site[4]. We use a different parametrisation of the beta regression model in terms of a mean parameter μ and a precision parameter γ (Ferrari and Cribari-Neto, 2004). The Beta distribution then takes the form

$$\mathcal{B}\text{eta}(y \mid \mu, \gamma) = \frac{\Gamma(\gamma)}{\Gamma(\mu\gamma)\Gamma((1-\mu)\gamma)} y^{\mu\gamma-1}(1-y)^{(1-\mu)\gamma-1},$$

where $\Gamma(\cdot)$ is the *Gamma* function, $\mu \in (0,1)$ and $\gamma > 0$. Then the log-likelihood of the Beta probit regression model for I observed CpGs becomes

$$\begin{aligned}\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \sum_{i=1}^{I} \Big[& \log\Gamma(\gamma) - \log\Gamma\left(\Phi(\mathbf{w}^\top \mathbf{x}_i)\gamma\right) - \log\Gamma\left((1-\Phi(\mathbf{w}^\top \mathbf{x}_i))\gamma\right) \\ & + (\Phi(\mathbf{w}^\top \mathbf{x}_i)\gamma - 1)\log y_i + \left((1-\Phi(\mathbf{w}^\top \mathbf{x}_i))\gamma - 1\right)\log|1-y_i| \Big].\end{aligned} \tag{5.3}$$

The gradient of (5.3) w.r.t. model parameters $\mathbf{w}$ is given by

$$\begin{aligned}\nabla_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \sum_{i=1}^{I} \gamma\phi(\mathbf{w}^\top \mathbf{x}_i) \Big[& \log y_i - \log|1-y_i| \\ & - \psi\left(\Phi(\mathbf{w}^\top \mathbf{x}_i)\gamma\right) + \psi\left((1-\Phi(\mathbf{w}^\top \mathbf{x}_i))\gamma\right)\mathbf{x}_i \Big],\end{aligned} \tag{5.4}$$

where $\psi(\cdot)$ is the *digamma* function. Note that in this formulation the precision parameter γ is assumed to be fixed and is the same across observations. We leave the extension of jointly optimising over the mean parameter μ and precision parameter γ as future work, since the main focus of this thesis is to model and analyse BS-seq data.

5.2 Bayesian probit regression model

Up until now we have focused on computing point estimates for the model parameters using maximum likelihood. However, a Bayesian approach might prove useful for quantifying the uncertainty in all model quantities and encoding our prior beliefs through the prior distribution. This is particularly important for single cell methylation data due to the inherent noisy measurements and the sparse CpG coverage, as a result of starting with small amounts of genomic DNA[5]. In addition, the Bayesian paradigm provides a natural way for performing

[4] A commonly used alternative statistic are M-values, which are logistically transformed β-values and often provide a better performance in detection rate of methylated and unmethylated CpG sites (Du et al., 2010). In addition, M-values are better suited for common statistical tests, since the methylation level of each CpG site is assumed to follow a Gaussian distribution. The BPRMeth model supports Gaussian distributed data as well, which results in performing basis function linear regression.

[5] In chapter 6 we show that scBS-seq data are extremely sparse and even the Bayesian approach is insufficient to infer informative methylation profiles. However, in this chapter the extended BPRMeth model will be applied on single cell chromatin accessibility data, which have substantially higher coverage than DNA methylation data.

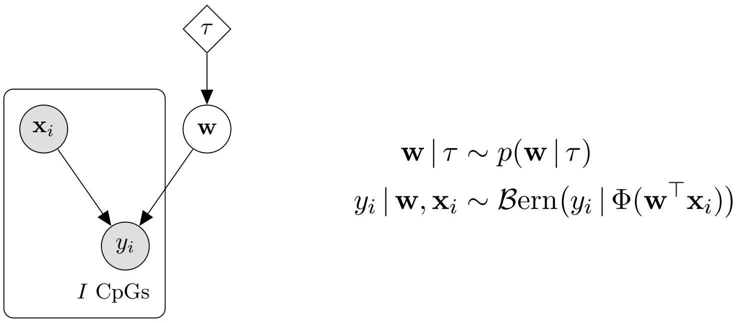

τ
$\mathbf{x}_i$
$\mathbf{w}$
y_i
I CpGs
$\mathbf{w} \,|\, \tau \sim p(\mathbf{w} \,|\, \tau)$
$y_i \,|\, \mathbf{w}, \mathbf{x}_i \sim \mathcal{B}\text{ern}(y_i \,|\, \Phi(\mathbf{w}^\top \mathbf{x}_i))$

Figure 5.3 Probabilistic graphical representation of the BPRMeth model.

model selection, and the computation of the marginal likelihood will enable us to perform differential analysis.

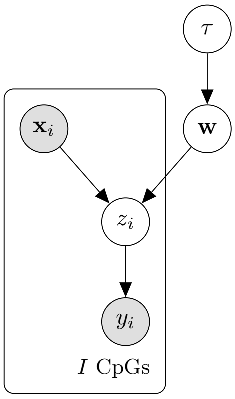
To complete the Bayesian formulation, we treat the parameters $\mathbf{w}$ as random variables and define a prior distribution $p(\mathbf{w} \,|\, \tau)$, where τ denotes the hyper-parameters of the prior. The probabilistic graphical representation of the Bayesian probit regression model is shown in figure 5.3. The posterior distribution over the model parameters then takes the form

$$p(\mathbf{w} \,|\, \mathbf{y}, \tau, \mathbf{X}) \propto p(\mathbf{w} \,|\, \tau) \prod_{i=1}^{I} \mathcal{B}\text{ern}(y_i \,|\, \Phi(\mathbf{w}^\top \mathbf{x}_i)). \tag{5.5}$$

Performing inference for this model in the Bayesian framework is complicated by the fact that no conjugate prior $p(\mathbf{w} \,|\, \tau)$ exists for the coefficients of the probit regression model. One approach to approximate the posterior distribution is to apply Metropolis-Hastings (see algorithm 3.2). The performance of Metropolis-Hastings depends heavily on the proposal distribution. A common choice is the multivariate Gaussian distribution $q(\mathbf{w}^* \,|\, \mathbf{w}) = \mathcal{N}(\mathbf{w}^* \,|\, \mathbf{w}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a fixed covariance matrix that needs to be tuned to achieve good mixing of the Markov chain and adequately explore the posterior distribution. However, identifying an optimal choice for $\mathbf{\Sigma}$ is often difficult and problem specific. One approach would be to use more advanced MCMC methods, such as adaptive Metropolis-Hastings or Hamiltonian Monte Carlo (Liang et al., 2010). Here we will use the data augmentation approach of Tanner and Wong (1987) to obtain efficient algorithms for approximating the intractable posterior distribution.

5.2.1 Data augmented BPRMeth

The probit regression model can be made amenable to Bayesian estimation thanks to a data augmentation strategy originally proposed by Albert and Chib (1993), which can be thought of as a specific application of data augmentation (Tanner and Wong, 1987). This strategy consists of introducing additional auxiliary latent variables z_i that follow a Gaussian distribution conditioned on the input $\mathbf{w}^\top \mathbf{x}_i$. The augmented model has the hierarchical structure shown in figure 5.4, where y_i is now deterministic conditional on the sign of the latent variable z_i.



$$
\begin{aligned}
\tau &\sim \mathcal{G}\text{amma}(\tau \,|\, \alpha_0, \beta_0) \\
\mathbf{w} \,|\, \tau &\sim \mathcal{N}(\mathbf{w} \,|\, \mathbf{0}, \tau^{-1}\mathbf{I}) \\
z_i \,|\, \mathbf{w}, \mathbf{x}_i &\sim \mathcal{N}(z_i \,|\, \mathbf{w}^\top \mathbf{x}_i, 1) \\
y_i \,|\, z_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}
\end{aligned}
$$

Figure 5.4 Probabilistic graphical representation of the data augmented BPRMeth model.

Hence, our original problem becomes a missing data problem where we have a Bayesian linear regression model on the latent variables z_i and the observations y_i are incomplete since we only observe whether $z_i > 0$ or $z_i \leq 0$. We introduce a conjugate Gaussian prior over the parameters $\mathbf{w} \sim \mathcal{N}(\mathbf{w} \,|\, \mathbf{0}, \tau^{-1}\mathbf{I})$, where the hyper-parameter τ — controlling the precision of the Gaussian prior — is considered as a random variable itself and is assumed to follow a Gamma distribution. Here the assumption of a zero prior mean effectively penalises weights moving away from the prior, and has a similar effect to the regularisation strategy used for the maximum likelihood formulation. This strategy reduces the necessary conditional distributions to a tractable form as either Gaussian, Gamma or one-dimensional truncated Gaussian distributions. From the graphical model, the joint distribution over all variables factorises as follows

$$
\begin{aligned}
p(\mathbf{y}, \mathbf{z}, \mathbf{w}, \tau \,|\, \mathbf{X}) &= p(\mathbf{y} \,|\, \mathbf{z}) \, p(\mathbf{z} \,|\, \mathbf{w}, \mathbf{X}) \, p(\mathbf{w} \,|\, \tau) \, p(\tau) \\
&= \left[\prod_{i=1}^{I} p(y_i \,|\, z_i) \, p(z_i \,|\, \mathbf{w}, \mathbf{x}_i) \right] p(\mathbf{w} \,|\, \tau) \, p(\tau),
\end{aligned}
\tag{5.6}
$$

with

$$
p(y_i \,|\, z_i) = \mathbb{I}(z_i > 0)^{y_i} + \mathbb{I}(z_i \leq 0)^{1 - y_i},
$$

where $\mathbb{I}(\cdot)$ is the indicator function, equal to one if the quantities inside the function are satisfied, and zero otherwise.

Binomial observation model

We can take a Bayesian approach for modelling bulk BS-seq data, by recasting the binomial likelihood as a Bernoulli observation model with additional observations. Assume that each CpG site $y = (s, \nu)$ follows a binomial distribution, i.e. $s \sim \mathcal{B}\text{inom}(s \,|\, \nu, \rho)$, where ν is the total number of trials, s denotes the number of successes and ρ is the probability of success. We can think of s as the total number of successes of ν independent Bernoulli experiments with outcomes $y_1^*, \ldots, y_\nu^*$, where now each y_t^* follows a Bernoulli distribution with $t \in \{1, \ldots, \nu\}$. It is simple to reconstruct the binary outcomes y_t^* from the binomial observations using

$$y_t^* = \begin{cases} 1 & \text{if } 1 \leq t \leq s \\ 0 & \text{if } s < t \leq \nu \end{cases}. \tag{5.7}$$

Using this approach of extending our observations to binary outcomes, we can apply the data augmentation strategy described above to perform inference for binomial observations[6].

5.2.2 Gibbs sampling for augmented BPRMeth

For the derivation of the Gibbs sampler we will assume that the precision hyper-parameter τ is held fixed. Then the joint posterior distribution of regression coefficients $\mathbf{w}$ and latent variables $\mathbf{z}$ becomes

$$p(\mathbf{z}, \mathbf{w} \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{w} \,|\, \mathbf{0}, \tau^{-1}\mathbf{I}) \prod_{i=1}^{I} p(y_i \,|\, z_i) \mathcal{N}(z_i \,|\, \mathbf{w}^\top \mathbf{x}_i, 1). \tag{5.8}$$

It is still difficult to normalise and sample directly from the joint posterior distribution, however, we can split the parameters into two blocks and run Gibbs sampling to create a Markov chain with samples drawn from the full conditionals $p(\mathbf{w} \,|\, \mathbf{z}, \mathbf{y}, \mathbf{X})$ and $p(\mathbf{z}|\mathbf{w}, \mathbf{y}, \mathbf{X})$, which are of standard forms. Focusing on the update for $\mathbf{w}$, since $\mathbf{w} \perp\!\!\!\perp \mathbf{y} \,|\, \mathbf{z}$, equation (5.8) simplifies to the posterior distribution for Bayesian linear regression. Using standard linear algebra results — e.g. see chapter 2 in Bishop (2006) — we obtain the following update for the regression coefficients

$$\begin{aligned} \mathbf{w} \,|\, \mathbf{z}, \mathbf{X} &\sim \mathcal{N}\left(\mathbf{w} \,|\, \mathbf{m}, \mathbf{S}\right), \\ \mathbf{m} &= \mathbf{S}\mathbf{X}^\top \mathbf{z}, \\ \mathbf{S} &= \left(\tau \mathbf{I} + \mathbf{X}^\top \mathbf{X}\right)^{-1}. \end{aligned} \tag{5.9}$$

Given the current sample of $\mathbf{w}$, we could easily draw each latent variable z_i from its full conditional, i.e. $z_i \sim \mathcal{N}(z_i \,|\, \mathbf{w}^\top \mathbf{x}_i, 1)$. However, since we also condition on y_i, we need to take into consideration this additional source of information. Using the fact that when observing $y_i = 1$ we have $z_i > 0$ and when observing $y_i = 0$ we have $z_i \leq 0$, the full conditional of the latent variables is given by a truncated Gaussian distribution

$$z_i \,|\, \mathbf{w}, y_i, \mathbf{x}_i \sim \begin{cases} \mathcal{TN}_+(\mathbf{w}^\top \mathbf{x}_i, 1) & \text{if } y_i = 1 \\ \mathcal{TN}_-(\mathbf{w}^\top \mathbf{x}_i, 1) & \text{if } y_i = 0 \end{cases}. \tag{5.10}$$

Here $\mathcal{TN}_+(\mathcal{TN}_-)$ denotes the Gaussian distribution truncated on the left (right) tail to zero to contain only positive (negative) values.

[6] Notice that this data augmentation approach cannot be applied for the Beta observation model. Due to this difficulty and since microarray data are not the main focus of this thesis, a Bayesian treatment of the Beta probit regression model is left as an interesting topic for future work.

Holmes and Held joint update scheme

A potential issue with the Albert and Chib (1993) data augmentation implementation is that there might be strong correlation between the model variables $\mathbf{w}$ and $\mathbf{z}$. Hence, the iterative updates for each variable might result in slow mixing in the Markov chain, requiring a larger number of simulations to effectively explore and summarise the posterior distribution. To reduce autocorrelation between the random variables, Holmes and Held (2006) proposed to jointly update $\mathbf{w}$ and $\mathbf{z}$ by using the factorisation

$$p(\mathbf{w}, \mathbf{z} \,|\, \mathbf{y}, \mathbf{X}) = p(\mathbf{w} \,|\, \mathbf{z}, \mathbf{X})\, p(\mathbf{z} \,|\, \mathbf{y}, \mathbf{X}). \tag{5.11}$$

Note that $p(\mathbf{w} \,|\, \mathbf{z}, \mathbf{X})$ remains exactly the same as presented in (5.9), however, the latent variable $\mathbf{z}$ is updated by marginalising out the regression coefficients $\mathbf{w}$. This marginalisation induces correlation between the latent variables z_i and using standard linear algebra we obtain

$$p(\mathbf{z} \,|\, \mathbf{y}, \mathbf{X}) \propto \mathcal{N}\left(\mathbf{z} \,|\, \mathbf{0}, \mathbf{I} + \mathbf{X}(\tau^{-1}\mathbf{I})\mathbf{X}^{\top}\right) \mathbb{I}(\mathbf{z}, \mathbf{y}), \tag{5.12}$$

where $\mathbb{I}(\mathbf{z}, \mathbf{y})$ is the indicator function that truncates the multivariate Gaussian distribution to the appropriate region.

Sampling directly from a truncated multivariate Gaussian distribution is difficult, however, we can use the Gibbs sampler to update the full conditional distributions in turn,

$$z_i \,|\, \mathbf{z}_{\neg i}, y_i, \mathbf{x}_i \sim \begin{cases} \mathcal{TN}_{+}(\mu_i, \upsilon_i) & \text{if } y_i = 1 \\ \mathcal{TN}_{-}(\mu_i, \upsilon_i) & \text{if } y_i = 0 \end{cases}, \tag{5.13}$$

where the means μ_i and variances υ_i are obtained from the marginal predictive densities, and can be efficiently calculated as follows (Henderson and Searle, 1981)

$$\begin{aligned} \mu_i &= \mathbf{m}^{\top}\mathbf{x}_i - w_i(z_i - \mathbf{m}^{\top}\mathbf{x}_i), \\ \upsilon_i &= 1 + w_i, \\ w_i &= h_i/(1 - h_i) \end{aligned} \tag{5.14}$$

where h_i denotes the i^{th} diagonal element of matrix $\mathbf{XSX}^{\top}$. After updating each z_i, the posterior mean $\mathbf{m}$ should be recalculated

$$\mathbf{m} = \mathbf{m}^{\text{old}} + \mathbf{SX}^{\top}(z_i - z_i^{\text{old}}), \tag{5.15}$$

where $\mathbf{m}^{\text{old}}$ and z_i^{old} denote values prior to updating z_i; more details can be found in Holmes and Held (2006). This approach results in better mixing in the chain, however, it comes with increased computational burden for each Gibbs iteration due to the correlation between the latent variables introduced by marginalising the coefficients $\mathbf{w}$.

5.2.3 Mean field variational inference for augmented BPRMeth

Using the data augmentation approach, the Bayesian probit regression for Bernoulli observations $\mathbf{y}$ is seen as having an underlying linear regression on latent variables $\mathbf{z}$ which we can easily handle. However, this does not come without paying a price, since for BS-seq experiments we are effectively introducing one latent variable per mapped CpG location. Performing genome-wide analysis using Gibbs sampling may be prohibitive, due to the cost of sampling millions of latent variables from truncated Gaussian distributions in each of the thousands iterations of the Gibbs sampler.

A more efficient approach is to perform deterministic approximation to the posterior distribution using mean field variational inference. Here we perform inference on the hyper-parameter τ as well. As explained in subsection 3.4.1, the mean field assumes that the variational distribution factorises over the latent variables

$$q(\mathbf{z}, \mathbf{w}, \tau) = q(\mathbf{z})\, q(\mathbf{w})\, q(\tau) \simeq p(\mathbf{z}, \mathbf{w}, \tau \,|\, \mathbf{y}, \mathbf{X}). \tag{5.16}$$

Applying equation (3.36) to our model, we obtain the following solutions for the optimised factors of the variational posterior[7]

$$\begin{aligned} q(\mathbf{z}) &= \prod_{i=1}^{I} \begin{cases} \mathcal{TN}_{+}\Big(z_i \,|\, \left\langle \mathbf{w}^\top \mathbf{x}_i \right\rangle_{q(\mathbf{w})}, 1\Big) & \text{if } y_i = 1 \\ \mathcal{TN}_{-}\Big(z_i \,|\, \left\langle \mathbf{w}^\top \mathbf{x}_i \right\rangle_{q(\mathbf{w})}, 1\Big) & \text{if } y_i = 0 \end{cases}, \\ q(\tau) &= \mathcal{G}\text{amma}\left(\tau \,|\, \alpha_0 + \frac{D}{2}, \beta_0 + \frac{1}{2}\left\langle \mathbf{w}^\top \mathbf{w} \right\rangle_{q(\mathbf{w})}\right), \\ q(\mathbf{w}) &= \mathcal{N}(\mathbf{w} \,|\, \mathbf{m}, \mathbf{S}), \end{aligned} \tag{5.17}$$

where

$$\begin{aligned} \mathbf{m} &= \mathbf{S}\mathbf{X}^\top \left\langle \mathbf{z} \right\rangle_{q(\mathbf{z})}, \\ \mathbf{S} &= \Big(\left\langle \tau \right\rangle_{q(\tau)} \mathbf{I} + \mathbf{X}^\top \mathbf{X} \Big)^{-1}. \end{aligned}$$

Notice the striking similarity between (5.17) and the Gibbs sampler updates, demonstrating the similarity of these two algorithms. Algorithm 5.1 provides a pseudo-code for inferring single-cell methylation profiles using variational inference.

To assess the performance of the proposed algorithms in terms of efficiency, we generate a synthetic genomic region with $I = 80$ CpG sites using $D = 3$ radial basis functions. The binary CpG methylation states are generated from cluster 1 (green profile) shown at the top right of figure 5.1. We run both Gibbs sampling algorithms for $S = 10{,}000$ iterations with a burn-in period of 5,000 samples, whereas the variational inference algorithm is run until convergence. As expected, MFVI (mean field variational inference) is substantially faster than

[7] Detailed mathematical derivation of mean field variational inference for the Bayesian probit regression model can be found in appendix A.1.

Algorithm 5.1 CAVI for BPRMeth model

```
initialize mean $\mathbf{m}$, variance $\mathbf{S}$ and Gamma parameters $\alpha_0, \beta_0$
Set $\alpha \leftarrow \alpha_0 + \frac{D}{2}$
Set $\beta \leftarrow \beta_0$
while ELBO has not converged do
    Set $\boldsymbol{\mu} \leftarrow \mathbf{Xm}$                                ▷ Mean of truncated Gaussian
    Set $\langle z_i \rangle_{q(z_i)} \leftarrow \begin{cases} \mu_i + \phi(-\mu_i)/(1-\Phi(-\mu_i)) & \text{if } y_i = 1 \\ \mu_i - \phi(-\mu_i)/\Phi(-\mu_i) & \text{if } y_i = 0 \end{cases}$
    Set $\mathbf{S} \leftarrow \left(\frac{\alpha}{\beta}\mathbf{I} + \mathbf{X}^\top \mathbf{X}\right)^{-1}$              ▷ Regression coefficient covariance
    Set $\mathbf{m} \leftarrow \mathbf{S}\mathbf{X}^\top \langle \mathbf{z} \rangle_{q(\mathbf{z})}$                 ▷ Regression coefficient mean
    Set $\beta \leftarrow \beta_0 + \frac{1}{2}\left(\mathbf{m}^\top \mathbf{m} + \mathrm{tr}(\mathbf{S})\right)$    ▷ Gamma distribution parameter
    Compute $\mathcal{L}\left(q(\mathbf{z}, \mathbf{w}, \tau)\right)$                 ▷ Using equation (A.11)
end while
```

Gibbs (see table 5.1), whereas the A&C (Albert & Chib) implementation is around five times faster than the H&H (Holmes & Held) extension. The efficiency of MFVI is mainly due to its quick convergence after a few tens of iterations, whereas one needs to run the Gibbs sampling simulation for the total number of iterations. The effective sample size — see equation (3.47) — provides an estimate of the equivalent number of independent samples obtained from the Markov chain. The H&H extension provides almost 2 times more ESS compared to the A&C implementation, nevertheless, it achieves this with high computational burden. That is, one could run the A&C chain for a larger number of iterations to achieve the same ESS with a faster CPU time. Figure A.1 shows marginal density plots together with the corresponding trace plots for each regression coefficient for the two Gibbs sampling algorithms, and figure A.2 shows corresponding marginal density plots when applying the MFVI algorithm.

Algorithm	Relative CPU time	Effective sample size
A&C	1595	2596
H&H	8367	4313
MFVI	1	—

Table 5.1 Efficiency of inference algorithms for data augmented BPRMeth. CPU times are computed relative to the MFVI implementation. The effective sample size is computed on the remaining 5000 samples after the burn-in period. There is no notion of effective sample size for MFVI, hence the corresponding column is empty. A&C: Albert & Chib, H&H: Holmes & Held, MFVI: Mean field variational inference.

5.3 Differential analysis using BPRMeth

In addition to quantifying uncertainty on model parameters, the Bayesian treatment of BPRMeth provides an appealing approach for performing differential analysis. Differential analysis refers to assessing whether epigenetic patterns are significantly different across two conditions of interest, e.g. across genomic regions or within a genomic region across cell sub-populations. To perform differential analysis we follow a similar approach to Stegle et al. (2010), in formulating the problem as a comparison of two models and using Bayes factors for performing model selection (see subsection 3.2.4). The first model (null hypothesis $\mathcal{H}_S$) assumes that the epigenetic profiles in both conditions are samples drawn from an identical shared distribution. The alternative model (alternative hypothesis $\mathcal{H}_I$) describes the epigenetic profiles in both conditions as samples drawn from two independent distributions.

Formally, under $\mathcal{H}_S$ we assume that a single latent function f generates observations from both conditions A and B. Whereas, under $\mathcal{H}_I$, each condition generates data from its own latent function f^A and f^B. Let $\mathcal{D}^k = \{\mathbf{x}^k, \mathbf{y}^k\}$ collectively denote the observed data for condition $k \in \{A, B\}$. Then, the Bayes factor between two competing models is given by

$$\begin{aligned} BF_{IS} &= \frac{\int p(\mathcal{D}^A, f^A \mid \mathcal{H}_I) df^A \int p(\mathcal{D}^B, f^B \mid \mathcal{H}_I) df^B}{\int p(\mathcal{D}^A, \mathcal{D}^B, f \mid \mathcal{H}_S) df} \\ &= \frac{p(\mathcal{D}^A \mid \mathcal{H}_I)\, p(\mathcal{D}^B \mid \mathcal{H}_I)}{p(\mathcal{D}^A, \mathcal{D}^B \mid \mathcal{H}_S)}. \end{aligned} \tag{5.18}$$

Since we generally work in the log domain, the log Bayes factor becomes

$$\log |BF_{IS}| = \log p(\mathcal{D}^A \mid \mathcal{H}_I) + \log p(\mathcal{D}^B \mid \mathcal{H}_I) - \log p(\mathcal{D}^A, \mathcal{D}^B \mid \mathcal{H}_S). \tag{5.19}$$

Here $p(\mathcal{D}^k \mid \mathcal{H}_I)$ denotes the marginal likelihood of data from condition k under model $\mathcal{H}_I$ and $p(\mathcal{D}^A, \mathcal{D}^B \mid \mathcal{H}_S)$ denotes the marginal likelihood of the 'concatenated' observations from both conditions under model $\mathcal{H}_S$. Calculating marginal likelihoods for the BPRMeth model is intractable, however, we can apply mean field variational inference and compute the evidence lower bound $\mathcal{L}\left(q(\mathbf{z}, \mathbf{w}, \tau)\right)$. This approach of using the evidence lower bound as a proxy for performing model selection was successfully applied before in Beal and Ghahramani (2003), although results will strongly depend on how close the evidence lower bound is to the marginal likelihood.

To assess the performance of the differential method, we generated N = 300 synthetic cells from three cell sub-populations. We assume a single genomic region, and each cell sub-population is generated from the profiles corresponding to the top right of figure 5.1. Next, we randomly select 150 pairs of cells to perform differential analysis, where we expect that pairs of cells belonging to the same cluster (e.g. green profile) will result in decisive evidence

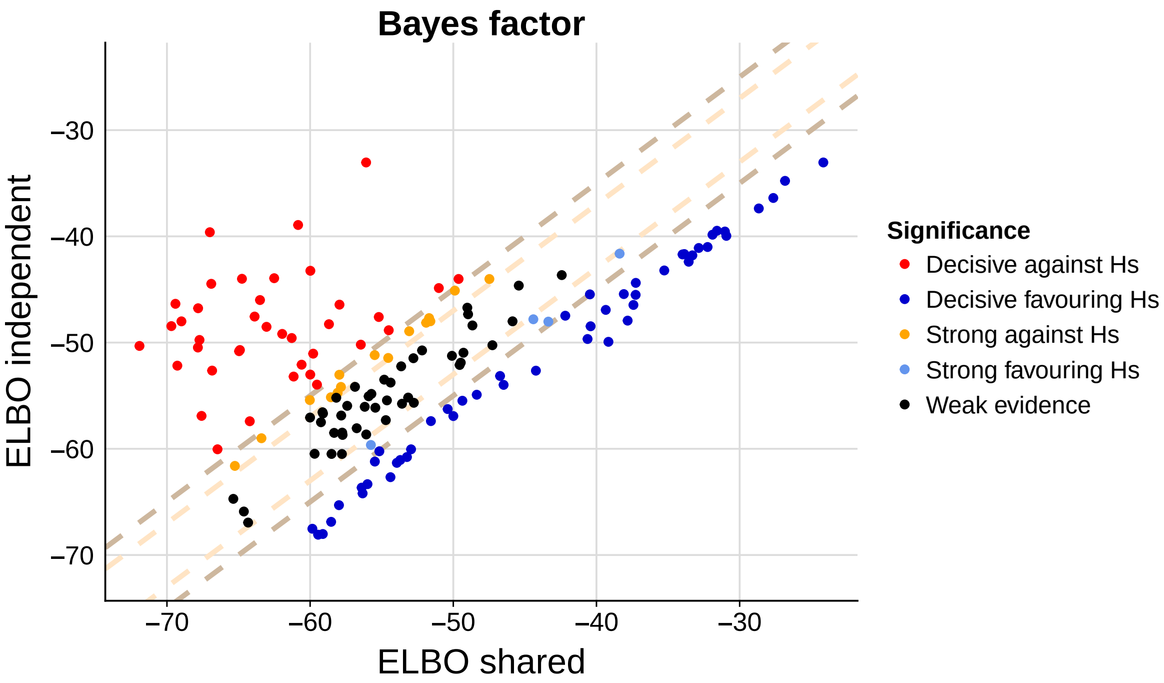


Figure 5.5 Differentially methylated cells using Bayes factors. The x-axis corresponds to the ELBO under model $\mathcal{H}_S$ and the y-axis to the product of the ELBOs under $\mathcal{H}_I$ for both conditions A and B.

in favour of $\mathcal{H}_S$. Figure 5.5 depicts the differential cells with evidence against or in favour of $\mathcal{H}_S$. Points towards the upper left corner correspond to differentially methylated cells, whereas points towards the lower right corner correspond to cells with similar methylation patterns. The scale of evidence for interpreting the significance of Bayes factors is shown in table 3.2. Figure 5.6 illustrates four comparisons across pairs of cells together with the corresponding Bayes factors. As expected, when comparing cells belonging to clusters 2 and 3, i.e. S-shape and inverse S-shape profiles (from figure 5.1), we obtain high Bayes factors which correspond to decisive evidence against $\mathcal{H}_S$ (see figure 5.6a). Comparing pairs of cells from the remaining cluster combinations results in strong (see figure 5.6b) or weak (see figure 5.6c) evidence against $\mathcal{H}_S$. Comparison of cells from the same sub-population (see figure 5.6d) results in negative Bayes factors, which are interpreted as evidence in favour of $\mathcal{H}_S$.

5.4 Application to single cell multi-omics study

As outlined in section 2.4, single cell multi-omics platforms have recently emerged as a powerful approach to simultaneously investigate the dynamic coupling between different biological layers at the single cell resolution. This section concerns the introduction of a novel multi-omics protocol termed scNMT-seq (single cell nucleosome, methylome and transcriptome sequencing) (Clark et al., 2018), that enables the joint analysis of these three molecular layers

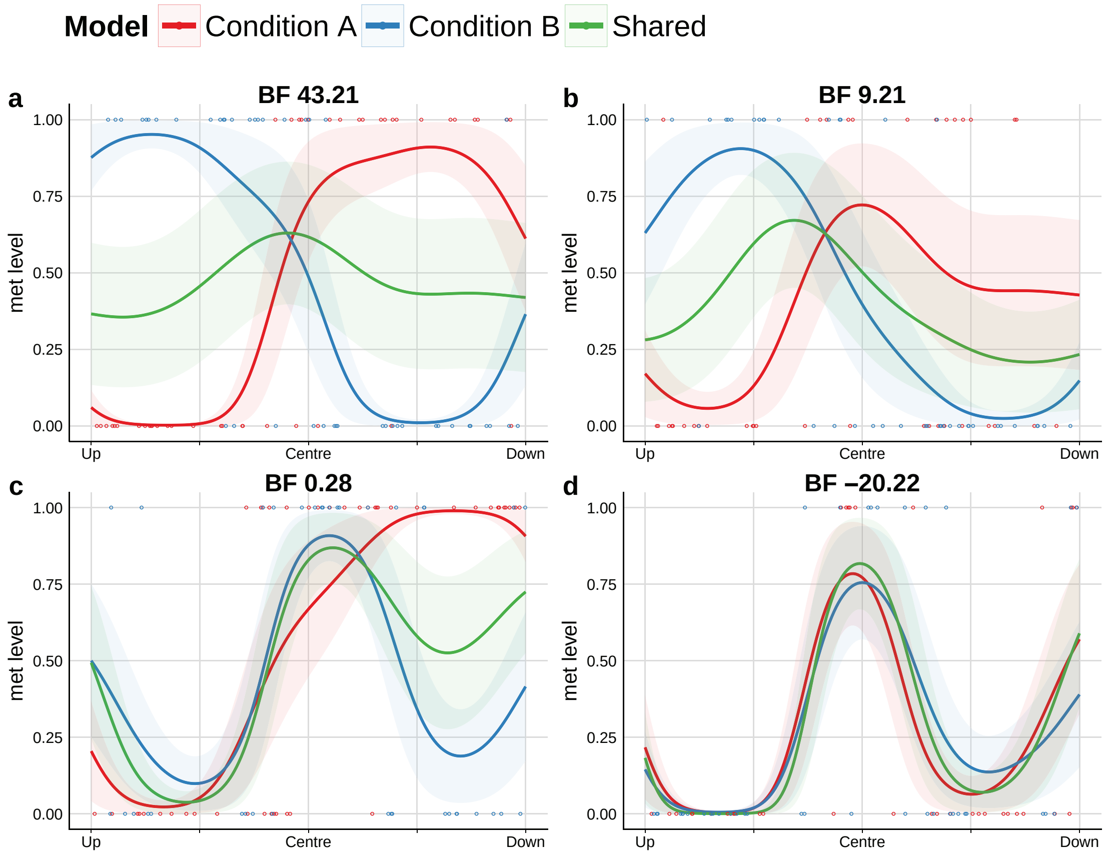


Figure 5.6 Methylation profiles for pairs of cells across conditions. The independent profiles, corresponding to the methylation patterns of each cell are denoted with red and blue coloured profiles. The shared profile, depicted with green line, is obtained by concatenating the observations from both conditions.

allowing us to obtain a more complete understanding of epigenetic dependencies and their associations with transcription[8].

Figure 5.7 illustrates the scNMT-seq protocol overview. Cells are first isolated into methyltransferase reaction mixtures using FACS sorting. Single cells are then lysed and accessible (or nucleosome depleted) DNA is labelled using GpC methyltransferase — M.CviPI enzyme. Subsequently, DNA and RNA are physically separated from the fully lysed cell in a similar fashion to G&T-seq (Macaulay et al., 2015) and scM&T-seq (Angermueller et al., 2016). The transcriptome is then profiled using a conventional Smart-seq2 protocol (Picelli et al., 2014), while chromatin accessibility and DNA methylation are measured using the scNOMe-seq protocol (Pott, 2017). The M.CviPI enzyme is used to methylate *exposed* cytosine residues in the GpC context (i.e. guanine followed by cytosine), whereas nucleosome protected DNA remains

[8] The scNMT-seq protocol was developed by Wolf Reik's, Oliver Stegle's and John Marioni's research groups with whom we established collaboration. Here we recap the Clark et al. (2018) paper to make the chapter self-contained, however, the data analysis focuses mostly on quantifying accessibility heterogeneity using the BPRMeth model and its applications for downstream analysis. The statistical analyses described in this section were performed by myself unless stated otherwise.

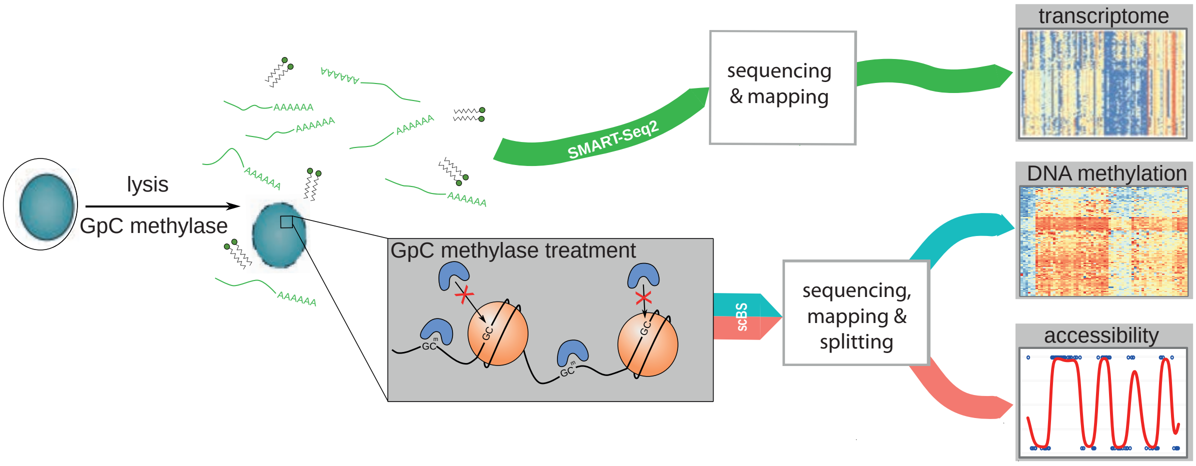


Figure 5.7 scNMT-seq protocol overview. Following cell lysis, accessible DNA is labelled using GpC methylase. RNA is separated and sequenced using scRNA-seq, whereas DNA undergoes scBS-seq; accessibility and methylation data are then separated using Bismark. Credits Stephen J. Clark.

unmodified. This allows us to distinguish between the two epigenetic states — since endogenous DNA methylation predominantly occurs in CpG dinucleotides — and after alignment of BS-seq reads, methylation and chromatin accessibility data can be separated using bioinformatics tools, such as Bismark (Krueger and Andrews, 2011). The only modification from conventional BS-seq analysis is an additional filtering step, where G-C-G and C-C-G positions are discarded due to inability to distinguish endogenous methylation from in vitro methylation and off-target effects of the enzyme, respectively. This filtering requirement reduces the number of genome-wide cytosines that can be assayed by approximately 50%. This results in even sparser epigenetic signal, nevertheless, a large proportion of regulatory genomic contexts, including promoters and enhancers, can in principle be assessed by scNMT-seq.

5.4.1 Data preprocessing

The original publication demonstrates the efficacy of scNMT-seq in robustly profiling these three biological layers on a batch of 70 serum grown EL16 mouse embryonic stem cells (ESCs). Here we focus on a second dataset used in the study, which has a larger degree of coordinated epigenetic and transcriptional heterogeneity than observed in ESCs. To do so, 43 serum grown ESCs were removed from LIF[9] for 3 days to initiate differentiation into embryoid bodies (EBs). Briefly, after library preparation and single cell sequencing of the embryoid body cells, BS-seq reads are aligned using Bismark (Krueger and Andrews, 2011), and RNA-seq libraries are aligned using HiSat2 (Kim et al., 2015). Subsequently, quality control is performed to discard single cells with poor quality. In total, 33 cells passed the quality control on both scRNA-seq and scBS-seq data using scNMT-seq.

[9] The leukaemia inhibitory factor (LIF) is a stem cell growth factor used for the *in vitro* culture of pluripotent mouse ESCs and affects cell growth by inhibiting differentiation (Nagy et al., 1993).

For the RNA-seq component only protein-coding genes are considered for further analysis. Following Lun et al. (2016) digital gene expression data are log-transformed and size-factor adjusted using a deconvolution approach that accounts for variation in cell size. Binary methylation states for individual CpG or GpC sites are obtained from the ratio of methylated read counts to total read counts (Angermueller et al., 2016). Subsequently, mean methylation or accessibility *rates* are computed for each genomic region by taking the average signal of CpG or GpC sites inside the region.

For the correlation analysis across cells (subsection 5.4.2), genes with low variability and expression are discarded according to the rationale of independent filtering (Bourgon et al., 2010). The top 50% of the most variable regions are considered for further analysis with the additional requirement of having coverage in at least 20 cells. A minimum coverage of 3 CpG / GpC sites are required per genomic feature. To correlate expression levels with different genomic contexts, such as enhancers or p300, each feature is associated to the closest gene within a 10 kb window. For promoter annotations, a small window of ±50 bp around TSS is used for accessibility data, whereas a larger window of ±2 kb is considered for methylation. Following Angermueller et al. (2016) correlation analysis is performed using the weighted Pearson correlation coefficient, which accounts for differences in coverage between cells. To test for non-zero correlation a two-tailed t-test is performed, and to control for false-discovery rate due to multiple testing, p-values are adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

5.4.2 Identifying genomic features with coordinated variability

The main goal of this analysis is to link all three molecular layers and reveal associations between the epigenome and transcriptome by leveraging the variation between multiple single cells. There are two approaches to perform association tests for single cell multi-omics. The *horizontal* analysis approach tests associations between molecular layers within individual cells (across all genes), which is similar to association studies using bulk data (see subsection 4.3.1). The *vertical* analysis approach — the main focus of this chapter — is to test feature-specific associations between molecular layers across all cells; enabling us to examine the potential of identifying genomic regions with coordinated variability across pairs of molecular layers. Figure 5.8 shows the correlation analysis across cells, which results in one association test per genomic feature. As expected, the majority of associations between methylation and transcription are negative, recapitulating the known relationship between these two layers. In contrast, the associations between chromatin accessibility and expression are less widespread, with a small number of mostly positive significant correlations. This might indicate that transcriptional changes are more dependent on DNA methylation rather than chromatin accessibility changes. Finally, a large number of significant associations is observed between methylation and accessibility, and as expected the majority of them tend to be negative, since open chromatin is related to hypo-methylated regions. Figure B.1 provides associations for additional genomic features.

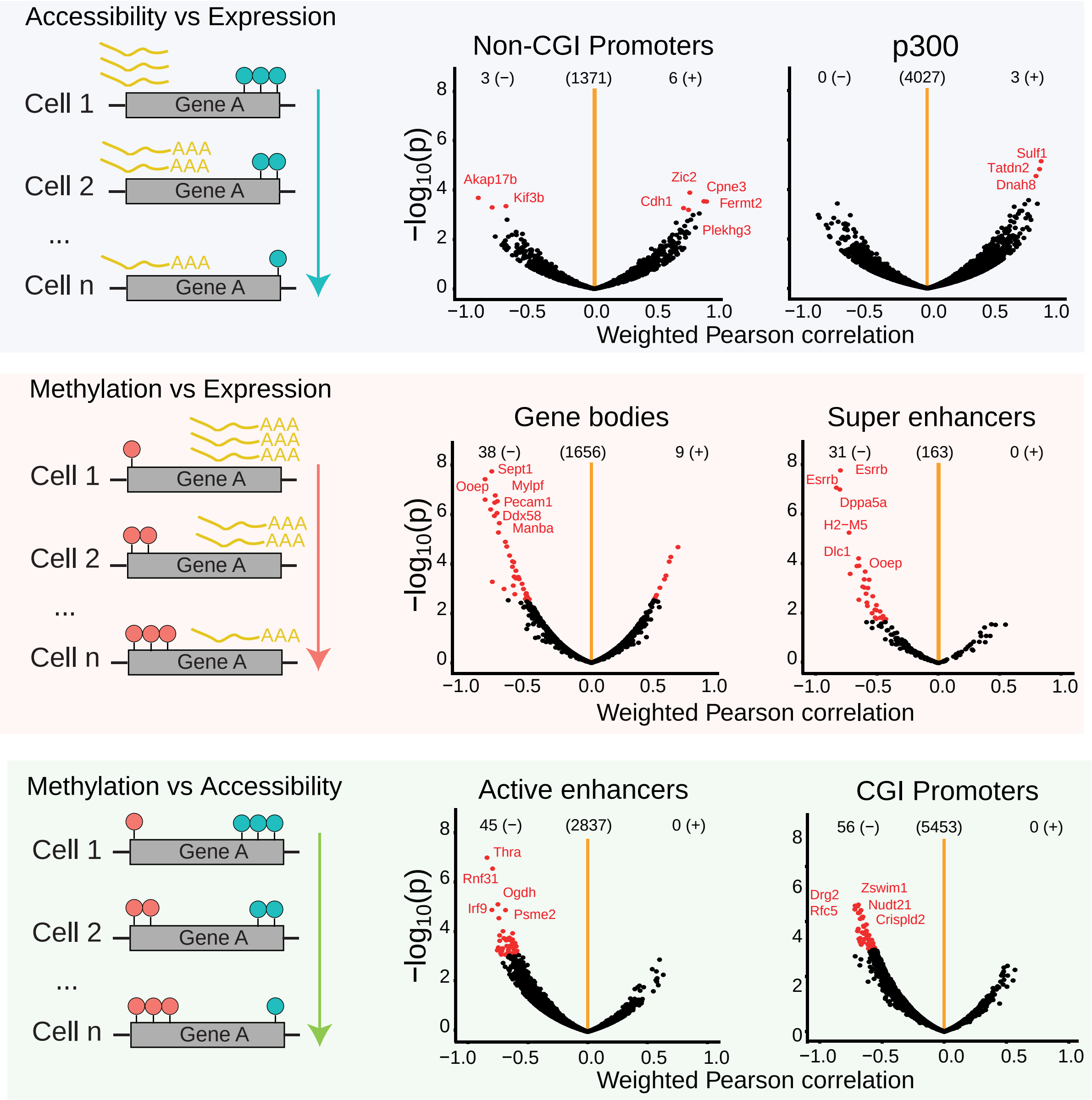


Figure 5.8 Correlation analysis across cells enables the discovery of novel associations at individual features. Weighted Pearson correlation (x-axis) and log10 p-value (y-axis) from association tests between pairs of molecular layers at individual features, stratified by genomic contexts. Significant associations (adjusted p-value < 0.1) are highlighted in red. The number of significant positive (+) and negative (-) associations and the number of tests (centre) are indicated at the top of each volcano plot. Joint analysis with Stephen J. Clark and Ricard Argelaguet.

5.4.3 Quantifying chromatin accessibility profiles

The association analysis across cells is based on computing average CpG or GpC methylation rates across genomic regions. However, inspection of chromatin accessibility data at the single GpC resolution reveals complex patterns due to the presence of nucleosomes, which

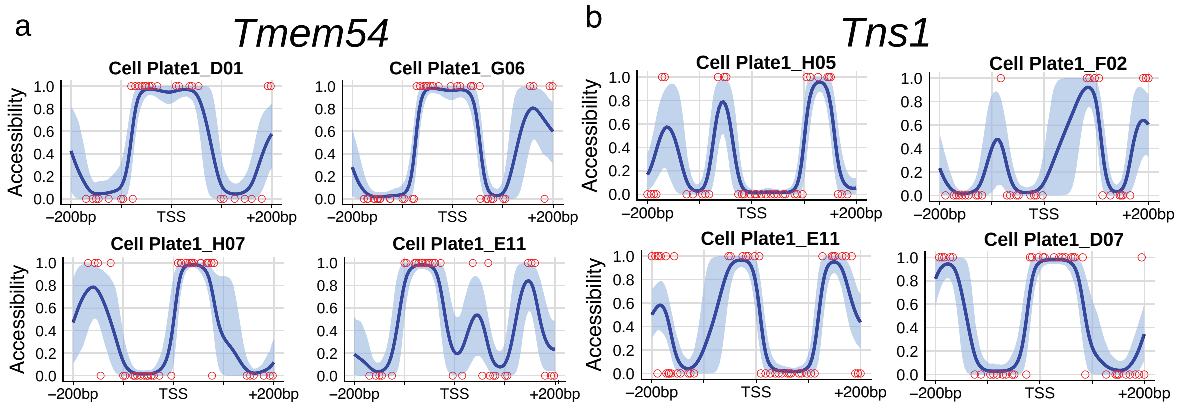


Figure 5.9 Single cell accessibility profiles around transcription start sites from four arbitrary cells in two example genes, (a) *Tmem54* and (b) *Tns1*. Each red dot represents a GpC site, with binary accessibility value (1 = accessible, 0 = inaccessible). Blue line represents the posterior mean of the inferred latent function, and the shading represents the corresponding 80% credible interval. Inference was performed using Gibbs sampling. We observe periodic patterns in the GpC accessibility data, which likely indicate positions of nucleosomes.

are not appropriately captured by averages calculated in pre-defined windows. Figure 5.9 shows the cell-to-cell variability of chromatin accessibility data and characteristic patterns of nucleosome positioning. In a similar fashion to chapter 4, we will apply the BPRMeth model — using a Bernoulli likelihood — to quantify the oscillatory patterns of DNA accessibility profiles[10]. The encoding of accessibility patterns as a simple average might explain the poor association between expression and accessibility. To explore whether chromatin accessibility profiles capture biologically meaningful information, we perform horizontal analysis (i.e. across genes) by predicting transcript abundance from accessibility profiles. To adequately reconstruct single cell profiles, we consider ±200 bp windows around TSS with a minimum coverage of 10 GpC sites. Indeed, figure 5.10 demonstrates that accessibility profiles are more predictive of gene expression than conventional accessibility rates.

Next, we exploit the inferred profiles to quantify the level of heterogeneity of chromatin accessibility around TSS. More specifically, we consider each gene independently and we are interested in measuring the variability of chromatin accessibility across cells, and subsequently linking accessibility heterogeneity to gene expression levels. To quantify heterogeneity we use the following generative reasoning. Genes with similar accessibility profiles across cells would be generated from the same latent cell cluster, which indicates high conservation or stability. If a gene requires additional latent clusters to adequately explain the accessibility profiles across all cells, it indicates higher heterogeneity. Hence, we formulate the quantification of accessibility heterogeneity as a model selection problem, where we need to identify the most likely number

[10] Notice that due to the extremely sparse CpG coverage (since in addition to single cell bisulfite sequencing we filter almost 50% of cytosines) inferring informative DNA methylation profiles from scNMT-seq is more challenging. We defer the analysis of single cell methylation data until chapter 6.

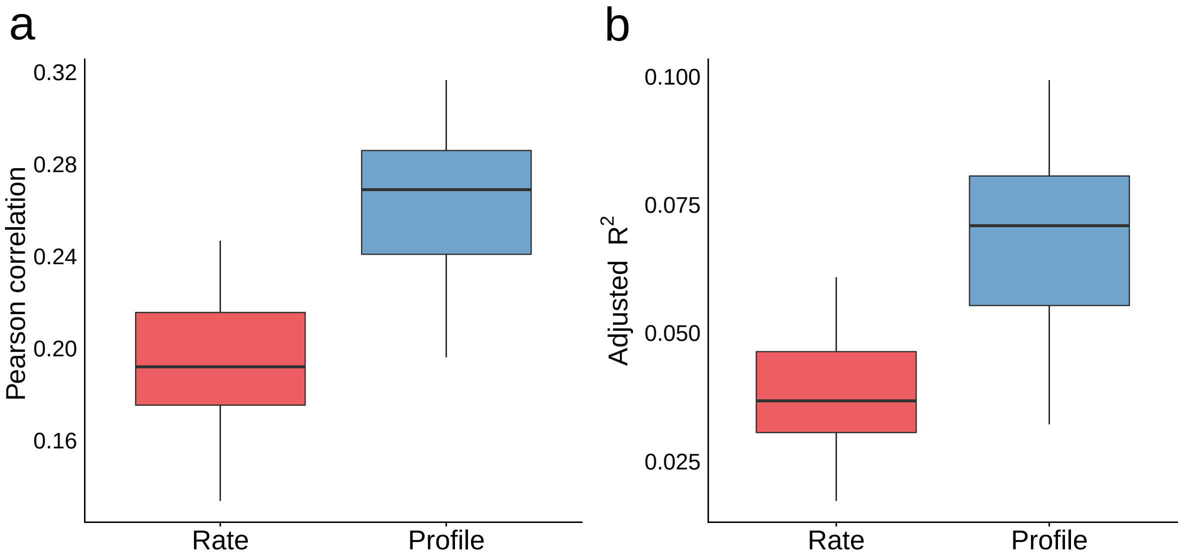


Figure 5.10 Accessibility profiles predict gene expression more accurately than accessibility rates. (a) Pearson's correlation between predicted and observed log-transformed gene expression levels using accessibility rates (red) and accessibility profiles (blue). (b) Adjusted R^2 to correct for the increased amount of parameters when using the BPRMeth model.

of clusters that generated the underlying profiles. For doing so, we cluster profiles using a mixture model and fit model parameters using EM, as shown in algorithm 4.1. We estimate the most likely number of clusters using the Bayesian information criterion (BIC) which is then used as a measure of cell-to-cell variation in the accessibility profile; the rationale being that homogeneous profiles will be grouped in a single cluster, while genes with heterogeneous profiles will be assigned a higher number of clusters. Figure 5.11a illustrates this process of quantifying chromatin accessibility heterogeneity for two example gene promoters: *Plekhg2* and *Tmem54*. To have an adequate number of cells during the clustering process, we consider genes that are covered in at least 40% of the cells with a minimum coverage of 10 GpCs per cell.

Subsequently, to identify the relationship between variability of accessibility profiles and transcript abundance, we stratify genes by the number of clusters estimated by the EM algorithm (see figure 5.11b). This revealed that genes with conserved accessibility profiles (fewer clusters) are associated with higher average expression levels. Examples of genes associated with two differentially expressed clusters are depicted in figure 5.12, whereas figure B.2 shows examples of genes with a single cluster corresponding to highly homogeneous accessibility profiles. In addition, we perform gene ontology analysis which revealed that highly homogeneous genes are enriched for gene ontology terms linked to house-keeping functions, such as regulation of gene expression, rRNA processing, splicing and translation (see figure 5.11d). Figure B.3 shows a more extensive list of enriched gene ontology terms for genes associated with highly conserved accessibility profiles. We should highlight that when considering accessibility rates, we lose

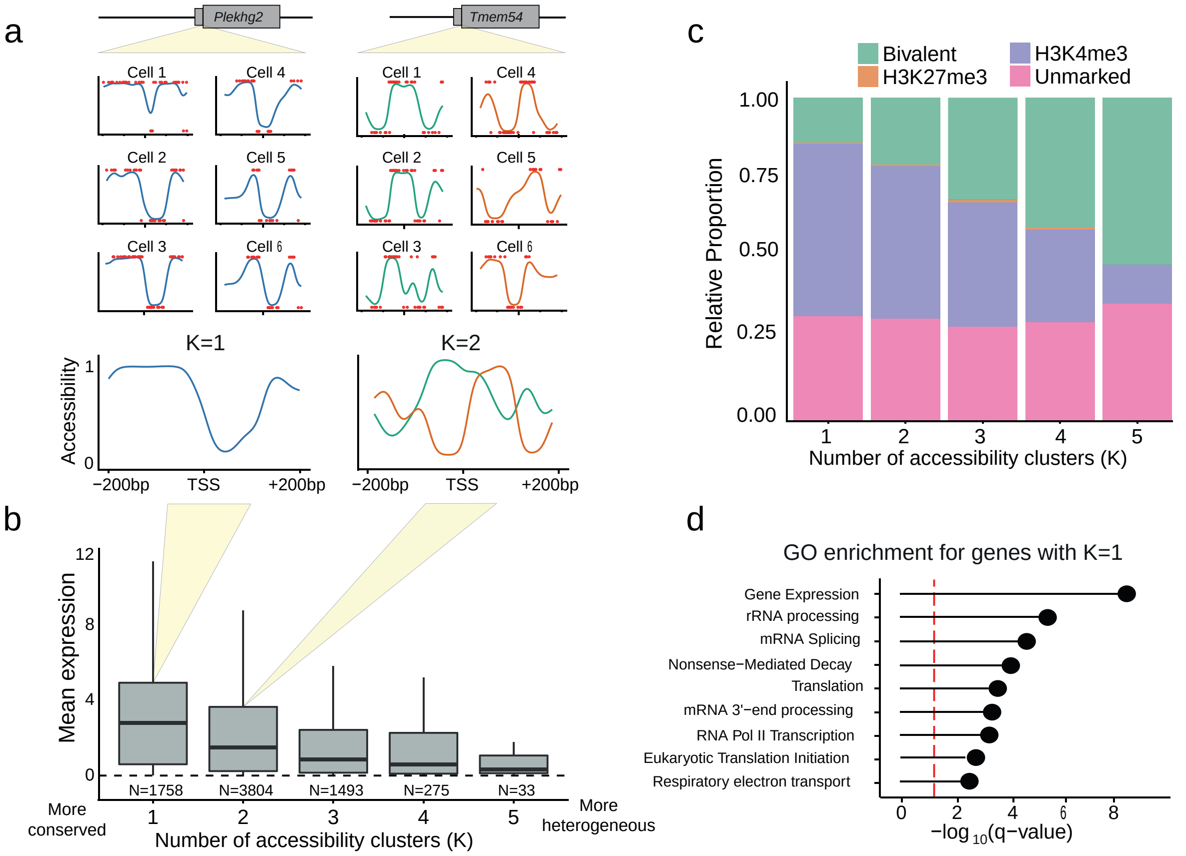


Figure 5.11 Modelling chromatin accessibility profiles at gene promoters in single cells. (a) Accessibility profiles for each cell and gene are fitted using BPRMeth, followed by clustering of profiles for each gene to estimate the most likely number of clusters. (b) Relationship between heterogeneity in the accessibility profile and gene expression. Boxplots show the distribution of average gene expression levels for genes with increasing numbers of accessibility clusters. Numbers below each boxplot correspond to the total number of genes assigned to each cluster. (c) Proportion of gene promoters marked with active H3K4me3 and/or repressive H3K27me3 histone marks stratified by number of accessibility clusters. Promoters with high levels of accessibility heterogeneity are associated with the presence of bivalent histone marks (both H3K4me3 and H3K27me3). (d) Gene ontology terms significantly enriched, using Fisher's exact test, in genes with most homogeneous accessibility profiles (K = 1). The p-values are adjusted for multiple testing using the Benjamini-Hochberg procedure.

both the association between accessibility variability and expression, and the enrichment for specific functions (see figure B.4); suggesting again that profiles capture biologically meaningful information. In contrast, genes with heterogeneous accessibility profiles (multiple clusters) are associated with low expression levels and are enriched for bivalent promoters containing both active H3K4me3 and repressive H3K27me3 ChIP-seq histone marks data, as figure 5.11c demonstrates. To account for differences in expression levels, i.e. remove the (potential) confounding effect of mean expression, we further stratify genes by expression groups. Figure B.5 shows that the increased bivalency of histone marks is independent of mean expression levels.

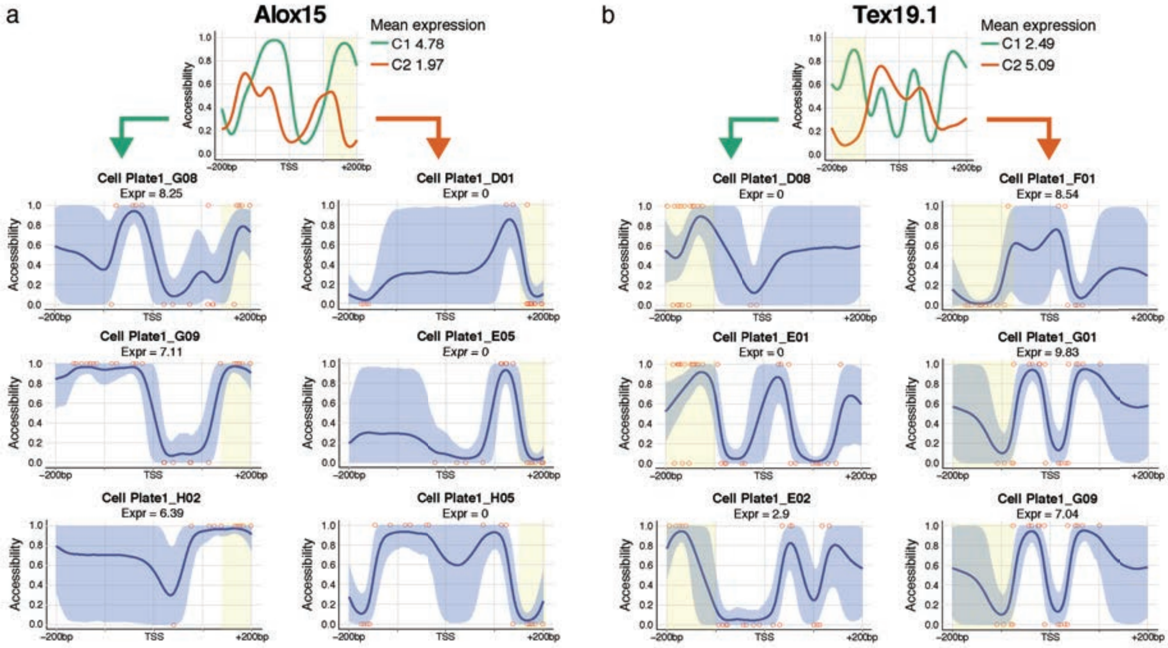


Figure 5.12 Accessibility profiles at gene promoters with two differentially expressed clusters. Shown are accessibility profiles for two representative genes with K=2 clusters that display cluster-driven changes in gene expression: (a) *Alox15* and (b) *Tex19.1*. The average profiles per gene and cluster (green and orange lines) are represented at the top, together with the corresponding mean expression levels. Representative examples of the single-cell profiles are shown at the bottom. Shading is used to highlight changes between clusters.

5.4.4 Epigenome dynamics along a developmental trajectory

An appealing application of scNMT-seq is that we can explore the epigenome dynamics during differentiation, using the RNA-seq component to infer the developmental trajectory from pluripotent to differentiated cell states (see figure 5.13). The transcriptome layer is used to order single cells along a putative developmental trajectory (pseudotime) with the destiny model (Haghverdi et al., 2016), using the top 500 genes with the most biological overdispersion as estimated by the scran package (Lun et al., 2016). Figure 5.13a shows the pseudotime ordering of single cells together with the expression level of the *Esrrb* gene — a marker gene primarily expressed in pluripotent cells (Festuccia et al., 2012). Subsequently, for each gene we tested whether the cluster assignments are associated with the cellular position in the differentiation trajectory using Spearman's rank coefficient, which identified a set of 15 genes that show a coherent dynamic pattern (see figure B.6). Figure 5.13b shows the dynamics of two representative genes, *Efhd1*, a gene that displays transition from a state with an open TSS to a state with a closed TSS; and *Rock2*, with a similar transition on the first nucleosome after TSS. Figure B.7 depicts additional examples of genes that show coordinated dynamic changes between accessibility profiles and pseudotime trajectory.

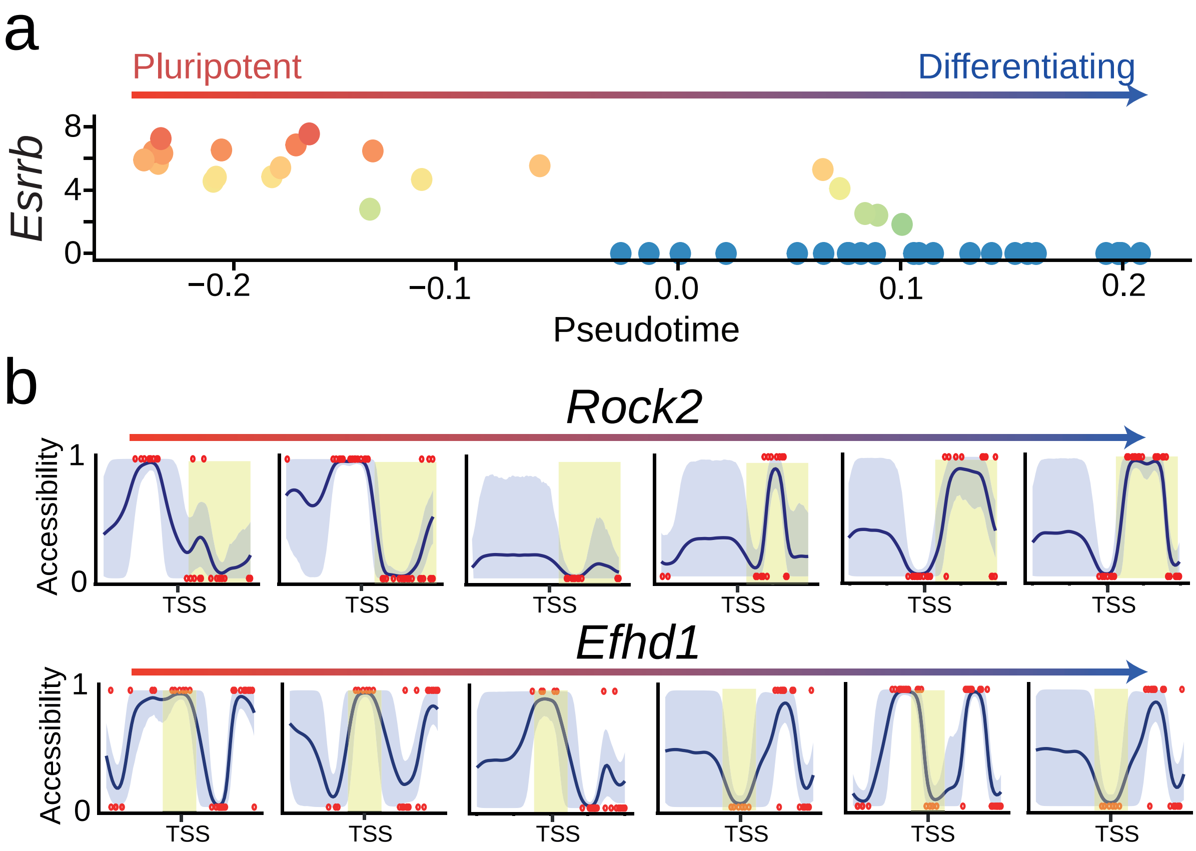


Figure 5.13 Exploring dynamics of epigenome during differentiation using scNMT-seq. (a) Embryoid body cells are ordered in a developmental trajectory inferred from the scRNA-seq data. The x-axis corresponds to the location of each cell in pseudotime (x axis) and y-axis denotes the expression level of the *Esrrb* gene. (b) Reconstructed dynamics of variation in chromatin accessibility profiles across pseudotime. Shown are profiles of representative cells for *Rock2* and *Efhd1*, where shading is used to highlight changes between cells. Joint analysis with Ricard Argelaguet.

5.5 Discussion

Given the continuing popularity of epigenetic assays and their rapid expansion in the clinical setting, rigorous analytical tools are required to interpret these high-dimensional data and investigate their role in major biological processes. BPRMeth is based on a generalised linear modelling approach making it a versatile tool that can be easily applied to different sequencing technologies. In this chapter the BPRMeth model, and its algorithmic implementation, were substantially extended both to accommodate different data types (including single cell sequencing and methylation array platforms), and to improve the scalability of the algorithm. Also, using a Bayesian formulation BPRMeth enables us to both quantify uncertainty in all model quantities, and perform differential analysis of epigenetic profiles — using Bayes factors — by formulating the problem as a comparison of two models. The Bayesian treatment is also important for partially overcoming the sparsity of single cell data through the introduction of structured informative priors. Indeed, in chapter 6 the BPRMeth model is used as a building block for Melissa which jointly imputes methylation states and clusters single cells thanks to hierarchical Bayesian formulation.

In addition, we described scNMT-seq a novel experimental method for parallel profiling of single cell DNA methylation, gene expression and chromatin accessibility. This single cell multi-omics technique — using bespoke statistical models — will expand our ability to investigate associations between the epigenome and transcriptome in heterogeneous cell types. Clearly the chromatin accessibility data show high spatial variability, likely due to nucleosome positioning patterns, hence standard approaches of computing simple averages across pre-defined windows are prone to fail. Utilising the extended BPRMeth model we were able to quantify cell-to-cell chromatin accessibility heterogeneity by reformulating the question as a model selection problem. This revealed that genes with conserved accessibility profiles are both associated with higher average expression levels, and enriched for gene ontology terms linked to house-keeping functions. The analysis in this chapter focused mostly in modelling each molecular layer independently. An extension of this analysis would be to develop integrative probabilistic methods that jointly model these different layers and propagate uncertainty for downstream analysis in a principled manner.

Chapter 6

Melissa: Bayesian clustering and imputation of single cell methylomes

Measurements of DNA methylation at the single cell level are promising to revolutionise our understanding of epigenetic control of gene expression (Kelsey et al., 2017). However, due to the small amounts of genomic DNA per cell, these protocols usually result in extremely sparse genome-wide CpG coverage (i.e. for most CpGs we have missing values), ranging from 5% in high throughput studies (Luo et al., 2017; Mulqueen et al., 2018) to 20% in low throughput ones (Angermueller et al., 2016; Smallwood et al., 2014). The sparsity of the data represents a major hurdle to effectively use single cell methylation assays to inform our understanding of epigenetic control of transcriptomic variability, or to distinguish individual cells based on their epigenomic state.

In this chapter[1] we address these problems by introducing Melissa (MEthyLation Inference for Single cell Analysis), which is a logical continuation of the previous material described in the thesis. Chapter 4 proposed BPRMeth to capture the local variability of (bulk) methylation patterns and identify genes that have similar methylation profiles. In addition to extending the model to single cell epigenomic assays, chapter 5 considered the quantification of cell-to-cell epigenetic heterogeneity independently per genomic region. Here we combine these strategies in a Bayesian hierarchical model which exploits the experimental design of assaying a large number of cells and the local variability of all genomic regions to both discover epigenetic differences and similarities among cell sub-populations, and transfer information across similar cells. In this way, Melissa can effectively use both the information of neighbouring CpGs and of other cells with similar methylation patterns in order to predict the methylation state of unassayed CpG sites. As an additional benefit, Melissa also provides a Bayesian clustering approach capable of identifying subsets of cells based solely on epigenetic state, to our knowledge the first clustering method tailored specifically to this rapidly expanding technology. We benchmark Melissa on

[1] Most of the material in this chapter have appeared before in Kapourani and Sanguinetti (2018b) and the manuscript was written by myself with Guido Sanguinetti providing feedback and editing.

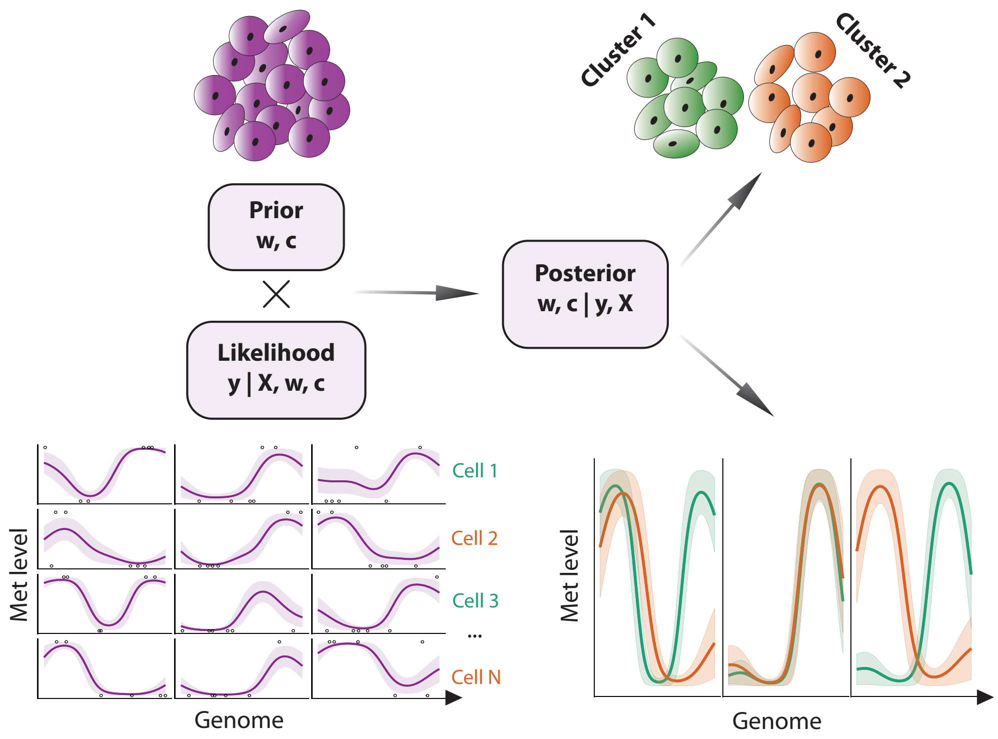


Figure 6.1 Melissa model overview. Melissa combines a likelihood computed from single cell methylation profiles fitted to each genomic region using a supervised regression approach (bottom left) and an unsupervised Bayesian clustering prior (top left). The posterior distribution provides a methylome-based clustering (top right) and imputation (bottom right) of single cells.

both simulated and real single cell BS-seq data, demonstrating that Melissa provides both state of the art imputation performance, and accurate and biologically meaningful clustering of cells.

6.1 Methods

Melissa[2] addresses the data sparsity issue by leveraging local correlations between neighbouring CpGs and similarity between individual cells (see figure 6.1). The starting point is the definition of a set of genomic regions (e.g. genes or enhancers). Within each region, Melissa postulates a latent profile of methylation, a function mapping each CpG within the region to a number in $[0, 1]$ which defines the probability of that CpG being methylated. To ensure spatial smoothness of the profile, Melissa uses as building block the BPRMeth model (with modified likelihood to account for single cell data). Local correlations are however often insufficient for regions with extremely sparse coverage, and these are quite common in scBS-seq data. Therefore, we share information across different cells by coupling the local GLM regressions through a shared prior distribution. In order to respect the (generally unknown) population structure that may be present within the cells assayed, we choose a (finite) Dirichlet mixture model prior.

[2] http://bioconductor.org/packages/Melissa, DOI: 10.5281/zenodo.2567427

The output of Melissa is therefore twofold: at each genomic region in each cell, we get a predicted profile of methylation, which can be used to impute missing data (i.e. unassayed CpGs). For each cell, we also get a discrete cluster membership probability, providing a methylome-based clustering of cells. This twofold output of Melissa reflects its methodological foundations as a hybrid between a global unsupervised model (Bayesian clustering of methylomes) and a local supervised learning model (GLM regression for every region). In this sense, Melissa is closer to a *mixture of experts* model (Bishop, 2006, chapter 14) than a standard mixture model.

6.1.1 Melissa model

The BPRMeth model is limited to sharing information across CpGs via local smoothing (which certainly helps in dealing with data sparsity), however, in our experience the coverage in scBS-seq data is insufficient to infer informative methylation profiles at many genomic regions (see below). We therefore propose Melissa to exploit the population structure of the experimental design and additionally transfer information across similar cells. To make the chapter self-contained we restate the main equations of the BPRMeth model, which is defined over a single cell. Briefly, for a specific genomic region m we model the observed methylation of CpG site i as

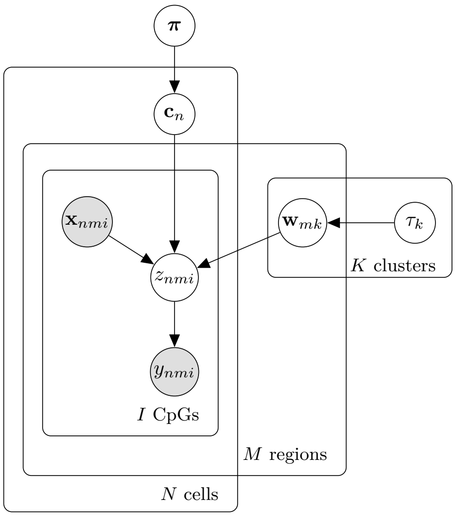
$$\begin{aligned}\mathbf{w}_m \mid \tau &\sim p(\mathbf{w}_m \mid \tau),\\ y_{mi} \mid \mathbf{w}_m, \mathbf{x}_{mi} &\sim \mathcal{B}\text{ern}\big(y_{mi} \mid \Phi(\mathbf{w}_m^\top \mathbf{x}_{mi})\big).\end{aligned}$$

Here $\mathbf{x}_{mi} \overset{\text{def}}{=} \mathbf{h}(x_{mi})$ denotes the basis function transformed CpG locations x_{mi}, $\mathbf{w}_m \in \mathbb{R}^D$ represents the regression coefficients, τ is the hyper-parameter of the prior, and Φ is the inverse probit function.

Assume now that we have $N(n = 1, \ldots, N)$ cells and each cell consists of $M(m = 1, \ldots, M)$ genomic regions, for example promoters, and we are interested in both partitioning the cells in K clusters and inferring the methylation profiles for each genomic region. To do so, we use a finite Dirichlet mixture model (FDMM) (McLachlan and Peel, 2004), where we assume that the methylation profile of the m^{th} region for each cell n is drawn from a mixture distribution with K components (where $K < N$). This way cells belonging to the same cluster will share the same methylation profile, although profiles will still differ across genomic regions. Let $\mathbf{c}_n$ be a latent variable comprising a 1-of-K binary vector with elements c_{nk} representing the component that is responsible for cell n, and π_k be the probability that a cell belongs to cluster k, i.e. $\pi_k = p(c_{nk} = 1)$. The conditional distribution of $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c_N}\}$ given $\boldsymbol{\pi}$ is

$$p(\mathbf{C} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{c_{nk}}. \tag{6.1}$$

Considering the FDMM as a generative model, the latent variables $\mathbf{c}_n$ will generate the latent observations $\mathbf{Z}_n \in \mathbb{R}^{M \times I_m}$, which in turn will generate the binary observations $\mathbf{Y}_n \in \{0, 1\}^{M \times I_m}$ depending on the sign of $\mathbf{Z}_n$, as shown in figure 5.4. The conditional distribution of $(\mathbf{Z}, \mathbf{Y})$



$$\begin{aligned}
\boldsymbol{\pi} &\sim \mathcal{D}\text{ir}(\boldsymbol{\delta}_0) \\
\mathbf{c}_n \,|\, \boldsymbol{\pi} &\sim \mathcal{D}\text{iscrete}(\boldsymbol{\pi}) \\
\tau_k &\sim \mathcal{G}\text{amma}(\alpha_0, \beta_0) \\
\mathbf{w}_{mk} \,|\, \tau_k &\sim \mathcal{N}(\mathbf{0}, \tau_k^{-1}\mathbf{I}) \\
z_{nmi} \,|\, \mathbf{w}_{mk}, \mathbf{x}_{nmi} &\sim \mathcal{N}(\mathbf{w}_{mk}^{\top}\mathbf{x}_{nmi}, 1) \\
y_{nmi} \,|\, z_{nmi} &= \begin{cases} 1 & \text{if } z_{nmi} > 0 \\ 0 & \text{if } z_{nmi} \leq 0 \end{cases}
\end{aligned}$$

Figure 6.2 Probabilistic graphical representation of the Melissa model.

given the latent variables $\mathbf{C}$ and the component parameters $\mathbf{W}$ becomes

$$p(\mathbf{Y}, \mathbf{Z} \,|\, \mathbf{C}, \mathbf{W}, \mathbf{X}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\prod_{m=1}^{M} p(\mathbf{y}_{nm} \,|\, \mathbf{z}_{nm})\, p(\mathbf{z}_{nm} \,|\, \mathbf{w}_{mk}, \mathbf{X}_{nm})\right]^{c_{nk}}, \tag{6.2}$$

where

$$p(\mathbf{y}_{nm} \,|\, \mathbf{z}_{nm}) = \mathbb{I}(\mathbf{z}_{nm} > 0)^{\mathbf{y}_{nm}} + \mathbb{I}(\mathbf{z}_{nm} \leq 0)^{(\mathbf{1}-\mathbf{y}_{nm})}.$$

To complete the model we introduce priors over the parameters. We choose a Dirichlet distribution over the mixing proportions, $p(\boldsymbol{\pi}) = \mathcal{D}\text{ir}(\boldsymbol{\pi} \,|\, \boldsymbol{\delta}_0)$, where for symmetry we choose the same parameter δ_{0_k} for each of the mixture components. We also introduce an independent Gaussian prior over the coefficients $\mathbf{W}$, that is,

$$p(\mathbf{W} \,|\, \boldsymbol{\tau}) = \prod_{m=1}^{M}\prod_{k=1}^{K} \mathcal{N}(\mathbf{w}_{mk} \,|\, \mathbf{0}, \tau_k^{-1}\mathbf{I}). \tag{6.3}$$

Finally, we introduce a prior distribution for the (hyper)-parameter $\boldsymbol{\tau}$, and assume that each cluster has its own precision parameter, $p(\tau_k) = \mathcal{G}\text{amma}(\tau_k \,|\, \alpha_0, \beta_0)$. Having defined our model we can now write the joint distribution over the observed and latent variables

$$p(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau} \,|\, \mathbf{X}) = p(\mathbf{Y} \,|\, \mathbf{Z})\, p(\mathbf{Z} \,|\, \mathbf{C}, \mathbf{W}, \mathbf{X})\, p(\mathbf{C} \,|\, \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\mathbf{W} \,|\, \boldsymbol{\tau})\, p(\boldsymbol{\tau}), \tag{6.4}$$

where the factorisation corresponds to the probabilistic graphical model shown in figure 6.2.

Importantly, Melissa is a hybrid between a global unsupervised clustering model and a local supervised prediction model encoded through the GLM regression coefficients $\mathbf{w}_{mk}$ for each genomic region and cluster. When considering Melissa as an imputation (or predictive) model, the training data are obtained by using only a subset of CpG tuples (x_{nmi}, y_{nmi}) for each region.

For example, from the observed I_{nm} CpGs in a given region and cell, Melissa will only see $I_{nm}/2$ random CpGs during training, and the remaining CpGs will be used as a held out test set to evaluate its prediction performance. Note that in any case, either using all CpGs or a subset during training, Melissa will additionally perform clustering at the global level which is encoded through the latent variables $\mathbf{c}_n$. Notice that both BPRMeth and Melissa do not explicitly model bisulfite conversion errors. Conversion errors are estimated to be relatively rare and below 1% according to Genereux et al. (2008), and we show in our simulation studies that Melissa is highly robust to the addition of noise mimicking possible conversion errors.

6.1.2 Mean field variational inference for Melissa

Similarly to BPRMeth exact computation of the posterior distribution for the Melissa model $p(\mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau} \mid \mathbf{Y}, \mathbf{X})$ is not analytically tractable; hence, we resort to approximate techniques. The most common method for approximate Bayesian inference is to perform MCMC, however, sampling methods require considerable computational resources and do not scale well when performing genome-wide analysis on hundreds or thousands of single cells. Variational methods can provide an efficient, approximate solution with better scalability in this case (see subsection 6.3.1 for a comparison between Gibbs sampling and variational inference for this model). Besides the computational advantages, the deterministic nature of the variational inference machinery makes it easier to assess convergence compared to MCMC methods (Beal, 2003). More specifically we use mean field variational inference (Blei et al., 2017), which as explained in subsection 3.4.1 assumes that the variational distribution factorises over the latent variables[3]

$$q(\mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau}) = q(\mathbf{Z})\, q(\mathbf{C})\, q(\mathbf{W})\, q(\boldsymbol{\pi})\, q(\boldsymbol{\tau}). \tag{6.5}$$

Next we iteratively update each factor q while holding the remaining factors fixed using the coordinate ascent variational inference (CAVI) algorithm. The procedure for performing CAVI for the Melissa model is summarised in algorithm 6.1.

Predictive density and model selection

Given an approximate posterior distribution we are in the position to predict the methylation state at unobserved CpG sites. The predictive density of a new observation $\mathbf{y}_*$, which is associated with latent variables $\mathbf{c}_*$, $\mathbf{z}_*$ and covariates $\mathbf{X}_*$, is given by

$$\begin{aligned} p(\mathbf{y}_* \mid \mathbf{X}_*, \mathbf{Y}) &= \sum_{c_*} \int \int p(\mathbf{y}_*, \mathbf{c}_*, \mathbf{z}_*, \boldsymbol{\theta} \mid \mathbf{X}_*, \mathbf{Y}) d\boldsymbol{\theta} d\mathbf{z}_* \\ &\simeq \sum_{k=1}^{K} \frac{\delta_k}{\sum_j \delta_j} \mathcal{B}\text{ern}\left(\mathbf{y}_* \middle| \Phi\left(\frac{\mathbf{X}_* \boldsymbol{\lambda}_k}{\sqrt{\mathbf{I} + \text{diag}(\mathbf{X}_* \mathbf{S}_k \mathbf{X}_*^T)}}\right)\right), \end{aligned} \tag{6.6}$$

where we collectively denote as $\boldsymbol{\theta}$ the relevant parameters being marginalised.

[3] Detailed mathematical derivations of the mean field variational inference can be found in appendix C.1.

Algorithm 6.1 CAVI for Melissa model

```
initialize Gaussian factor $\boldsymbol{\lambda}, \mathbf{S}$; Dirichlet factor $\boldsymbol{\delta_0}$; and Gamma factor $\alpha_0, \beta_0$.
Update $\alpha_k \leftarrow \alpha_0 + MD/2$
Update $\beta_k \leftarrow \beta_0$
while ELBO has not converged do
    Set $\gamma_{nmk} = (\mathbf{z}_{nm} - \mathbf{X}_{nm}\mathbf{w}_{mk})$    ▷ Variational E-step
    Update $r_{nk} \propto \langle \log \pi_k \rangle_{q(\pi_k)} + \sum_m \left\langle -\frac{1}{2}\gamma_{nmk}^{\top}\gamma_{nmk} \right\rangle_{q(\mathbf{z}_{nm}, \mathbf{w}_{mk})}$
                                                                             ▷ Variational M-step
    Update $\delta_k \leftarrow \delta_{0_k} + \sum_n r_{nk}$    ▷ Dirichlet distribution parameter
    Update $\beta_k \leftarrow \beta_0 + \frac{1}{2}\sum_m \left\langle \mathbf{w}_{mk}^{\top}\mathbf{w}_{mk} \right\rangle_{q(\mathbf{w}_{mk})}$    ▷ Gamma distribution parameter
    Update $\mu_{nmi} \leftarrow \sum_k r_{nk} \left\langle \mathbf{w}_{mk}^{\top}\mathbf{x}_{nmi} \right\rangle_{q(\mathbf{w}_{mk})}$    ▷ Mean of truncated Gaussian
    Set $\langle z_{nmi} \rangle_{q(z_{nmi})} = \begin{cases} \mu_{nmi} + \phi(-\mu_{nmi})/(1 - \Phi(-\mu_{nmi})) & \text{if } y_{nmi} = 1 \\ \mu_{nmi} - \phi(-\mu_{nmi})/\Phi(-\mu_{nmi}) & \text{if } y_{nmi} = 0 \end{cases}$
    Update $\mathbf{S}_{mk} \leftarrow \left(\frac{\alpha_k}{\beta_k}\mathbf{I} + \sum_n r_{nk}\mathbf{X}_{nm}^{\top}\mathbf{X}_{nm}\right)^{-1}$    ▷ Regression coefficient covariance
    Update $\boldsymbol{\lambda}_{mk} \leftarrow \mathbf{S}_{mk}\sum_n r_{nk}\mathbf{X}_{nm}^{\top} \langle \mathbf{z}_{nm} \rangle_{q(\mathbf{z}_{nm})}$    ▷ Regression coefficient mean
    Update $\mathcal{L}\left(q(\mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau})\right)$    ▷ Compute ELBO
end while
```

One of the most appealing aspects of variational approximations within mixture models is the possibility of directly performing model selection, i.e. determining the number of clusters within the optimisation procedure. It has been repeatedly observed (Corduneanu and Bishop, 2001) that, when fitting variationally a mixture model with a large number of components, the variational procedure will prune away components with no support in the data, hence effectively determining an appropriate number of clusters in an automatic fashion. We can gain some intuition as to why this happens in the following way. We can rewrite the KL divergence as

$$\mathrm{KL}\left[q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \,|\, \mathbf{X})\right] = \log p(\mathbf{X}) - \langle \log p(\mathbf{X} \,|\, \boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})} + \mathrm{KL}\left[q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right], \qquad (6.7)$$

where $\log p(\mathbf{X})$ can be ignored since is constant with respect to $q(\boldsymbol{\theta})$. To minimize this objective function the variational approximation will both try to increase the expected log likelihood of the data $\log p(\mathbf{X} \,|\, \boldsymbol{\theta})$ while minimizing its KL divergence with the prior distribution p($\boldsymbol{\theta}$). Hence, using variational inference we have an automatic trade-off between fitting the data and model complexity (Bishop, 2006, chapter 10); giving the possibility to automatically determine the number of clusters without resorting to cross-validation techniques. Visualising the model selection property of the Melissa model is rather involved due to the existence of many cells and genomic regions. Nevertheless, by focusing on a single cell we provide a concrete example of model selection by generating M = 300 genomic regions from K = 3 clusters (generated from the methylation profiles corresponding to the top right of figure 5.1). Then, we set the

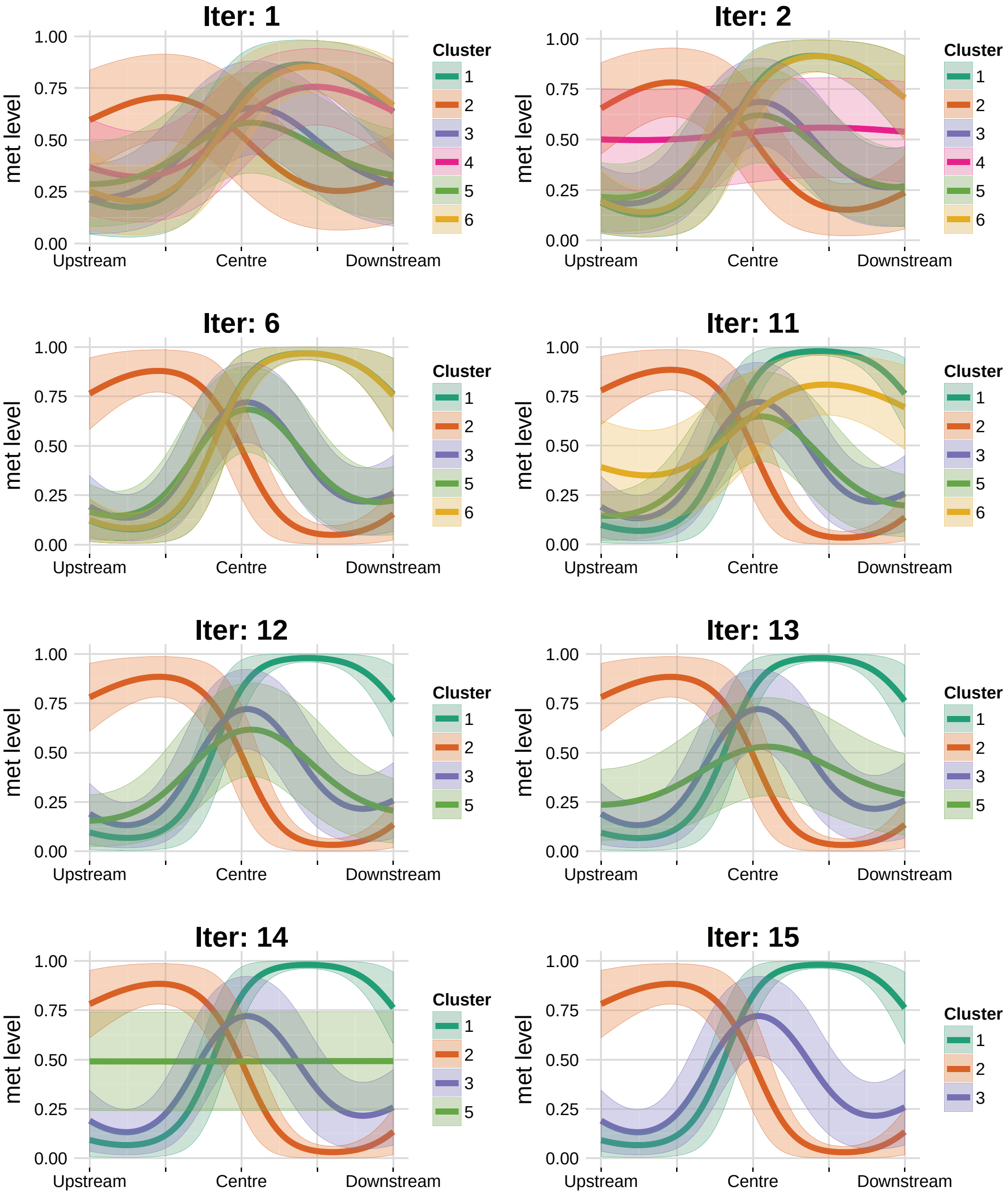


Figure 6.3 Variational inference automatically performs model selection for Melissa. Essentially, components that did not explain the data were returned back to their prior values, i.e. constant functions with 0.5 methylation level corresponding to $\mathbf{w} = \mathbf{0}$ for all basis function coefficients.

initial number of clusters to K = 6 and let the variational optimisation to prune away inactive clusters. Figure 6.3 shows the state of the algorithm during different iterations; where only after 15 iterations it automatically recovered the correct number of clusters. Figure 6.4 shows the trajectory of the evidence lower bound over each iteration during the optimisation procedure, including small bumps when the model discards mixture components.

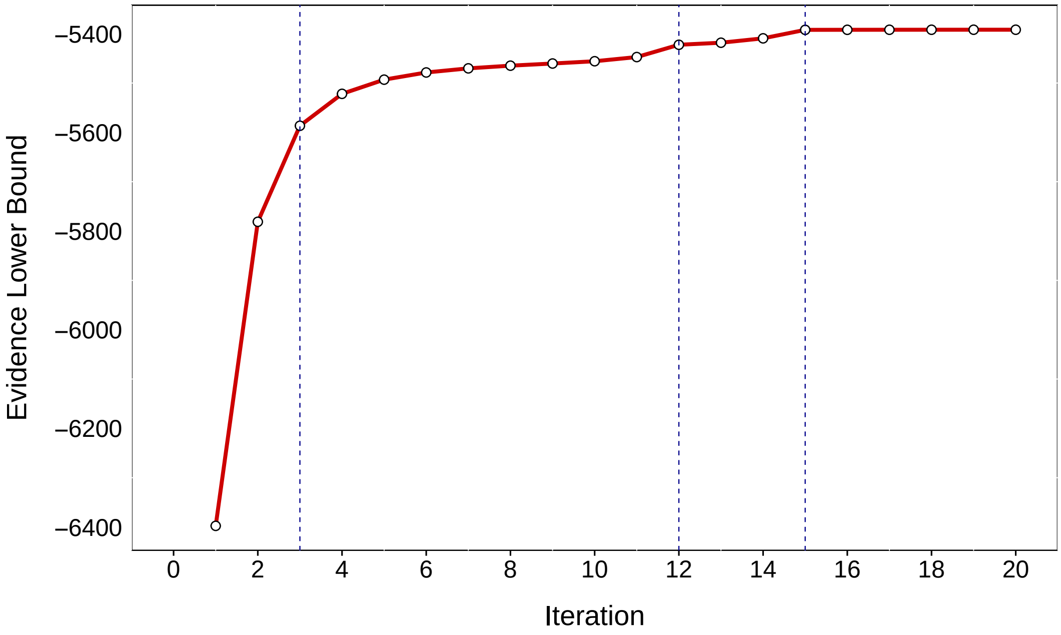


Figure 6.4 Evidence lower bound during model optimisation for the synthetic data shown in figure 6.3. Initially the model had six components. Each vertical blue line indicates the iteration time when a mixture component was pruned away; note the bumps in ELBO when the model discards components.

6.2 Experiment design and data preprocessing

We benchmark the ability of our model to cluster and impute CpG methylation states at the single cell level both on simulated and mouse embryonic stem cell (ESC) datasets. To assess test prediction performance we consider different metrics, including the F-measure (equation D.7), the area under the receiver operating characteristics curve (AUC) and precision recall curves (Powers, 2011). As a measure of clustering performance we use the Adjusted Rand Index (ARI, equation D.6) (Hubert and Arabie, 1985), which compares the true cluster assignment used to generate the synthetic data and the predicted cluster membership given by the model.

To benchmark the performance of *Melissa* in predicting CpG methylation states, we compare it against seven different imputation strategies. As a baseline approach, we compute the average methylation rate separately for each cell and region (*Rate*), that is, the average is taken over all CpG sites forming a genomic region. We also use the *BPRMeth* model (Kapourani and Sanguinetti, 2016, 2018a), where we account for the binary nature of the observations, which we train independently across cells and regions. Note that BPRMeth shares information across CpG sites inside each genomic region, however, it does not transfer information across cells. To share information across cells, but not across neighbouring CpGs inside the region, we constrain Melissa to infer constant functions, i.e. learn average methylation rate (*Melissa rate*). We also use a Gaussian mixture model (*GMM*) that takes as input M-values (Du et al., 2010) instead of average methylation rates across the region. The transformation from average methylation rates to M-values is obtained by

$$M\text{-value} = \log_2 \left(\frac{\text{rate} + 0.01}{1 - \text{rate} + 0.01} \right). \tag{6.8}$$

Additionally, as a fully independent baseline we use a Random Forest classifier trained on individual cells and regions, where the input features are the observed CpG locations and the response variable is the CpG methylation state: methylated or unmethylated (*RF*). This is essentially the method of Zhang et al. (2015), however, without using additional annotation data or DNA sequence patterns. Finally, we compare *Melissa* to the deep learning method *DeepCpG* (Angermueller et al., 2017) that uses the information of neighbouring CpGs to predict the methylation state of each target CpG site. It should be noted that DeepCpG is designed to predict individual missing CpGs, rather than missing regions, and requires always information about neighbouring CpGs[4]. This means that during prediction *DeepCpG* will potentially have access to more data than competing methods, potentially providing it with an unfair advantage; to partly address this issue we also present results when *DeepCpG* had access to sub-sampled data (labelled *DeepCpG Sub* in the figures below). In general, *DeepCpG* should be thought of as complementary to *Melissa* and comparisons should be evaluated with caution (see subsection 6.3.4 below).

6.2.1 Assessing Melissa on a simulation study

To generate realistic simulated single-cell methylation data, we first use the BPRMeth package (Kapourani and Sanguinetti, 2018a) to infer five prototypical methylation profiles from the GM12878 lymphoblastoid cell line. The bulk BS-seq data for the GM12878 cell line are publicly available from the ENCODE project (Dunham et al., 2012). Based on these profiles we simulate single cell methylation data (i.e. binary CpG methylation states) for M = 100 genomic regions, where each CpG is generated by sampling from a Bernoulli distribution with probability of success given by the latent function evaluation at the specific site. To mimic the inherent noise introduced by bisulfite conversion error, Gaussian noise $\mathcal{N}(\mu = 0, \sigma = 0.05)$ is introduced to the probability of success prior to generating each binary CpG site. This process can be thought of as generating methylation data for a specific single cell. Next, we generate K = 4 cell sub-populations by randomly shuffling the genomic regions across clusters, so now each cell sub-population has its own methylome landscape. In total we generated N = 200 cells with the following cell sub-population proportions: 40%, 25%, 20% and 15%. Finally, to account for different levels of similarity between cell sub-populations, we simulate 11 different datasets by varying the proportion of similar genomic regions between clusters.

[4] This different training approach makes the *DeepCpG* model incompatible with the simulation setting presented in subsection 6.3.1.

6.2.2 Assessing Melissa on subsampled ENCODE data

To faithfully simulate methylation data that resemble scBS-seq experiments, we generate an additional synthetic dataset by subsampling the bulk ENCODE RRBS data from two different cell lines (GM12878 and H1-hESC). To retain the structure of missing data observed in scBS-seq experiments (due to read length), we directly subsample the raw FASTQ (Cock et al., 2009) files which essentially lead to discarding individual reads rather than individual CpGs. From each cell line we generate 40 pseudo-single cells by randomly keeping 10% of the mapped reads from the bulk experiment, resulting in 80 cells when combining both synthetic datasets. This process is performed for chromosomes 1 to 6 to alleviate the computational burden. Subsequently, the same preprocessing steps detailed in subsection 6.2.3 were performed, with the only difference that for this study we consider only ±5 kb regions around the transcription start site (TSS). Each model, except DeepCpG, used 20%, 50% and 80% of the CpGs as training set, and the remaining CpGs were used as a test set to evaluate imputation performance. The DeepCpG model used chromosomes 1 and 3 as training set, chromosome 5 as validation set and the remaining chromosomes as test set.

6.2.3 Real scBS-seq data and preprocessing

Two mouse embryonic stem cells (ESCs) datasets are used to validate the performance of the Melissa model. The first dataset presented in Angermueller et al. (2016), after quality assessment consists of 75 single cells out of which 14 cells are cultured in 2i medium (*2i ESCs*) and the remaining 61 cells are cultured in serum conditions (*serum ESCs*). The Bismark (Krueger and Andrews, 2011) processed data, with reads mapped to the GRCm38 mouse genome, are downloaded from the Gene Expression Omnibus under accession GSE74535. The second dataset of Smallwood et al. (2014) consists 32 cells out of which 12 cells are *2i ESCs* and the remaining 20 cells are *serum ESCs*, and the Bismark processed data, with reads mapped to the GRCm38 mouse genome, are publicly available under accession number GSE56879. For both datasets the observed data that are used as input to Melissa are binary methylation states: unmethylated CpGs are encoded with zero and methylated CpGs with one.

Since Melissa considers regions for a specific genomic context, we use the BPRMeth package to filter CpGs that do not fall inside these regions. Then we create a simple data structure where each cell is a encoded as a list, and each entry of the list — corresponding to a specific genomic region — is a matrix with two columns: the (relative) CpG location and the methylation state. We apply the competing methods on six different genomic contexts: protein coding promoters with varying genomic windows: ±1.5 kb, ±2.5 kb and ±5 kb around TSS, active enhancers, super enhancers and Nanog regulatory regions. Due to the sparse CpG coverage, for the three genomic contexts, except promoters, we filtered regions with smaller than 1 kb annotation length and specifically for Nanog regions we took a window of ±2.5 kb around the centre of the genomic annotation. In addition, we only considered regions that were covered in at least

50% of the cells with a minimum coverage of 10 CpGs and had between cell variability; the rationale being that homogeneous regions across cells do not provide additional information for identifying cell sub-populations. The sparsity level of the two scBS-seq datasets across different genomic contexts is shown in table C.3. It should be noted, that imputation performance is evaluated only on genomic regions that pass the filtering threshold. We run the model with $K = 6$ and $K = 5$ clusters for the Angermueller et al. (2016) and Smallwood et al. (2014) datasets, respectively, and we use a broad prior over the model parameters.

DeepCpG training

The DeepCpG method takes a different imputation approach: it is trained on a specific set of chromosomes and predicts methylation states on the remaining chromosomes where it imputes each CpG site sequentially by using as input a set of neighbouring CpGs. This approach makes it difficult to fairly compare with the rival methods, since for each CpG the input features to DeepCpG are all the neighbouring sites, whereas the competing models have access to a subset of the data and they make predictions in one pass for the whole region. Since we only had access to CpG methylation data and in order to make it comparable with the considered methods, we trained the CpG module of DeepCpG (termed *DeepCpG CpG* in Angermueller et al. (2017)).

For the Angermueller et al. (2016) dataset, chromosomes 3 and 17 were used as training set, chromosomes 12 and 14 as validation set and the remaining chromosomes as test set. For the Smallwood et al. (2014) dataset, chromosomes 3, 17 and 19 were used as training set, chromosomes 12 and 14 as validation set and the remaining chromosomes as test set. The chosen chromosomes had at least 3 million CpGs used as training set; a sensible size for the DeepCpG model as suggested by the authors. A neighbourhood of K = 20 CpG sites to the left and the right for each target CpG was used as input to the model. During testing time, even if a given genomic region did not contain at least 40 CpGs, the DeepCpG model used additional CpGs outside this window to predict methylation states; hence using more information compared to the rival models. In total the DeepCpG model took around three to four days per dataset for training and prediction on a cluster equipped with NVIDIA Tesla K40ms GPUs.

6.3 Results

6.3.1 Benchmarking Melissa on simulated data

We benchmark the ability of our model to cluster and impute CpG methylation states at the single cell level both on simulated and mouse ESC datasets. For the simulated dataset presented in subsection 6.2.1, we compare *Melissa* against all methods except *DeepCpG*, since it is not applicable in the settings of this simulation. Applying the competing methods to the synthetic data we observe that *Melissa* yields a substantial improvement in prediction accuracy compared

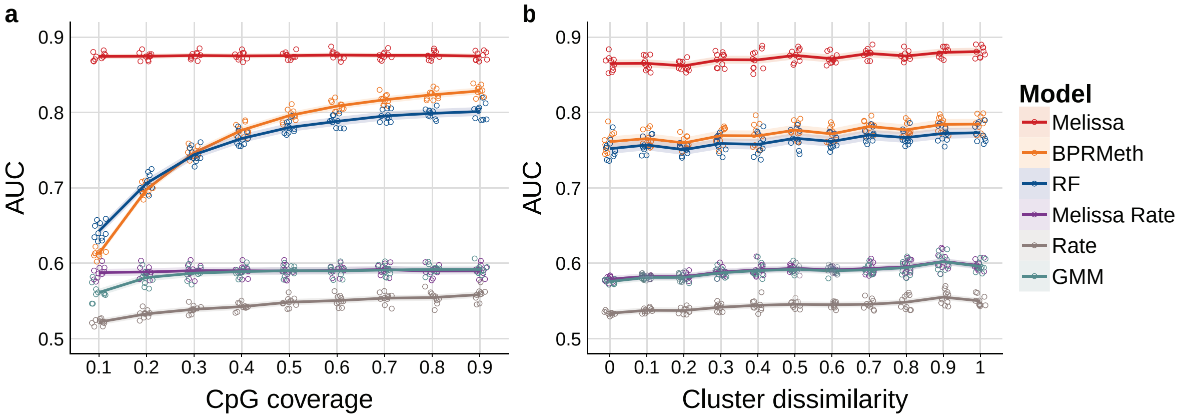


Figure 6.5 Melissa robustly imputes CpG methylation states on synthetic data. (**a**) Imputation performance in terms of AUC as we vary the proportion of covered CpGs used for training. Higher values correspond to better imputation performance. For each CpG coverage setting a total of 10 random splits of the data to training and test sets was performed. Each coloured circle corresponds to a different simulation. The plot shows also the LOESS curve for each method as we increase CpG coverage. (**b**) Imputation performance measured by AUC for varying proportions of similar genomic regions between clusters. Values closer to zero correspond to highly similar cell sub-populations, whereas values closer to one correspond to well separated cell sub-populations. In (**a**) cluster dissimilarity was set to 0.5 and in (**b**) CpG coverage was set to 0.4.

to all other models (see figures 6.5 and C.1). Notably, *Melissa* is robust across different settings of the data, such as CpG coverage proportion in each region (figure 6.5a), or different levels of dissimilarity across clusters (figure 6.5b). Due to its ability to transfer information across cells and neighbouring CpGs, our model robustly maintains its prediction accuracy even at a very sparse coverage level of 10%. The *BPRMeth* and *RF* models perform poorly at low CpG coverage settings, becoming comparable to *Melissa* when using the majority of the CpGs for training set. Importantly, *Melissa* still performs better at 90% CpG coverage, demonstrating that the clustering acts as an effective regularisation for imputing unassayed CpG sites. As expected, *Melissa Rate* and *GMM* have very similar performance (due to the very similar model structure); for both methods performance is significantly weaker than *Melissa* across the full range of simulation settings, since they are not expressive enough to capture spatial correlations between CpGs. Finally, the naive *Rate* method has the worst imputation performance of all methods, by a considerable margin. The imputation performance of all methods is relatively insensitive to the degree of cluster dissimilarity (figure 6.5b).

Next we consider the clustering performance of *Melissa*. Since most of the rival methods do not have a notion of clustering, we compare *Melissa* to clustering using methylation rates for binary data (*Melissa Rate*) or Gaussian data (*GMM*) using M-values (Du et al., 2010). As a performance metric we use the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). Figure 6.6a shows ARI values comparing the three models for varying CpG coverage (with cluster dissimilarity level at 0.5). *Melissa* performs perfectly in all settings, demonstrating its

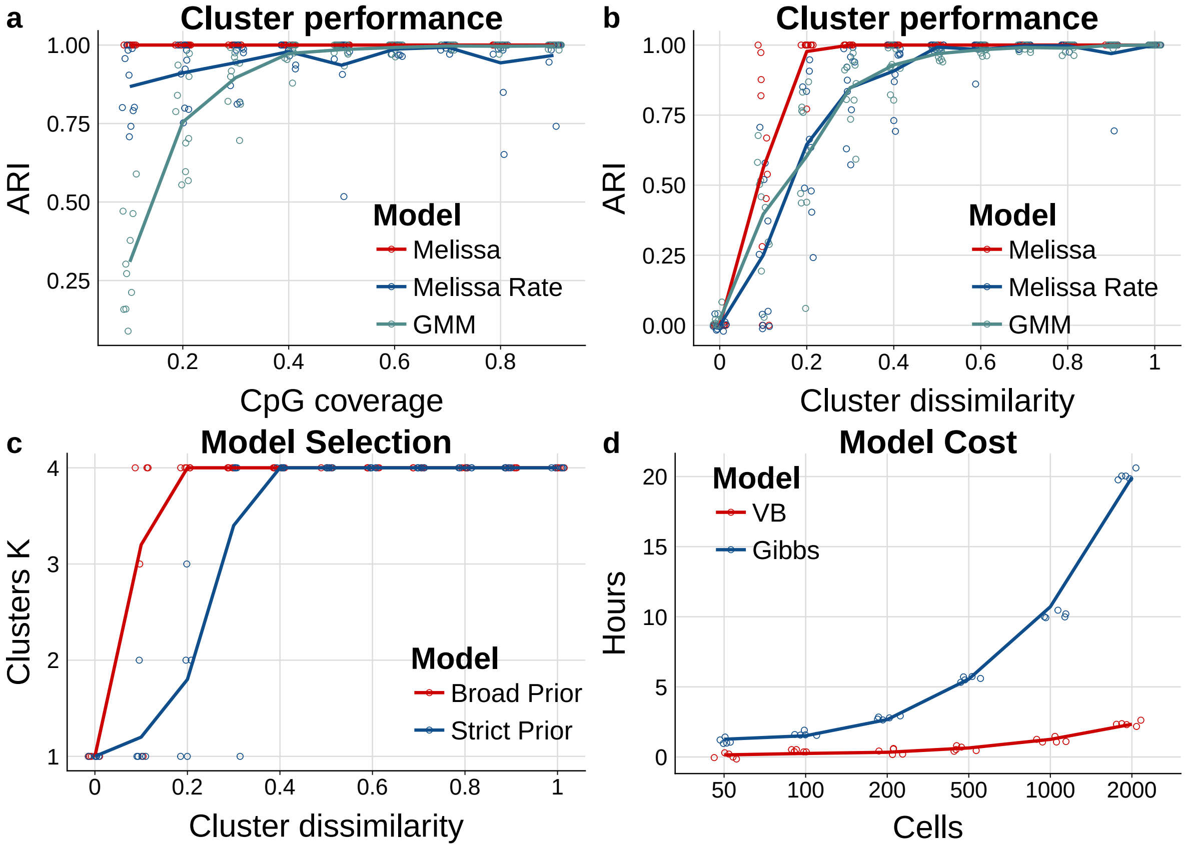


Figure 6.6 Melissa efficiently and accurately clusters cell sub-populations. (**a**) Clustering performance measured by ARI as we vary CpG coverage. Higher values correspond to better agreement between predicted and true cluster assignments. For each CpG coverage setting a total of 10 random splits of the data to training and test sets was performed. Each coloured circle corresponds to a different simulation. The plot shows also the LOESS curve for each method as we increase CpG coverage. (**b**) Clustering performance (ARI) for varying proportions of similar genomic regions between clusters. (**c**) Predicted number of clusters using two different prior settings: a broad and a strict prior as we vary cluster dissimilarity. Initial number of clusters was set to $K = 10$. Melissa identifies the correct number of clusters in most parameter settings ($K = 4$); notably when there is no dissimilarity across clusters (i.e. we have one global cell sub-population), Melissa prunes away all components and keeps only one cluster ($K = 1$). (**d**) Running times for varying number of cells for the variational Bayes (VB) and Gibbs sampling implementations for the Melissa model, where each cell consists of $M = 200$ genomic regions.

power and sensitivity in identifying robustly the cell sub-population structure. When varying the level of cluster dissimilarity (see figure 6.6b), the model is still able to retain its high clustering performance. As expected, for settings with low variability between clusters (i.e. cell sub-populations are difficult to distinguish), the performance drops; however, *Melissa* is consistently superior to the *Melissa Rate* and *GMM* models rapidly reaching near-perfect clustering accuracy.

Subsequently, we test *Melissa*'s ability to perform model selection, that is, to identify the appropriate number of cell sub-populations. To do so, we run the model on simulated data by setting the initial number of clusters to K = 10 and letting the variational optimisation prune away inactive clusters (Corduneanu and Bishop, 2001). We use both broad (red line) and shrinkage (blue line) priors. Figure 6.6c shows that the variational optimisation automatically recovered the correct number of mixture components for almost all parameter settings. As expected, in settings with high between cluster similarity, the model with shrinkage prior returns fewer clusters, since the data complexity term in equation (6.7) is penalizing more the variational approximation compared to the gain in likelihood from explaining the data.

Finally, we assess the scalability of *Melissa* with respect to the number of single cells. Figure 6.6d compares the variational inference (VB, red line) with the Gibbs sampling (blue line) algorithm, demonstrating the good scalability of variational inference where we can analyse thousands of single cells in acceptable running times. The maximum number of iterations for variational inference was set to 400 and the Gibbs algorithm was run for 3000 iterations. Both algorithms are implemented in the R programming language and were run on a machine utilising at most 16 CPU cores.

6.3.2 Benchmarking Melissa on subsampled bulk ENCODE data

The results in subsection 6.3.1 convincingly showed a substantial advantage of *Melissa* over competing methods in terms of both imputation performance and clustering. However, conditioned on some seed profiles learnt from bulk experiments, the simulation is conducted on data which is directly sampled from the generative *Melissa* model (with some additional noise to account for conversion errors), which could conceivably introduce an unfair bias in the comparison. Additionally, since the synthetic data are simulated as separate independent regions, comparison with the deep learning method DeepCpG (Angermueller et al., 2017) is not possible, since DeepCpG requires the information of a large number of neighbouring CpGs to predict the methylation state of each target site.

To address these limitations we generated an additional benchmark dataset by directly sub-sampling bulk ENCODE experiments as explained in subsection 6.2.2. In addition, this simulation study produces observations with a more similar structure to scBS-seq experiments, since the uneven read coverage better captures the structure of missing data observed in single cell epigenomic experiments. Table 6.1 shows the results of this study when imputing CpGs falling in genomic regions of ±5 kb around transcription start sites across different sparsity levels. Consistently with the simulation study in the previous subsection, *Melissa* performs significantly better than competitors at imputation tasks. DeepCpG has a strong performance with comparable (but systematically lower) accuracy to *Melissa* on this dataset across all CpG coverage settings (notice that training of DeepCpG is slightly different, see subsection 6.2.3). The results are consistent across all different metrics considered in this paper (see figures C.2 and C.3). Finally, Melissa could easily separate both cell sub-populations for all settings considered in this study.

Model	AUC 20% coverage	AUC 50% coverage	AUC 80% coverage
Melissa	**0.965 (0.008)**	**0.965 (0.006)**	**0.968 (0.006)**
DeepCpG	0.946 (0.008)	0.946 (0.008)	0.946 (0.008)
BPRMeth	0.882 (0.008)	0.912 (0.01)	0.911 (0.008)
RF	0.818 (0.007)	0.888 (0.011)	0.895 (0.011)
Melissa rate	0.865 (0.008)	0.867 (0.009)	0.865 (0.008)
Rate	0.808 (0.011)	0.828 (0.009)	0.822 (0.008)

Table 6.1 Melissa robustly imputes CpG methylation states on subsampled ENCODE data. Imputation performance in terms of AUC as we vary the proportion of covered CpGs used for training. Higher values correspond to better imputation performance. For each CpG coverage setting a total of 10 random splits of the data to training and test sets was performed; shown are the mean AUC value together with two standard deviations of the estimate in parenthesis. Note that DeepCpG was trained once on two chromosomes, hence, the values do not change as we vary the CpG coverage.

6.3.3 Melissa accurately predicts methylation states on real data

To assess *Melissa*'s performance on real scBS-seq data we considered two mouse embryonic stem cell (ESC) datasets from the Angermueller et al. (2016) and Smallwood et al. (2014) studies. The mouse ESCs were cultured either in 2i medium (*2i ESCs*) or serum conditions (*serum ESCs*), hence we expect methylation heterogeneity between cell sub-populations. In addition, in *serum ESCs* there is evidence of additional CpG methylation heterogeneity (Ficz et al., 2013), making these data suitable for the model selection task to infer cell sub-population structure. The analysis on both datasets was performed on six different genomic contexts: protein coding promoters with varying genomic windows: ±1.5 kb, ±2.5 kb and ±5 kb around TSS, active enhancers, super enhancers and Nanog regulatory regions (see subsection 6.2.3 for details on data preprocessing).

We first applied *Melissa* on the Angermueller et al. (2016) dataset which consists of 75 single cells (14 *2i ESCs* and 61 *serum ESCs*). Figure 6.7a shows a direct comparison of the imputation performance of all the methods across a variety of genomic contexts. *Melissa* is better or comparable to rival methods in terms of AUC (figure 6.7a), and substantially more accurate in terms of F-measure (figure C.4), demonstrating its ability to capture local CpG methylation patterns. *DeepCpG* also performs strongly on most genomic regions, indicating that a flexible deep learning method is effective in capturing patterns of methylation. Similar results were obtained by considering different metrics (see figures C.5 and C.6). Boxplots show performance distributions across 10 independent training / test splits of the data, except for *DeepCpG*, where the high computational costs prevented such investigation. Interestingly, methods based on methylation rates performed poorly at promoters, underlining the importance of methylation profiles in distinguishing epigenetic state near transcription start sites and identifying meaningful cell sub-populations. For all models the imputation performance (in

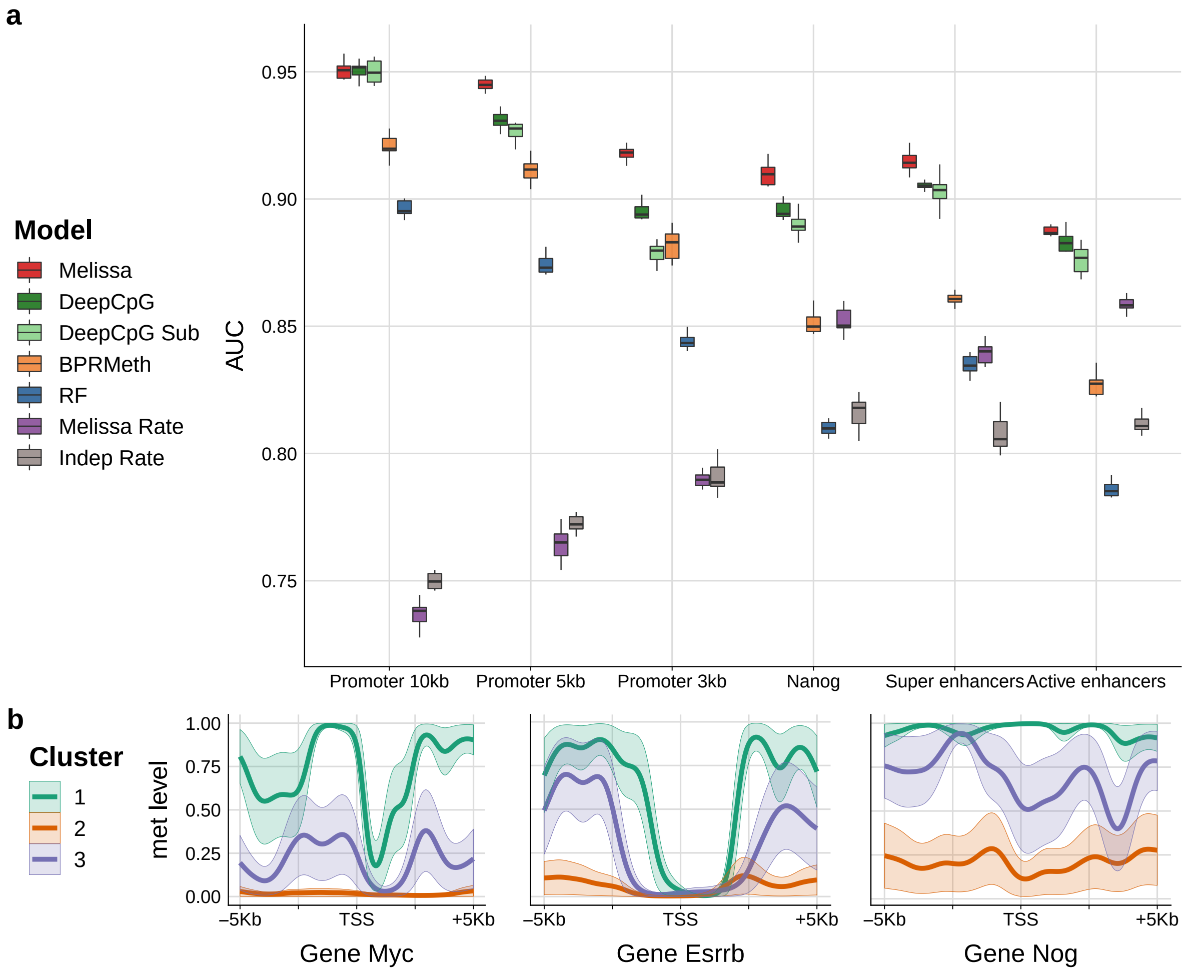


Figure 6.7 Imputation performance and clustering of mouse ESCs (Angermueller et al., 2016) based on genome wide methylation profiles. **(a)** Prediction performance on test set for imputing CpG methylation states in terms of AUC. Higher values correspond to better imputation performance. Each coloured boxplot indicates the performance using 10 random splits of the data in training and test sets; due to high computational costs, DeepCpG was trained only once and the boxplots denote the variability across ten random subsamplings of the test set. **(b)** Example promoter regions with the predicted methylation profiles for three developmental genes: *Myc, Esrrb* and *Nog*. Each coloured profile corresponds to the average methylation pattern of the cells assigned to each sub-population, in our case Melissa identified $K = 3$ clusters.

terms of AUC) at active enhancers was lower, indicating high methylation variability across cells and nearby CpG sites as shown in Smallwood et al. (2014).

In terms of clustering performance *Melissa* confirms that the data support the existence of a sub-population of serum cells as suggested in Ficz et al. (2013), by returning three clusters in almost all contexts. Further insights on the biological significance of the clusters obtained can be gleaned by inspecting the inferred methylation profiles at relevant regions. Figure 6.7b shows posterior methylation profiles for three developmental genes for each cell

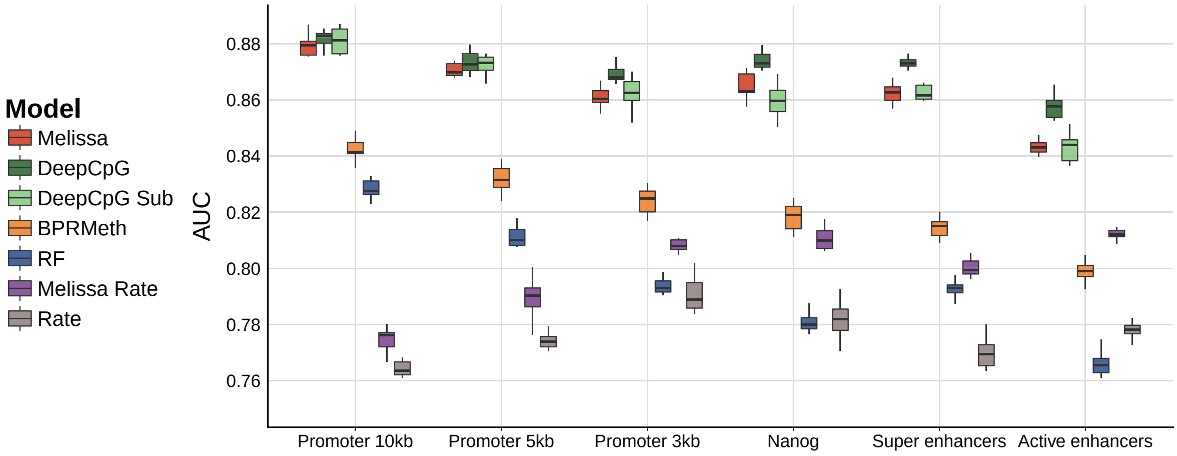


Figure 6.8 Imputation performance of mouse ESCs dataset (Smallwood et al., 2014) based on genome wide methylation profiles. Shown is the prediction performance, in terms of AUC, for imputing CpG methylation states. Each coloured boxplot indicates the performance using 10 random splits of the data in training and test sets; due to high computational costs, DeepCpG was trained only once and the boxplots denote the variability across ten random subsamplings of the test set.

sub-population (figure C.7 shows additional methylation profiles of developmental genes). Each colour corresponds to a different cell sub-population, with orange profiles corresponding to *2i ESCs* which are globally hypo-methylated. The green and purple profiles correspond to serum cells, which, as expected, present an increased level of methylation overall. However, *Melissa* identifies a clear sub-population structure within these serum cells: the purple cluster clearly represents a sub-population of cells which has only incompletely transitioned towards the final differentiated state (high global methylation punctuated by hypo-methylated CpG islands). Interestingly, 2i cells can be easily separated from serum cells based on methylation rate alone, due to the global hypo-methylation of 2i cells, however, the sub-population structure within serum cells appears to be determined by changes in profiles.

As a second real dataset we analysed the smaller Smallwood et al. (2014) study which consists of only 32 cells (12 *2i ESCs* and 20 *serum ESCs*). The imputation performance in terms of AUC across genomic contexts is shown in figure 6.8. Melissa retains its high prediction accuracy and is comparable with DeepCpG across most contexts (see figures C.8 – C.10 for performance on different metrics), even though the full DeepCpG model has slightly better performance on this dataset. This suggests that the small number of cells in this dataset did not allow an effective sharing of information. In terms of clustering performance, *Melissa* identifies three clusters in the vast majority of settings, once again underlying the emergence of epigenomically distinct populations within serum cells (see figures C.11 and C.12 for example methylation profiles across genomic contexts).

6.3.4 A note on the comparison with DeepCpG

Melissa and DeepCpG models reported substantially better imputation performance compared to the rival methods and show comparable performance when analysed on real datasets, demonstrating their flexibility in capturing complex patterns of methylation. However, the two methods have significantly different computational performances. In our experiments, Melissa's runtime was less than six hours for all genomic contexts running on a small server machine utilising at most ten CPU cores (see tables C.1 and C.2). By contrast, DeepCpG required around three to four days to analyse each dataset on a GPU cluster equipped with high end NVIDIA Tesla K40ms GPUs, and had very high memory requirements. These computational overheads effectively make DeepCpG out of reach for smaller research groups. On the other hand, Melissa operates on a set of genomic contexts of interest (e.g. promoters), while DeepCpG is designed for genome-wide imputation; computational performance of both methods will therefore depend on specific choices, such as the size/ number of the regions of interest for Melissa, or the number of training chromosomes for DeepCpG.

In addition to the differences in scope between the two methods, one should also be cautious when directly comparing prediction performances due to the different design of the DeepCpG model. DeepCpG is trained on a specific set of chromosomes and considers each CpG site independently; hence it does not have a notion of genomic region to be trained on, and will in any case utilize information from neighbouring CpGs within or outside the region, information that Melissa and the rival methods do not have access to.

6.4 Discussion

Single cell DNA methylation measurements are rapidly becoming a major tool to understand epigenetic gene regulation in individual cells. Newer platforms are rapidly expanding the scope of the technology in terms of assaying large numbers of cells (Luo et al., 2017), however, all technologies are plagued by intrinsically low coverage in terms of numbers of CpGs assayed.

In this chapter, we proposed Melissa as a way of addressing the low coverage issue by sharing information between CpGs with a local smoothing and between cells with a Bayesian clustering prior. On both synthetic and real data, Melissa achieved state of the art imputation performance over a panel of competing methods, including DeepCpG (Angermueller et al., 2017) and random forests. While achieving comparable or superior performance to black-box methods, such as neural networks and random forests, Melissa is more transparent and needs minimal tuning: all the results shown, on both synthetic and real data, were obtained with the same settings of the algorithm. Additionally, as all Bayesian methods, Melissa outputs are probability distributions that fully quantify the uncertainty on the model's prediction, and which are more easily usable for further experimental design compared to the point-estimates provided by black-box approaches. Melissa does not require additional annotation data as in Zhang et al. (2015) or Ernst and Kellis (2015), and does not exploit sequence information like DeepCpG, but an extension leveraging side data would be easily accomplished within the Bayesian framework and would represent an interesting extension for future research. By using a Bayesian clustering prior, Melissa has the added benefit of simultaneously uncovering the population structure within the assay, as we demonstrated in the real data examples; Melissa can therefore be a useful tool in uncovering epigenetic diversity among cells.

While Melissa accounts for heterogeneity in the cell population structure, it does not allow for heterogeneity at the single gene level: each cluster has a single methylation profile within each region, and all variability at the single locus level is attributed to noise. This rigidity limits the usefulness of Melissa as a tool to investigate intrinsic stochasticity in methylation at the single locus level. Relaxing the modelling assumptions to accommodate methylation variability in Melissa is an interesting topic for future research. Another area where Melissa could be fruitfully applied is the integrative study of multiple high-throughput features in single cells. With the advent of novel technologies measuring gene expression and multiple epigenomic features in individual cells (Clark et al., 2018), interpretable Bayesian models like Melissa are likely to play an important role in furthering our understanding of epigenetic control of gene expression in single cells.

Chapter 7

Discussion

High throughput sequencing platforms have enabled the profiling of epigenetic marks at unprecedented resolution, and this wealth of experimental data is promising to revolutionise our understanding of complex biological processes. To do so, bespoke computational methods that incorporate prior biological knowledge are indispensable to interpret the high-dimensional data and elucidate the regulatory role of the epigenome. Epigenetic marks make reversible modifications to the DNA and often have local effects in regulating gene expression (Richards and Hawley, 2011), hence, rigorous models are essential for capturing this local heterogeneity and effectively using it for downstream analysis. This thesis contributes to the development of flexible statistical models and algorithms to capture and quantify spatial correlations of epigenetic marks, mostly focusing on DNA methylation data generated from BS-seq experiments.

DNA methylation is implicated in diverse biological processes of direct clinical relevance, with the most notable example being cancer (Baylin and Jones, 2011). When it comes to analysing BS-seq data, the most common approach is to take simple averages across genomic regions of interest. This simplistic encoding of DNA methylation however cannot capture local spatially correlated patterns and does not exploit the richness of (the expensive to generate) BS-seq data which provide single nucleotide resolution. Chapter 4 concerned the development of BPRMeth, a generic modelling approach based on a generalised linear model of basis function regression to quantitate spatially distributed methylation profiles from bulk sequencing experiments. The rich representation of methylation patterns enabled us to build a powerful predictive model for gene expression, achieving correlations twice as large as previously reported across different ENCODE cell lines. In addition, methylation profiles were clustered based on a mixture modelling approach which identified prototypical profiles that explained most of the methylation variability across promoter regions. Reassuringly, some of these patterns recapitulated existing biological knowledge, such as U-shape methylation profiles that are associated with highly expressed genes (Edgar et al., 2014). This pattern might also explain findings that methylation at gene body is often positively correlated with active transcription (Lou et al., 2014).

BPRMeth is based on a GLM approach making it a versatile tool that can be readily applied on different sequencing technologies. In chapter 5 the BPRMeth model, and its algorithmic implementation, were substantially extended to provide a flexible environment for analysing and modelling spatial patterns of DNA methylation and similarly structured data from a variety of experimental platforms, including methylation arrays. In addition, the Bayesian formulation enabled differential analysis of epigenetic profiles using Bayes factors, by formulating the problem as a comparison of two models. Although not extensively evaluated, this approach has potentially the statistical power to identify differences between conditions or cells due to leveraging the spatial correlations of nearby epigenetic sites. The current approach however ignores variability between replicates (or cells within the same population), and extension on the lines of Stegle et al. (2010) is left as an interesting topic for future work.

Chapter 5 also concerned the scNMT-seq study, a novel single cell multi-omics protocol for parallel profiling of DNA methylation, chromatin accessibility and gene expression, which provides a powerful approach to investigate the coupling between the epigenome and transcriptome. Utilising the extended BPRMeth model, we quantified the cell-to-cell chromatin accessibility heterogeneity by reformulating the question as a model selection problem in mixture models; where the goal is to identify the most likely number of clusters that best explain the variability of accessibility profiles across cells. This approach revealed that genes with conserved accessibility profiles (fewer clusters) were both associated with higher gene expression levels and enriched for gene ontology terms linked to house-keeping functions, such as rRNA processing, splicing and translation. Intriguingly, when considering accessibility rates we lost both the association with gene expression and the enrichment for specific functions, suggesting again that profiles capture biologically meaningful information.

Single cell bisulfite sequencing experiments have enabled the characterisation of epigenetic heterogeneity and its dynamics on small sub-population of cells. However, due to inherent limitations of the technology the resulting methylation data are extremely sparse, effectively limiting the analysis repertoire to a semi-quantitative level. To tackle the sparse coverage issue chapter 6 introduced Melissa, a Bayesian hierarchical model that shares information between CpGs with local smoothing and between cells with a Bayesian clustering prior. This methodological foundation of Melissa as a hybrid between a local supervised approach and a global unsupervised model, enabled the accurate imputation of unassayed CpG sites and the methylome-based clustering of single cells, respectively. While achieving comparable or superior prediction performance to black-box methods, such as deep learning (Angermueller et al., 2017) and random forest models, on both synthetic and real mouse ESCs, Melissa is additionally more transparent and provides outputs that fully quantify the uncertainty in the model predictions. The clustering of single cell methylomes is based on a Bayesian finite Dirichlet mixture model which as we extensively demonstrated could effectively perform model selection and identify biologically meaningful cell sub-populations. A more principled modelling approach however would be to introduce a Dirichlet process prior for robustly identifying the cell sub-population structure (Blei and Jordan, 2006). Finally, thanks to a fast variational inference strategy Melissa has good scalability and can provide an effective modelling tool for the increasingly large single cell methylation studies which will become prevalent in coming years. These results provide the basis to attempt to answer several biologically and methodologically interesting questions. We discuss here briefly some of the most promising directions for future work.

7.1 Future work

This thesis focused mostly on the modelling aspect of spatial correlations of epigenomic marks, demonstrating that higher order epigenomic features can potentially capture biologically meaningful information. The methylation profiles for example might implicitly capture chromatin accessibility, transcription factor binding or histone modification information, assuming that these biological processes affect in a certain way the methylome landscape. Hence, interesting questions arise: Do methylation profiles correlate well with histone modification or transcription factor binding data obtained from ChIP-seq experiments? When incorporating these features as additional covariates, can we predict more accurately transcript abundance? Do these epigenetic 'signatures' affect different cell types in different ways? To leverage their full potential, the proposed statistical models should also be applied on targeted biological experiments. For instance, in bulk studies one application would be to quantify differences in the shape of methylation profiles when knocking out certain DNA methyltransferase enzymes. Regarding the analysis of single cell methylomes, datasets with rich sub-population structure and higher heterogeneity would prove useful in assessing the discriminative power of epigenomic marks in identifying cell subtypes.

The BPRMeth model makes a strong smoothness modelling assumption, where neighbouring CpGs have similar methylation levels and the observed differences, originating from technical or biological variability, are attributed to binomial noise in the case of bulk BS-seq experiments. Hence BPRMeth cannot capture drastic shifts or uncoordinated variability of methylation levels between nearby CpGs. To do so, one either should increase the resolution of the inferred profiles or devise a more rigorous method that takes into account these shifts; however, both approaches are prone to overfitting and would require a larger number of observations. In addition, prior to increasing the modelling complexity, one should assess the biological implications (and the frequency) of these methylation shifts on BS-seq data.

Modelling epigenome dynamics from single cell studies

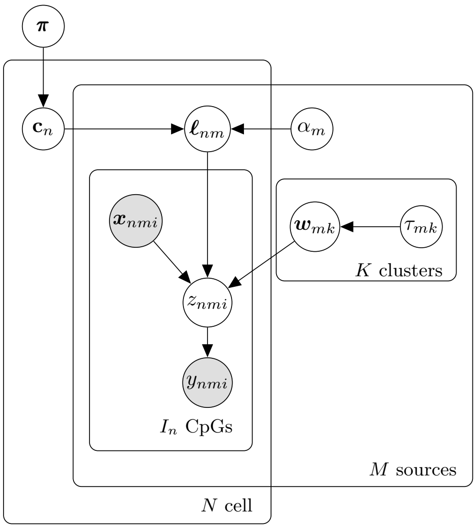
A recent breakthrough in epigenetic research has been the identification of biomarkers that can predict chronological age from DNA methylation, termed 'epigenetic clock' (Horvath, 2013), which is shown to capture aspects of biological age (Marioni et al., 2015). Based on this finding, could we leverage the scBS-seq data to create a pseudo-temporal trajectory of single cells, e.g. during early embryo development? Although the methylation dynamics might not change significantly in these smaller time-scales, information from additional molecular layers, such as scRNA-seq, could be effectively incorporated to perform pseudotime inference, in a similar fashion to the recently proposed MATCHER method (Welch et al., 2017). In single-cell multi-omics studies, the scRNA-seq component alone could be used to create a pseudotime trajectory of single cells, as we described in subsection 5.4.4. The temporal dimension could then be added as an additional layer to the spatial information across the genome, thus jointly modelling the spatio-temporal patterns of DNA methylation. A promising application of this approach would be to identify groups of genes that show coordinated temporal and spatial methylation variability, suggesting that similar regulatory mechanisms affect the epigenome of these genes.

Modelling local heterogeneity of single cell methylomes

In chapter 6 we introduced the Melissa model to capture heterogeneity in the cell population structure. This global clustering of cells however does not allow for variability at the single region level, that is, each cluster has a single methylation profile within each region, and all variability at the region level is attributed to noise. To investigate the local heterogeneity of methylation patterns we need to relax the modelling assumptions of Melissa. A promising approach is for each cell $n \in \{1, \dots, N\}$ to consider each gene / region as a different source $m \in \{1, \dots, M\}$. Subsequently we assume a separate local clustering for each source — encoded through $\boldsymbol{\ell}_{mn} \in \{1, \dots, K+1\}$ — but these source-specific clusterings adhere loosely to an overall consensus methylome clustering of single cells — encoded through $\mathbf{c}_n \in \{1, \dots, K\}$. This approach is based on the Bayesian Consensus Clustering (BCC) model proposed by (Lock and Dunson, 2013) for performing integrative clustering of heterogeneous datasets. Each source has an additional background cluster, denoted by $\mathbf{b}$, which is used to model genes that do not agree with the overall clustering, and the relationship between the source-specific clusterings and overall clustering is given by

$$p(\boldsymbol{\ell}_{mn} \,|\, \mathbf{c}_n, \alpha_m) = \begin{cases} \alpha_m & \text{if } \mathbf{c}_n = \boldsymbol{\ell}_{mn} \\ 1 - \alpha_m & \text{if } \boldsymbol{\ell}_{mn} = \mathbf{b} \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

where $\alpha_m \in [0.5, 1]$ is called the adherence parameter and controls the adherence of each data source to the overall clustering. The probabilistic graphical representation of the extended Melissa model is shown in figure 7.1. The introduction of the two additional latent variables will enable us both to identify genes that shape the cell sub-population structure ($\alpha_m \simeq 1$), and genes that have distinct methylation profiles ($\alpha_m \simeq 0.5$) and do not adhere to the overall structure. An EM strategy is implemented for this model and there is ongoing work for deriving a fast variational inference algorithm and applying it on simulated and real datasets.



$$
\begin{aligned}
\boldsymbol{\pi} &\sim \mathcal{D}irichlet(\boldsymbol{\delta}_0)\\
\alpha_m &\sim \mathcal{TB}eta(\zeta_0, \xi_0)\\
\tau_k &\sim \mathcal{G}amma(\alpha_0, \beta_0)\\
\mathbf{c}_n \,|\, \boldsymbol{\pi} &\sim \mathcal{D}iscrete(\boldsymbol{\pi})\\
\boldsymbol{\ell}_{mn} \,|\, \mathbf{c}_n, \alpha_m &\sim p(\boldsymbol{\ell}_{mn} \,|\, \mathbf{c}_n, \alpha_m)\\
\mathbf{w}_{mk} \,|\, \tau_k &\sim \mathcal{N}(\mathbf{0}, \tau_k^{-1}\mathbf{I})\\
z_{nmi} \,|\, \mathbf{w}_{mk}, \mathbf{x}_{nmi} &\sim \mathcal{N}(\mathbf{w}_{mk}^{\top}\mathbf{x}_{nmi}, 1)\\
y_{nmi} \,|\, z_{nmi} &= \begin{cases} 1 & \text{if } z_{nmi} > 0 \\ 0 & \text{if } z_{nmi} \leq 0 \end{cases}
\end{aligned}
$$

Figure 7.1 Probabilistic graphical representation of extended Melissa model.

Integrative modelling of multi-omics data

It is increasingly clear that biological processes are highly coordinated both in space and time by a complex and still incompletely understood interaction network of regulatory molecular layers. Most computational studies, including contributions of this thesis, rely on developing interpretable models for each molecular layer independently, and post-hoc identifying dependencies between different layers. Examples of this approach include associating transcript abundance with DNA methylation (Kapourani and Sanguinetti, 2016), histone modifications (Karlic et al., 2010) or chromatin accessibility heterogeneity (Clark et al., 2018), predicting epigenetic marks from DNA sequence motifs (Whitaker et al., 2015), and predicting histone modifications from transcription factor binding patterns (Benveniste et al., 2014). However, effective integrative models that have the potential to directly or indirectly capture the coordinated variability and dynamic coupling between different layers are still under-represented in the literature.

The joint integration of molecular profiles from different (although potentially dependent) modalities will enable us to define cellular identity as a superposition of *basis vectors*, each determining a different aspect of cellular organisation and function (Wagner et al., 2016). The advent of single cell multi-omics platforms will potentially provide the discriminative power to identify rare cell subtypes, that cannot be uncovered when sequencing each layer independently, and derive molecular mechanisms from cellular heterogeneity at different layers. In addition to the computational complexity, incorporation of domain knowledge for data integration is crucial, since the variability in one 'omics' dimension might act as a confounder for another and the dynamics of the multiple 'omics' layers might operate on different time scales (Kelsey et al., 2017). In conclusion, integration of heterogeneous biological data via interpretable statistical models is essential for analysing biological systems and furthering our understanding of epigenetic control of gene expression and cell identity.

Appendix A

Supplementary information for BPRMeth model

A.1 Mean field variational inference derivation

The joint distribution over all variables for the augmented BPRMeth model is

$$p(\mathbf{y}, \mathbf{z}, \mathbf{w}, \tau \,|\, \mathbf{X}) = p(\mathbf{y} \,|\, \mathbf{z})\, p(\mathbf{z} \,|\, \mathbf{w}, \mathbf{X})\, p(\mathbf{w} \,|\, \tau)\, p(\tau), \tag{A.1}$$

where the factorisation corresponds to the probabilistic graphical model shown in figure 5.4. The mean field approximates the posterior distribution by assuming that the variational distribution factorises over the variables

$$q(\mathbf{z}, \mathbf{w}, \tau) = q(\mathbf{z})\, q(\mathbf{w})\, q(\tau) \simeq p(\mathbf{z}, \mathbf{w}, \tau \,|\, \mathbf{y}, \mathbf{X}). \tag{A.2}$$

In addition to this factorisation, we have an *induced factorisation*: the latent variables z_i are independent given the observations y_i and regression coefficients $\mathbf{w}_i$, hence,

$$q(\mathbf{z}, \mathbf{w}, \tau) = q(\mathbf{w})\, q(\tau) \prod_i q(z_i). \tag{A.3}$$

A.1.1 Deriving optimised factors

Below we derive the optimised factors of the variational posterior using equation (3.36).

Factor $q(\mathbf{z})$: The logarithm of the optimised factor $q(z_i)$ assuming that the corresponding $y_i = 1$ is given by

$$\begin{aligned}
\log q^*(z_i) &= \left\langle \log \left| p(\mathbf{y} \,|\, \mathbf{z})\, p(\mathbf{z} \,|\, \mathbf{w}, \mathbf{X})\, \underbrace{p(\mathbf{w} \,|\, \tau)\, p(\tau)}_{\text{const}} \right| \right\rangle_{q(\mathbf{w}, \tau)} + \text{const} \\
&= \log p(y_i \,|\, z_i) + \left\langle \log \mathcal{N}(z_i \,|\, \mathbf{w}^\top \mathbf{x}_i, 1) \right\rangle_{q(\mathbf{w})} + \text{const} \\
&= y_i \log \mathbb{I}(z_i > 0) + \underbrace{(1 - y_i) \log \mathbb{I}(z_i \leq 0)}_{0,\ \text{since}\ y_i = 1} - \frac{1}{2} \left\langle \left(z_i - \mathbf{w}^\top \mathbf{x}_i\right)^2 \right\rangle_{q(\mathbf{w})} + \text{const} \\
&= y_i \log \mathbb{I}(z_i > 0) - \frac{1}{2} z_i^2 + z_i \left\langle \mathbf{w}^\top \right\rangle_{q(\mathbf{w})} \mathbf{x}_i + \text{const.}
\end{aligned} \tag{A.4}$$

We complete the square using equation (D.4) and by exponentiating this quantity and setting $\mu_i = \left\langle \mathbf{w}^\top \right\rangle_{q(\mathbf{w})} \mathbf{x}_i$ we obtain

$$q^*(z_i) \propto \mathbb{I}(z_i > 0) \exp\left(-\frac{1}{2} z_i^2 + z_i \mu_i\right). \tag{A.5}$$

We observe that the optimized factor $q(z_i)$ is an un-normalised truncated Gaussian distribution

$$q^*(z_i) = \begin{cases} \mathcal{TN}_+\left(z_i \mid \mu_i, 1\right) & \text{if } y_i = 1 \\ \mathcal{TN}_-\left(z_i \mid \mu_i, 1\right) & \text{if } y_i = 0 \end{cases}. \tag{A.6}$$

Factor $q(\tau)$: The logarithm of the optimised factor is given by

$$\begin{aligned} \log q^*(\tau) &= \left\langle \log \Big| \underbrace{p(\mathbf{y} \mid \mathbf{z})\, p(\mathbf{z} \mid \mathbf{w}, \mathbf{X})}_{\text{const}}\, p(\mathbf{w} \mid \tau)\, p(\tau) \Big| \right\rangle_{q(\mathbf{z}, \mathbf{w})} + \text{const} \\ &= \langle \log p(\mathbf{w} \mid \tau) \rangle_{q(\mathbf{w})} + \log p(\tau) + \text{const} \\ &= \underbrace{\frac{D}{2} \log \tau - \frac{\tau}{2} \left\langle \mathbf{w}^\top \mathbf{w} \right\rangle_{q(\mathbf{w})}}_{\text{Gaussian PDF}} + \underbrace{(\alpha_0 - 1) \log \tau - \beta_0 \tau}_{\text{Gamma PDF}} \\ &= \underbrace{(\alpha_0 + \frac{D}{2} - 1)}_{\alpha \text{ parameter}} \log \tau - \underbrace{\left(\beta_0 + \frac{1}{2} \left\langle \mathbf{w}^\top \mathbf{w} \right\rangle_{q(\mathbf{w})}\right)}_{\beta \text{ parameter}} \tau, \end{aligned} \tag{A.7}$$

which is the logarithm of the un-normalised Gamma distribution, leading to

$$\begin{aligned} q^*(\tau) &= \mathcal{G}\text{amma}(\tau \mid \alpha, \beta), \\ \alpha &= \alpha_0 + \frac{D}{2}, \\ \beta &= \beta_0 + \frac{1}{2} \left\langle \mathbf{w}^\top \mathbf{w} \right\rangle_{q(\mathbf{w})}. \end{aligned} \tag{A.8}$$

Factor $q(\mathbf{w})$: The logarithm of the optimised factor is given by

$$\begin{aligned} \log q^*(\mathbf{w}) &= \left\langle \log \Big| \underbrace{p(\mathbf{y} \mid \mathbf{z})}_{\text{const}}\, p(\mathbf{z} \mid \mathbf{w}, \mathbf{X})\, p(\mathbf{w} \mid \tau)\, \underbrace{p(\tau)}_{\text{const}} \Big| \right\rangle_{q(\mathbf{z}, \tau)} + \text{const} \\ &= \langle \log \mathcal{N}(\mathbf{z} \mid \mathbf{X}\mathbf{w}, \mathbf{I}) \rangle_{q(\mathbf{z})} + \langle \log p(\mathbf{w} \mid \tau) \rangle_{q(\tau)} + \text{const} \\ &= \left\langle -\frac{1}{2} \Big(\mathbf{z} - \mathbf{X}\mathbf{w}\Big)^\top \Big(\mathbf{z} - \mathbf{X}\mathbf{w}\Big) \right\rangle_{q(\mathbf{z})} - \frac{1}{2} \langle \tau \rangle_{q(\tau)} \mathbf{w}^\top \mathbf{w} + \text{const} \\ &= \mathbf{w}^\top \mathbf{X}^\top \langle \mathbf{z} \rangle_{q(\mathbf{z})} - \frac{1}{2} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} - \frac{1}{2} \langle \tau \rangle_{q(\tau)} \mathbf{w}^\top \mathbf{w} + \text{const} \\ &= \mathbf{w}^\top \mathbf{X}^\top \langle \mathbf{z} \rangle_{q(\mathbf{z})} - \frac{1}{2} \mathbf{w}^\top \left(\langle \tau \rangle_{q(\tau)} \mathbf{I} + \mathbf{X}^\top \mathbf{X}\right) \mathbf{w} + \text{const}. \end{aligned} \tag{A.9}$$

Because this is a quadratic form, the distribution $q^*(\mathbf{w})$ is a Gaussian distribution and we again complete the square to identify the mean and the covariance matrix

$$\begin{aligned} q^*(\mathbf{w}) &= \mathcal{N}(\mathbf{w} \mid \mathbf{m}, \mathbf{S}) \\ \mathbf{m} &= \mathbf{S}\mathbf{X}^\top \langle \mathbf{z} \rangle_{q(\mathbf{z})} \\ \mathbf{S} &= \left(\langle \tau \rangle_{q(\tau)} \mathbf{I} + \mathbf{X}^\top \mathbf{X}\right)^{-1} \end{aligned} \tag{A.10}$$

A.1.2 Computing expectations

The factor $q(\mathbf{w})$ is a Gaussian distribution $\mathcal{N}(\mathbf{w}\,|\,\mathbf{m},\mathbf{S})$, hence its expected value is $\langle \mathbf{w}\rangle_{q(\mathbf{w})} = \mathbf{m}$. The factor $q(\tau)$ is a Gamma distribution $\mathcal{G}\mathrm{amma}(\tau\,|\,\alpha,\beta)$, hence its expected value is $\langle \tau\rangle_{q(\tau)} = \frac{\alpha}{\beta}$. The factor $q(\mathbf{w})$ is a Gaussian distribution $\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S})$ hence we have

$$\left\langle \mathbf{w}^\top\mathbf{w}\right\rangle_{q(\mathbf{w})} = \mathrm{tr}\left(\left\langle \mathbf{w}^\top\mathbf{w}\right\rangle_{q(\mathbf{w})}\right) = \mathrm{tr}\left(\mathbf{m}\mathbf{m}^\top + \mathbf{S}\right) = \mathbf{m}^T\mathbf{m} + \mathrm{tr}(\mathbf{S}).$$

The factor $q(z_i)$ is a truncated Gaussian distribution, hence, form standard results its expected value is given by

$$\langle z_i\rangle_{q(z_i)} = \begin{cases} \mu_i + \phi_i/(1-\Phi_i) & \text{if } y_i = 1 \\ \mu_i - \phi_i/\Phi_i & \text{if } y_i = 0 \end{cases}.$$

where $\phi_i \overset{\mathsf{def}}{=} \phi(-\mu_i)$ and $\Phi_i \overset{\mathsf{def}}{=} \Phi(-\mu_i)$.

A.1.3 Evidence lower bound

The evidence lower bound (ELBO) is given by

$$\begin{aligned}
\mathcal{L}\left(q(\mathbf{z},\mathbf{w},\tau)\right) &= \int q(\mathbf{z},\mathbf{w},\tau)\log\left|\frac{p(\mathbf{y},\mathbf{z},\mathbf{w},\tau\,|\,\mathbf{X})}{q(\mathbf{z},\mathbf{w},\tau)}\right| d\mathbf{z}\,d\mathbf{w}\,d\tau \\
&= \langle \log p(\mathbf{y}\,|\,\mathbf{z})\rangle_{q(\mathbf{z})} + \langle \log p(\mathbf{z}\,|\,\mathbf{w},\mathbf{X})\rangle_{q(\mathbf{z},\mathbf{w})} + \langle \log p(\mathbf{w}\,|\,\tau)\rangle_{q(\mathbf{w},\tau)} + \langle \log p(\tau)\rangle_{q(\tau)} \\
&\quad - \langle \log q(\mathbf{z})\rangle_{q(\mathbf{z})} - \langle \log q(\mathbf{w})\rangle_{q(\mathbf{w})} - \langle \log q(\tau)\rangle_{q(\tau)}\,.
\end{aligned} \tag{A.11}$$

Notice that the terms involving expectations of $\log q(\cdot)$ simply represent the negative entropies H of those distributions. The various terms in the ELBO are evaluated as follows

$$\begin{aligned}
\langle \log p(\mathbf{y}\,|\,\mathbf{z})\rangle_{q(\mathbf{z})} &= \sum_i^I \langle y_i \log \mathbb{I}(z_i > 0) + (1-y_i)\log\mathbb{I}(z_i \le 0)\rangle_{q(z_i)} \\
&= \sum_i^I \begin{cases} \int_{-\infty}^{\infty} q(z_i)\log\mathbb{I}(z_i > 0)\,dz_i & \text{if } y_i = 1 \\ \int_{-\infty}^{\infty} q(z_i)\log\mathbb{I}(z_i \le 0)\,dz_i & \text{if } y_i = 0 \end{cases} = 0, \\
\langle \log p(\mathbf{z}\,|\,\mathbf{w},\mathbf{X})\rangle_{q(\mathbf{z},\mathbf{w})} &= \langle \log \mathcal{N}(\mathbf{z}\,|\,\mathbf{X}\mathbf{w},\mathbf{I})\rangle_{q(\mathbf{z},\mathbf{w})} \\
&= -\frac{I}{2}\log 2\pi - \frac{1}{2}\left\langle \mathbf{z}^\top\mathbf{z}\right\rangle_{q(\mathbf{z})} + \left\langle \boldsymbol{\mu}^\top\mathbf{z}\right\rangle_{q(\mathbf{z})} - \frac{1}{2}\mathrm{tr}\left(\mathbf{X}^\top\mathbf{X}\left(\mathbf{m}\mathbf{m}^\top + \mathbf{S}\right)\right), \\
\langle \log p(\mathbf{w}\,|\,\tau)\rangle_{q(\mathbf{w},\tau)} &= \left\langle -\frac{D}{2}\log 2\pi - \frac{D}{2}\log\tau^{-1} - \frac{1}{2\tau^{-1}}\mathbf{w}^\top\mathbf{w}\right\rangle_{q(\mathbf{w},\tau)} \\
&= -\frac{D}{2}\log 2\pi + \frac{D}{2}\left(\psi(\alpha) - \log\beta\right) - \frac{\alpha}{2\beta}\left(\mathbf{m}^\top\mathbf{m} + \mathrm{tr}(\mathbf{S})\right), \\
\langle \log p(\tau)\rangle_{q(\tau)} &= \alpha_0\log\beta_0 + (\alpha_0 - 1)\left(\psi(\alpha) - \log\beta\right) - \beta_0\frac{\alpha}{\beta} - \log\Gamma(\alpha_0), \\
\langle \log q(\mathbf{z})\rangle_{q(\mathbf{z})} &= \sum_i^I \left\langle \log\left(\mathcal{TN}_+(z_i\,|\,\mu_i,1)^{y_i}\mathcal{TN}_-(z_i\,|\,\mu_i,1)^{(1-y_i)}\right)\right\rangle_{q(z_i)} \\
&= -\frac{I}{2}\log 2\pi - \frac{1}{2}\left\langle \mathbf{z}^\top\mathbf{z}\right\rangle_{q(\mathbf{z})} + \left\langle \boldsymbol{\mu}^\top\mathbf{z}\right\rangle_{q(\mathbf{z})} - \frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu} - \\
&\qquad \sum_{i=1}^I \Big\{ y_i\log\left(1-\Phi_i\right) + (1-y_i)\log\left(\Phi_i\right) \Big\}, \\
\langle \log q(\mathbf{w})\rangle_{q(\mathbf{w})} &= -\frac{1}{2}\log|\mathbf{S}| - \frac{D}{2}(1 + \log 2\pi), \\
\langle \log q(\tau)\rangle_{q(\tau)} &= -\log\Gamma(\alpha) + (\alpha-1)\psi(\alpha) + \log\beta - \alpha.
\end{aligned} \tag{A.12}$$

A.2 Supplementary figures for augmented BPRMeth

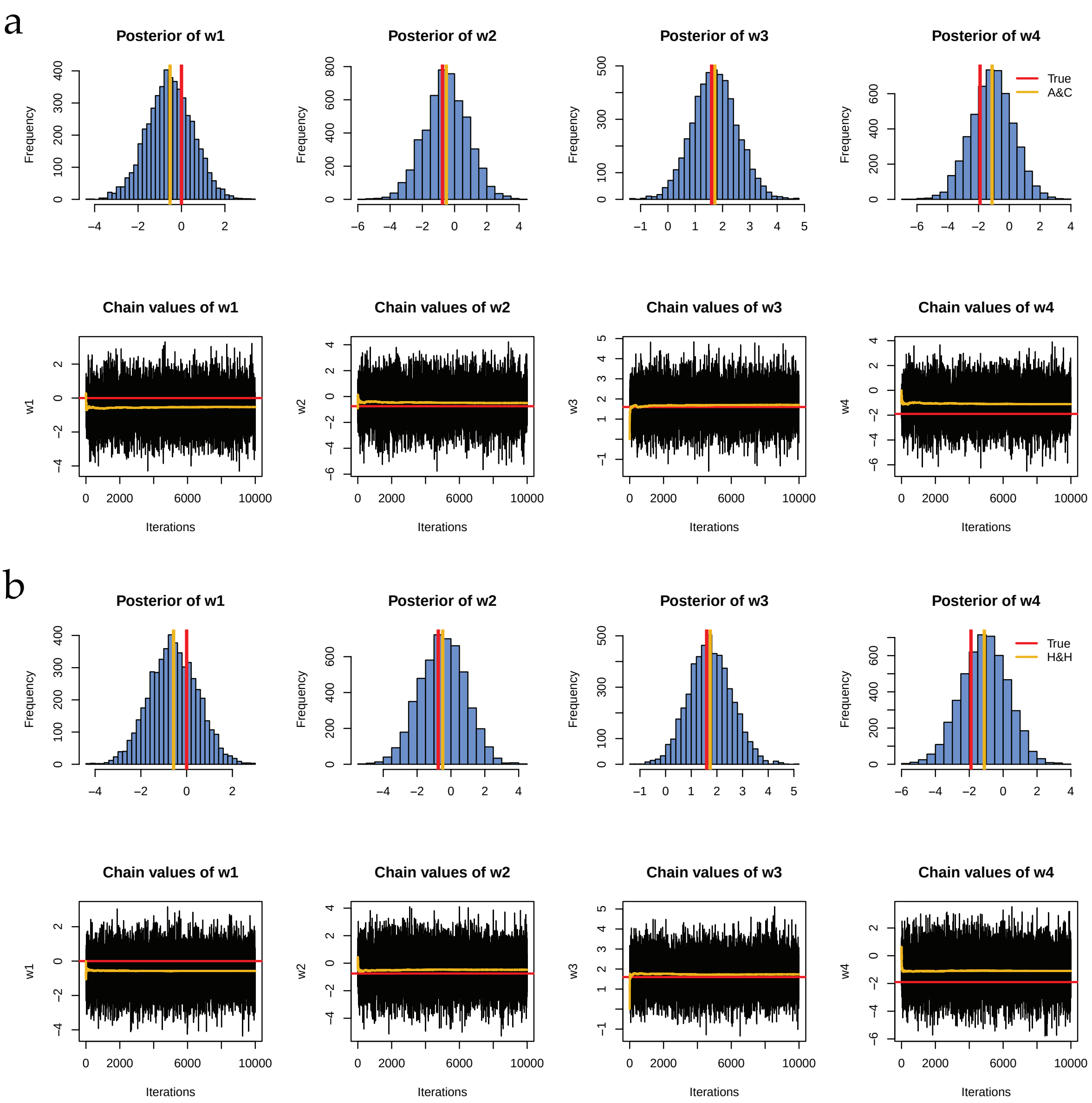


Figure A.1 Marginal density and trace plots for each inferred parameter using the Gibbs sampling algorithm. The trace plot depicts the iteration number on the x-axis, and the value of the draw at each iteration on the y-axis. (a) Albert and Chib (1993) algorithm implementation for augmented BPRMeth model and (b) Holmes and Held (2006) extension using a joint update scheme.

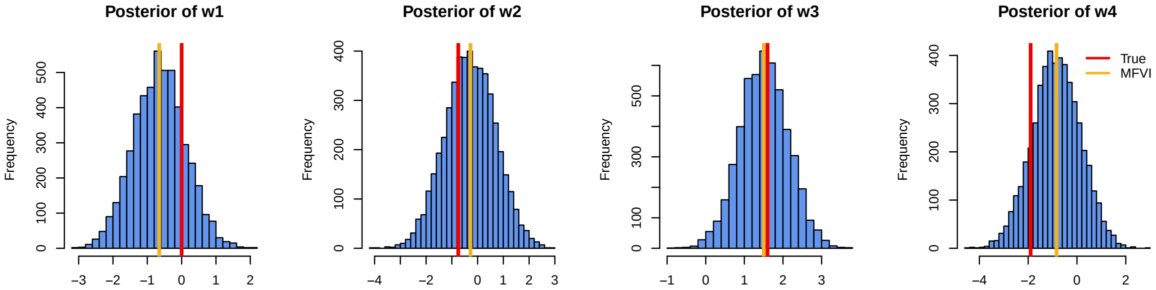
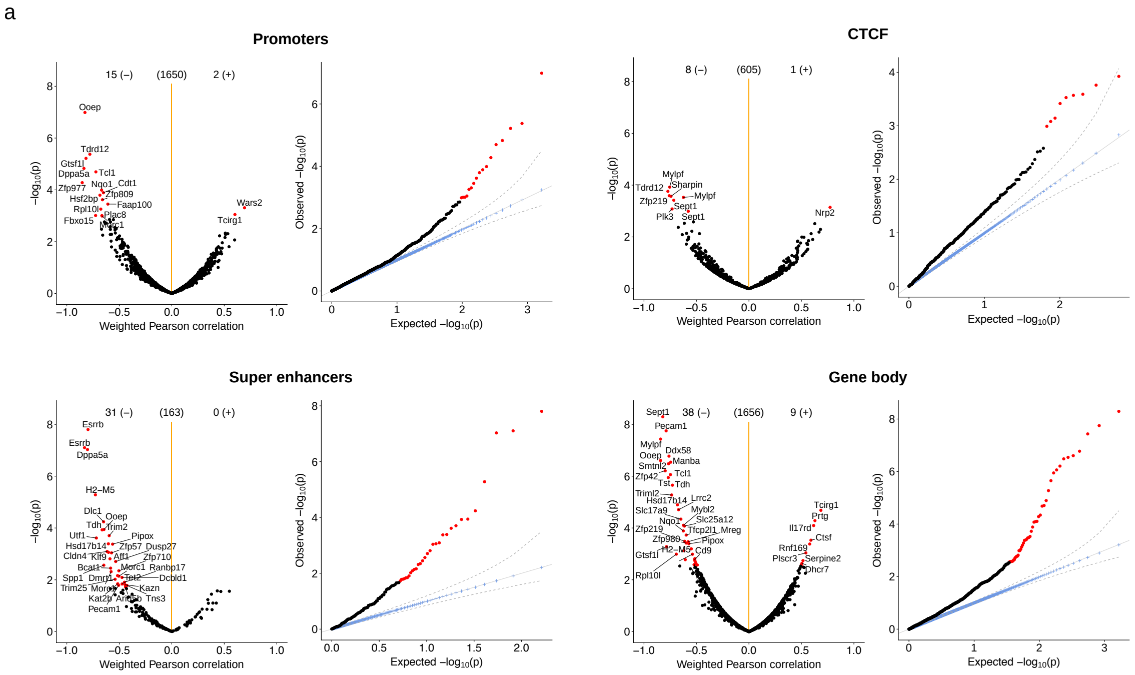


Figure A.2 Marginal density plots for each inferred parameter using the mean field variational inference algorithm. Shown are the actual parameters that generated the data (red colour), the posterior mean (yellow colour), and 5,000 samples from the inferred posterior distribution depicted as histograms.

Appendix B

Supplementary information for scNMT-seq study



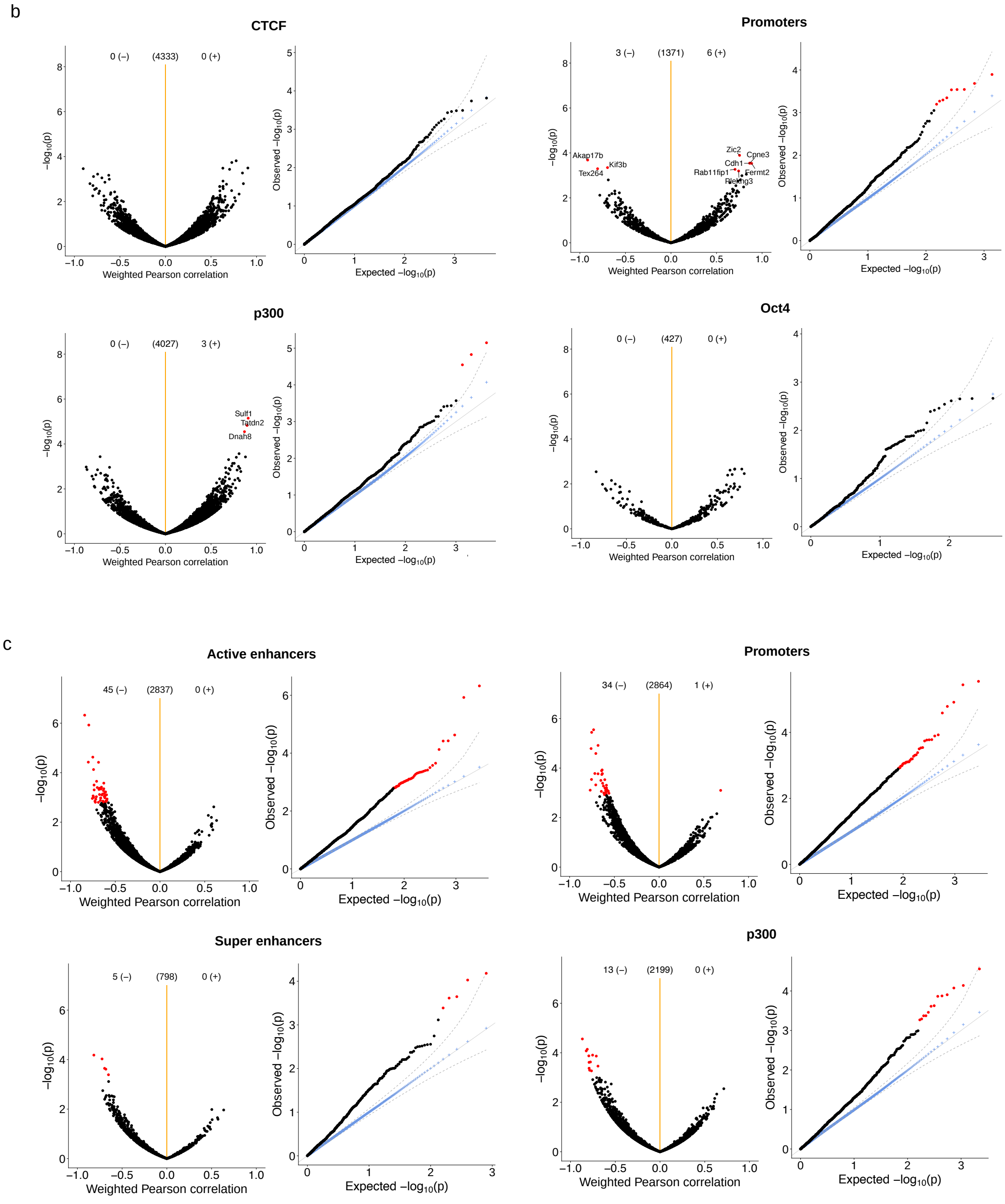


Figure B.1 Association tests between molecular layers in selected genomic contexts. Shown are correlation analysis across cells (one test per feature) between (a) methylation and expression, (b) accessibility and expression and (c) methylation and accessibility. Volcano plots display the Pearson correlation coefficients r and adjusted p-values using the Benjamini-Hochberg procedure. The orange vertical line indicates the position of $r = 0$. Red dots denote features that pass threshold of statistical significance (adjusted p-value < 0.01). Q-Q plots show the distribution of observed p-values (black and red dots), the uniform distribution (grey lines, with solid line showing the mean and the dashed line showing the 95% confidence interval), and p-values obtained after 100 permutations of both features and samples (blue crosses). Joint analysis with Stephen J. Clark and Ricard Argelaguet.

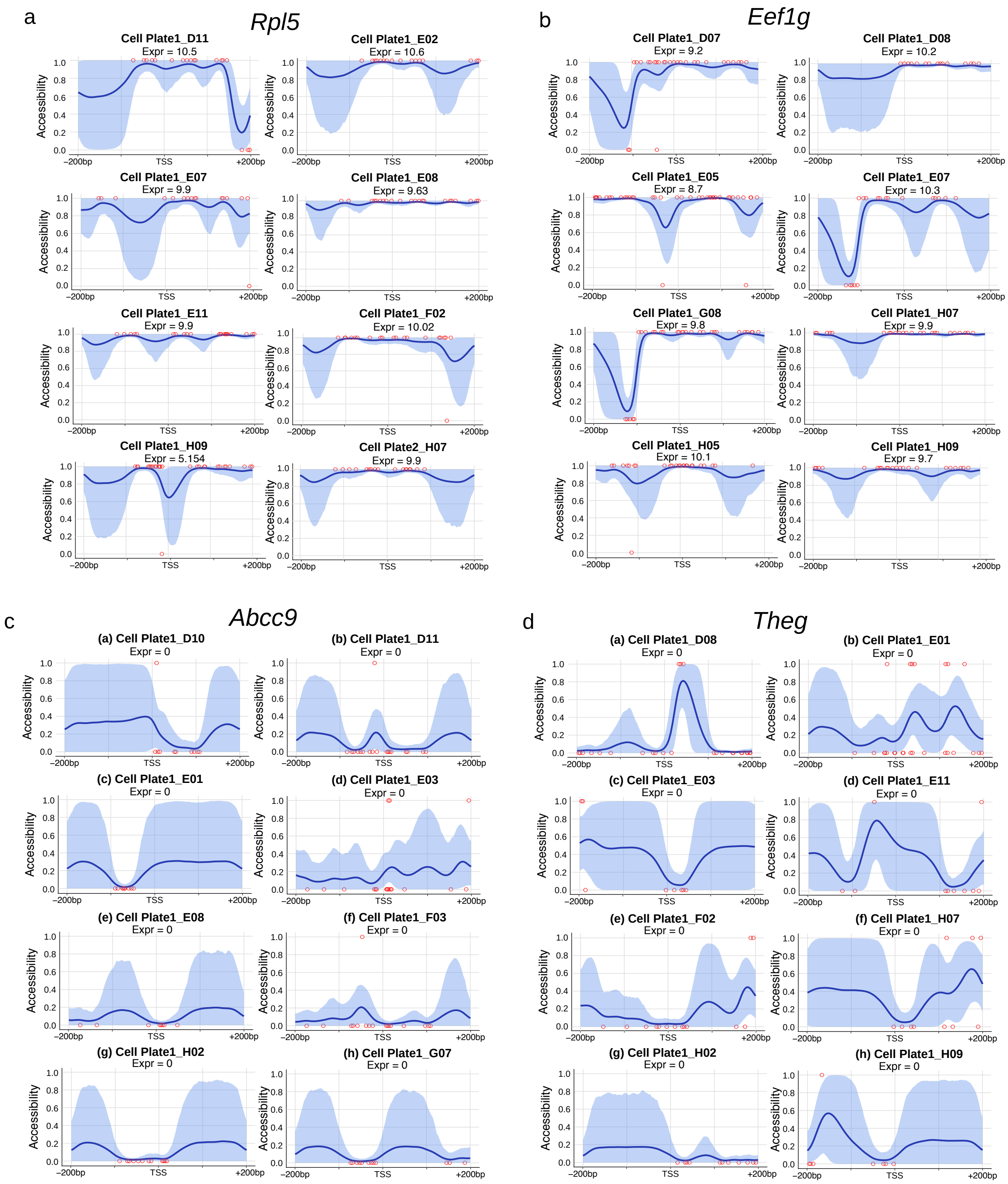


Figure B.2 Inferred single-cell accessibility profiles from genes with different gene expression regimes. Shown are profiles of representative cells for highly accessible and expressed housekeeping genes: (a) *Rpl5* and (b) *Eef1g*, and for non-accessible and non-expressed genes: (c) *Abcc9* and (d) *Theg*. Each red dot represents a GpC site, with binary accessibility value (1 = accessible, 0 = inaccessible).

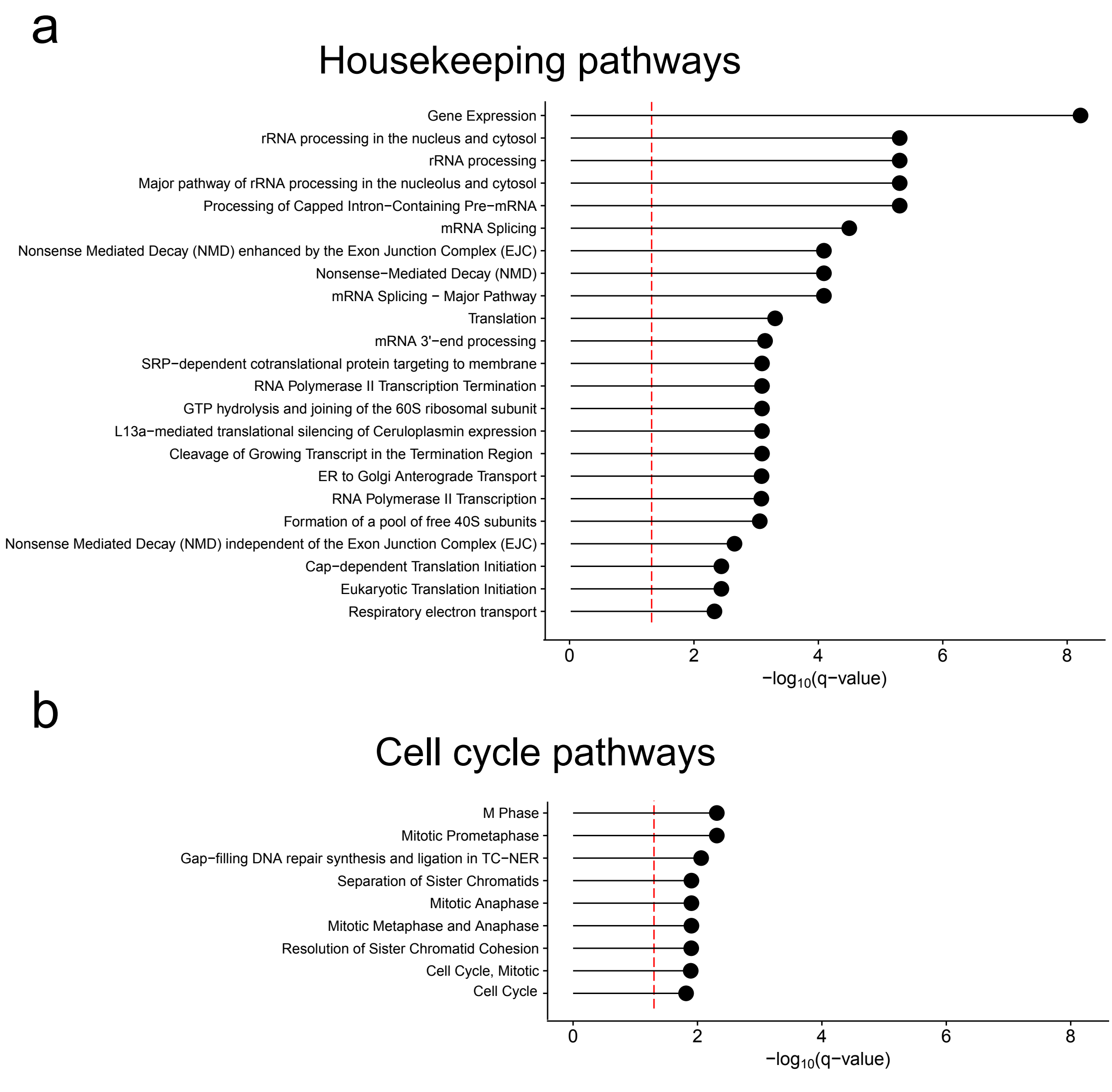


Figure B.3 Extensive list of significantly enriched gene ontology terms, using Fisher's exact test, for genes with the most homogeneous accessibility profiles (K = 1). The p-values are adjusted for multiple testing using the Benjamini-Hochberg procedure. Joint analysis with Ricard Argelaguet.

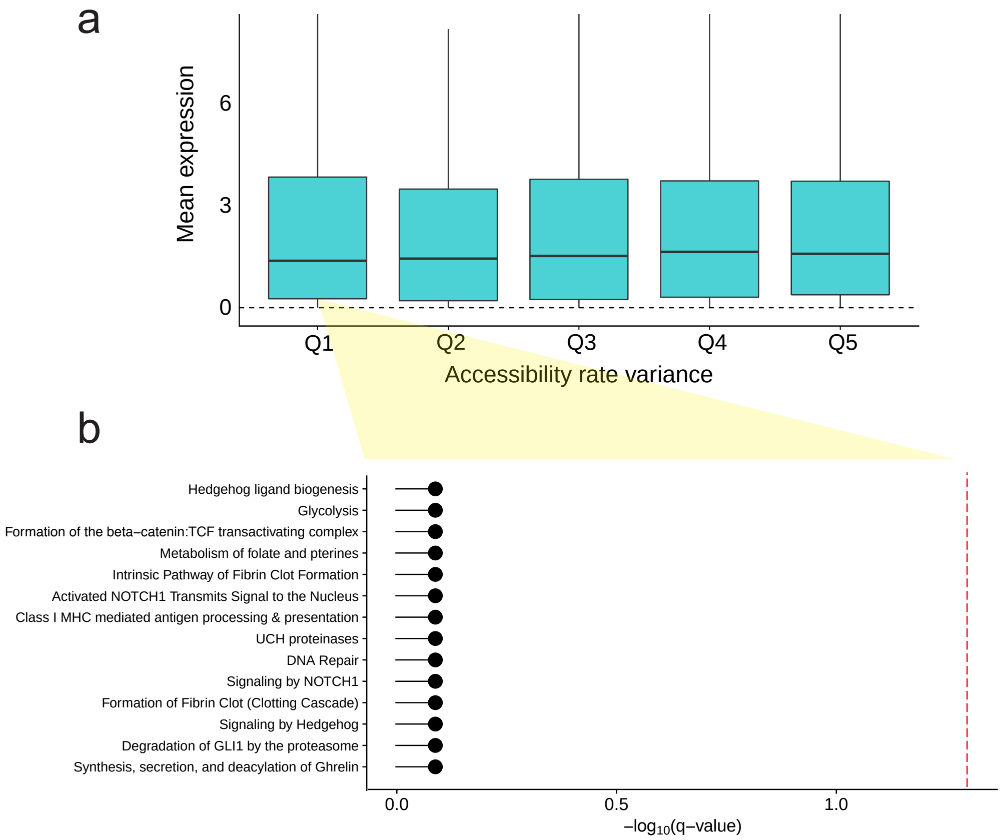


Figure B.4 Quantifying the variability of chromatin accessibility using conventional rates. (a) There is no association between variability in accessibility rate and gene expression; and (b) there is no significant enrichment of gene ontology terms for genes associated with the most conserved promoters.

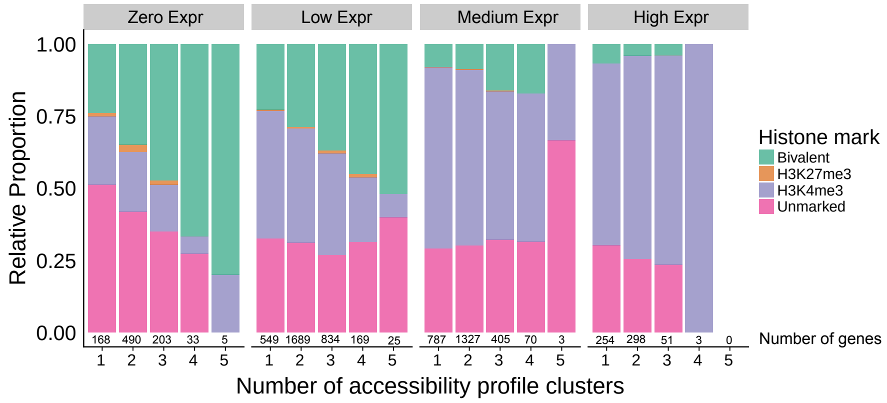


Figure B.5 Presence of bivalent histone marks (H3K4me3 and H3K27me3) is associated with high cell-to-cell variability in accessibility profiles. Genes with one cluster (K = 1) correspond to a more homogeneous chromatin pattern than genes with multiple clusters. The results are overlapped with ChIP-seq histone marks data. The x-axis denotes the number of clusters (i.e. heterogeneity), and the y-axis displays the relative proportion of each histone mark. To account for differences in mean expression levels, genes are split in four different expression groups (‘Zero Expr’ for average log normalised counts equal to 0, ‘Low Expr’ between 0 and 2, ‘Medium Expr’ between 2 and 6 and ‘High Expr’ higher than 6).

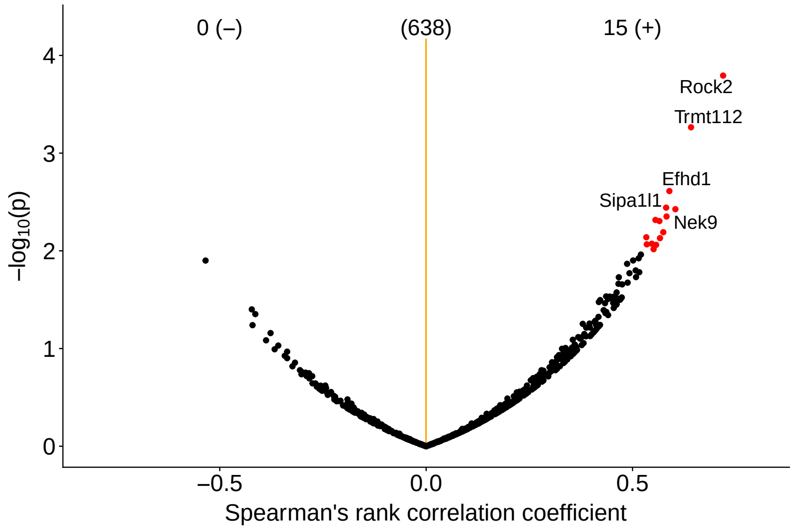


Figure B.6 Association analysis between promoter accessibility profile and development trajectory. For each gene, the cluster assignments are associated with the cellular position in the differentiation trajectory. Shown is a volcano plot of Spearman's rank coefficient in the x-axis with the corresponding log-transformed p-values in the y-axis. Red dots denote genes that pass statistical significance threshold (p-value < 0.01).

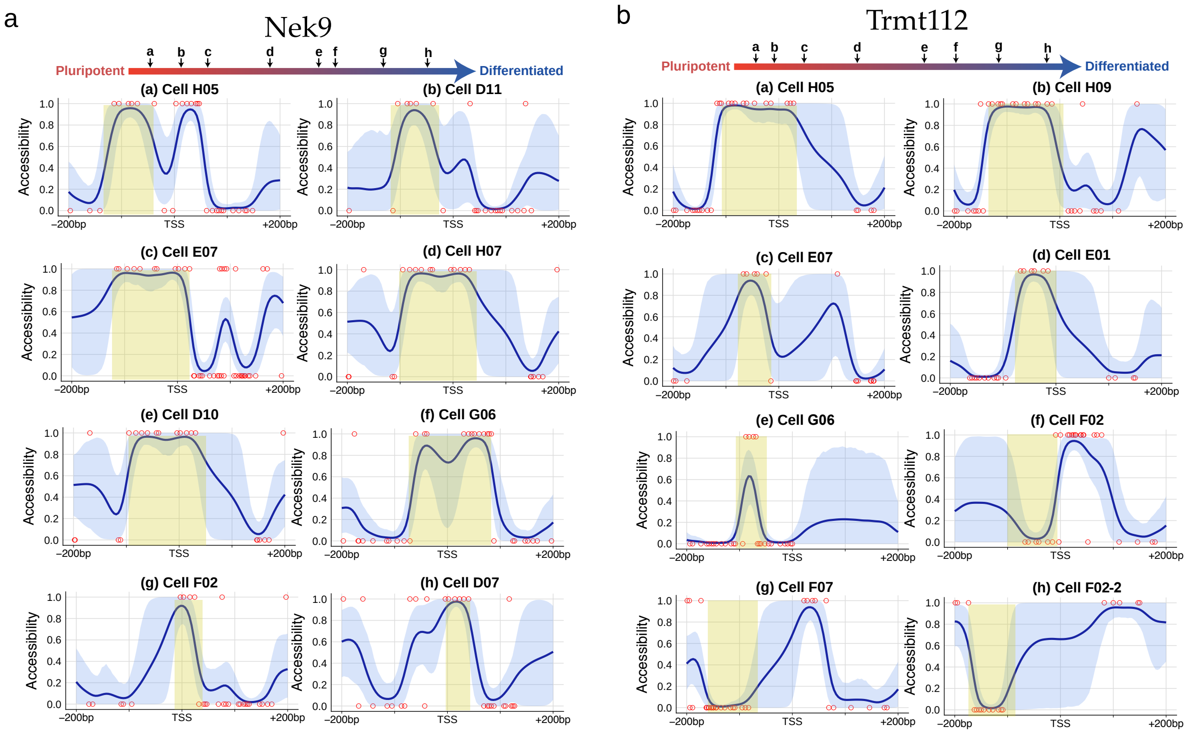


Figure B.7 Dynamics of variation in chromatin accessibility profiles along the developmental trajectory. Shown are profiles of representative cells for the (a) *Nek9* and (b) *Trmt112* genes. Each red dot represents a GpC site, with binary accessibility value (1 = accessible, 0 = inaccessible). Yellow shading is used to highlight the relevant region of dynamic changes. Joint analysis with Ricard Argelaguet.

Appendix C

Supplementary information for Melissa model

C.1 Mean field variational inference derivation

The joint distribution over the observed and latent variables is

$$p(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau} \,|\, \mathbf{X}) = p(\mathbf{Y} \,|\, \mathbf{Z})\, p(\mathbf{Z} \,|\, \mathbf{C}, \mathbf{W}, \mathbf{X})\, p(\mathbf{C} \,|\, \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\mathbf{W} \,|\, \boldsymbol{\tau})\, p(\boldsymbol{\tau}), \tag{C.1}$$

where the factorisation corresponds to the probabilistic graphical model shown in figure 6.2. We assume that the variational approximation to our posterior distribution factorises over the latent variables (mean-field variational inference)

$$q(\mathbf{Z}, \mathbf{C}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau}) = q(\mathbf{Z})\, q(\mathbf{C})\, q(\mathbf{W})\, q(\boldsymbol{\pi})\, q(\boldsymbol{\tau}). \tag{C.2}$$

C.1.1 Deriving optimised factors

Below we derive the optimised factors of the variational posterior using equation (3.36).

Factor $q(\mathbf{C})$ The logarithm of the optimised factor $q(\mathbf{C})$ is given by

$$\begin{aligned} \log q^*(\mathbf{C}) &= \left\langle \log \left| \underbrace{p(\mathbf{Y} \,|\, \mathbf{Z})}_{\text{const}}\, p(\mathbf{Z} \,|\, \mathbf{C}, \mathbf{W}, \mathbf{X})\, p(\mathbf{C} \,|\, \boldsymbol{\pi})\, \underbrace{p(\boldsymbol{\pi})\, p(\mathbf{W} \,|\, \boldsymbol{\tau})\, p(\boldsymbol{\tau})}_{\text{const}} \right| \right\rangle_{q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\pi}, \boldsymbol{\tau})} + \text{const} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} c_{nk} \, \log \rho_{nk} + \text{const}, \end{aligned} \tag{C.3}$$

where

$$\log \rho_{nk} = \sum_{m=1}^{M} \left\langle -\frac{1}{2} \Big(\mathbf{z}_{nm} - \mathbf{X}_{nm} \mathbf{w}_{mk} \Big)^T \Big(\mathbf{z}_{nm} - \mathbf{X}_{nm} \mathbf{w}_{mk} \Big) \right\rangle_{q(\mathbf{z}_{nm}, \mathbf{w}_{mk})} + \langle \log \pi_k \rangle_{q(\pi_k)} . \tag{C.4}$$

Taking the exponential on both sides and requiring that this distribution be normalised we obtain

$$q^*(\mathbf{C}) = \prod_{n=1}^{N} \prod_{k=1}^{K} r_{nk}^{c_{nk}}, \qquad \text{where} \quad r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K} \rho_{nj}}. \tag{C.5}$$

Factor $q(\boldsymbol{\tau})$ The logarithm of the optimised factor $q(\boldsymbol{\tau})$ is given by

$$\begin{aligned}
\log q^*(\boldsymbol{\tau}) &= \left\langle \log \left| \underbrace{p(\mathbf{Y}\,|\,\mathbf{Z})\,p(\mathbf{Z}\,|\,\mathbf{C},\mathbf{W},\mathbf{X})\,p(\mathbf{C}\,|\,\boldsymbol{\pi})\,p(\boldsymbol{\pi})}_{\text{const}}\,p(\mathbf{W}\,|\,\boldsymbol{\tau})\,p(\boldsymbol{\tau}) \right| \right\rangle_{q(\mathbf{Z},\mathbf{C},\mathbf{W},\boldsymbol{\pi})} + \text{const} \\
&= \langle \log p(\mathbf{W}\,|\,\boldsymbol{\tau}) \rangle_{q(\mathbf{W})} + \log p(\boldsymbol{\tau}) + \text{const} \qquad \text{(C.6)} \\
&= \sum_{k=1}^{K}\sum_{m=1}^{M} \langle \log p(\mathbf{w}_{mk}\,|\,\tau_k) \rangle_{q(\mathbf{w}_{mk})} + \sum_{k=1}^{K} \log p(\tau_k) + \text{const.}
\end{aligned}$$

Here we observe that the right hand side comprises a sum over k, i.e. each τ_k is independent of each other, hence

$$\begin{aligned}
\log q^*(\tau_k) &= \sum_{m=1}^{M} \langle \log p(\mathbf{w}_{mk}\,|\,\tau_k) \rangle_{q(\mathbf{w}_{mk})} + \log p(\tau_k) + \text{const} \\
&= \underbrace{\frac{MD}{2}\log\tau_k - \frac{\tau_k}{2}\sum_{m=1}^{M}\left\langle \mathbf{w}_{mk}^T\mathbf{w}_{mk} \right\rangle_{q(\mathbf{w}_{mk})}}_{\text{Gaussian PDF}} + \underbrace{(\alpha_0 - 1)\log\tau_k - \beta_0\tau_k}_{\text{Gamma PDF}} \qquad \text{(C.7)} \\
&= \underbrace{(\alpha_0 + \frac{MD}{2} - 1)}_{\alpha_k \text{ parameter}}\log\tau_k - \underbrace{\left(\beta_0 + \frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{w}_{mk}^T\mathbf{w}_{mk} \right\rangle_{q(\mathbf{w}_{mk})}\right)}_{\beta_k \text{ parameter}}\tau_k,
\end{aligned}$$

which is the logarithm of the (un-normalised) Gamma distribution, leading to

$$\begin{aligned}
q^*(\tau_k) &= \mathcal{G}\text{amma}(\tau_k\,|\,\alpha_k,\beta_k) \\
\alpha_k &= \alpha_0 + \frac{MD}{2} \qquad \text{(C.8)} \\
\beta_k &= \beta_0 + \frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{w}_{mk}^T\mathbf{w}_{mk} \right\rangle_{q(\mathbf{w}_{mk})}.
\end{aligned}$$

Note that the update for the α hyperparameter depends only on the total number of genomic regions and the number of basis functions used to estimate the underlying methylation profiles. On the other hand the β hyperparameter is updated at each CAVI iteration, since it depends on the expected value of the regression coefficients. The expected value of the $\mathcal{G}$amma distribution is $E = \alpha/\beta$, and the inverse of this quantity is the variance parameter for the prior Gaussian distribution of the coefficients $\mathbf{w}$. Large values of E result in small variance Gaussian priors, hence the model is substantially penalised when weights are moving away from prior mean $\mu_0 = 0$; as a consequence the model will tend to prune away clusters, that is, set all weights $\mathbf{w}_{mk} = \mathbf{0}$. This may strongly affect the model in the initial iterations of CAVI, which will affect the β parameter but not the α parameter of the $\mathcal{G}$amma distribution, potentially leading to convergence to a suboptimal local maximum. Hence, one should be cautious when setting the initial values for these parameters; in the current implementation of Melissa we set $\alpha_0 = 0.5$ and $\beta_0 = \sqrt{a_k}$.

Factor $q(\boldsymbol{\pi})$ The logarithm of the optimised factor is given by

$$\begin{aligned}\log q^*(\boldsymbol{\pi}) &= \left\langle \log \left| \underbrace{p(\mathbf{Y}\,|\,\mathbf{Z})\, p(\mathbf{Z}\,|\,\mathbf{C}, \mathbf{W}, \mathbf{X})}_{\text{const}}\; p(\mathbf{C}\,|\,\boldsymbol{\pi})\, p(\boldsymbol{\pi})\; \underbrace{p(\mathbf{W}\,|\,\boldsymbol{\tau})\, p(\boldsymbol{\tau})}_{\text{const}} \right| \right\rangle_{q(\mathbf{Z},\mathbf{W},\mathbf{C},\boldsymbol{\tau})} + \text{const} \\ &= \log p(\boldsymbol{\pi}) + \langle \log p(\mathbf{C}\,|\,\boldsymbol{\pi}) \rangle_{q(\mathbf{C})} + \text{const} \\ &= \underbrace{\log C(\boldsymbol{\delta}_0)}_{\text{const}} + \sum_{k=1}^{K} \log \pi_k^{\delta_{0_k}-1} + \sum_{k=1}^{K}\sum_{n=1}^{N} \underbrace{\langle c_{nk} \rangle_{q(c_{nk})}}_{r_{nk}} \log \pi_k + \text{const} \\ &= \sum_{k=1}^{K} \log \pi_k^{\delta_{0_k}-1} + \sum_{k=1}^{K}\sum_{n=1}^{N} r_{nk} \log \pi_k + \text{const}.\end{aligned} \tag{C.9}$$

Taking the exponential on both sides we observe that $q(\boldsymbol{\pi})$ is a Dirichlet distribution

$$\begin{aligned}q^*(\boldsymbol{\pi}) &= \prod_{k=1}^{K} \pi_k^{(\delta_{0_k} + \sum_{n=1}^{N} r_{nk} - 1)} \\ &= \mathcal{D}\text{ir}(\boldsymbol{\pi}\,|\,\boldsymbol{\delta}),\end{aligned} \tag{C.10}$$

where $\boldsymbol{\delta}$ has components δ_k given by $\delta_k = \delta_{0_k} + \sum_{n=1}^{N} r_{nk}$.

Factor $q(\mathbf{W})$ The logarithm of the optimised factor is given by

$$\begin{aligned}\log q^*(\mathbf{w}_{mk}) &= \left\langle \log \left| \underbrace{p(\mathbf{Y}\,|\,\mathbf{Z})}_{\text{const}}\; p(\mathbf{Z}\,|\,\mathbf{C}, \mathbf{W}, \mathbf{X})\; \underbrace{(\mathbf{C}\,|\,\boldsymbol{\pi})\, p(\boldsymbol{\pi})}_{\text{const}}\; p(\mathbf{W}\,|\,\boldsymbol{\tau})\; \underbrace{p(\boldsymbol{\tau})}_{\text{const}} \right| \right\rangle_{q(\mathbf{Z},\mathbf{C},\boldsymbol{\pi},\boldsymbol{\tau})} + \text{const} \\ &= \sum_{n=1}^{N} \langle c_{nk} \rangle_{q(c_{nk})} \; \langle \log \mathcal{N}(\mathbf{z}_{nm}\,|\,\mathbf{X}_{nm}\mathbf{w}_{mk}, \mathbf{I}_{nm}) \rangle_{q(\mathbf{z}_{nm})} + \langle \log p(\mathbf{w}_{mk}\,|\,\tau_k) \rangle_{q(\tau_k)} + \text{const} \\ &= \sum_{n=1}^{N} r_{nk} \left\langle -\frac{1}{2}\Big(\mathbf{z}_{nm} - \mathbf{X}_{nm}\mathbf{w}_{mk}\Big)^T \Big(\mathbf{z}_{nm} - \mathbf{X}_{nm}\mathbf{w}_{mk}\Big) \right\rangle_{q(\mathbf{z}_{nm})} - \frac{1}{2} \langle \tau_k \rangle_{q(\tau_k)} \mathbf{w}_{mk}^T \mathbf{w}_{mk} + \text{const} \\ &= \sum_{n=1}^{N} r_{nk} \left\{ \mathbf{w}_{mk}^T \mathbf{X}_{nm}^T \; \langle \mathbf{z}_{nm} \rangle_{q(\mathbf{z}_{nm})} - \frac{1}{2} \mathbf{w}_{mk}^T \mathbf{X}_{nm}^T \mathbf{X}_{nm} \mathbf{w}_{mk} \right\} - \frac{1}{2} \langle \tau_k \rangle_{q(\tau_k)} \mathbf{w}_{mk}^T \mathbf{w}_{mk} + \text{const} \\ &= \mathbf{w}_{mk}^T \sum_{n=1}^{N} r_{nk} \mathbf{X}_{nm}^T \; \langle \mathbf{z}_{nm} \rangle_{q(\mathbf{z}_{nm})} - \frac{1}{2} \mathbf{w}_{mk}^T \left\{ \langle \tau_k \rangle_{q(\tau_k)} \mathbf{I} + \sum_{n=1}^{N} r_{nk} \mathbf{X}_{nm}^T \mathbf{X}_{nm} \right\} \mathbf{w}_{mk} + \text{const}.\end{aligned} \tag{C.11}$$

Because this is a quadratic form, the distribution $q(\mathbf{w}_{mk})$ is a Gaussian distribution and we can complete the square, using equation (D.4), to identify the mean and the covariance matrix

$$\begin{aligned}q^*(\mathbf{w}_{mk}) &= \mathcal{N}(\mathbf{w}_{mk}\,|\,\boldsymbol{\lambda}_{mk}, \mathbf{S}_{mk}) \\ \boldsymbol{\lambda}_{mk} &= \mathbf{S}_{mk} \sum_{n=1}^{N} r_{nk} \mathbf{X}_{nm}^T \; \langle \mathbf{z}_{nm} \rangle_{q(\mathbf{z}_{nm})} \\ \mathbf{S}_{mk} &= \left(\langle \tau_k \rangle_{q(\tau_k)} \mathbf{I} + \sum_{n=1}^{N} r_{nk} \mathbf{X}_{nm}^T \mathbf{X}_{nm} \right)^{-1}.\end{aligned} \tag{C.12}$$

Factor $q(\mathbf{Z})$ The logarithm of the optimised factor assuming that the corresponding $y_{nmi} = 1$ is given by

$$\begin{aligned}
\log q^*(z_{nmi}) &= \left\langle \log \left| p(\mathbf{Y} \,|\, \mathbf{Z})\, p(\mathbf{Z} \,|\, \mathbf{C}, \mathbf{W}, \mathbf{X})\, \underbrace{p(\mathbf{C} \,|\, \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\mathbf{W} \,|\, \boldsymbol{\tau})\, p(\boldsymbol{\tau})}_{\text{const}} \right| \right\rangle_{q(\mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau})} + \text{const} \\
&= \log p(y_{nmi} \,|\, z_{nmi}) + \left\langle \log \prod_{k=1}^{K} \mathcal{N}(z_{nmi} \,|\, \mathbf{w}_{mk}^T \mathbf{x}_{nmi}, 1)^{c_{nk}} \right\rangle_{q(\mathbf{c}_n, \mathbf{w}_m)} + \text{const} \\
&= y_{nmi} \log \mathbf{1}(z_{nmi} > 0) + \underbrace{(1 - y_{nmi}) \log \mathbf{1}(z_{nmi} \leq 0)}_{0,\ \text{since}\ y_{nmi}=1} + \\
&\qquad \sum_{k=1}^{K} r_{nk} \ \left\langle \log \mathcal{N}(z_{nmi} \,|\, \mathbf{w}_{mk}^T \mathbf{x}_{nmi}, 1) \right\rangle_{q(\mathbf{w}_{mk})} + \text{const} \\
&= \log \mathbf{1}(z_{nmi} > 0) - \frac{1}{2} z_{nmi}^2 \underbrace{\sum_{k=1}^{K} r_{nk}}_{1} + z_{nmi} \sum_{k=1}^{K} r_{nk} \ \left\langle \mathbf{w}_{mk}^T \right\rangle_{q(\mathbf{w}_{mk})} \mathbf{x}_{nmi} + \text{const.}
\end{aligned} \tag{C.13}$$

Exponentiating this quantity and setting $\mu_{nmi} = \sum_k r_{nk} \left\langle \mathbf{w}_{mk}^T \right\rangle_{q(\mathbf{w}_{mk})} \mathbf{x}_{nmi}$ we obtain

$$q^*(z_{nmi}) \propto \mathbf{1}(z_{nmi} > 0) \exp\left(-\frac{1}{2} z_{nmi}^2 + z_{nmi} \mu_{nmi} \right). \tag{C.14}$$

We observe that the optimized factor $q(z_{nmi})$ is an un-normalised truncated Normal distribution

$$q^*(z_{nmi}) = \begin{cases} \mathcal{TN}_+ (z_{nmi} \,|\, \mu_{nmi}, 1) & \text{if } y_{nmi} = 1 \\ \mathcal{TN}_- (z_{nmi} \,|\, \mu_{nmi}, 1) & \text{if } y_{nmi} = 0 \end{cases}. \tag{C.15}$$

C.1.2 Evidence lower bound

The evidence lower bound (ELBO) is given by

$$\begin{aligned}
\mathcal{L}(q) = \sum_{\mathbf{C}} \int \int \int \int & q(\mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau}) \ln \left\{ \frac{p(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau} | \mathbf{X})}{q(\mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau})} \right\} d\mathbf{Z}\, d\boldsymbol{\pi}\, d\mathbf{W}\, d\boldsymbol{\tau} \\
= \ & \langle \ln p(\mathbf{Y}, \mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau} | \mathbf{X}) \rangle_{q(\mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau})} - \langle \ln q(\mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau}) \rangle_{q(\mathbf{Z}, \mathbf{C}, \boldsymbol{\pi}, \mathbf{W}, \boldsymbol{\tau})} \\
= \ & \langle \ln p(\mathbf{Y} | \mathbf{Z}) \rangle_{q(\mathbf{Z})} + \langle \ln p(\mathbf{Z} | \mathbf{C}, \mathbf{W}, \mathbf{X}) \rangle_{q(\mathbf{Z}, \mathbf{C}, \mathbf{W})} + \langle \ln p(\mathbf{C} | \boldsymbol{\pi}) \rangle_{q(\mathbf{C}, \boldsymbol{\pi})} + \langle \ln p(\boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi})} \\
& + \langle \ln p(\mathbf{W} | \boldsymbol{\tau}) \rangle_{q(\mathbf{W}, \boldsymbol{\tau})} + \langle \ln p(\boldsymbol{\tau}) \rangle_{q(\boldsymbol{\tau})} - \langle \ln q(\mathbf{Z}) \rangle_{q(\mathbf{Z})} - \langle \ln q(\mathbf{C}) \rangle_{q(\mathbf{C})} \\
& - \langle \ln q(\boldsymbol{\pi}) \rangle_{q(\boldsymbol{\pi})} - \langle \ln q(\mathbf{W}) \rangle_{q(\mathbf{W})} - \langle \ln q(\boldsymbol{\tau}) \rangle_{q(\boldsymbol{\tau})}.
\end{aligned} \tag{C.16}$$

We can derive the expectations in a similar fashion to section C.1.1. The ELBO $\mathcal{L}(q)$ is used to assess convergence of the coordinate ascent variational inference (CAVI) algorithm (Blei et al., 2017).

C.1.3 Predictive density

The predictive density of a new observation $\mathbf{y}_*$ which will be associated with a latent variable $\mathbf{c}_*$, latent observation $\mathbf{z}_*$ and covariates $\mathbf{X}_*$ is given by

$$\begin{aligned}
p(\mathbf{y}_*|\mathbf{X}_*,\mathbf{Y},\mathbf{X}) &= \sum_{c}\int p(\mathbf{y}_*,\mathbf{c}_*,\mathbf{z}_*,\boldsymbol{\pi},\mathbf{W},\boldsymbol{\tau}|\mathbf{X}_*,\mathbf{Y},\mathbf{X})\ d\boldsymbol{\pi}\ d\boldsymbol{\tau}\ d\mathbf{W}\ d\mathbf{z}_* \\
&= \sum_{c}\int p(\mathbf{y}_*|\mathbf{z}_*)p(\mathbf{z}_*|\mathbf{c}_*,\mathbf{W},\mathbf{X}_*)p(\mathbf{c}_*|\boldsymbol{\pi})p(\boldsymbol{\pi},\mathbf{W},\boldsymbol{\tau}|\mathbf{Y},\mathbf{X})\ d\boldsymbol{\pi}\ d\boldsymbol{\tau}\ d\mathbf{W}\ d\mathbf{z}_* \\
&= \sum_{k=1}^{K}\int p(\mathbf{y}_*|\mathbf{z}_*)p(\mathbf{z}_*|\mathbf{W}_k,\mathbf{X}_*)\pi_k\ q(\boldsymbol{\pi})q(\mathbf{W}_k)q(\tau_k)\ d\boldsymbol{\pi}\ d\tau_k\ d\mathbf{W}_k\ d\mathbf{z}_* \\
&= \sum_{k=1}^{K}\frac{\delta_k}{\hat{\delta}}\int p(\mathbf{y}_*|\mathbf{z}_*)\mathcal{N}(\mathbf{z}_*|\mathbf{X}_*\mathbf{W}_k,\mathbf{I}_n)\mathcal{N}(\mathbf{W}_k|\boldsymbol{\lambda}_k,\mathbf{S}_k)\ d\mathbf{W}_k\ d\mathbf{z}_* \\
&= \sum_{k=1}^{K}\frac{\delta_k}{\hat{\delta}}\int p(\mathbf{y}_*|\mathbf{z}_*)\mathcal{N}\left(\mathbf{z}_*|\mathbf{X}_*\boldsymbol{\lambda}_k,\ \mathbf{I}_n+\operatorname{diag}\big(\mathbf{X}_*\mathbf{S}_k\mathbf{X}_*^T\big)\right)\ d\mathbf{z}_* \\
&= \sum_{k=1}^{K}\frac{\delta_k}{\hat{\delta}}\begin{cases}\int_0^{\infty}\mathcal{N}\left(\mathbf{z}_*|\mathbf{X}_*\boldsymbol{\lambda}_k,\ \mathbf{I}_n+\operatorname{diag}\big(\mathbf{X}_*\mathbf{S}_k\mathbf{X}_*^T\big)\right)\ dz & \text{where } \mathbf{y}_*=1\\ \int_{-\infty}^{0}\mathcal{N}\left(\mathbf{z}_*|\mathbf{X}_*\boldsymbol{\lambda}_k,\ \mathbf{I}_n+\operatorname{diag}\big(\mathbf{X}_*\mathbf{S}_k\mathbf{X}_*^T\big)\right)\ dz & \text{where } \mathbf{y}_*=0\end{cases} \\
&= \sum_{k=1}^{K}\frac{\delta_k}{\hat{\delta}}\Phi\left(\rho\right)^{\mathbf{y}_*}\left(1-\Phi\left(\rho\right)\right)^{(1-\mathbf{y}_*)} \\
&= \sum_{k=1}^{K}\frac{\delta_k}{\hat{\delta}}\mathcal{B}\text{ern}\left(\mathbf{y}_*\middle|\Phi\left(\rho\right)\right),
\end{aligned} \tag{C.17}$$

where $\hat{\delta}=\sum_k\delta_k$, $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal distribution and

$$\rho=\frac{\mathbf{X}_*\boldsymbol{\lambda}_k}{\left(\mathbf{I}_n+\operatorname{diag}\big(\mathbf{X}_*\mathbf{S}_k\mathbf{X}_*^T\big)\right)^{1/2}}. \tag{C.18}$$

C.2 Supplementary figures

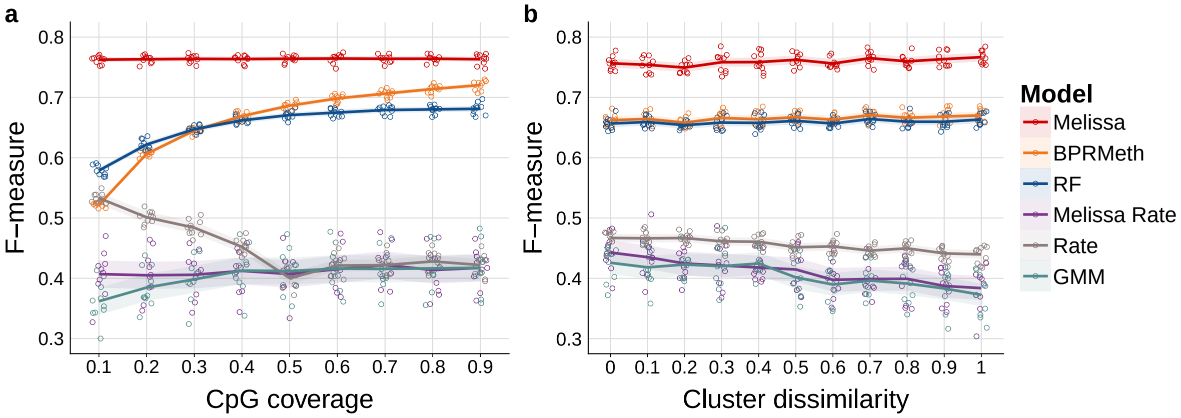


Figure C.1 Melissa robustly imputes CpG methylation states on synthetic data. (**a**) Imputation performance in terms of F-measure as we vary the proportion of covered CpGs used for training. Higher values correspond to better imputation performance. For each CpG coverage setting a total of 10 random splits of the data to training and test sets was performed. Each coloured circle corresponds to a different simulation. The plot shows also the LOESS curve for each method as we increase CpG coverage. (**b**) Imputation performance measured by F-measure for varying proportions of similar genomic regions between clusters. Values closer to zero correspond to highly similar cell sub-populations, whereas values closer to one correspond to well separated cell sub-populations. In (**a**) cluster dissimilarity was set to 0.5 and in (**b**) CpG coverage was set to 0.4.

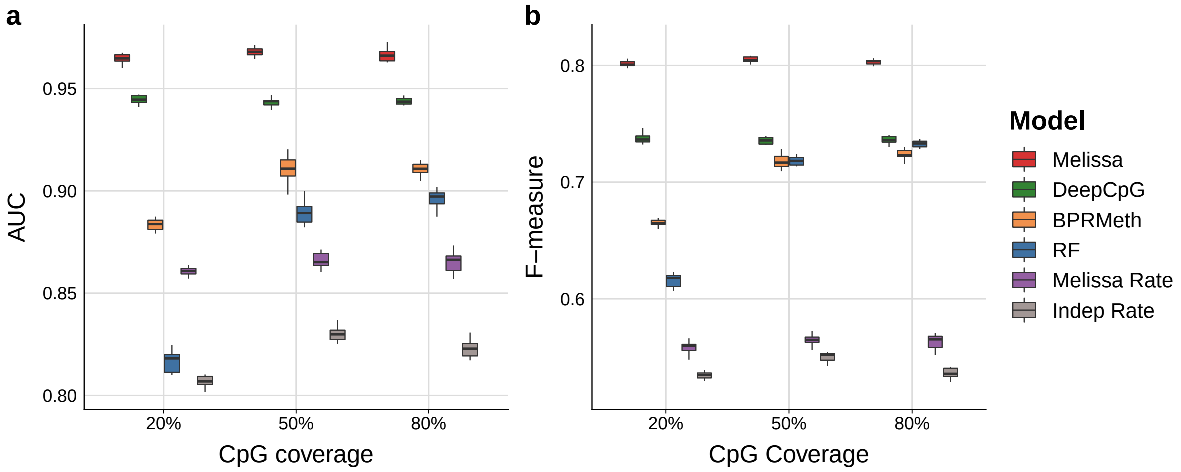


Figure C.2 Melissa robustly imputes CpG methylation states on the subsampled ENCODE methylation data. Imputation performance in terms of (**a**) AUC and (**b**) F-measure for varying levels of CpG coverage for pre-defined 10kb regions around TSS. For each CpG coverage setting a total of 10 random splits of the data to training and test sets was performed. Each dot corresponds to a different simulation.

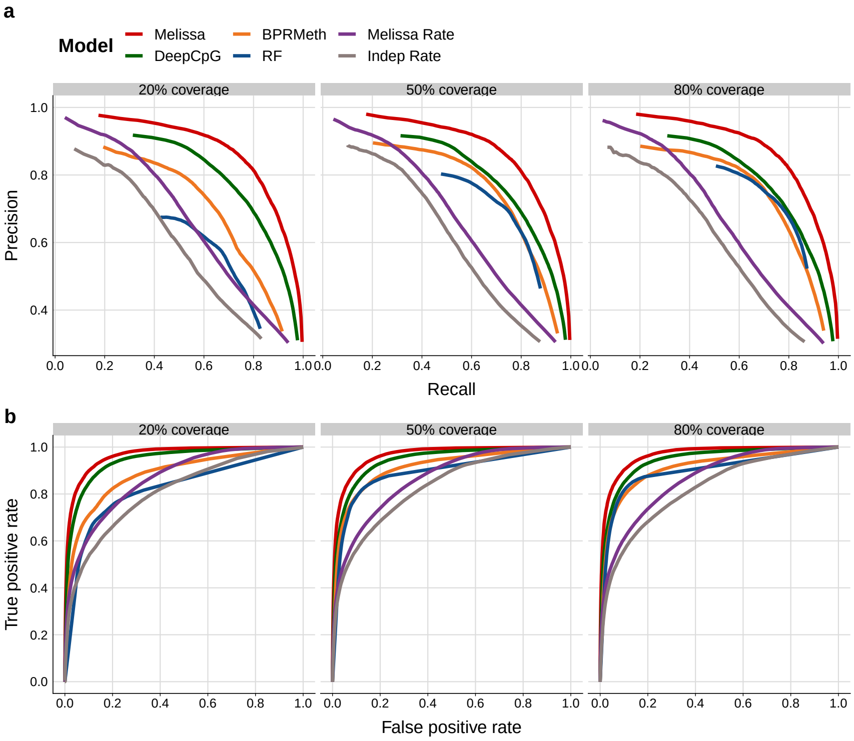


Figure C.3 (**a**) Precision recall curves and (**b**) receiver operating characteristic curves on varying CpG coverage levels for imputing CpG methylation states for the subsampled ENCODE methylation data.

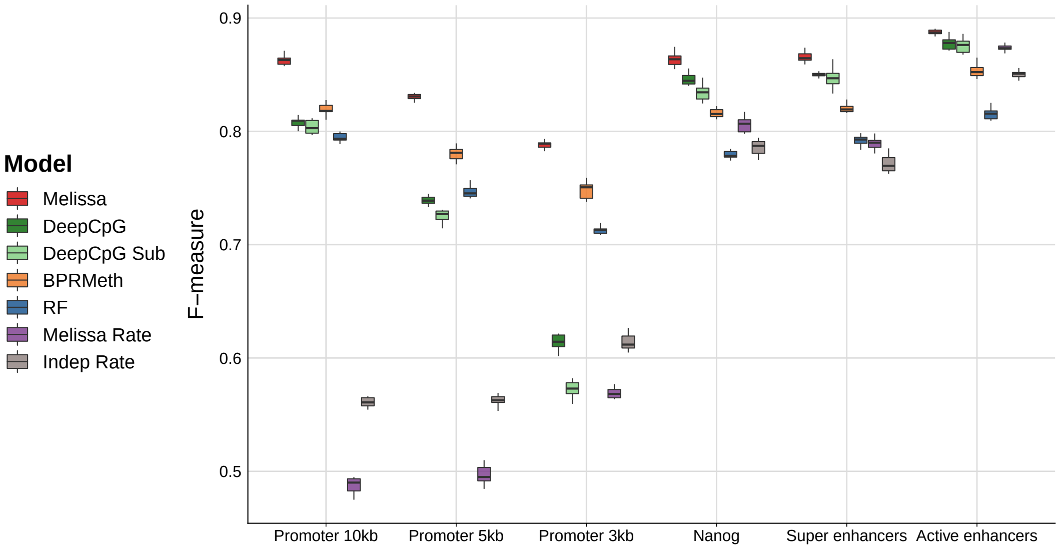


Figure C.4 Prediction performance using the F-measure metric for imputing CpG methylation states of the Angermueller et al. (2016) dataset. Higher values correspond to better imputation performance. Each coloured boxplot indicates the performance using 10 random splits of the data in training and test sets; due to high computational costs, DeepCpG was trained only once and the boxplots denote the variability across ten random subsamplings of the test set. Shown is the prediction performance for alternative genomic contexts: promoters (±3kb, ±5kb and ±10kb regions), active enhancers, super enhancers and Nanog regulatory regions.

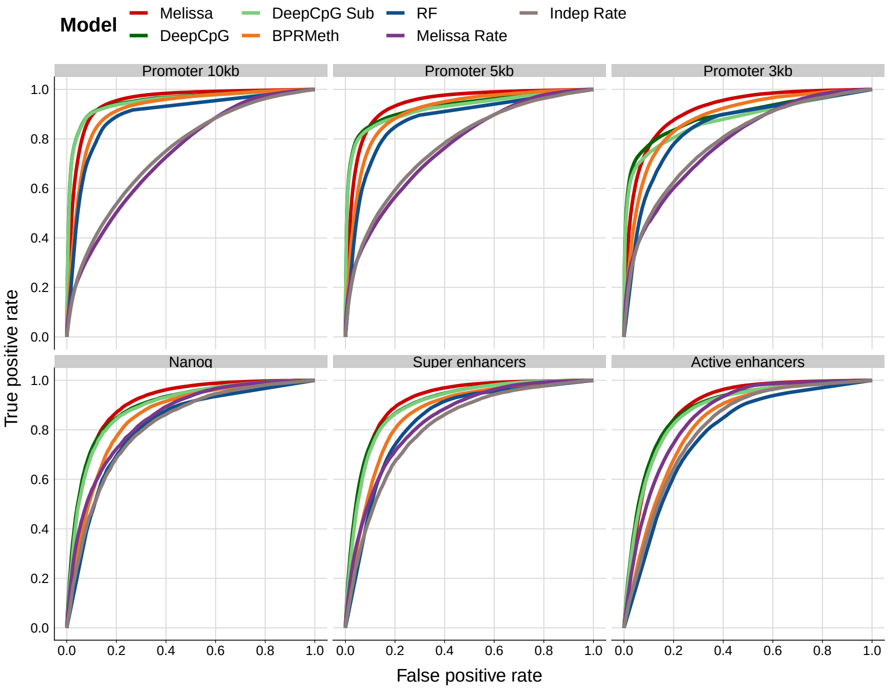


Figure C.5 Receiver operating characteristic curves for imputing CpG methylation states of the Angermueller et al. (2016) dataset.

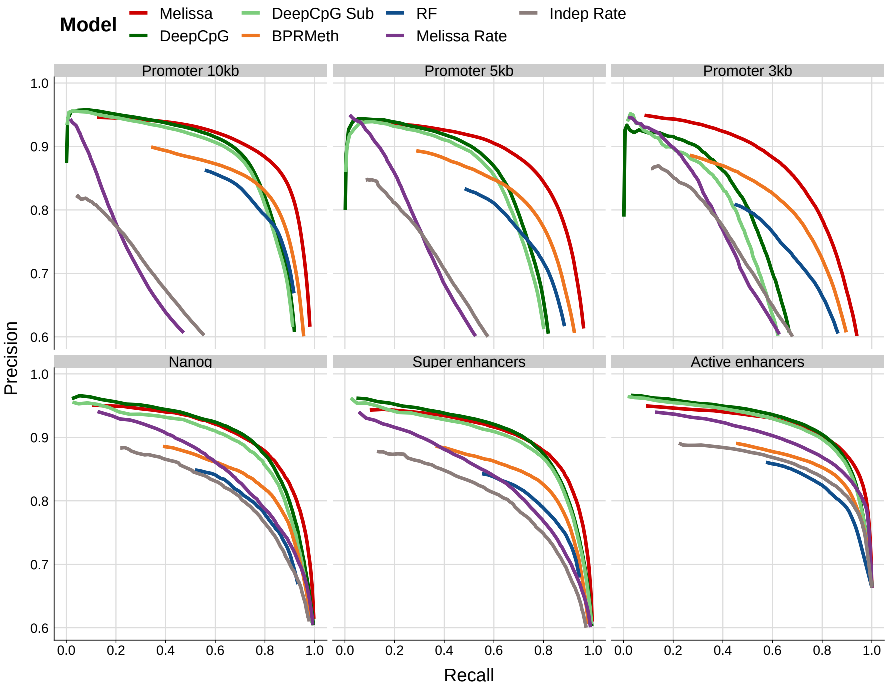


Figure C.6 Precision recall curves for imputing CpG methylation states of the Angermueller et al. (2016) dataset.

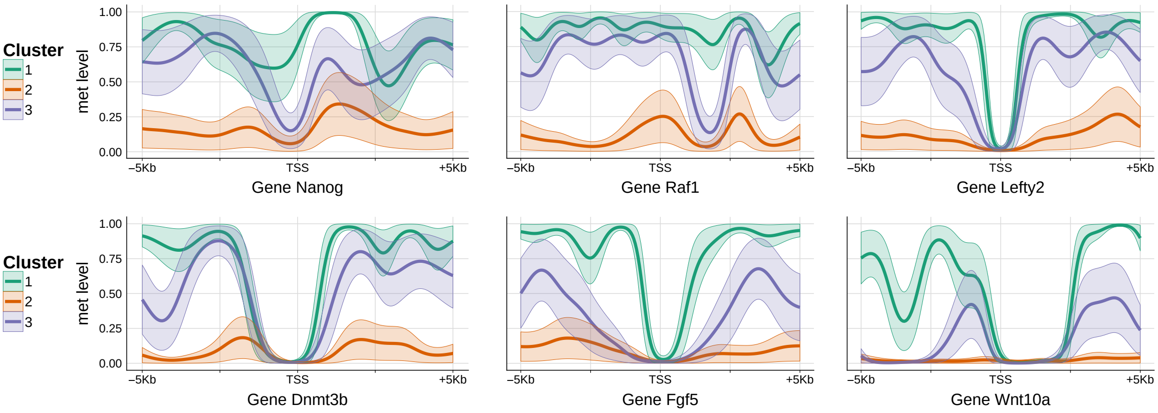


Figure C.7 Example methylation profiles for different promoter regions of developmental genes with window length ±$5kb$ for the Angermueller et al. (2016) dataset. Melissa identified three cell subpopulations.

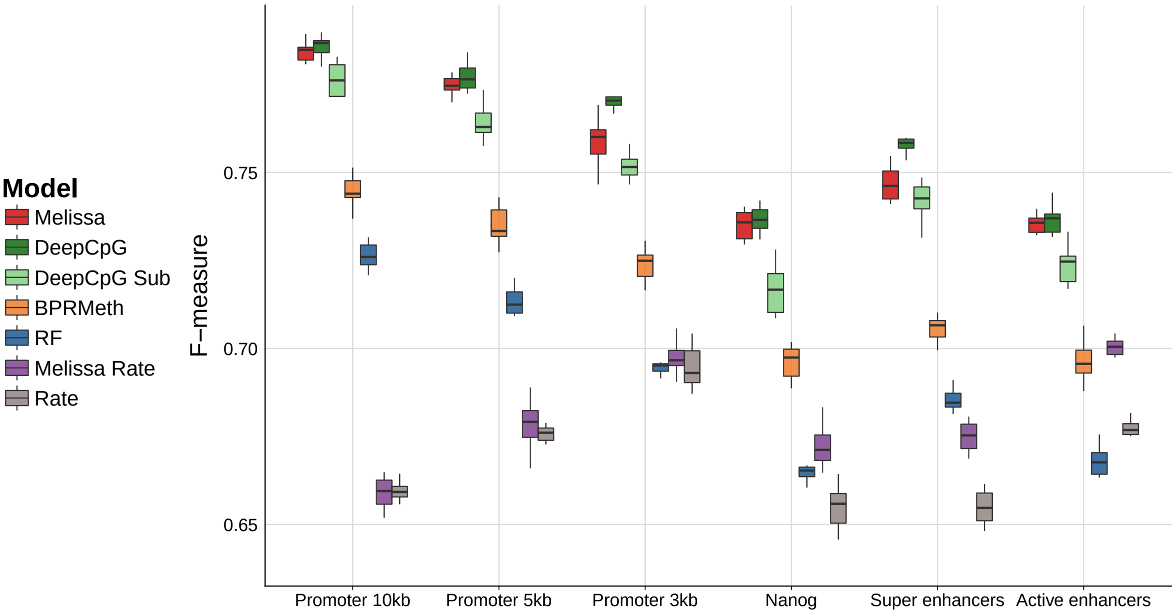


Figure C.8 Prediction performance using the F-measure metric for imputing CpG methylation states of the Smallwood et al. (2014) dataset. Higher values correspond to better imputation performance. Each coloured boxplot indicates the performance using 10 random splits of the data in training and test sets; due to high computational costs, DeepCpG was trained only once and the boxplots denote the variability across ten random subsamplings of the test set. Shown is the prediction performance for alternative genomic contexts: promoters (±3kb, ±5kb and ±10kb regions), active enhancers, super enhancers and Nanog regulatory regions.

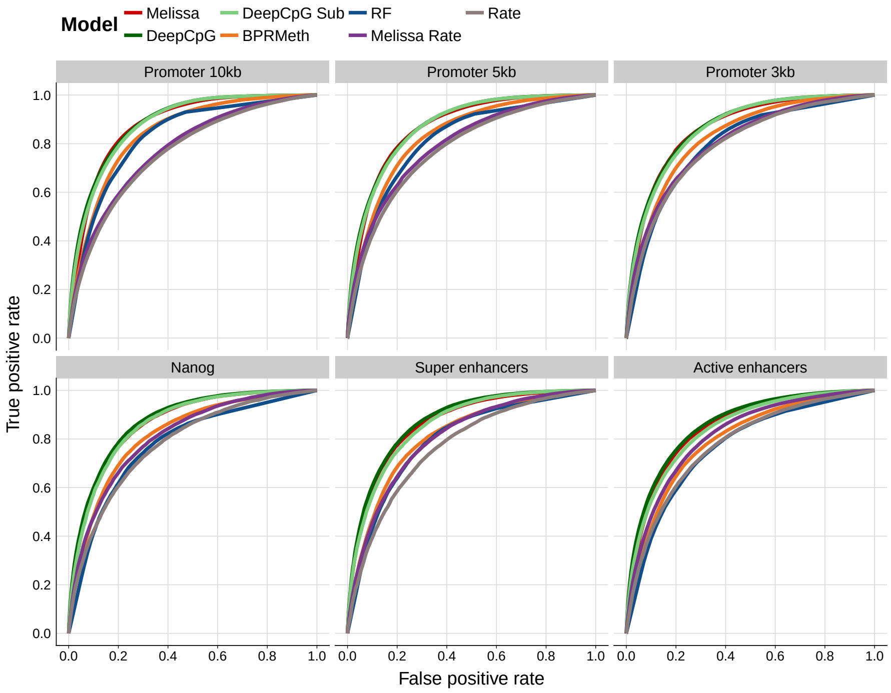


Figure C.9 Receiver operating characteristic curves for imputing CpG methylation states of the Smallwood et al. (2014) dataset.

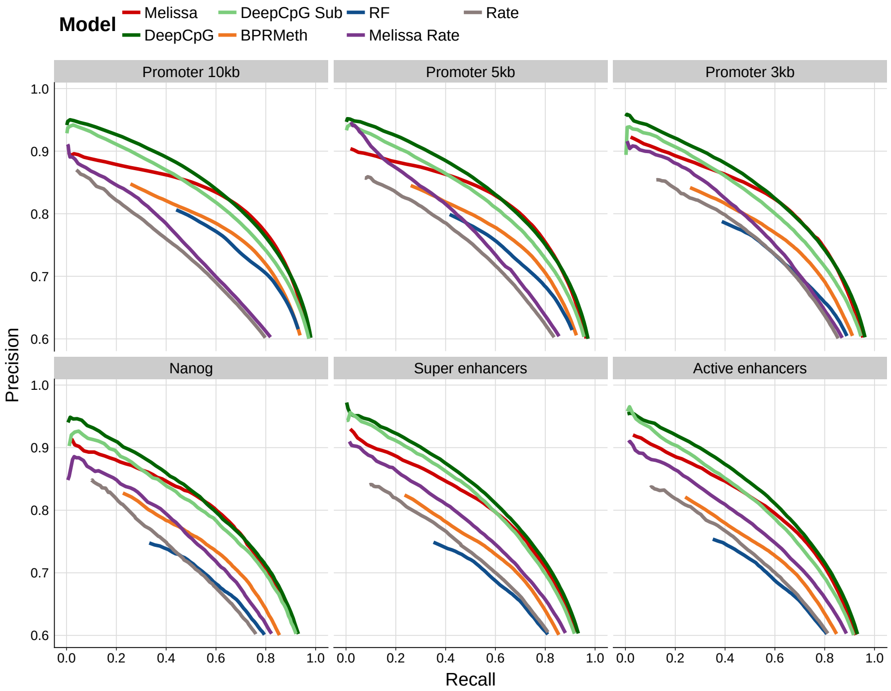


Figure C.10 Precision recall curves for imputing CpG methylation states of the Smallwood et al. (2014) dataset.

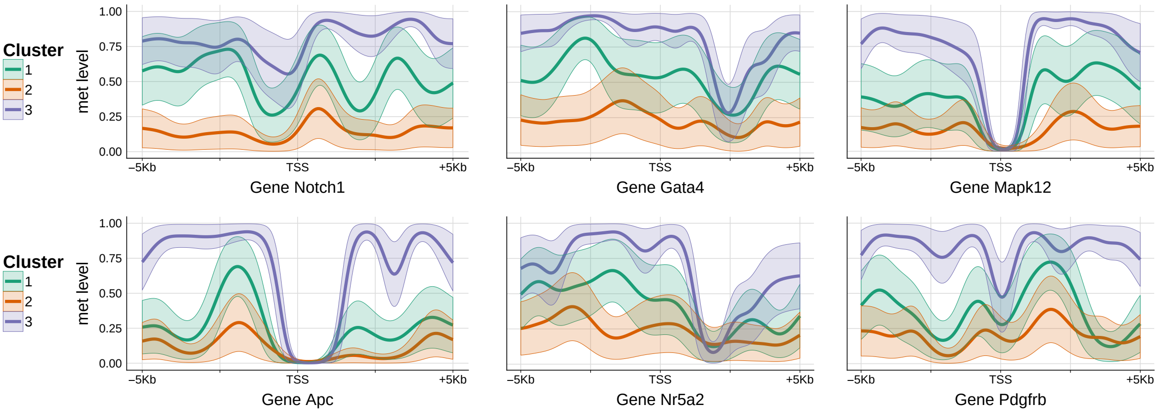


Figure C.11 Example profiles for different promoter regions with window length ±5kb for the Smallwood et al. (2014) dataset. Melissa identified three cell sub-populations.

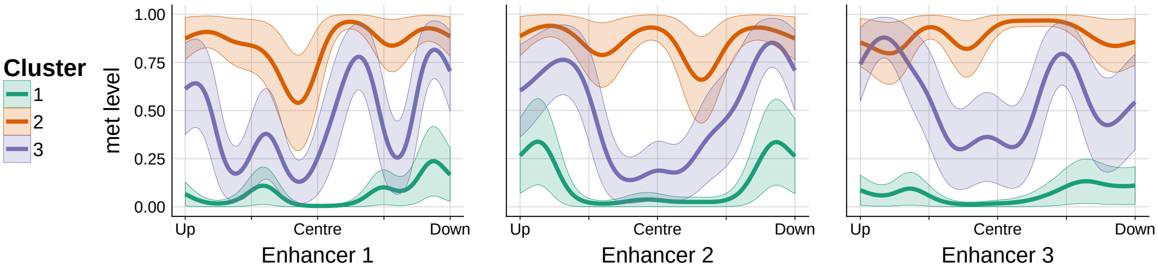


Figure C.12 Example profiles for different enhancer regions for the Smallwood et al. (2014) dataset. Melissa identified three cell sub-populations.

C.3 Supplementary tables

Genomic context	CpGs (in millions)	Time (in hours)
Promoter 10kb	6	5.6
Promoter 5kb	2.1	2.9
Promoter 3kb	0.62	1.31
Nanog	0.18	0.61
Super enhancers	0.5	1
Active enhancers	0.7	1.76

Table C.1 Melissa training time for the Angermueller et al. (2016) mouse ESC dataset. Across different genomic contexts are shown the total number of CpGs used for training set and the time required (in hours) for running Melissa to impute and cluster single cells. As a comparison, the DeepCpG model took about three to four days to train on around four million CpGs.

Genomic context	CpGs (in millions)	Time (in hours)
Promoter 10kb	4.13	4
Promoter 5kb	1.54	2.21
Promoter 3kb	0.98	1.83
Nanog	0.26	0.9
Super enhancers	0.29	0.9
Active enhancers	0.85	2

Table C.2 Melissa training time for the Smallwood et al. (2014) mouse ESC dataset. Across different genomic contexts are shown the total number of CpGs used for training set and the time required (in hours) for running Melissa to impute and cluster single cells. As a comparison, the DeepCpG model took about three to four days to train on around four million CpGs.

Genomic context	Smallwood study	Angermueller study
Promoter 10kb	21%	17%
Promoter 5kb	23%	20%
Promoter 3kb	24%	24%
Nanog	19%	17%
Super enhancers	19%	12%
Active enhancers	25%	17%

Table C.3 Sparsity level of the two scBS-seq data after filtering across different genomic regions.

Appendix D

Miscellaneous

D.1 Probit function

The probit function is defined as the inverse of the cumulative distribution function of the standard normal distribution denoted by Φ, and is given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt. \tag{D.1}$$

D.2 Root mean square error

The root mean square error (RMSE) measures the square root of the average squared difference between the observed $\mathbf{y}$ and estimated values $\hat{\mathbf{y}}$,

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2}. \tag{D.2}$$

D.3 Pearson correlation coefficient

The sample Pearson correlation coefficient r measures the linear correlation between samples from two random variables X and Y,

$$r = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N} (x_n - \bar{x})^2} \sqrt{\sum_{n=1}^{N} (y_n - \bar{y})^2}}. \tag{D.3}$$

D.4 Completing the square

When we are given a quadratic form defining the exponent terms in a Gaussian distribution and we need to determine the corresponding mean and covariance, we make use of the fact that the exponent in a Gaussian distribution $\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be written as

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const}, \tag{D.4}$$

where *const* denotes terms that are independent of $\mathbf{x}$, and we have made use of the symmetry of $\boldsymbol{\Sigma}$.

D.5 M-value

The transformation from average methylation rates to M-values is obtained by

$$M\text{-value} = \log_2\left(\frac{\text{rate} + 0.01}{1 - \text{rate} + 0.01}\right). \tag{D.5}$$

D.6 Adjusted rand index

The adjusted rand index (ARI) is a measure of the similarity between the true cluster assignment and the predicted cluster membership returned from the model,

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{\alpha_i}{2}\sum_j \binom{\beta_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{\alpha_i}{2} + \sum_j \binom{\beta_j}{2}\right] - \left[\sum_i \binom{\alpha_i}{2}\sum_j \binom{\beta_j}{2}\right]/\binom{n}{2}}. \tag{D.6}$$

D.7 F-measure

The F-measure or F$_1$-score is the harmonic mean of precision and recall,

$$F\text{-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \tag{D.7}$$

where

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}, \tag{D.8}$$

and

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}. \tag{D.9}$$

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169. (page 39)

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723. (page 37)

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679. (pages 82, 85, and 130)

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2015). *Molecular Biology of the Cell*. Garland Science, 6th edition. (pages 6, 7, and 8)

Aldiri, I. and Vetter, M. L. (2012). PRC2 during vertebrate organogenesis: a complex in transition. *Developmental biology*, 367(2):91–99. (page 13)

Aleksic, J., Carl, S. H., and Frye, M. (2014). Beyond library size: a field guide to NGS normalization. *bioRxiv*, 006403. (page 18)

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences*, 51(5):786–794. (page 13)

Amir, R. E., den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23(2):185–188. (page 16)

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106. (page 18)

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the American Statistical Association*, 65(331):1248–1255. (page 30)

Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43. (pages 46 and 55)

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., and Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature methods*, 13(3):229–32. (pages 24, 79, 80, 90, 92, 101, 110, 111, 115, 116, 145, 146, 147, and 150)

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, 18(1):67. (pages 109, 111, 114, 118, and 122)

Angus, J., Beal, M., Li, J., Rangel, C., and Wild, D. (2010). Inferring transcriptional networks using prior biological knowledge and constrained state-space models. *Learning and Inference in Computational Systems Biology*, pages 117–152. (page 33)

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124. (page 24)

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, 17(9):507. (page 1)

Avner, P. and Heard, E. (2001). X-chromosome inactivation: counting, choice and initiation. *Nature Reviews Genetics*, 2(1):59. (page 15)

Bai, L., Charvin, G., Siggia, E. D., and Cross, F. R. (2010). Nucleosome-depleted regions in cell-cycle-regulated promoters ensure reliable gene expression in every cell cycle. *Developmental cell*, 18(4):544–555. (page 23)

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press. (pages 36 and 41)

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837. (page 19)

Baylin, S. B. and Jones, P. a. (2011). A decade of exploring the cancer epigenome - biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734. (pages 15, 74, and 121)

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. PhD thesis, University College London. (page 105)

Beal, M. J. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian statistics*, 7:453–464. (page 88)

Beaumont, M. A. and Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261. (page 33)

Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–396. (page 12)

Bell, O., Tiwari, V. K., Thomä, N. H., and Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, 12(8):554–564. (pages 11 and 12)

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press. (page 51)

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 57(1):289–300. (page 92)

Benveniste, D., Sonntag, H.-J., Sanguinetti, G., and Sproul, D. (2014). Transcription factor binding predicts histone modifications in human cell lines. *Proceedings of the National Academy of Sciences*, 111(37):13367–13372. (pages 13, 29, 75, and 125)

Berger, J. O. and Wolpert, R. L. (1988). *The likelihood principle*. Institute of Mathematical Statistics, 2nd edition. (page 33)

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236. (page 54)

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., and Others (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, 98(4):288–295. (page 23)

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development*, 16(1):6–21. (pages 14 and 15)

Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–398. (pages 9 and 10)

Bird, A., Taggart, M., Frommer, M., Miller, O. J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99. (page 14)

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, 8(7):1499–1504. (page 14)

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306. (page 33)

Bishop, C., Bishop, C. M., and Others (1995). *Neural networks for pattern recognition*. Oxford university press. (page 63)

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. (pages 28, 31, 32, 34, 38, 39, 41, 45, 49, 52, 84, 103, and 106)

Biswas, M., Voltz, K., Smith, J. C., and Langowski, J. (2011). Role of histone tails in structural stability of the nucleosome. *PLoS computational biology*, 7(12):e1002279. (page 13)

Blackwood, E. M. and Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63. (page 9)

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, 1(1):121–144. (page 123)

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877. (pages 48, 105, and 142)

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. (page 47)

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719. (pages 20, 21, and 23)

Bock, C., Beerman, I., Lien, W. H., Smith, Z. D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D. J., and Meissner, A. (2012). DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Molecular Cell*, 47(4):633–647. (pages 21, 59, and 64)

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551. (page 92)

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for experimenters*. John Wiley and Sons. (page 28)

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press. (page 50)

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., and Others (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091):349–353. (page 13)

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527. (page 18)

Breiman, L. E. O. (2001). Random Forests. *Machine Learning*, 45(1):5–32. (page 65)

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095. (page 19)

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455. (page 57)

Brownell, J. E. and Allis, C. D. (1995). An activity gel assay detects a single, catalytically active histone acetyltransferase subunit in Tetrahymena macronuclei. *Proceedings of the National Academy of Sciences*, 92(14):6364–6368. (page 13)

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015a). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology*, 109(1):21–29. (page 23)

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490. (page 24)

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. a., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–60. (page 18)

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M. B. (2013). Identification of active regulatory regions from DNA methylation data. *Nucleic acids research*, 41(16):e155. (page 21)

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304. (page 12)

Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209. (page 5)

Chen, T., Ueda, Y., Dodge, J. E., Wang, Z., and Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and cellular biology*, 23(16):5594–5605. (page 14)

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335. (page 51)

Clark, S. J., Argelaguet, R., Kapourani, C. A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., and Reik, W. (2018). ScNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1):1–9. (pages 3, 24, 25, 77, 89, 90, 119, and 125)

Clark, S. J., Harrison, J., and Molloy, P. L. (1997). Sp1 binding is inhibited by m Cp m CpG methylation. *Gene*, 195(1):67–71. (page 9)

Clark, S. J., Lee, H. J., Smallwood, S. A., Kelsey, G., and Reik, W. (2016). Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome biology*, 17(72):1–10. (pages 22 and 79)

Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., and Kelsey, G. (2017). Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nature protocols*, 12(3):534–547. (page 22)

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research*, 38(6):1767–1771. (page 110)

Colgan, D. F. and Manley, J. L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes & development*, 11(21):2755–2766. (page 8)

Corduneanu, A. and Bishop, C. M. (2001). Variational Bayesian Model Selection for Mixture Distributions. *In Artificial Intelligence and Statistics*, pages 27–34. (pages 106 and 114)

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13. (page 31)

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., and Others (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936. (page 9)

Crick, F. H. C. (1958). On protein synthesis. *Symp Soc Exp Biol*, 12(138):63. (page 6)

Crick, F. H. C. (1968). The origin of the genetic code. *Journal of molecular biology*, 38(3):367–379. (page 6)

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563. (page 6)

Cross, S. H., Charlton, J. A., Nan, X., and Bird, A. P. (1994). Purification of CpG islands using a methylated DNA binding column. *Nature Genetics*, 6(3):236–244. (page 14)

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914. (page 24)

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581. (page 5)

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31. (page 38)

Dawid, A. P. (2004). Probability, Causality and the Empirical World: A Bayes - de Finetti - Popper - Borel Synthesis. *Statistical Science*, 19(1):44–57. (page 31)

De Finetti, B. (1974). *Theory of Probability: A critical introductory treatment.* Wiley Chichester. (page 31)

Deaton, A. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10):1010–1022. (pages 14 and 65)

Deichmann, U. (2016). Epigenetics: The origins and evolution of a fashionable topic. *Developmental biology*, 416(1):249–254. (page 10)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38. (pages 42, 43, and 65)

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, 49(4):1057–1072. (page 37)

Diebolt, J. and Robert, C. P. (1994). Distributions of Finite Mixture Estimation through Bayesian Sampling. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 56(2):363–375. (page 44)

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380. (page 12)

Dolzhenko, E. and Smith, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*, 15(1):1–8. (page 21)

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., and Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9):R53. (pages 13, 29, 72, and 75)

Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences*, 110(14):5294–5300. (page 7)

Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, 11(1):587. (pages 81, 109, and 112)

Dunham, I., Kundaje, A., and Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74. (pages 1, 7, 17, 60, and 109)

Eckersley-Maslin, M. A., Alda-Catalinas, C., and Reik, W. (2018). Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, pages 1–15. (pages 7, 12, and 15)

Edgar, R., Tan, P. P. C., Portales-Casamar, E., and Pavlidis, P. (2014). Meta-analysis of human methylomes reveals stably methylated sequences surrounding CpG islands associated with high gene expression. *Epigenetics & chromatin*, 7(1):28. (pages 75 and 121)

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574. (pages 8 and 73)

Elgar, G. and Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in genetics*, 24(7):344–352. (page 7)

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186. (page 28)

Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825. (page 13)

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–76. (page 119)

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588. (page 44)

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Reports*, 10(8):1386–1397. (page 22)

Feller, W. (1968). *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, 3rd edition. (page 27)

Felsenfeld, G. (2014). A brief history of epigenetics. *Cold Spring Harbor perspectives in biology*, 6(1):a018200. (page 11)

Felsenfeld, G. and Groudine, M. (2003). Controlling the double helix. *Nature*, 421(6921):448–453. (page 11)

Feng, H., Conneely, K. N., and Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69. (page 21)

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230. (pages 28 and 44)

Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815. (page 81)

Festuccia, N., Osorno, R., Halbritter, F., Karwacki-Neisius, V., Navarro, P., Colby, D., Wong, F., Yates, A., Tomlinson, S. R., and Chambers, I. (2012). Esrrb is a direct Nanog target gene that can substitute for Nanog function in pluripotent cells. *Cell stem cell*, 11(4):477–490. (page 97)

Ficz, G., Hore, T. A., Santos, F., Lee, H. J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y. L., Walter, J., Cook, S. J., Andrews, S., Branco, M. R., and Reik, W. (2013). FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell*, 13(3):351–359. (pages 115 and 116)

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2):155–181. (page 37)

Franklin, R. E. and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171(4356):740–741. (page 5)

Franks, A., Airoldi, E., and Slavov, N. (2017). Post-transcriptional regulation across human tissues. *PLoS computational biology*, 13(5):e1005535. (page 8)

Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403. (page 34)

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics, 2nd edition. (pages 27, 34, and 64)

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67. (page 65)

Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome research*, 19(1):92–105. (page 8)

Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831. (pages 20 and 59)

Gaidatzis, D., Burger, L., Murr, R., Lerch, A., Dessus-Babus, S., Schübeler, D., and Stadler, M. B. (2014). DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. *PLoS genetics*, 10(2):e1004143. (page 21)

Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477. (page 18)

Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282. (page 14)

Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium (Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections)*, volume 7. Perthes et Besser. (page 29)

Gawad, C., Koh, W., and Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188. (page 17)

Gelfand, A. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal Of The American Statistical Association*, 85(410):398–409. (pages 47 and 54)

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC. (pages 35, 37, and 57)

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741. (page 54)

Genereux, D. P., Johnson, W. C., Burden, A. F., Stöger, R., and Laird, C. D. (2008). Errors in the bisulfite conversion of DNA: modulating inappropriate-and failed-conversion frequencies. *Nucleic acids research*, 36(22):e150. (page 105)

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome research*, 17(6):669–681. (pages 6 and 13)

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, 7(4):473–483. (page 57)

Ghahramani, Z. (2004). Unsupervised learning. In *Advanced lectures on machine learning*, pages 72–112. Springer. (pages 29 and 43)

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459. (page 27)

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press. (page 52)

Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of mathematical physics*, 4(2):294–307. (page 39)

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351. (page 16)

Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., Wen, L., and Tang, F. (2017). Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell research*, 27(8):1–22. (page 24)

Guo, H., Zhu, P., Wu, X., Li, X., Wen, L., and Tang, F. (2013). Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, pages 2126–2135. (pages 22 and 77)

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., and Others (2014). The DNA methylation landscape of human early embryos. *Nature*, 511(7511):606–610. (page 22)

Haberland, M., Montgomery, R. L., and Olson, E. N. (2009). The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nature Reviews Genetics*, 10(1):32–42. (page 13)

Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848. (page 97)

Hansen, K. D., Langmead, B., and Irizarry, R. A. (2012). BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome biology*, 13(10):R83. (pages 21 and 59)

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., Mcdonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R. A., and Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775. (pages 21, 59, and 64)

Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, 405(6785):486–489. (page 15)

Harris, R. A., Wang, T., Coarfa, C., Nagarajan, R. P., Hong, C., Downey, S. L., Johnson, B. E., Fouse, S. D., Delaney, A., Zhao, Y., and Others (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*, 28(10):1097–1105. (page 20)

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrica*, 57(1):97–109. (page 53)

Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60. (page 85)

Hestenes, R. and Stiefel, E. (1952). Methods of Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436. (page 64)

Hicks, S. C., Teng, M., and Irizarry, R. A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-Seq data. *bioRxiv*, 25528. (page 19)

Hnisz, D., Day, D. S., and Young, R. A. (2016). Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell*, 167(5):1188–1200. (page 11)

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67. (page 34)

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347. (pages 30, 47, and 50)

Holliday, R. and Pugh, J. E. (1975). DNA modification mechanisms and gene activity during development. *Science*, 187(4173):226–232. (pages 10 and 14)

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1 A):145–168. (pages 85 and 130)

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):3156. (page 123)

Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175(1):315–332. (page 14)

Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., and Peng, J. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell research*, 26(3):304–19. (page 24)

Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z., and Fan, G. (2016). Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome biology*, 17(1):88. (page 24)

Huang, Y. and Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome biology*, 18(123):1–11. (page 33)

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218. (pages 108 and 112)

Illingworth, R. S. and Bird, A. P. (2009). CpG islands – A rough guide. *FEBS Letters*, 583(11):1713–1720. (page 14)

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S., and Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2):178–86. (page 75)

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, 21(7):1160–1167. (page 18)

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166. (page 19)

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37. (page 50)

Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(March):245–254. (pages 10 and 14)

Janknecht, R. and Hunter, T. (1996). A growing coactivator network. *Nature*, 383(6595):22–23. (page 9)

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press, Oxford, 3rd edition. (page 37)

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193. (page 42)

Jenuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532):1074–80. (page 13)

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. a., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., Kim, S.-B., Yang, L., Ko, M., Chen, R., Göttgens, B., Lee, J.-S., Gunaratne, P., Godley, L. a., Darlington, G. J., Rao, A., Li, W., and Goodell, M. a. (2014). Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nature genetics*, 46(1):17–23. (page 75)

Ji, H. and Liu, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nature biotechnology*, 28(4):337–340. (page 35)

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502. (page 19)

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7):484–92. (pages 14, 15, and 59)

Jordan, M. I. (2003). An introduction to probabilistic graphical models. *University of California, Berkeley*. (pages 39 and 41)

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233. (page 47)

Kærn, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464. (page 28)

Kapourani, C. A. and Sanguinetti, G. (2016). Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics*, 32(17):i405–i412. (pages 2, 17, 22, 29, 59, 108, and 125)

Kapourani, C. A. and Sanguinetti, G. (2018a). BPRMeth: a flexible Bioconductor package for modelling methylation profiles. *Bioinformatics*, 34(14):2485–2486. (pages 2, 23, 77, 108, and 109)

Kapourani, C. A. and Sanguinetti, G. (2018b). Melissa: Bayesian clustering and imputation of single cell methylomes. *bioRxiv*, 312025:1–16. (pages 3, 23, 29, and 101)

Karemaker, I. D. and Vermeulen, M. (2018). Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends in biotechnology*, In Press:1–13. (pages 21 and 22)

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7):2926–2931. (pages 13 and 125)

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795. (pages 36 and 37)

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., and Others (2014). Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences*, 111(17):6131–6138. (page 7)

Kellum, R. and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, 64(5):941–950. (page 12)

Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., and Jones, P. A. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome research*, 22(12):2497–2506. (pages 23 and 78)

Kelsey, G., Stegle, O., and Reik, W. (2017). Single-cell epigenomics: Recording the past and predicting the future. *Science*, 358(6359):69–75. (pages 25, 101, and 125)

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360. (pages 18 and 91)

Kohli, R. M. and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. *Nature*, 502(7472):472–479. (page 15)

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: principles and techniques.* The MIT Press. (pages 2, 38, 39, and 41)

Kornberg, R. D. and Thomas, J. O. (1974). Chromatin structure: oligomers of the histones. *Science*, 184(4139):865–868. (page 13)

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705. (page 13)

Krivega, I. and Dean, A. (2012). Enhancer and promoter interactions - long distance calls. *Current opinion in genetics & development*, 22(2):79–85. (page 9)

Krueger, F. and Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572. (pages 20, 91, and 110)

Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–151. (page 20)

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86. (page 43)

Kunkel, T. A. (2004). DNA replication fidelity. *Journal of Biological Chemistry*, 279(17):16895–16898. (page 6)

Laird, C. D., Pleasant, N. D., Clark, A. D., Sneeden, J. L., Hassan, K. M. A., Manley, N. C., Vary, J. C., Morgan, T., Hansen, R. S., and Stöger, R. (2004). 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. (page 14)

Laird, P. W. (2003). The power and the promise of DNA methylation markers. *Nature reviews. Cancer*, 3(April):253–266. (page 74)

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203. (page 20)

Landau, D. A., Clement, K., Ziller, M. J., Boyle, P., Fan, J., Gu, H., Stevenson, K., Sougnez, C., Wang, L., Li, S., Kotliar, D., Zhang, W., Ghandi, M., Garraway, L., Fernandes, S. M., Livak, K. J., Gabriel, S., Gnirke, A., Lander, E. S., Brown, J. R., Neuberg, D., Kharchenko, P. V., Hacohen, N., Getz, G., Meissner, A., and Wu, C. J. (2014). Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell*, 26(6):813–825. (page 22)

Lander, E. S., Linton, L., and International Human Genome Sequencing (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. (pages 1 and 16)

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., and Others (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–1831. (page 19)

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25. (page 18)

Lappalainen, T. and Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics*, 18(7):441–451. (pages 10 and 23)

Latchman, D. S. (1997). Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312. (page 8)

Lauritzen, S. L. (1996). *Graphical models.* Clarendon Press. (page 2)

Lawrence, N. D., Girolami, M., Rattray, M., and Sanguinetti, G. (2010). *Learning and inference in computational systems biology.* The MIT Press. (pages 27, 28, and 46)

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. (page 63)

Lee, J. T. (2011). Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nature reviews Molecular cell biology*, 12(12):815–826. (pages 12 and 23)

Levene, P. A. (1917). The structure of yeast nucleic acid. *Journal of Biological Chemistry*, 31(3):591–598. (page 5)

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldestein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M.-A., Soltis, P. S., Xu, X., Yang, H., and Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333. (page 1)

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20. (page 8)

Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, 69(6):905–914. (page 16)

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, 3(9):662–673. (page 15)

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365. (page 15)

Li, N., Ye, M., Li, Y., Yan, Z., Butcher, L. M., Sun, J., Han, X., Chen, Q., Wang, J., and Others (2010). Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, 52(3):203–212. (page 20)

Liang, F., Liu, C., and Carroll, R. (2010). *Advanced Markov chain Monte Carlo methods: learning from past samples*. John Wiley & Sons. (pages 55 and 82)

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., and Others (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293. (page 24)

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271):315–322. (page 20)

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer. (pages 47 and 53)

Lock, E. F. and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics*, 29(20):2610–2616. (page 124)

Lock, L. F., Takagi, N., and Martin, G. R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1):39–46. (page 15)

Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., Chan, L. L., Lam, V., So, W.-Y., Wang, Y., Lok, S., Wang, J., Ma, R., Tsui, S., Chan, J., Chan, T.-F., and Yip, K. Y. (2014). Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome biology*, 15(7):408. (pages 75 and 121)

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19(4):474–482. (page 65)

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, 389(6648):251–260. (page 11)

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5(2122). (pages 92 and 97)

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067. (page 56)

Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., Bui, B., Sejnowski, T. J., Harkins, T. T., Mukamel, E. A., Behrens, M. M., and Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351):600–604. (pages 22, 101, and 118)

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., and Voet, T. (2015). G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522. (pages 24 and 90)

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends in Genetics*, 33(2):155–168. (page 24)

Mackay, D. J. C. (1999). Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068. (page 35)

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press. (pages 43 and 56)

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214. (page 19)

Madhani, H. D., Francis, N. J., Kingston, R. E., Kornberg, R. D., Moazed, D., Narlikar, G. J., Panning, B., and Struhl, K. (2008). Epigenomics: a roadmap, but to where? *Science*, 322(5898):43–44. (page 11)

Mann, M. and Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology*, 21(3):255–261. (page 8)

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141. (pages 16 and 17)

Marin, J. M., Mengersen, K., and Robert, C. P. (2005). Bayesian Modelling and Inference on Mixtures of Distributions. *Handbook of Statistics*, 25:459–507. (page 44)

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq : An assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517. (page 17)

Marioni, R. E., Shah, S., McRae, A. F., Chen, B. H., Colicino, E., Harris, S. E., Gibson, J., Henders, A. K., Redmond, P., Cox, S. R., and Others (2015). DNA methylation age of blood predicts all-cause mortality in later life. *Genome biology*, 16(1):25. (page 123)

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Embryogenesis: demethylation of the zygotic paternal genome. *Nature*, 403(6769):501–502. (page 15)

Mayo, T. R., Schweikert, G., and Sanguinetti, G. (2015). M 3 D: A kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics*, 31(6):809–816. (pages 21, 59, and 60)

Mazutis, L., Gilbert, J., Ung, W. L., Weitz, D. A., Griffiths, A. D., and Heyman, J. A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. *Nature protocols*, 8(5):870. (page 19)

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. CRC press, 2nd edition. (pages 28 and 30)

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions.* John Wiley & Sons, 2nd edition. (page 44)

McLachlan, G. and Peel, D. (2004). *Finite mixture models.* John Wiley & Sons. (pages 44, 46, 65, and 103)

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering.* Applied Statistics. (page 44)

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877. (page 20)

Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770. (page 20)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092. (pages 47, 53, and 54)

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341. (pages 47 and 51)

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46. (page 16)

Miescher-Rüsch, F. (1871). *Ueber die chemische Zusammensetzung der Eiterzellen [On the chemical composition of pus cells].* (page 5)

Min, I. M., Waterfall, J. J., Core, L. J., Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J., and Lis, J. T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes & development*, 25(7):742–754. (page 72)

Minka, T. P. (1999). Expectation Propagation for Approximate Bayesian Inference. *Statistics*, 17(2):362–369. (page 47)

Mitchell, P. J. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–378. (page 8)

Miura, F., Enomoto, Y., Dairiki, R., and Ito, T. (2012). Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic acids research*, 40(17):e136. (page 22)

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–19. (page 8)

Mohandas, T., Sparkes, R. S., and Shapiro, L. J. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*, 211(4480):393–396. (page 15)

Moran, S., Arribas, C., and Esteller, M. (2016). Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, 8(3):389–399. (pages 23 and 80)

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628. (page 17)

Mukherjee, S. (2017). *The gene: An intimate history.* Bodley Head. (page 10)

Müller, F. (2016). *Analyzing DNA methylation signatures of cell identity.* PhD thesis, Saarland University. (pages 12 and 18)

Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J., and Adey, A. C. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nature Biotechnology*, 36(5):428—-431. (pages 22 and 101)

Murphy, K. P. (2012). *Machine Learning: A probabilistic perspective.* The MIT Press. (pages 30, 34, 35, 37, and 63)

Mutskov, V. J., Farrell, C. M., Wade, P. A., Wolffe, A. P., and Felsenfeld, G. (2002). The barrier function of an insulator couples high histone acetylation levels with specific protection of promoter DNA from methylation. *Genes & development*, 16(12):1540–1554. (page 16)

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64. (page 24)

Nagy, A., Rossant, J., Nagy, R., Abramow-Newerly, W., and Roder, J. C. (1993). Derivation of completely cell culture-derived mice from early-passage embryonic stem cells. *Proceedings of the National Academy of Sciences*, 90(18):8424–8428. (page 91)

Nanney, D. L. (1958). Epigenetic control systems. *Proceedings of the National Academy of Sciences*, 44(7):712–717. (page 10)

Nature Editorial (2016). Daunting data - The power of big data must be harnessed for medical progress. *Nature*, 539:467–468. (page 1)

Nau, R. F. (2001). De Finetti was right: probability does not exist. *Theory and Decision*, 51(2-4):89–124. (page 31)

Neal, R. M. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. *Technical Report*, 1:1–144. (page 51)

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265. (page 44)

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384. (page 29)

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., and Others (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920. (page 67)

Nocedal, J. and Wright J., S. (1999). *Numerical Optimization*, volume 2nd. Springer. (pages 46 and 64)

Norris, J. R. (1998). *Markov chains.* Cambridge university press, 2nd edition. (page 52)

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769. (page 28)

Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H., and Nakatani, Y. (1996). The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell*, 87(5):953–959. (page 9)

O'Hagan, A. (1987). Monte Carlo is fundamentally unsound. *The Statistician*, 36(2):247–249. (page 47)

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257. (page 14)

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics*, 19(3):219–220. (page 14)

Ong, C.-T. and Corces, V. G. (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 12(4):283–293. (page 9)

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. *Current Biology*, 10(8):475–478. (page 15)

Paisley, J., Blei, D., and Jordan, M. (2012). Variational Bayesian inference with stochastic search. *arXiv*, 1206.6430. (page 50)

Palazzo, A. F. and Gregory, T. R. (2014). The case for junk DNA. *PLoS genetics*, 10(5):e1004351. (page 7)

Parisi, G. (1988). *Statistical field theory*. Addison-Wesley. (pages 47 and 50)

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–80. (page 19)

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., and Others (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, 110(44):17921–17926. (page 9)

Pastor, W. A., Pape, U. J., Huang, Y., Henderson, H. R., Lister, R., Ko, M., McLoughlin, E. M., Brudno, Y., Mahapatra, S., Kapranov, P., and Others (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, 473(7347):394–397. (page 15)

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–401. (page 18)

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419. (page 18)

Pearl, J. (1988). Probabilistic reasoning in intelligent systems: Networks of plausible reasoning. (pages 39 and 40)

Pennisi, E. (2012). ENCODE project writes eulogy for junk DNA. *Science*, 337(6099):1159–1161. (page 7)

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature Methods Supplement*, 6(11):S22–S32. (page 17)

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, 9(1):171. (page 90)

Pierson, E. and Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16(241):1–10. (pages 19 and 29)

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21(3):447–455. (page 23)

Ponting, C. P. and Hardison, R. C. (2011). What fraction of the human genome is functional? *Genome research*, 21(11):1769–1776. (page 7)

Ponts, N., Harris, E. Y., Prudhomme, J., Wick, I., Eckhardt-Ludka, C., Hicks, G. R., Hardiman, G., Lonardi, S., and Le Roch, K. G. (2010). Nucleosome landscape and control of transcription in the human malaria parasite. *Genome research*, 20(2):228–238. (page 23)

Pott, S. (2017). Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife*, 6:e23203. (pages 24 and 90)

Powers, D. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63. (page 108)

Pradhan, S., Bacolla, A., Wells, R. D., and Roberts, R. J. (1999). 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. (page 14)

Prendergast, G. C. and Ziff, E. B. (1991). Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science*, 251(4990):186–189. (page 9)

Ptashne, M. (2013). Epigenetics: Core misconcept. *Proceedings of the National Academy of Sciences*, 110(18):7101–7103. (pages 11 and 13)

Ptashne, M. and Gann, A. (2002). *Genes and Signals*. Cold Spring Harbor Laboratory Press. (page 8)

Ptashne, M., Hobert, O., and Davidson, E. (2010). Questions over the scientific basis of epigenome project. *Nature*, 464(7288):487. (page 11)

Qin, L.-X. and Self, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):526–533. (page 65)

Rackham, O. J. L., Langley, S. R., Oates, T., Vradi, E., Harmston, N., Srivastava, P. K., Behmoaras, J., Dellaportas, P., Bottolo, L., and Petretto, E. (2017). A Bayesian approach for analysis of whole-genome bisulfite sequencing data identifies disease-associated changes in DNA methylation. *Genetics*, 205(4):1443–1458. (pages 21 and 59)

Rands, C. M., Meader, S., Ponting, C. P., and Lunter, G. (2014). 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS genetics*, 10(7):e1004525. (page 7)

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. (page 50)

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian Processes for Machine Learning*. MIT Press. (page 28)

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., and Others (2017). Science forum: the human cell atlas. *Elife*, 6:e27041. (page 1)

Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21. (page 15)

Richards, J. E. and Hawley, R. (2011). *The central dogma of molecular biology: how cells orchestrate the use of genetic information*. The Human Genome. (pages 6, 8, and 121)

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 59(4):731–792. (page 44)

Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*, 14(1):9–25. (pages 10 and 14)

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 3(22):41–46. (page 39)

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature communications*, 9(284):1–17. (page 19)

Robert, C. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer. (pages 47, 51, 54, and 57)

Robinson, M. D., Kahraman, A., Law, C. W., Lindsay, H., Nowicka, M., Weber, L. M., and Zhou, X. (2014). Statistical methods for detecting differentially methylated loci and regions. *Frontiers in genetics*, 5(324):1–7. (page 21)

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140. (page 18)

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):R25. (page 18)

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., and Others (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393. (page 19)

Rozowsky, J., Euskirchen, G., Auerbach, R. K., Zhang, Z. D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M. B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75. (page 19)

Russo, V. E. A., Martienssen, R. A., and Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press. (page 10)

Ruthenburg, A. J., Allis, C. D., and Wysocka, J. (2007). Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Molecular cell*, 25(1):15–30. (page 13)

Sabatti, C. and James, G. M. (2005). Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746. (page 33)

Sanger, F., Nicklen, S., and Coulson, a. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467. (page 16)

Sanguinetti, G., Lawrence, N. D., and Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781. (page 33)

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801. (page 31)

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417. (page 14)

Scarano, E., Iaccarino, M., Grippo, P., and Parisi, E. (1967). The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. *Proceedings of the National Academy of Sciences*, 57(5):1394–1400. (page 14)

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press. (page 64)

Schübeler, D. (2015). Function and information content of DNA methylation. *Nature*, 517(7534):321–326. (pages 14 and 59)

Schultz, M. D., Schmitz, R. J., and Ecker, J. R. (2012). 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. *Trends in Genetics*, 28(12):583–585. (pages 20 and 21)

Schwartzman, O. and Tanay, A. (2015). Single-cell epigenomics: techniques and emerging applications. *Nature Reviews Genetics*, 16(12):716–726. (pages 17, 22, and 77)

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. (page 37)

Schweikert, G., Cseke, B., Clouaire, T., Bird, A., and Sanguinetti, G. (2013). MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. *BMC genomics*, 14(1):826. (page 19)

Selega, A. (2018). *Computational methods for RNA integrative biology.* PhD thesis, The University of Edinburgh. (page 6)

Sexton, T. and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, 160(6):1049–1059. (page 12)

Shachter, R. D. (1998). Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the 14th conference on uncertainty in artificial intelligence*, pages 480–487. Morgan Kaufmann Publishers Inc. (page 41)

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630. (page 17)

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145. (page 16)

Si, Y., Liu, P., Li, P., and Brutnell, T. P. (2013). Model-based clustering for RNA-seq data. *Bioinformatics*, 30(2):197–205. (page 18)

Siegmund, K. D. (2011). Statistical approaches for the analysis of DNA methylation microarray data. *Human Genetics*, 129(6):585–595. (page 80)

Silva, A. J., Ward, K., and White, R. (1993). Mosaic methylation in clonal tissue. *Developmental biology*, 156(2):391–398. (page 14)

Smallwood, S. a., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–20. (pages 22, 77, 101, 110, 111, 115, 116, 117, 147, 148, 149, and 150)

Smiraglia, D. J., Rush, L. J., Frühwald, M. C., Dai, Z., Held, W. A., Costello, J. F., Lang, J. C., Eng, C., Li, B., Wright, F. A., and Others (2001). Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Human molecular genetics*, 10(13):1413–1419. (page 21)

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1):1–25. (page 35)

Song, J. J., Lee, H.-J., Morris, J. S., and Kang, S. (2007). Clustering of time-course gene expression data using functional data analysis. *Computational biology and chemistry*, 31(4):265–274. (page 65)

Speybroeck, L. (2002). From epigenesis to epigenetics - The Case of C. H. Waddington. *Annals of the New York Academy of Sciences*, 981(1):61–81. (page 10)

Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626. (page 8)

Sproul, D., Nestor, C., Culley, J., Dickson, J. H., Dixon, J. M., Harrison, D. J., Meehan, R. R., Sims, A. H., and Ramsahoye, B. H. (2011). Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. *Proceedings of the National Academy of Sciences*, 108(11):4364–4369. (page 21)

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., and Others (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480:490–495. (page 21)

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17(3):355–367. (pages 88 and 122)

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(January 2014):133–145. (pages 17 and 19)

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474. (page 29)

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press. (page 31)

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842. (page 65)

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779. (page 1)

Swendsen, R. H. and Wang, J.-S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical review letters*, 58(2):86–88. (page 42)

Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J., and Jones, P. A. (2014). Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome research*, 24(9):1421–1432. (page 23)

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872. (page 9)

Takahashi, K. and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676. (page 9)

Tamaru, H. and Selker, E. U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. *Nature*, 414(6861):277–283. (page 16)

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., and Others (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382. (page 18)

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540. (pages 42, 55, and 82)

Tanner, M. A. and Wong, W. H. (2010). From EM to data augmentation: the emergence of MCMC Bayesian computation in the 1980s. *Statistical science*, 25(4):506–516. (pages 42 and 47)

Teschendorff, A. E., Zhuang, J., and Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics*, 27(11):1496–1505. (page 23)

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110(46):18602–18607. (page 19)

Thienpont, B., Steinbacher, J., Zhao, H., D'Anna, F., Kuchnio, A., Ploumakis, A., Ghesquière, B., Van Dyck, L., Boeckx, B., Schoonjans, L., and Others (2016). Tumour hypoxia causes DNA hypermethylation by reducing TET activity. *Nature*, 537(7618):63–68. (page 15)

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., and Others (2012). The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82. (page 23)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288. (page 34)

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728. (page 53)

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3):562–578. (page 67)

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375. (page 18)

Tsompana, M. and Buck, M. J. (2014). Chromatin accessibility: a window into the genome. *Epigenetics & chromatin*, 7(33):1–16. (page 23)

Turner, B. M. (2000). Histone acetylation and an epigenetic code. *Bioessays*, 22(9):836–845. (page 13)

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS computational biology*, 11(6):e1004333. (pages 19, 33, and 35)

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods*, 14(6):565–571. (pages 19 and 33)

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50. (page 41)

Vanderkraats, N. D., Hiken, J. F., Decker, K. F., and Edwards, J. R. (2013). Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes. *Nucleic Acids Research*, 41(14):6816–6827. (page 60)

Varley, K. E., Gertz, J., Bowling, K. M., Parker, S. L., Reddy, T. E., Pauli-behn, F., Cross, M. K., Williams, B. a., Stamatoyannopoulos, J. a., Crawford, G. E., Absher, D. M., Wold, B. J., and Myers, R. M. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome research*, 23(3):555–567. (pages 14, 59, and 75)

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., and Others (2001). The sequence of the human genome. *Science*, 291(5507):1304–1351. (page 16)

Villota-Salazar, N. A., Mendoza-Mendoza, A., and González-Prieto, J. M. (2016). Epigenetics: from the past to the present. *Frontiers in Life Science*, 9(4):347–370. (page 10)

Voigt, P., Tee, W.-W., and Reinberg, D. (2013). A double take on bivalent promoters. *Genes & development*, 27(12):1318–1338. (page 13)

Waddington, C. H. (1939). *An introduction to modern genetics.* Routledge. (page 10)

Waddington, C. H. (1942). The epigenotype. *Endeavour*, 1:18–20. (page 10)

Waddington, C. H. (1957). *The strategy of the genes.* Allen and Unwin, London, 1st edition. (pages 10 and 11)

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145. (page 125)

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63. (page 17)

Waterland, R. A. and Jirtle, R. L. (2003). Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Molecular and cellular biology*, 23(15):5293–5300. (page 15)

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356):737–738. (page 5)

Welch, J. D., Hartemink, A. J., and Prins, J. F. (2017). MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology*, 18(1):138. (page 124)

Whitaker, J. W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. *Nature Methods*, 12(3):265–272. (pages 16 and 125)

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319. (page 9)

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6(Apr):661–694. (page 50)

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390. (page 28)

Woodcock, C. L. and Ghosh, R. P. (2010). Chromatin higher-order structure and dynamics. *Cold Spring Harbor perspectives in biology*, 2(5):a000596. (page 11)

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103. (page 66)

Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., and E (2010). Dnmt3a-Dependent Nonpromoter DNA Methylation Facilitates Transcription of Neurogenic Genes. *Science*, 329:444–448. (page 75)

Xu, J., Pope, S. D., Jazirehi, A. R., Attema, J. L., Papathanasiou, P., Watts, J. A., Zaret, K. S., Weissman, I. L., and Smale, S. T. (2007). Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 104(30):12377–12382. (page 21)

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but Did It Work?: Evaluating Variational Inference. *arXiv*, 1802.02538. (page 47)

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., and Others (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142. (pages 18 and 29)

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829. (page 17)

Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. (2015). Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome biology*, 16(1):14. (pages 109 and 119)

Zhang, Y., Liu, T., Meyer, C. a., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9):R137. (page 19)

Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., and Others (2018). Single-cell DNA methylome sequencing of human preimplantation embryos. *Nature genetics*, 50(1):12–19. (page 22)

Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., and Others (2013). Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477–481. (page 20)

Index