Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht - Karls - University
Heidelberg

Presented by

M.Sc. Max Frank
born in: Vienna, Austria
Oral examination: 8th May, 2023

Modeling epigenetic heterogeneity across time and genome in single-cell multi-omics experiments

Referees: Prof. Dr. Henrik Kaessmann
Dr. Arnaud Krebs

"*Develop a habit of pushing yourself a bit out of your comfort zone every day.*"

Dave MacLeod

This work was carried out at the European Molecular Biology Laboratory in Heidelberg from October 2018 to January 2024 under the supervision of Dr. Oliver Stegle.

Abstract

The genomic sequence of an organism is nearly identical in all its cells and over its lifetime. Epigenomic marks, however, such as DNA methylation and chromatin accessibility, are subject to drastic changes across different tissues and throughout organism development. Recent advancements, notably the development of multi-omics single-cell technologies, allow for simultaneous interrogation of DNA methylation, chromatin accessibility, and transcriptomes within individual cells. This offers unique opportunities to gain insight into mechanisms by which the epigenome shapes gene expression and influences cell fate. However, analyzing these datasets poses major challenges: Typically, smaller numbers of cells can be assayed per experiment than conventional single-cell RNAseq with lower coverage due to small amounts of input material. This means that classical statistical methods are underpowered to detect subtle changes in DNA methylation and chromatin accessibility. Furthermore, current tests can only detect differences between discrete and pre-defined cell populations, whereas single-cell approaches allow for studying continuous processes in organismal lineage development.

To address this, I propose computational methods for decomposing single-cell epigenetic heterogeneity across developmental time and genomic loci. This thesis introduces new concepts, leveraging pseudotemporal ordering of cells to conduct statistical inferences upon epigenetic changes. At the core of these developments is GPmeth, a Gaussian process framework designed to model highly sparse single-cell methylation and accessibility information by enforcing smooth variation across pseudotime and genomic coordinates and thus effectively sharing information between cells and genomic positions. Importantly, this model does not rely on averaging methylation signals across fixed genomic windows but can identify differentially methylated/accessible regions in a data-driven way. Testing GPmeth against other models without dynamic aggregation of methylation data revealed increased sensitivity to detect even subtle epigenetic changes.

Application of GPmeth to scNMT-seq data from mouse embryonic stem cells undergoing gastrulation revealed over 3000 enhancer elements that exhibited dynamic changes in chromatin accessibility or DNA methylation rates during germ layer formation. The detailed spatiotemporal model allowed for a precise definition of differentially methylated regions, validated by transcription factor binding motif analysis. Furthermore, the clustering of temporal epigenetic patterns identified lineage-specific enhancers in an unsupervised manner.

I expect GPmeth to be a valuable tool for studying time-resolved epigenetic regulation in several emerging multimodal single-cell datasets.

Zusammenfassung

Die Genomsequenz eines Organismus ist in allen seinen Zellen und über sein gesamtes Leben hinweg nahezu identisch. Epigenomische Marker wie DNA-Methylierung und die Zugänglichkeit von Chromatin variiren jedoch drastisch zwischen verschiedenen Geweben und während der Entwicklung des Organismus. Jüngste Fortschritte, insbesondere die Entwicklung von Multi-Omics-Einzelzelltechnologien, ermöglichen die gleichzeitige Messung von DNA-Methylierung, Chromatin-Zugänglichkeit und Genexpression innerhalb einzelner Zellen. Dies bietet neue Möglichkeiten, Einblicke in die Mechanismen zu gewinnen, durch die das Epigenom die Genexpression prägt und das die Entwicklung von Zellen beeinflusst. Die Analyse dieser Datensätze stellt jedoch große Herausforderungen dar: Verglichen mit herkömmlichem Einzelzell-RNAseq, kann typischerweise pro Experiment eine geringere Anzahl von Zellen mit geringerer Abdeckung untersucht werden. Dies bedeutet, dass klassische statistische Methoden zum Testen von DNA-Methylierungs- und Chromatin-Zugänglichkeitsunterschieden nicht ausreichen, um subtile Veränderungen zu erkennen. Dazu kommt, dass aktuelle Tests nur Unterschiede zwischen diskreten und vordefinierten Zellpopulationen testen, während Einzelzellansätze die Untersuchung kontinuierlicher Prozesse der Entwicklung der Abstammungslinie von Organismen ermöglichen.

Deshalb führe ich hier rechnerische Methoden zur Zerlegung der epigenetischen Heterogenität einzelner Zellen über die Entwicklungszeit und die genomischen Loci ein. Diese Arbeit stellt neue Konzepte vor, die die pseudotemporale Ordnung von Zellen nutzen, um statistische Rückschlüsse auf epigenetische Veränderungen zu ziehen. Im Mittelpunkt dieser Entwicklungen steht GPmeth, ein Gaußsches Prozess-Framework, das darauf ausgelegt ist, äußerst spärliche Einzelzell-Methylierungs- und Chromatin Zugänglichkeitsinformationen zu modellieren, indem eine kontinuierliche Variation über Pseudozeit und Genomkoordinaten hinweg erzwungen wird, und so, Informationen effektiv über Zellen und Genompositionen hinweg ausgetauscht werden. Wichtig ist, dass dieses Modell keine festgesetzten Genomfenster vorraussetzt, sondern differenziell methylierte/zugängliche Regionen auf datengesteuerte Weise identifizieren kann. Im Vergleich zu anderen Modellen ohne dynamische Aggregation von Methylierungsdaten, hat GPmeth erhöhte Sensitivität zur Identifikation subtiler epigenetischer Veränderungen.

Die Anwendung von GPmeth auf scNMT-seq-Daten aus embryonalen Stammzellen von Mäusen während des Gastrulationsprozesses, ergab über 3000 Enhancer-Elemente, die dynamische Veränderungen in der Zugänglichkeit von Chromatin oder den DNA-Methylierungsraten zeigten. Das detaillierte räumlich-zeitliche Modell ermöglichte eine präzise Definition unterschiedlich methylierter Regionen, validiert durch die

Analyse von Transkriptionsfaktor-Bindungsmotiven. Darüber hinaus identifizierte die Clusteranalyse der Modell-Resultate bekannte Abstammungsspezifische Enhancer.

Ich erwarte, dass GPmeth ein wertvolles Werkzeug zur Untersuchung der zeitaufgelösten epigenetischen Regulation in mehreren neu entstehenden multimodalen Einzelzelldatensätzen sein wird.

Acknowledgments

Writing this thesis was a *struggle*. So when I say it would not have been possible without the help of others, this is not just a platitude.

I would like to thank my supervisor, Oliver Stegle, for introducing me to the world of Gaussian processes, and for giving me the time to finish this project when I needed more than we thought.

I also want to thank my TAC members Arnaud Krebs, Henrik Kaessmann, and Judith Zaugg for the great feedback and advice throughout my PhD.

Thank you, Danila, for the great discussions and your unwavering enthusiasm for Science. You're an inspiration for everyone around you.

I also want to thank the rest of the EMBL regulars: Luca, Danai, Danila, and Manu. Coffee in the Stegle corner was never boring because of you.

Thank you to Hana, my main collaborator on the Sarcoma project. Even though we were both a bit out of our element here, I really enjoyed working on a project together. You made me realize that I only want to work on shared projects in the future. Also, thank you to Sophie and Kendra for being excellent collaborators and Caroline for being an exceptional student.

Thank you, Ricard, for our collaborations at the beginning of my PhD and when teaching courses at EMBL.

I am very grateful to my fellow PhD students. Paul, thank you for making the Predoc course more fun. Let's look at some genomes soon. Elisa, thank you for being such a great scientist and friend. Anna, for forcing me to party at least sometimes. Gilberto, for the philosophy/science conversations at those parties. Ana, for the dinners, the football, and the running. Jesus and Henrik for the dinners. The SOLA team, Fergus, Wolfi, Kai, Ale, Vale, and Filipe.

I am grateful for the bouldering crew, Thorsten, Dewi, Ed, and Higor. Climbing is the thing that kept me sane towards the end of Thesis writing.

Thank you to Thorsten and Luca; our training during the lockdown was by far the best part of the whole Covid thing. Amazing what a set of rings can do.

Thank you to all members of the Tschanz family for all the nice holidays during my PhD, no matter if in Greece or Niederhelfenschwil.

I want to thank my amazing brother and sister Paul and Hannah. Paul, you are the humblest smart person I know. Thank you for our gaming sessions and late-night talks. Hannah, thank you for being a joy to be around. I can't wait to see what you do in the future.

To my parents, Anita and Mike, I don't know how to thank you properly. You have always been there for me without a thought. I think I appreciate this more and more the older I get. Anita, thank you for always believing in me and helping me find motivation during tough times. Mike, you are the reason I'm a scientist. Thank you for teaching me to strive to understand things deeply.

And lastly, Aline, my partner. You make my life better in every way. And you gave me the reason and the strength to finish this PhD. I am sorry it took so long. Thank you for everything. This thesis is for you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ARD	Automatic relevance determination
ASM	Allele specific methylation
ATAC-seq	Assay for transposase-accessible chromatin sequencing
AUC	Area under the ROC Curve
BF	Bayes factor
BH	Benjamini-Hochberg
BS-seq	Bisulfite sequencing
BY	Benjamini–Yekutieli
CCA	Canonical correlation analysis
CI	Confidence interval
CITE-seq	Cellular indexing of transcriptomes and epitopes by sequencing
CW	Change window
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferases
DPT	Diffusion pseudotime
ELBO	Evidence lower bound
FDR	False discovery rate
GO	Gene ontology
GP	Gaussian process
KL	Kullback-Leibler
LLR	Log-likelihood ratio
LOWESS	Locally weighted scatterplot smoothing
MELISSA	MEthyLation Inference for Single-cell Analysis
MMRC	Maximum methylation rate change
MOFA	Multi-omics factor analysis
NOME-seq	Nucleosome occupancy and methylome sequencing
PBAT	Post-bisulfite adaptor tagging
PCA	Principle component analysis
PCR	Polymerase chain reaction
RBF	Radial basis function
RNA	Ribonucleic acid

ROC	Receiver operating characteristic curve
SE	Squared exponential
TF	Transcription factor
TSS	Transcription start site
UMAP	Uniform manifold approximation and projection
VI	Variational inference
WNN	Weighted nearest neighbors

1 Introduction

The advent of massively parallel sequencing technologies over the last two decades has allowed researchers to study the genomic and transcriptomic landscape of many organisms in increasing detail (Metzker, 2010; Reuter *et al.*, 2015). Being able to measure gene expression of all genes, or to call all mutations at once, has turned biology into a data science discipline (Wang *et al.*, 2009). The extension of genome-wide sequencing to single-cell technologies added an additional dimension of cellular state to these experiments. It allowed us to look at organisms and cells as dynamic systems with complex interplay where any changes to the system can have surprising emergent properties. One of the ultimate goals of systems biology is to understand how all components of a system work together to produce a phenotype (Aderem, 2005).

In this context, a system of interest can be whole organisms, specialized tissues, or individual cells. Since cells are the fundamental building blocks of all higher organisms, understanding their inner workings is crucial to biology as a field of research. One of the guiding frameworks that biologists use to think about cellular systems is called the central dogma of molecular biology (Crick, 1970). It states that there is a flow of information from DNA to RNA to protein. This follows the molecular pathway in which genes are transcribed into messenger RNAs (mRNA) in the cell nucleus, which are then transported to the cytoplasm and translated into amino acid chains that fold into mature proteins. Given a full understanding of the system, it should be possible to predict RNA expression from DNA sequence and protein abundance from RNA expression, ultimately determining the system's phenotypic characteristics. While this might seem like a simple task at first glance, countless aspects of biological systems make it more difficult to generate accurate predictions about them (Kim and Wysocka, 2023). For example, the same DNA stretch can encode multiple proteins due to alternative splicing and posttranslational modifications (Pan *et al.*, 2008). Predicting RNA expression from sequence information is complicated by the fact that transcription, the process of transcribing DNA into messenger RNA molecules, is regulated by a myriad of factors that act on the DNA in the nucleus (Lee and Young, 2013). In fact, it turns out that the flow of information in cells is not linear from DNA to protein but cyclical since proteins called transcription factors (TFs) play a critical role in controlling transcription (Reményi *et al.*, 2004).

Furthermore, DNA contains more information than is encoded purely in its sequence. DNA molecules in a cell carry chemical modification that can be inherited through cell divisions and even across generations. One of the key modifications is DNA methylation, which modifies cytosine, one of the four bases that make up DNA. Additionally, the spatial organization of DNA in the nucleus plays a major regulatory role. DNA is organized into chromatin, a complex of a DNA molecule wrapped around eight histone proteins forming nucleosomes. The density of packing nucleosomes together directly impacts the accessibility of the DNA and, therefore, transcription (Venkatesh and Workman, 2015). Furthermore, nucleosomes can also carry chemical modifications, which influence transcription in both direct and indirect ways (Millán-Zambrano *et al.*, 2022).

Taken together, these modes of gene regulation that are not a direct result of DNA sequence are called epigenetics. Epigenetic information is a key factor that determines cell identity. Cells can transmit this information to daughter cells during mitosis, allowing them to form coordinated groups of cells that make up tissues and organs. It also plays a key role in establishing cell identities during the development of organisms, where cells differentiate to fulfill different roles.

In this work, I will introduce a modeling approach to track DNA methylation and chromatin accessibility that is continuously changing during organism development. My approach makes use of recent advances in single-cell multimodal sequencing technology (Clark *et al.*, 2018) that enable the measurement of gene expression and epigenetics in the same cell. I apply my modeling framework to study how DNA methylation and chromatin accessibility change during embryonic development and to assess their impact on gene regulation.

1.1 Epigenetics and gene regulation

Mammalian cells carry out their functions using molecular machines called proteins. In mice, the amino acid sequences of these proteins are encoded by an estimated 25,000 genes (Blake *et al.*, 2020). Cells can have drastically different relative protein abundances depending on cell type, tissue of origin, and cellular environment (Gianansanti *et al.*, 2022). A protein's abundance is a direct result of the rate of production (translation) of the protein and the rate of degradation over time. The rate of protein translation is influenced by the rate of expression (transcription) of its corresponding gene. Gene expression is tightly controlled by the cell to ensure its proper function.

1.1.1 Principles of gene regulation

For a gene to be transcribed into mRNA, RNA polymerase II has to be recruited to the transcription start site (TSS) of that gene (Sainsbury *et al.*, 2015). This typically happens in conjunction with different co-factors, forming a so-called transcription pre-initiation complex (Fig 1.1). This protein complex binds to the promoter, a

short region of DNA that is located in close proximity to the TSS. Typically, genes have a single promoter whose sequence can influence the transcription rate of a gene. However, gene regulation in eukaryotes typically involves additional genomic elements that can be located further away from the TSS (Panigrahi and O'Malley, 2021.) These elements include enhancers, silencers, and insulators. Their mode of action involves the recognition of their sequence by DNA-binding proteins called TFs (Kim and Wysocka, 2023). Although these elements are distal in sequence space, the 3D organization of DNA puts them in close spatial proximity to the promoter of a regulated gene. Thus, TFs bound to an enhancer can form a physical protein complex with proteins bound to promoter elements. Because of the dynamics of DNA organization in the nucleus, one enhancer can regulate multiple genes, and a single gene can be under the control of multiple enhancers.

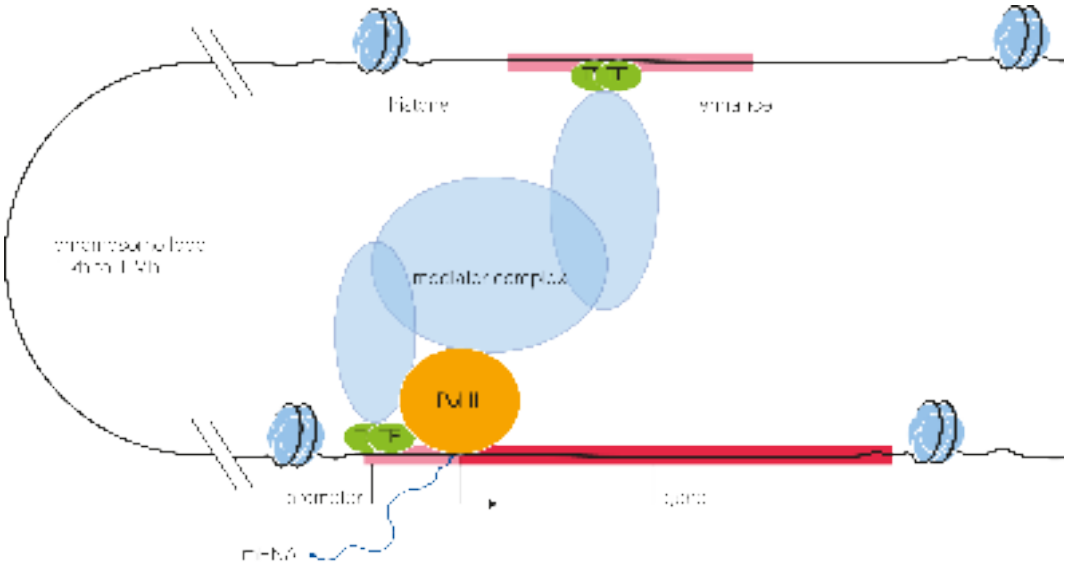


Figure 1.1 | Transcription initiation complex. Mammalian transcription is carried out by RNA Polymerase II (Pol II). In most cases, Pol II requires the binding of TFs to regulatory elements. These TFs can bind close to the transcription start site (TSS) of the gene to promoters or to distal regulatory elements called enhancers. Bound TFs and different co-factors form the pre-initiation complex that is essential for regulating transcription. *Figure generated by Max Frank.*

Gene-enhancer associations are highly cell-type specific and are, therefore, likely to have major roles in establishing cell identity. Mapping the network of gene-enhancer links and understanding how this network is regulated is key to understanding organism development. As mentioned above, TF binding is necessary to induce transcription. TFs possess DNA binding domains that recognize 6-10 base pair (bp) long DNA sequences called binding motifs (Lambert *et al.*, 2018). About 8% of human genes are estimated to function as TFs. This subset of genes can be considered a set of "master regulators" for the expression profile of the rest of the genome. TF binding is mediated not only by DNA sequence but also by the following epigenetic marks. *DNA methylation* of enhancers can have both a positive and a negative impact on

TF binding (see Section 1.1.2). The *accessibility* of the binding motif (i.e., whether a nucleosome covers the motif sequence) is a prerequisite for most TFs to bind. However, there are pioneer TFs that can bind to histone-bound DNA (see Section 1.1.3). *Histone marks* also play a role in TF binding, both directly and indirectly, by establishing the 3D architecture of DNA (see Section 1.1.4). Furthermore, many TFs do not bind DNA in isolation but require co-factors to bind cooperatively. Different epigenetic marks will be discussed in more detail in the next Sections.

1.1.2 DNA methylation

DNA methylation involves the deposition of a methyl group to the DNA (Mattei *et al.*, 2022). The most commonly studied form of DNA methylation is the addition of a methyl group to the 5th position of the pyrimidine ring of a cytosine base (C), converting it to 5-methylcytosine (5mC). In the rest of this text, DNA methylation will be used synonymously with 5mC. The conversion of C to 5mC is carried out by enzymes called methyltransferases almost exclusively in the context of CpG dinucleotides (Fig 1.2). CpGs are generally depleted in mammalian genomes, except for local exceptions called CpG islands (Bird *et al.*, 1985). While the majority (75%) of CpG sites in mammalian genomes are methylated, methylation rates are typically low in CpG islands (Moore *et al.*, 2013).

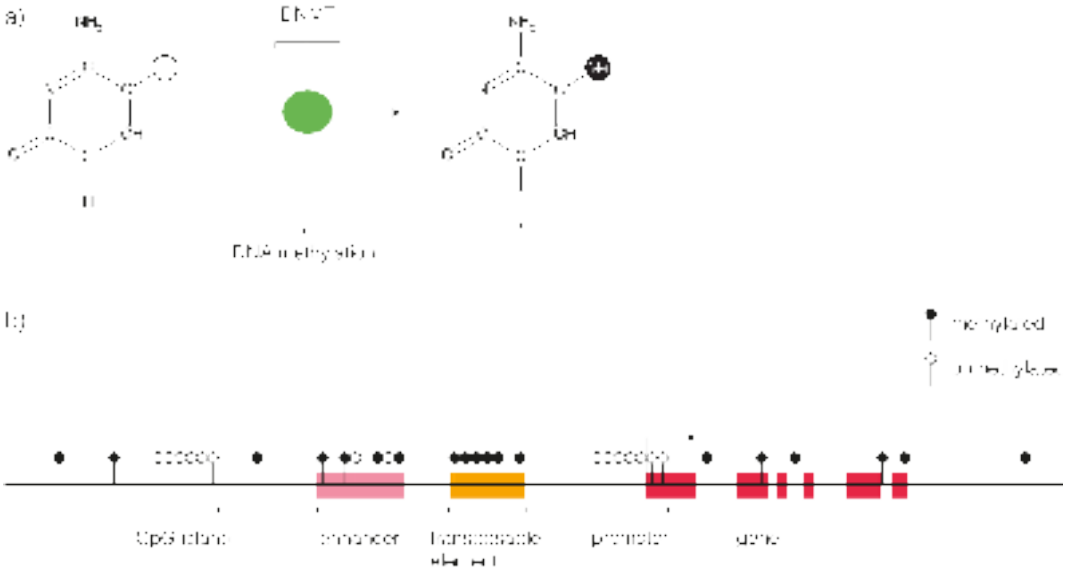


Figure 1.2 | Overview of mammalian 5mC DNA methylation. *a)* 5mC DNA methylation adds a methyl group to the carbon atom in the 5th position of the pyrimidine ring. *b)*Mammalian DNA methylation in different chromosomal contexts. CpG islands are regions in the genome with a high density of CpG sites. In enhancer and promoter regions, DNA methylation is correlated with gene expression, suggesting a regulatory role. It also plays a role in silencing otherwise harmful stretches of DNA, such as transposable elements. *Figure generated by Max Frank.*

1.1.2.1 DNA methylation in different genomic contexts

DNA methylation has major impacts on gene expression, most famously in the complete repression of transcription of inactivated X-chromosomes. In XX cells of female mammals, only one X-chromosome is transcriptionally active to ensure dosage compensation of gene products (Galupa and Heard, 2018). CpG islands on the inactivated X-chromosome are typically highly methylated, ensuring lasting repression of transcription. However, DNA methylation is not the only factor in X inactivation and might only be an additional safeguard ensuring its stability.

DNA methylation is also paramount in permanently silencing parts of the genome that would cause genomic instability if active. These areas include transposable elements and structural elements of chromosomes, such as telomeres and centromeres. Transposable elements comprise a large part of mammalian genomes and can change their position in the genome and duplicate themselves if transcribed (Pourrajab and Hekmatimoghaddam, 2021). Repressions of these elements might be the function that initially contributed to the evolution of DNA methylation as a regulatory mechanism (Yoder *et al.*, 1997).

DNA methylation is also enriched in the gene bodies of highly transcribed genes. While methylation clearly does not function as a repressor in this context, it is not fully understood what exact function it serves (Jones, 2012). One suggested role is the control of alternative splicing, supported by the fact that exons have higher methylation rates than introns. It could also be involved in preventing transcription from intragenic promoters (Dahlet *et al.*, 2020) .

Gene promoters can be roughly categorized by the presence or absence of a CpG island in their sequence (Saxonov *et al.*, 2006). In CpG-poor promoters, which make up 30-40% of most mammalian genomes, the influence of DNA methylation on transcription is unclear. Promoters containing CpG islands are generally more highly expressed than genes with CpG-poor promoters (Larsen *et al.*, 1992). However, it seems that while low methylation rates in these promoters are a requirement for high rates of transcription, it is not sufficient since transcriptionally silent genes with demethylated CpG island promoters are also frequently found. This hints at a more intricate regulatory mechanism involving enhancers and multiple TFs (Weber *et al.*, 2007).

In enhancers, DNA methylation is likely to play a regulatory role, evidenced by the fact that methylation rates are highly variable in active enhancer elements (Ziller *et al.*, 2013; Schultz *et al.*, 2015). Enhancers are unlikely to be found in CpG island regions. While most of the mammalian genome outside CpG islands is fully methylated, active enhancers tend to have intermediate methylation rates between 10-50% (Stadler *et al.*, 2011). An open question is how enhancer methylation influences the binding of TFs. Several TFs preferentially bind to unmethylated DNA, such as CTCF (Wang *et al.*, 2012b; Maurano *et al.*, 2015), CREB1 (Kaluscha *et al.*, 2022) and NRF1 (Domcke

et al., 2015). This suggests that DNA methylation can serve as a repressive mark for these TFs. However, a subset of TFs, called pioneer factors, bind to DNA in a repressed epigenetic state and initiate a cascade of events that contributes to the activation of an enhancer (Iwafuchi-Doi and Zaret, 2014). Some of these factors have been shown to bind preferentially to methylated DNA. Examples are p53 (Kribelbauer *et al.*, 2017), which is often mutated in human cancers, and Oct4 (Yin *et al.*, 2017), which is a marker for stem cells. Recent evidence in mouse embryonic stem cells showed that the majority of TFs are not sensitive to DNA methylation, therefore asking the question of whether there is a more indirect effect of DNA methylation on enhancer activity or if it is a consequence of it (Kreibich *et al.*, 2023).

1.1.3 Chromatin accessibility

In the nucleus of eukaryotic cells, DNA is organized in a complex arrangement called chromatin (Kouzarides, 2007). The central chromatin components are nucleosomes, octamers of histone proteins that DNA wraps around twice. Typically, 147bp of DNA are nucleosome-wrapped with linker regions of about 80bp (Luger *et al.*, 1997). These chains of DNA-wrapped nucleosomes can further condense into tightly packed fibers of heterochromatin that are hard to access for DNA-binding proteins (Fig 1.3). Heterochromatin is typically marked by specific histone modifications such as absence of Histone acetylation and H3K9me, as well as DNA methylation (Allshire and Madhani, 2018). Large parts of the human genome, such as centromeres and telomeres, are permanently in a heterochromatic state and, thus, called constitutive heterochromatin. Other parts of the genome are subject to active regulation and only become heterochromatin depending on cell state. Thus, they are called facultative heterochromatin. In female mammals, the inactivated X chromosome is completely heterochromatic, while the other X chromosome has more accessible stretches of DNA called open chromatin or euchromatin (Galupa and Heard, 2018). On a global scale, euchromatin makes up only 2-3% of the total genomic sequence but harbors the majority of regulatory elements that can be bound by TFs (Klemm *et al.*, 2019).

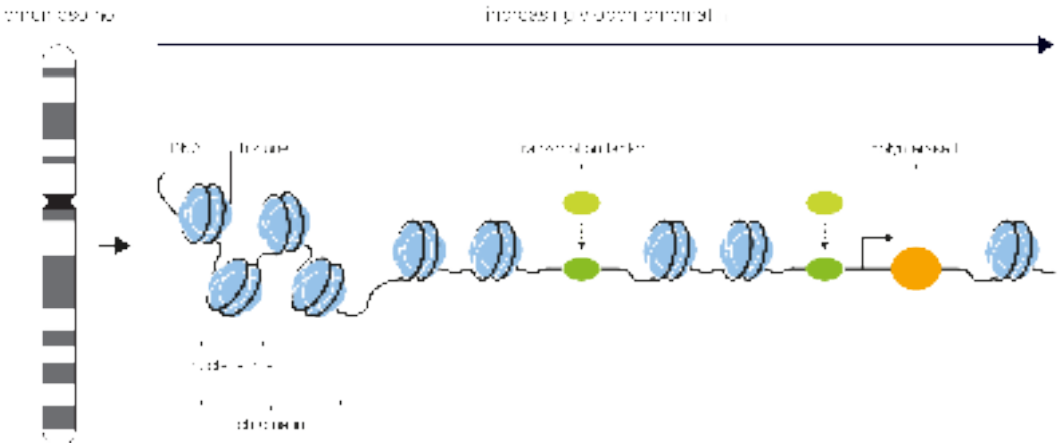


Figure 1.3 | Overview of mammalian chromatin accessibility. Mammalian DNA strands are organized as chromatin by wrapping around nucleosomes. These nucleosomes can be tightly packed, leading to inaccessible heterochromatin (*left*). Euchromatin is more loosely packed, and nucleosomes can be displaced by TFs or other binding proteins, allowing access to the DNA sequence. *Figure generated by Max Frank.*

In euchromatic regions, DNA accessibility is determined by the positioning of nucleosomes. Nucleosomes can dynamically detach and attach from DNA and are in constant competition for binding with other regulatory proteins (Jiang and Pugh, 2009). There are several mechanisms of nucleosome replacement by TFs that cause local changes in DNA accessibility (Venkatesh and Workman, 2015). The simplest mechanism is TF binding during a short period of DNA exposure due to natural nucleosome turnover (Workman, 2006). Other mechanisms involve the interaction of TFs with the nucleosomes themselves, whereby different histone modifications play a role (see below). Pioneer TFs may be able to bind nucleosome-bound DNA and displace the nucleosome. It is important to notice that all these events happen on very short timescales and with some degree of stochasticity (Lammers *et al.*, 2020). Therefore, measurements of DNA accessibility in cell populations will yield an average of all DNA states, weighted by occupancy time and strength of the different DNA binding proteins. Different methods to measure DNA accessibility will be discussed in more detail below.

1.1.4 Histone modifications

Histone Modifications refer to post-translational modifications occurring on histones within nucleosomes (Fig 1.4). These modifications can influence gene expression considerably, mostly operating indirectly by regulating the binding of TFs or chromatin remodeling proteins. The result of this regulation can both induce or inhibit gene expression. (Millán-Zambrano *et al.*, 2022). While early studies focused on modifications at the histone tail (Fig 1.1.4), more recent studies also investigate the function of such modification at the globular domains of histones (Millán-Zambrano *et al.*, 2022). Several types of histone modifications exist, each capable of eliciting drastically

different effects on gene regulation. Among these modifications are methylation, acetylation, phosphorylation, ubiquitination, and sumoylation. Notably, histone acetylation and histone methylation are the most thoroughly studied. Nomenclature for histone modification consists of defining the modified histone within the nucleosome, followed by the amino acid within the histone, and finally, the modification itself. Acetylation of the lysine at the 27th N-terminal position on histone 3, therefore, would be denoted as H3K27ac.

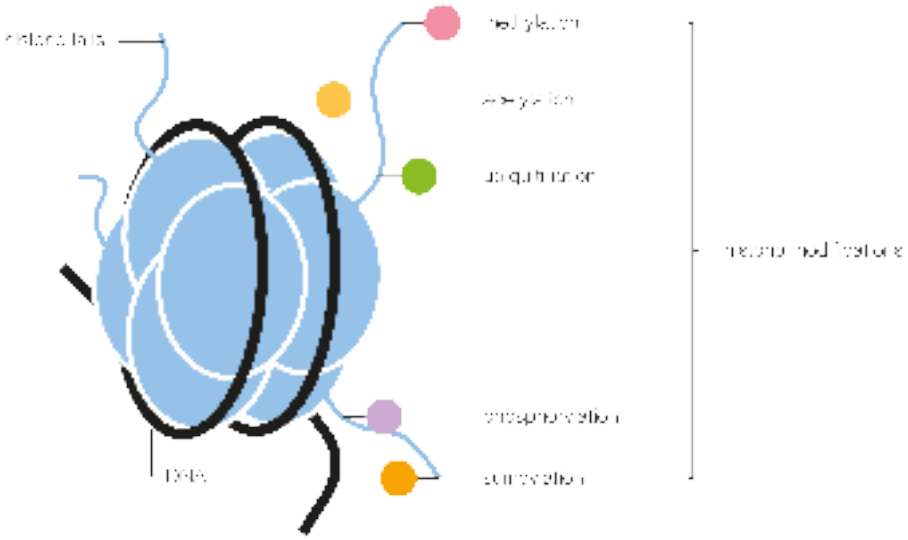


Figure 1.4 | Overview of mammalian histone modifications. Histone modifications refer to post-translational modifications of the histone within nucleosomes. Known histone modifications include methylation, acetylation, ubiquitination, phosphorylation, and sumoylation. These modifications can occur at different amino acids of the histone tails, leading to a complex modification pattern. *Figure generated by Max Frank.*

Histone methylation involves adding one or multiple methyl groups to residues, resulting in mono-, di-, or trimethylation (me1, me2, me3) (Greer and Shi, 2012). These modifications are added by histone methyltransferases (HMTs), while histone demethylases remove them (Rice *et al.*, 2003). Histone 3 lysine 4 trimethylation (H3K4me3) is a well-studied modification enriched at active gene promoters (Talbert *et al.*, 2019). While it is thought to facilitate transcription by promoting the recruitment of transcriptional machinery (Vermeulen *et al.*, 2007), its necessity for transcription remains debated (Henikoff, MillanHenikoff and Shilatifard, 2011; Millán-Zambrano *et al.*, 2022). In mammals, H3K4me3 can persist during transcriptionally quiescent states, potentially contributing to epigenetic memory and influencing gene expression patterns and developmental capacity in embryos (Zhang *et al.*, 2016).

Additional modifications include H3K4me1, which is associated with enhancers, and H3K27me3 and H3K9me3, which are linked to heterochromatin and transcriptional repression, respectively (Millán-Zambrano *et al.*, 2022). H3K4me1 is enriched at enhancers but not highly correlated with their activity, potentially priming enhancers for future activities. H3K27me3 is associated with silenced heterochromatin, repression

of enhancers and promoters, and serves as epigenetic memory. H3K9me3 is a hallmark of constitutive heterochromatin and is, therefore, also associated with transcriptional repression. Inactivated X chromosomes are enriched in H3K9me2, while actively transcribed gene bodies are typically marked by H3K36me3 (Barski *et al.*, 2007). Histone methylation was shown to be a central regulator of embryonic development in animals, playing important roles in maintaining pluripotency in stem cells and differentiation of tissues (Jambhekar *et al.*, 2019).

Histone acetylation is generally associated with transcriptional activity. It occurs at active promoters, enhancers, and accessible chromatin regions and is added by histone acetyltransferases and removed by histone deacetylases (Grunstein, 1997). The most prominent histone acetylation is H3K27ac, which is often used to verify the activity of enhancers (Wang *et al.*, 2008). H3K27ac might directly influence TF binding since several TFs showed altered binding patterns after a knockout of histone deacetylases in mouse embryonic stem cells (Cusack *et al.*, 2020).

1.1.5 Epigenetic regulation during embryonic development

While epigenetic marks are mostly stable in a given cell type over the span of a mammalian organism, they undergo drastic changes during organism development (Lee *et al.*, 2014). Epigenetics is a key factor in the differentiation of omnipotent stem cells that all carry the same genetic code into highly specialized cells that make up the tissues and organs of an adult mammal (Rulands *et al.*, 2018; Meissner, 2010). Embryonic development in mammals is a highly conserved process, which makes it amenable to study in model organisms (Solnica-Krezel and Sepich, 2012). The most widely used model organisms are mice (*Mus musculus*) (Hanna *et al.*, 2018).

Mouse and human embryonic development begins with fertilization, at which the genetic code of a sperm and an ovarian cell fuses to form a zygote. This is followed by blastulation, during which the zygote divides and forms a blastocyst that will implant itself into the mother's uterus. The embryonic part of the blastocyst contains Epiblast cells, a disc-shaped collection of pluripotent stem cells that eventually give rise to the embryo (Gilbert, 2000). The process of forming the three main germ layers, Mesoderm, Endoderm, and Ectoderm, from Epiblast cells is called gastrulation. It starts with the linear invagination of the Epiblast cells to form the primitive streak, establishing the bilateral symmetry of the embryo. Cells invaginated to form the primitive streak will give rise to the Mesoderm and Endoderm germ layers. The Mesoderm further differentiates during organogenesis to give rise to the heart, kidneys, circulatory system, bones, and muscular tissues. The Endoderm layer gives rise to the lungs, intestines, thyroid, pancreas, and bladder. Cells that did not invaginate to form the primitive streak form the Ectoderm, which gives rise to the nervous system as well as the eyes and ears and the outermost layer of skin (Stern, 2004).

Since embryonic development involves drastic changes in epigenetic regulation and gene and protein expression, it provides an attractive system to understand gene regulation that undergoes a natural perturbation of this magnitude. Furthermore, several clinically relevant dysfunctions in early embryonic development have epigenetic origins (Bergman and Cedar, 2013). Therefore, extensive research has been done, historically mainly focusing on individual components of the system. The advance of high-throughput sequencing technologies has made it possible to simultaneously track thousands of genes during development (Cao *et al.*, 2019; Pijuan-Sala *et al.*, 2019), as well as assess epigenetic marks on a genome-wide scale (Lee *et al.*, 2014; Rulands *et al.*, 2018; Wu *et al.*, 2016; Atlasi and Stunnenberg, 2017).

DNA methylation and accessibility, as well as histone modifications, have been studied with bulk sequencing methods (see Section 1.2), which revealed that there are two main waves of genome-wide demethylation and remethylation (Lee *et al.*, 2014; Smith *et al.*, 2012; Wang *et al.*, 2014) (Fig 1.5).

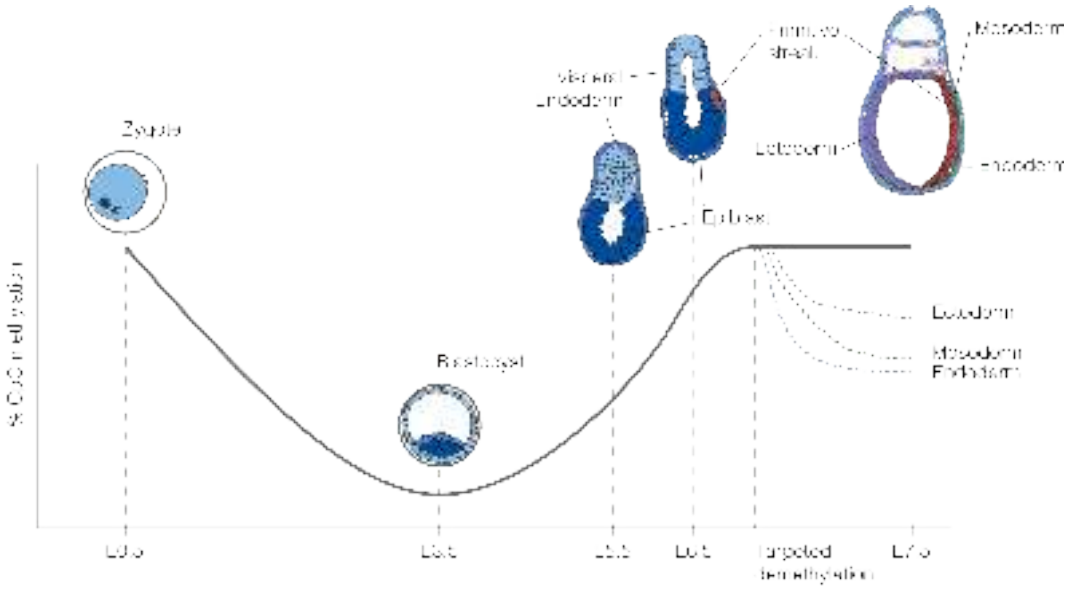


Figure 1.5 | Genome-wide changes in DNA methylation during embryonic development. The grey line represents the genome-wide methylation rate during the first days of embryonic development (E0.5 to E6.5), measured from fertilization. After fertilization, there is a rapid wave of demethylation up to the Blastocyst stage at E3.5, where the methylation rate drops to around 20%. After this, the Blastocyst implants into the uterus, and methylation levels are increased to about 80% during the Epiblast stage around E6.5. Most somatic tissues will maintain this methylation level throughout the organism's lifetime. Changes in methylation after this stage are highly localized and target regulatory regions involved in tissue differentiation. *Figure generated by Max Frank.*

The first methylation wave involves rapid and progressive demethylation, resulting in only approximately 20% of CpGs in the genome remaining methylated at the blastocyst stage, which is thought to allow cells to achieve a pluripotent state (Wang *et al.*, 2014). DNA methylation is only retained at transposable elements and constitutive

heterochromatic regions. This also corresponds with high chromatin accessibility (Wu *et al.*, 2016) and a lack of topologically associated domains as measured by Hi-C experiments (Ke *et al.*, 2017; Du *et al.*, 2017). Histone marks are still present in this phase and are hypothesized to influence the gene expression of pluripotent stem cells (Tee and Reinberg, 2014). An interesting set of developmental genes carry both activating H3K4me3 and repressive H3K27me3 marks at their so-called bivalent promoters (Bernstein *et al.*, 2006). These promoters generally remain unmethylated and are thought to be poised for quick transcription initiation. After implantation, a global remethylation wave occurs, leading to a global hypermethylation state.

The blastocyst that forms after implantation consists of relatively homogeneous cells, making it suitable for bulk sequencing to obtain accurate characterizations (Smith *et al.*, 2012). However, studying germ layer specification, which involves the development of distinct cell lineages, is extremely challenging without single-cell technologies. Despite the difficulties, some studies have manually dissected each germ layer and performed bulk sequencing (Xiang *et al.*, 2020; Auclair *et al.*, 2014). These studies have revealed that the initially homogeneous epigenetic landscape in the Epiblast gives way to a more dynamic landscape, where regulatory elements are activated in a lineage-specific manner.

Recently, the development of single-cell multi-modal technologies has provided new opportunities to study cell fate commitment events during gastrulation (Kelsey *et al.*, 2017; Clark *et al.*, 2018). These technologies allow the unambiguous assignment of epigenomes to transcriptomes (gene expression profiles) at the single-cell level, enabling a more comprehensive understanding of the processes involved.

1.2 Techniques for epigenetic profiling

Methods for epigenetic profiling have been available for bulk tissues for quite a while. With the advent of single-cell RNA sequencing and the increased ability to study heterogeneous populations of cells as well as dynamically changing biological processes, the need for epigenetic measurements in individual cells became clear. The following Sections will discuss currently available methods to profile DNA methylation and accessibility on a single-cell level.

1.2.1 DNA methylation

Single-cell DNA methylation profiling protocols have been developed based on bulk methods, particularly bisulfite sequencing (BS-seq). BS-seq involves treating DNA with sodium bisulfite, which converts unmethylated cytosine residues to uracil (and later to thymine after PCR amplification), leaving methylated cytosine intact. The resulting C-to-T transitions can be detected by DNA sequencing (Frommer *et al.*, 1992). Care must be taken in the alignment of bisulfite-converted reads since there are now mismatches with the reference genomes at all unmethylated cytosine positions. In

principle, this technique could be extended without many adaptations to a well-based single-cell sequencing approach. However, conventional BS-seq has limitations due to DNA degradation caused by purification steps and bisulfite treatment, making it challenging to use with low amounts of DNA. To overcome this issue, a modified protocol called post-bisulfite adaptor tagging (PBAT) was developed, which includes multiple rounds of 3' random primer amplification. By performing bisulfite treatment before adaptor ligation, the loss of adapter-tagged molecules is minimized, enabling the use of single-cell BS-seq (scBS-seq) with low-input material (Smallwood *et al.*, 2014).

1.2.2 DNA accessibility

The main protocols for profiling DNA accessibility in bulk tissues are the assay for transposase-accessible chromatin sequencing (ATAC-seq) (Buenrostro *et al.*, 2013), DNase I hypersensitive sites sequencing (DNase-seq) (Song and Crawford, 2010) and Nucleosome Occupancy and Methylome-sequencing (NOMe-seq) (Kelly *et al.*, 2012). All three techniques rely on the introduction of DNA-modifying proteins to the cell. These proteins will modify the accessible parts of DNA in a way that is ultimately detectable via sequencing readout. Figure 1.6 overviews the experimental protocols for NOMe-seq, ATAC-seq, and DNase-seq.

ATAC-seq relies on the Tn5 transposase to cut accessible sites but with the advantage of directly ligating sequencing adapters to the cleavage sites. This means that cleaved fragments can be amplified via PCR and ultimately sequenced. ATAC-seq is currently the standard protocol for quantifying DNA accessibility in bulk populations. It has also been adapted as a droplet-based single-cell protocol, allowing it to measure accessibility in large populations of cells at low cost (Buenrostro *et al.*, 2015).

In the case of DNase-seq, at low concentrations, DNase I will cleave accessible DNA, making it amenable to sequencing (Song and Crawford, 2010). This technique has been widely used in the past and has also been adapted as a single-cell protocol (Jin *et al.*, 2015).

NOME-seq is a technique that measures both DNA accessibility and DNA methylation. It relies on the GpC methyltransferase M.CviPI, which will methylate accessible cytosines in the GpC sequence context (Kelly *et al.*, 2012). Note that this differs from endogenous methylation, which occurs mainly in the CpG context. Bisulfite sequencing can then be used as a readout for both endogenous DNA methylation and DNA accessibility. Since bisulfite sequencing has been successfully adapted to single-cell applications, NOME-seq can also be utilized to assay single-cell methylation and accessibility (Pott, 2017).

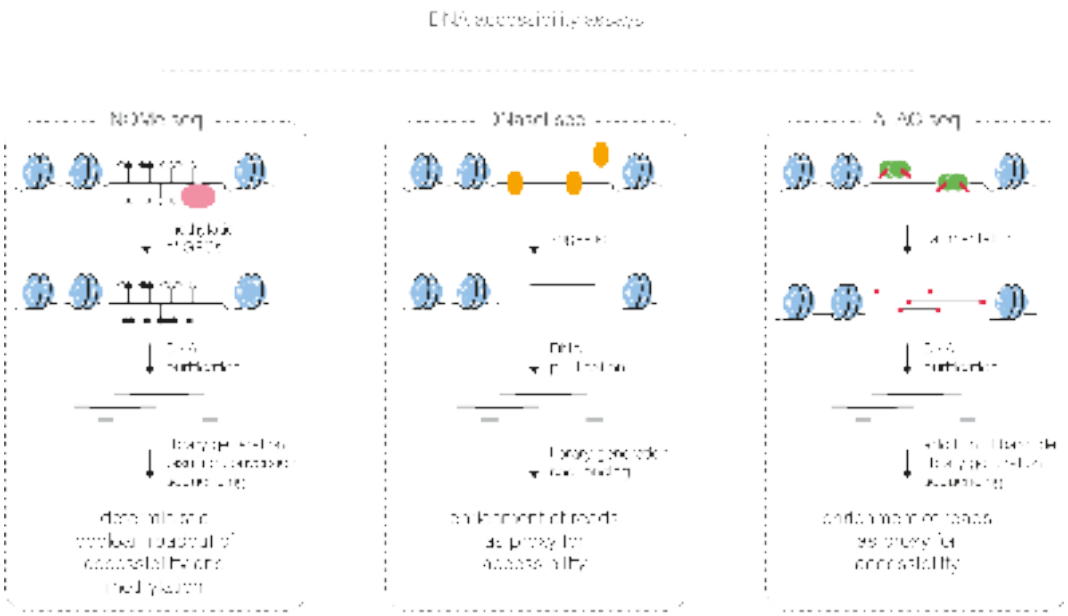


Figure 1.6 | Overview of different techniques for chromatin accessibility profiling. The three columns depict the steps involved in profiling chromatin accessibility with NOMe-seq, DNase-seq and ATAC-seq (from *left to right*). CpG sites (circles) are endogenously methylated (black) or unmethlyated (white), while all accessible GpC sites (hexagons) become methylated (black) after M.CviPI (pink) treatment. Histones are denoted in blue; DNaseI in yellow; and Tn5 transposase in green, with its sequencing adapters in red. *Figure generated by Max Frank, adapted from Nordström* et al.*, 2019.*

There are several advantages and disadvantages to NOMe-seq compared to the standard ATAC-seq protocol (Nordström *et al.*, 2019). ATAC-seq is a cheaper droplet-based protocol that is able to profile many more cells, albeit at lower sequencing depths per cell. Being well-based, the cost of profiling more than a few thousand cells with NOMe-seq can become prohibitive. However, NOMe-seq generates a DNA accessibility readout at high resolution that is only limited by the density of GpC sites in the genome (which is every 16bp on average) as opposed to cleavage fragment sizes. Furthermore, NOME-seq provides a deterministic boolean output at any covered site since a CpG/GpC is either methylated or not. With count-based methods, inaccessible regions cannot be distinguished directly from regions with low coverage. Furthermore, NOMe-seq has the advantage of DNA methylation as an additional output, making it a valuable tool to study the interplay of those two genomic layers.

1.3 Single-cell multi-modal profiling

In the previous Chapter, I discussed NOMe-seq as an example of a protocol that can profile two molecular layers within a single cell. Such techniques will be referred to as multi-modal or multi-omics techniques. In this Chapter, I will discuss the advantages of such techniques and the challenges of applying them at the single-cell level.

As discussed previously, one goal of cell biology is the ability to make predictions about cells along the information flow of the central dogma (Crick, 1970). However, gene

regulation is a complex and cyclical process that involves multiple layers of control, including TFs, DNA accessibility, histone modifications, and DNA methylation, among others. Profiling only a single layer at a time will thus not give a comprehensive picture of the regulatory processes we need to capture to truly understand how cells work.

Multi-modal experiments in bulk have been useful in determining coarse regulatory differences between conditions or different tissue types in the past. These experiments can use uni-modal techniques on different subsets of the same sample, thus making the extension to a multi-modal technique relatively straightforward (Ritchie *et al.*, 2015a). The analysis of these data typically involves the discovery of marginal associations between the different modalities (see Section 1.4) to find candidates for causal regulatory links. However, bulk assays only provide averages over cell populations and, therefore, fail to capture heterogeneity within these cell populations. To understand regulatory mechanisms, it is precisely this heterogeneity that is crucial. Furthermore, gene regulation can be best studied when cells are not in a steady state but undergoing dynamic changes due to changes in their environment. These systems can involve a deliberate stimulus such as drug administration or naturally occurring changes, e.g., during embryonic development.

The application of multi-omics techniques to single-cell data can be done in the same manner as for bulk techniques by aliquoting samples and subjecting them to uni-modal assays (Stuart and Satija, 2019). However, there is no direct link between cells in these experiments, and these datasets have to be integrated post-hoc, which can be a challenge, depending on the features measured. If two unimodal techniques have a (sub)set of shared features, this integration is possible with different computational methods that perform so-called horizontal integration. However, if there are no shared features, this integration becomes a very challenging task called diagonal integration. Figure 1.7 overviews different analysis scenarios for single-cell multi-omics experiments. For a comprehensive review of multi-omics analysis strategies, see Argelaguet *et al.*, 2021.

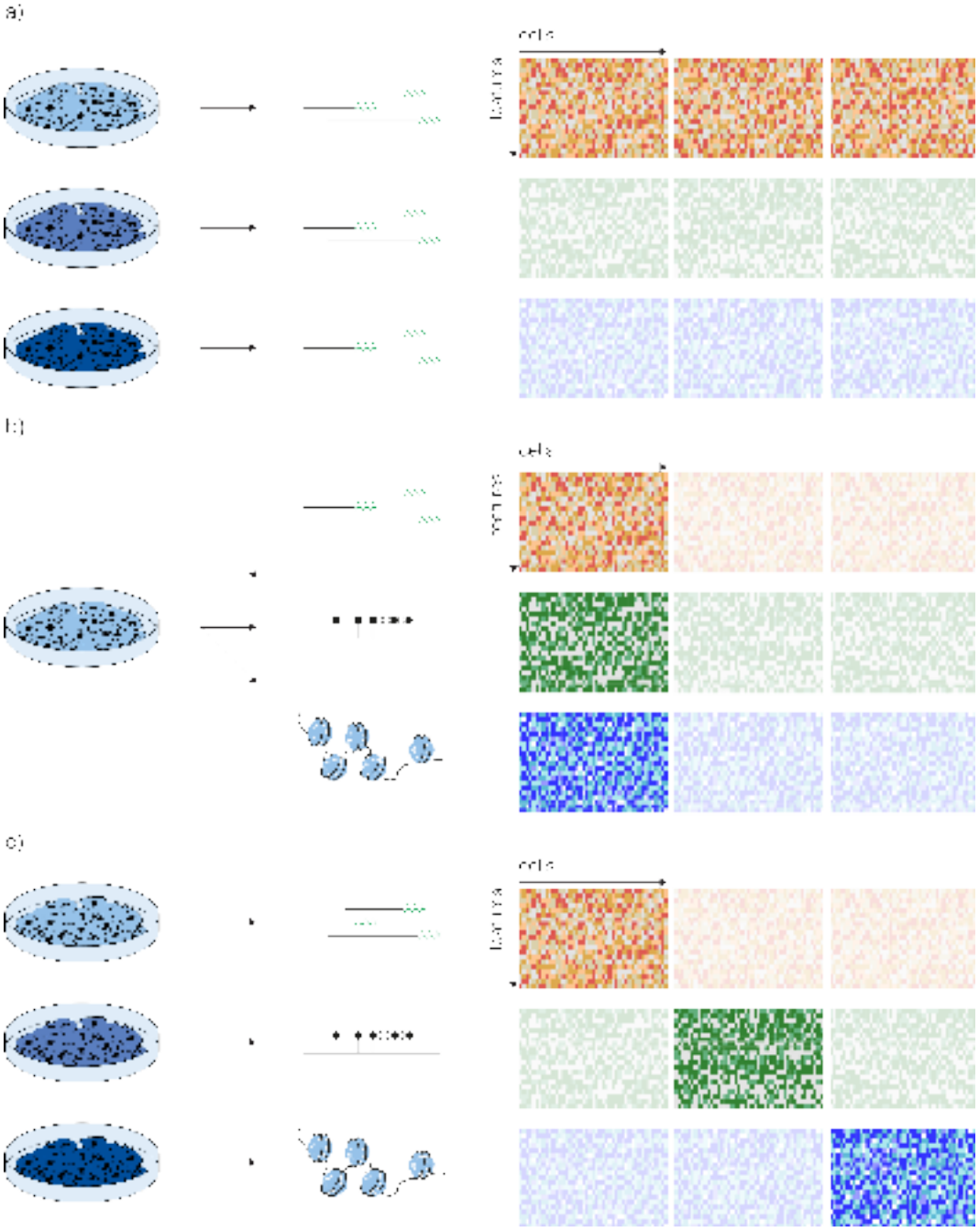


Figure 1.7 | Different integration strategies for multi-omics data. *a)* Horizontal integration is performed when the same set of features is observed in different experiments. *b)* Vertical integration is performed when different sets of features are observed in the same cell. *c)* Diagonal integration matches features and cells that do not overlap between experiments. *Figure generated by Max Frank, adapted from Argelaguet* et al.*, 2021.*

A better experimental setup for elucidating the regulatory relationships between different modalities is one that profiles two or more modalities in one cell. Note that the term multi-modal in this context is often used to distinguish assays measuring multiple modalities in the same cell from multi-omics experiments that integrate uni-modal assays in-silico (Argelaguet *et al.*, 2021). These techniques are only emerging in recent years and face a host of technical challenges. A fundamental problem for these

assays is the limited amount of genetic, transcriptomic, and proteomic input material present in a cell. Low coverage in unimodal assays can usually be compensated by increasing the number of cells assayed, which allows the merging of similar cells to more completely covered meta-cells. In multi-modal experiments, to establish links between two features in different modalities, both features of interest have to be covered in the same cell. Another problem is that most assays involve the destruction of their input material, preventing the application of subsequent assays.

Figure 1.8 gives an overview of single-cell multimodal assays that combine measurements of the transcriptome with genomic-, epigenomic- and proteomic assays. The following paragraphs will give a brief overview of a selection of these assays.

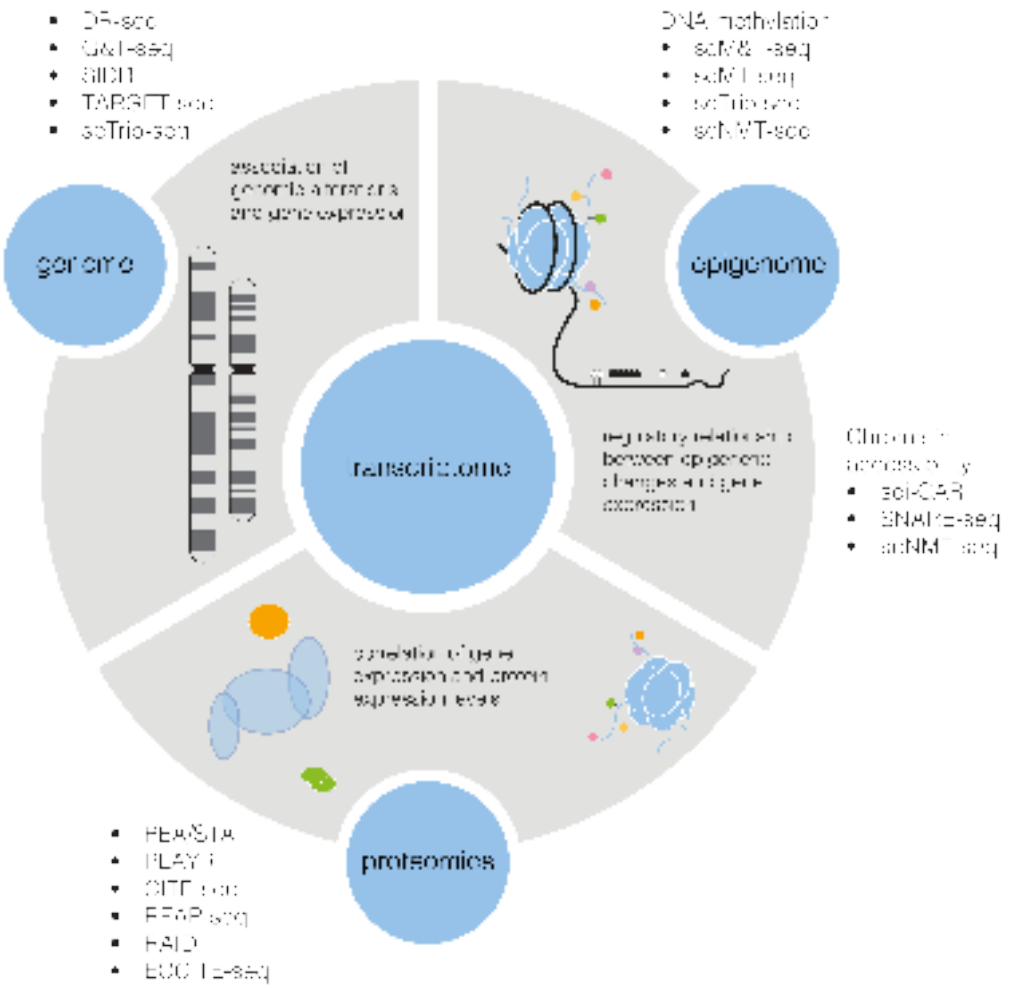


Figure 1.8 | Overview of single-cell multimodal assays. Different technologies are placed with respect to the molecular data they assay in addition to the transcriptome. Techniques that assay more than two molecular layers, such as scNMT-seq are listed multiple times. Each combination of modalities provides opportunities to study different biological processes as described within the grey boxes. *Figure generated by Max Frank, adapted from Lee* et al.*, 2020.*

Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) (Stoeckius *et al.*, 2017) is a technique that can quantify transcriptomes and the concentration of a limited number of cell-surface proteins in parallel. It does so by incubating cells with antibodies that bind specific surface proteins and carry unique oligonucleotide sequences. After washing away unbound antibodies, the oligonucleotides carried by bound antibodies can be identified in a sequencing readout in conjunction with RNA quantification. CITE-seq has been successfully applied to distinguish subpopulations of immune cells that could not be distinguished by RNA-seq alone (Hao *et al.*, 2021). The technique is limited by the availability of antibodies for surface proteins of interest and its inability to assay intracellular proteins.

Another set of multi-modal droplet-based technologies assays DNA accessibility and nuclear RNA concentrations in parallel (Chen *et al.*, 2019; Ma *et al.*, 2020). This method is also commercially available under the name 10X multiome kit. Here, nuclei are incubated with the Tn5 transposase in bulk, as for a single-cell ATAC-seq workflow. Transposed nuclei are then microfluidicly paired with gel beads containing two types of oligonucleotide sequences, or oligos. One oligo will contain cell barcodes and a sequence that can bind to the poly-A tail of mature mRNA molecules. The other oligo contains barcodes and generic sequences that will pair with Tn5-cleaved DNA fragments. Cells can then be pooled again, and two fractions are used for ATAC-seq and RNA-seq. This allows the profiling of large numbers of cells at low cost, with the drawback of reduced complexity at the single-cell level. This assay has recently been applied to a dataset of mouse embryonic stem cells undergoing gastrulation (Argelaguet *et al.*, 2022).

Other techniques make use of the physical separation of the input material needed to assay different modalities. One of the first applications that made use of this was single-cell genome and transcriptome sequencing (scG&T-seq) (Macaulay *et al.*, 2015). Here, a cell's mature messenger RNA (mRNA) is separated from its DNA using biotinylated oligo-dT primers that bind the mRNA poly-A tail and can be removed from the solution with magnetic beads. The RNA fraction can then be subjected to standard single-cell RNA-seq protocols, while the DNA fraction can be used for genetic or epigenetic assays. This protocol has various applications in the study of gene-regulatory mechanisms. In the original study, the authors could directly measure the impact of DNA copy number variations in a subpopulation of cells on gene expression of genes on the same chromosome.

1.3.1 scNMT-seq

Single-cell nucleosome, methylation, and transcription sequencing (scNMT-seq) (Clark *et al.*, 2018) combines the ideas of the scG&T-seq (Macaulay *et al.*, 2015) and NOMe-seq protocols introduced in Section 1.3 (Fig 1.3.1). Cells are separated into wells and incubated with a GpC methyltransferase as in NOMe-seq, methylating accessible GpC sites. Then DNA is separated from mRNA as in the scG&T protocol. The mRNA

fraction is assayed using the Smart-seq2 protocol (Picelli *et al.*, 2014), a full-length scRNA-seq protocol with high coverage of the transcriptome. The DNA fraction is subjected to single-cell bisulfite sequencing, which carries information about both DNA methylation and DNA accessibility.

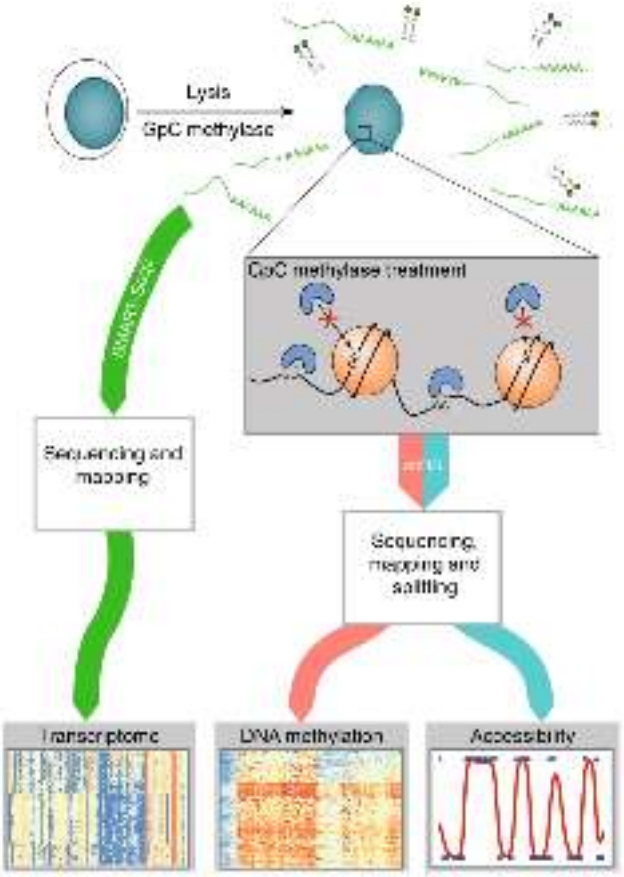


Figure 1.9 | Overview of the scNMT-seq protocol. scNMT-seq involves the lysis of cells and the methylation of accessible GpC sites by introducing a GpC methyltransferase. The cytoplasmic fraction of mRNA molecules is assayed with the Smart-seq2 protocol. The nuclear portion is assayed with single-cell bisulfite sequencing using the NOMe-seq protocol. Separately evaluating CpG and GpG methylation allows the readout of endogenous methylation and chromatin accessibility. *Figure adapted from Clark* et al.*, 2018*

In the analysis of scNOMe-seq data, special care has to be taken when interpreting ambiguous signals from the analysis. Cytosines in the CGC sequence context could be methylated by both endogenous and exogenous M.CvPI methyltransferases. Therefore, these sites are commonly excluded from the analysis altogether.

scNMT-seq thus provides a lot of parallel information about the cells assayed. The biggest limitation of this technique is the cost associated with scaling this assay up to large numbers of cells. While other droplet-based single-cell methods can compensate for the lack of coverage in individual cells by scaling to hundreds of thousands of cells, the analysis of scNMT data relies on maximizing the information extracted from every cell assayed. This work contributes to the analysis of sparse multi-modal

single-cell data that need careful consideration when designing algorithms. In the next Section, I will discuss general approaches to analyzing multi-modal single-cell data.

1.4 Integrative analysis of single-cell multi-modal data

As mentioned before, the term multi-modal data in the context of single-cell experiments describes multiple layers of molecular data extracted from the same cell. Each modality will consist of a set of features corresponding to a molecular layer in the cell, resulting in a set of two or more cell-by-feature matrices. This allows the linkage of these molecular layers via cell identity, which is also referred to as vertical integration (Fig 1.7). Vertical integration typically aims to elucidate cellular processes that determine the causative links between the assayed modalities. Since each cell provides only a snapshot of these regulatory interactions, finding causative links directly can be challenging. Instead, in the first step, researchers are often interested in the co-variation of features, which can help to narrow down the space of causal links and, in conjunction with prior biological knowledge, can produce hypotheses about cellular mechanisms (Macaulay *et al.*, 2017; Argelaguet *et al.*, 2021).

Vertical integration methods can be categorized as local or global approaches. Local analyses focus on specific associations between molecular features across different layers, aiming to detect interactions between them. In contrast, global integration utilizes the full range of measurements to identify broader cellular states, such as cell cycle phase and pluripotency potential, in an unsupervised manner.

1.4.1 Global analysis

The power of global analysis is already apparent in uni-modal single-cell experiments, where it can be applied to define cell types or order cells along a continuous biological process. This is done with dimensionality reduction techniques, the most widely used of which is principle component analysis (PCA) (Luecken and Theis, 2019). PCA transforms high-dimensional data (in this case, high-dimensional in feature space) onto a lower-dimensional space while maximizing the variance explained by the orthogonal remaining axes. This transformation is linear, making principle components highly interpretable but limited to detecting only certain (linear) sources of variation. Other methods with similar matrix-factorization ideas have been successfully applied to single-cell data, including Bayesian matrix-factorization that forms the basis of Multi-omics Factor Analysis (MOFA) (Argelaguet *et al.*, 2019a; Argelaguet *et al.*, 2018a), as well as canonical correlation analysis (CCA) (Butler *et al.*, 2018) and others. All these techniques can be combined with clustering algorithms to find groups of similar cells or cell types without prior biological knowledge.

Global integration of multi-modal data utilizes unsupervised dimensionality reduction methods with the same principle as PCA to define cellular states resulting from

interactions between multiple genomic features and across modalities (Heumos *et al.*, 2023). MOFA is a Bayesian matrix factorization tool designed for this task (Argelaguet *et al.*, 2018b; Argelaguet *et al.*, 2019a). It finds orthogonal axes of variation called factors that are shared between modalities as well as those that are only present in a single modality. CCA (Butler *et al.*, 2018) can also be used for vertical integration, however, with the limitation that it will prioritize sources of variation shared between all modalities. Another approach is to generate a multi-modal nearest neighbor graph that weighs the utility of each modality in each cell. This approach is called weighted-nearest neighbor (WNN) (Hao *et al.*, 2021) analysis and outputs a graph that can be used for downstream visualization and clustering. These methods have been shown to be able to resolve cellular states that would not have been resolved by either modality alone. For example, WNN has been applied to CITE-seq data of peripheral mononuclear blood cells to identify lymphoid subpopulations not detected by scRNA-seq alone (Hao *et al.*, 2021). Furthermore, there is also great potential to identify regulatory links with global analysis. For example, MOFA identified a broad set of lineage-defining enhancers during mouse gastrulation that regulates the expression of important marker genes using scNMT-seq data.

However, the potential of these global methods to discover novel biology in an unsupervised manner is complicated by several challenges. One challenge is that most integration methods are designed with Gaussian likelihoods, meaning there is an assumption of normality for the cell by feature matrices. This is often a valid assumption for count data, e.g., after log transformation, but is not appropriate for other modalities such as methylation, which is binary at the CpG level (Du *et al.*, 2010). Some methods, like MOFA, allow alternative likelihoods to address this problem, but model performance can suffer compared to Gaussian likelihoods. Different modalities can also have vastly different amounts of technical noise associated with them, making assessing co-variation challenging.

Another problem is the definition of features as input. There are clearly defined features for transcriptomics and proteomics (assuming alternative splicing and post-translational modifications are ignored). Epigenetic measurements, however, often lack defined features since the epigenomic landscape is not exhaustively mapped. Therefore, these methods often rely on prior biological knowledge to define features of interest. For example, DNA methylation features are often predefined by chromatin marks for potential enhancers such as H3K27ac (Wang *et al.*, 2008). All methylation measurements within the boundaries of a feature are then aggregated to produce the input matrix. If the boundaries of the features are incorrect, this leads to the inclusion of uninformative CpG sites or the exclusion of informative CpG sites and, therefore, a decrease in signal-to-noise ratio. In some cases, such as ATAC-seq, features are defined in a data-dependent manner (Yan *et al.*, 2020). For example, one standard way of defining features for ATAC-seq is to sum up the signal of all cells and then run a peak-calling algorithm, identifying regions with increased accessibility. This has the

problem of potentially missing accessible DNA only present in rare cell populations. To overcome this, another strategy is to first separate cell populations based on the signal of tiled genomic windows and then run peak calling separately on clusters of cells. Data-dependent feature identification makes it harder to compare the results of different experiments because they work with a different set of features.

In general, global analysis of multi-modal data is a powerful way to get an overview of a biological system of interest and to find major axes of biological variation. However, to study detailed regulatory interactions, it is often necessary to develop models that are specifically tailored to the mechanism of regulation of interest and that test links between individual features across modalities directly. This approach to integration is called local integration.

1.4.2 Local analysis

Local analysis is a different paradigm for the analysis of multi-modal data. Here, interactions between different features in two or more modalities are explicitly investigated in a supervised manner. Often, one can restrict the search space of these methods by considering biological priors such as the genomic proximity of the features. For example, when looking for gene-enhancer interactions, tests are often restricted to certain distances of the regulatory element to the transcription start site of the gene. In cis-expression quantitative trait loci (cis-eQTL) mapping, the same principle is applied to filter out genes that are too far away from a variant of interest (Nica and Dermitzakis, 2013). This is sensible and necessary since there would be a combinatorial explosion of tests without filters, adding to the multiple testing burden.

When pairwise interactions between elements are tested, the test can be tailored to the data modalities. For example, for the investigation of cis-eQTLs, linear mixed models can be used specifically designed to deal with sparse information from single-cell data (Cuomo *et al.*, 2020).

Local analysis is often key to getting a detailed understanding into biological mechanisms. They can often include prior biological knowledge. For example cell-type identity can be validated by the expression of known marker genes.

1.4.3 Combining global and local analyses

Another powerful approach is the combination of global with local analyses. This is routinely done in single-cell RNA-seq assays, where global analysis will aid in the unsupervised identification of cell types via clustering, followed by differential expression analysis between these cell types for individual genes (Luecken and Theis, 2019). This allows the discovery of novel cell types or states and potential marker genes that define their identity. Recently, several approaches have adapted this pipeline to remove the need for clustering in this pipeline, testing for genes that co-vary with any of the global axes of variation (Dann *et al.*, 2022; Ahlmann-Eltze and Huber, 2024).

This is an exciting avenue for experiments that study temporal or spatial biological processes because these tests can be much more powerful in detecting continuous changes in gene expression. For example, in developmental studies, cell identities change gradually over time, which means that a clustering algorithm would introduce arbitrary boundaries between states. One caveat of these approaches is that care must be taken with the dual use of information since the input data for differential expression testing will be the same (albeit a subset) data used to infer global cell state.

In the case of finding cis-eQTLs in single-cell data, combining global and local analysis can involve the integration of principal components from PCA to remove global effects, similar to stratifying a population of human subjects. Furthermore, the unsupervised clustering results in the global analysis might be used to find eQTLs that are cell-type specific (Cuomo *et al.*, 2022).

The same principles can be applied to multi-modal approaches (Argelaguet *et al.*, 2021; Heumos *et al.*, 2023). However, they have the advantage that one modality can be used for inferring cell states with global analysis, followed by the local analysis of other modalities to find individual features that co-vary with the global state. This approach avoids the double-dipping problem of uni-modal approaches. In general, scRNA-seq is often best suited for inferring cell state since it has a fixed set of features and is a fairly robust assay. RNA expression also sits in the middle of the information flow paradigm of the central dogma, making it a good anchor to which to compare most modalities. Epigenetic single-cell assays are often technically more complicated with less well-defined features, requiring more refined local analysis.

An example of this type of analysis would be detecting DNA methylation changes within tumor subpopulations. With a multimodal assay, RNA expression could be used to classify cells as tumor-surrounding healthy cells or into subgroups of tumor cells. Then, a test could be applied that finds genomic regions that are differentially methylated between these groups of cells (Fan *et al.*, 2022). Since DNA methylation has a very different noise model, testing for differential methylation requires a tailored test (Kapourani *et al.*, 2021). I will discuss methods for detecting epigenomic changes in the next Chapter.

1.5 Statistical methods to detect epigenomic changes

Methods for detecting epigenomic changes between conditions vary broadly based on the input type. As discussed above, ATAC-seq provides a signal in the form of peaks of chromatin accessibility (Buenrostro *et al.*, 2013). The insertion frequency of the Tn5 transposase and the resulting fragment sizes limit the resolution of this technique. Resulting resolutions are typically on the 100bp scale, sufficient to detect nucleosome positioning, but smaller events, such as TF binding, are harder to detect (Bentsen *et al.*, 2020). Furthermore, the output of ATAC-seq after peak finding is a

sample by peak (in case of bulk assays) or a cell by peak count matrix (in case of single-cell assays), allowing the adaptation of computational methods designed for RNA-seq. There are several frameworks to analyze (single-cell) ATAC seq data, such as ArchR (Granja *et al.*, 2021), Signac (Stuart *et al.*, 2021) and SnapATAC2 (Zhang *et al.*, 2024). These packages typically provide standard tests like the Wilcoxon Test or logistic regression. Additionally, one can utilize popular purpose-built RNA-seq differential testing packages such as DEseq2 (Love *et al.*, 2014) or limma (Ritchie *et al.*, 2015b).

In the case of bisulfite sequencing-based assays, such as NOME-seq, the output data will initially be a matrix of samples by CpG/GpC (in the case of bulk assays) or cell by CpG/GpC (in the case of single-cell assays). In the single-cell case, this matrix will be very sparse, with typical coverage of bisulfite sequencing in single-cells ranging from 0.01-20% (Angermueller *et al.*, 2016), and most entries will be either 0 or 1, indicating methylation or no methylation. In the case of bulk, coverage is often higher due to the increased amount of input material. Entries in the matrix will range from 0 to 1, indicating the fraction of methylated cells at this position in a given sample.

When testing for differential DNA methylation or DNA accessibility, researchers are often not interested in the changes of single bases but are looking for segments or regions in the genome that are changing. We will call these segments differentially methylated regions (DMRs) or differentially accessible regions (DARs). This has biological and technical reasons. Biologically, regulatory epigenetic processes will typically involve a change across multiple bases in the genome. For example, repositioning one nucleosome will make 147bp accessible at once. DNA-binding proteins that methylate or demethylate DNA will affect multiple nearby cytosines. Therefore, the signal of close-by CpG or GpC sites will be correlated (Mayo *et al.*, 2015). From the technical side, testing all CpG or GpC sites in a mammalian genome comes with an enormous multiple-testing burden, reducing the statistical power of these tests. Therefore, methods that test for epigenetic changes with base-resolution data must solve two problems:

- The definition of region boundaries within which epigenetic change occurs
- The statistical assessment of the significance and magnitude of change within those regions

Most methods for detecting epigenomic changes with base resolution were designed for bulk bisulfite-sequencing data. Recently, some specific models for single-cell approaches have also been developed. The next Section will overview the existing landscape of available methods.

1.5.1 Bulk methods

1.5.1.1 Models that compute statistics on fixed genomic windows

These methods implicitly assume the methylation rate to be constant within these windows and compute statistics on the aggregated counts of all CpG sites within a region. This is a robust and fast way to test for differentially methylated regions (DMR), but has the disadvantage that a correct list of candidate regions must be known a priori. These methods include: IMA (Wang *et al.*, 2012a), COHCAP (Warden *et al.*, 2013), DMAP (Stockwell *et al.*, 2014), methylSig (Park *et al.*, 2014) and methylKit (Akalin *et al.*, 2012)

1.5.1.2 Models that compute statistics on individual CpG sites

These methods often use variations of Fisher's exact test (Fisher, 1922) or beta-binomial regression to compute significance for loci. DMR can then be computed by aggregating nearby significant CpG sites. This has the issue that there is no proper FDR control on the region level. Aggregation can broadly be characterized by aggregation with heuristics and aggregation by smoothing of methylation rates.

Methods that use aggregation heuristics have more complicated FDR control models than models that test fixed windows. Since the methylation data is both used to define and test DMR's, care must be taken not to "double dip". A fundamental problem with these methods is that, given the sparsity of single-cell data, it is often impossible to calculate any meaningful statistics on individual CpG sites, which renders most aggregation heuristics invalid. Furthermore, they are usually designed with replication in mind, prohibiting application to most single-cell datasets. Examples of this type of methodology are Methylpy (Schultz *et al.*, 2015) and DSS (Feng and Wu, 2019).

Methods that compute statistics on smoothed estimates of methylation rates make use of the fact that information can be shared between neighboring CpG sites. Thus, they average the methylation signal with some smoothing or clustering operation before calculating statistics. These methods include BSsmooth (Hansen *et al.*, 2012), Metilene (Jühling *et al.*, 2016), and BiSeq (Hebestreit *et al.*, 2013).

1.5.2 Single-cell methods

Although it might seem straightforward to adapt the bulk methodologies for detecting DMRs for single-cell data, many of these methods contain heuristics for filtering low-coverage data on the individual CpG level that often does not allow their use in sparse single-cell methylation data. One strategy to overcome this sparsity problem is to impute the methylation state of CpG sites that have missing information. The two main methods designed to impute single-cell DNA methylation at base resolution are deepCpG (Angermueller *et al.*, 2017) and MELISSA (Kapourani and Sanguinetti, 2019) (see Section 1.5.2.1). Another approach is the aggregation of multiple CpG

sites in regions of interest that are defined *a-priori*. This approach makes it feasible to adapt the use of bulk methods such as Fisher's exact test or DSS (Feng and Wu, 2019). It is also used in the scMET method (see Section 1.5.2.2).

1.5.2.1 Imputation models

DeepCpG DeepCpG (Angermueller *et al.*, 2017) is a neural network-based method that uses DNA sequence information in conjunction with neighboring CpG site methylation states of all cells in the dataset. It was tested on single-cell BS-seq of mouse embryonic stem cells by subsampling the data. It was able to predict global methylation states with high accuracy and precision (area under the receiver-operating characteristic curve (AUC) > 0.85). This was a significant improvement compared to other methods that did not use DNA sequence information or only used information from the same cell. When assessing the imputation performance in different genomic contexts, prediction power was the highest in promoter regions and exons. This is expected since these elements are typically either fully methylated or unmethylated. In enhancer regions, marked by H3K27ac or H3K4me1, performance dropped to AUCs of 0.6-0.8, while still outperforming simpler methods. This is also expected since these regions are associated with increased methylation variability.

MELISSA MEthyLation Inference for Single-cell Analysis (MELISSA) (Kapourani and Sanguinetti, 2019) is an imputation tool for single-cell BS-seq data that leverages the combination of a global analysis with a local analysis. It consists of a Bayesian hierarchical model that jointly learns smooth representations of methylation rate in genomic regions of interest and clusters cells based on the genome-wide patterns of these representations. Thus, it shares information between neighboring CpG sites, thanks to the smoothing of methylation rates in genomic coordinates, and between grouped cells in global methylation space. It had improved precision and accuracy compared to similar methods that did not share information between cells in simulated data and was able to cluster cells correctly. When evaluated on single-cell BS-seq of mouse embryonic stem cell data, it had a similar performance to DeepCpG without the DNA sequence information. It performed slightly worse than DeepCpG with DNA sequence, with considerably less computational complexity. Prediction performance also decreased in the context of active enhancers, similar to DeepCpG.

1.5.2.2 Differential testing

To my knowledge, there is currently only one tool for differential methylation testing explicitly designed for single-cell BS-seq data, called scMET (Kapourani *et al.*, 2021).

scMET scMET is a Bayesian framework that tests for both differential mean and variability of methylation between groups of cells in single-cell BS-seq data. It fixed genomic regions of interest as an input to overcome sparsity issues within single-cell data. For each region of interest, it fits a beta-binomial model of methylation rate

that has an explicit overdispersion component describing biological variability. Thus, it can also be used to detect highly variable regions in heterogeneous populations of cells, which can be used for downstream global analyses. Interestingly, when comparing differential mean methylation analysis between groups of cells, there was no substantial power benefit of this model versus a Fishers-exact test or the beta-binomial model implemented by DSS. Interestingly, both scMET and Fishers-exact test show massively varying false positive rates depending on the number of cells per group and the average number of CpG sites per region, indicating that neither test is calibrated.

scMET has the limitation that it requires fixed genomic regions as an input to the model, making it reliant on accurate annotations of regulatory elements in the genome that are expected to change their methylation rate. Furthermore, it assumes that within a genomic region of interest, methylation rate is constant within a cell which might not always be the case. A model that is able to take imperfect or no prior annotation of genomic regions as input would not have these shortcomings but is challenging to implement since it would have to combine region finding/refinement in parallel with differential testing.

Another limitation of all current models that test for differential methylation is that they are designed to test for methylation changes between groups of samples or groups of cells. However, in many dynamic biological systems, methylation changes will happen continuously over space or time. With the emerging availability of multi-modal single-cell assays, these continuous changes could, in theory, be tracked with great resolution. However, the current methods are not designed to model methylation in these cases.

1.6 Aims of this Thesis

In the previous Section, I outlined the limitations of current methods to model epigenetic heterogeneity in single-cell experiments. These limitations are especially pronounced if the observed biological system is undergoing continuous changes, such as a developmental process. As introduced in Section 1.4.1, global analysis strategies can be used to identify continuous biological processes in single-cell RNA sequencing datasets without the need for large numbers of experiments. This allows for the study of continuous gene expression changes across developmental processes. Single-cell multimodal technologies further open up the possibility of studying the epigenetic changes that go hand in hand with transcriptomic changes. Therefore, multimodal single-cell technologies have the potential to study the gene regulatory landscape that determines cell fate during development. However computational tools that facilitate these investigations are not well established.

In this thesis, I want to combine the concept of global analysis that has been well-established for single-cell RNA sequencing with a local analysis approach in the highly

sparse epigenetic modalities of multimodal single-cell experiments. Section 1.7.1 will explain how pseudotime inference can be used to assign every cell a position along a developmental trajectory. To then model epigenetic changes across these developmental trajectories, I aimed to develop a local model that could describe continuous, non-linear changes in DNA methylation and chromatin accessibility of regulatory genomic regions. This is complicated by the highly sparse readout that single-cell epigenomic profiling technologies produce (see Section 1.3). I therefore wanted to develop a model that makes use of the idea of sharing information between cells and neighboring genomic loci while still providing rigorous statistics for differential testing. In this thesis, I developed a Gaussian process (GP) model that satisfies these criteria, called GPmeth. I introduce GPs in Section 1.8 and describe the detailed considerations and derivation of the model in Chapter 2.

1.7 Biological motivation of the GPmeth model

1.7.1 Modeling continuous changes in single-cell RNA-seq studies

1.7.1.1 Global analysis

As discussed in Section 1.4, single-cell analysis provides the opportunity to discover structure in populations of cells in an unsupervised manner. This often involves dimensionality reduction, which was introduced in Section1.4.1. This type of global analysis typically leads to one of two scenarios, depending on the biological system that is studied.

1. Firstly, clearly distinct populations of cells that can be clustered into cell types (Fig 1.10, bottom). A typical example of this is the study of blood cells in adult humans, where different immune cell types, such as B-cells and T-cells, are clearly distinguished by clustering. This analysis can also lead to the discovery of previously unknown cell types.

2. Secondly, single-cell assays also have the potential to reveal a snapshot of a developmental process (Fig 1.10, top). Even though a single experiment will typically assay cell populations that were extracted at a single time point during a developmental process, there is often enough natural variation in the differentiation speed of individual cells that developmental trajectories can be faithfully reconstructed. For single-cell RNA-seq (scRNAseq) data, many algorithms have been developed that detect axes of continuous variation (Saelens *et al.*, 2019; Andrews *et al.*, 2021). These methods are generally referred to as pseudotime analyses.

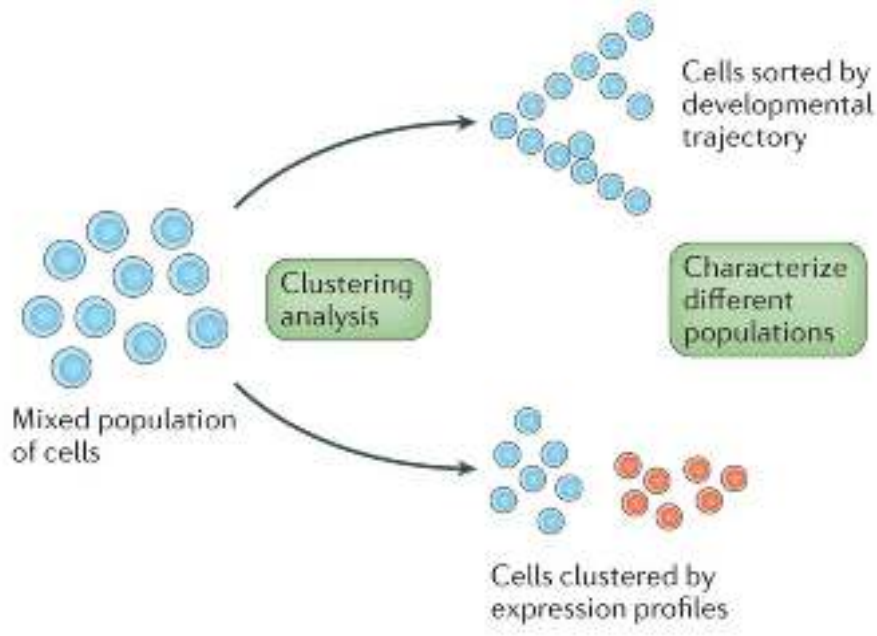


Figure 1.10 | Unsupervised global analysis strategies for single-cell RNA-seq studies. Unsupervised global analysis allows the assignment of cells to clusters that represent distinct cell states or types or onto a position along a developmental trajectory. *Figure adapted from Stegle* et al.*, 2015* with permission of the authors.

1.7.1.2 Psedudotime analysis

The aim of pseudotime analysis is to map cells along a developmental/differentiation trajectory. Most pseudotime analysis tools employ two main approaches. The first method involves utilizing dimensionality reduction techniques to uncover a low-dimensional 'manifold' where the cells are situated, similar to the dimensionality reduction for discovering cell types. Cells are then ordered in pseudotime based on a neighborhood graph in this manifold. Well-known methods employing this strategy are Monocle (Cao *et al.*, 2019) and DPT (Haghverdi *et al.*, 2016). The second approach revolves around employing unsupervised clustering to group cells, followed by connecting the clusters and projecting individual cells onto the resulting branches. TSCAN (Ji and Ji, 2016) and Mpath (Chen *et al.*, 2016) are examples of methods that follow this approach. Cluster-based pseudotime methods exhibit higher accuracy in scenarios where there is an uneven distribution of cells along the trajectory, such as when certain cell states are more prevalent or consistently captured compared to others, or in large-scale developmental hierarchies. On the other hand, manifold approaches excel when there is a uniform sampling of cells throughout the transition and when examining intricate details of individual transitions. The main goals of scRNA-seq pseudotime analyses are to establish lineage relationships during organismal development. This can uncover which cell types give rise to which differentiated tissues and the transition states that cells have to go through.

1.7.1.3 Local analysis

After global analysis, the next step is to identify genes that are differentially expressed across the studied cell population through local analysis to reveal biologically relevant changes. The appropriate statistical test for this depends on the scenario that was encountered in the global analysis.

If there are clearly distinct cell populations in the data, the application of statistics that are developed to compare bulk experiments is appropriate. These include methods such as the Wilcoxon test (Wilcoxon, 1945), or DEseq (Love *et al.*, 2014).

To find temporally variable genes along a differentiation trajectory, there are again two main approaches for differential testing (Fig 1.11). The first approach is to define cutoff values in pseudotime to group cells into discrete temporal stages. This approach enables the application of differential testing methods to detect marker genes for cell-type clusters mentioned above. However, if changes in gene expression are truly continuous, grouping cells will necessarily sacrifice some statistical power for the test. This can be seen in the upper panel of Fig 1.11: There are some cells close to the arbitrarily introduced cutoff point that have intermediate expression values. These cells will decrease the difference in means between the two populations.

The problem can be overcome by using a model that takes pseudotime as a continuous covariate. The first choice of models to work with continuous covariates are linear models, but pseudotemporal trajectories of gene expression are not guaranteed to be linear. Thus, nonlinear models have been developed that calculate test statistics by model comparisons (Fig 1.11).

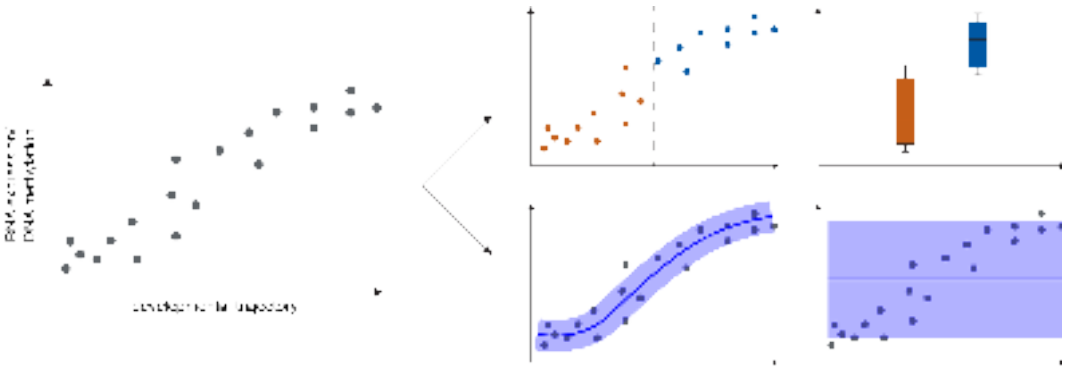


Figure 1.11 | Differential testing with continuous covariates. This figure shows the difference between discrete tests (*top row*) and continuous tests (*bottom row*) for changes in RNA expression with a continuous covariate. *Figure generated by Max Frank.*

Examples of these models are GPfates (Lönnberg *et al.*, 2017) , GPcounts (BinTayyash *et al.*, 2021), Monocle (Trapnell *et al.*, 2014; Qiu *et al.*, 2017), and tradeSeq (Van den Berge *et al.*, 2020). GPfates implement a GP model with a branching kernel that can pinpoint the divergence of gene expression for trajectory branching events. GPcounts implements a GP model with a zero-inflated negative-binomial likelihood

that is better suited to count-based scRNA-seq data and also has a branching kernel. Monocle fits an additive model to the count data to find genes associated with branching events. TradeSeq fits generalized additive models to gene expression data and provides a range of tests for different biological questions.

The same principles outlined above also apply to single-cell multi-modal data. Global analysis of one modality can reveal developmental trajectories and order cells along pseudotime. Then, other modalities can be queried for features that change along a pseudotemporal axis. Compared to the previous scenario, this also has the advantage of avoiding the use of the same information for global and local analysis. However, detecting epigenetic changes comes with a particular set of challenges (see Section 1.5). The two main challenges are:

1. Epigenetic modalities can have fundamentally different noise models compared to count-based readouts that one gets with scRNA-seq. The techniques that provide base-pair resolution of DNA methylation and chromatin accessibility, such as scNMT-seq will provide a binary signal.

2. There is no clearly defined set of features to test for. With RNA-seq, genes can serve as a clearly defined unit of aggregation for reads. In contrast, for epigenetics there is no gold-standard database of regulatory elements that can be tested for.

The GP model that I will propose in this thesis aims to address these challenges.

1.7.2 Modeling continuous changes in single-cell DNA methylation measurements

Measuring DNA methylation and chromatin accessibility at base-pair resolution in individual cells or even on individual molecules (Krebs *et al.*, 2017) has advantages over coarser techniques such as scATAC-seq. With these techniques, it is possible to not only quantify nucleosome occupancy of DNA but also the much smaller footprint of TFs, which allows the study of the cooperative binding of multiple TFs (Sönmezer *et al.*, 2021) or the effects of DNA methylation on TF binding (Kreibich *et al.*, 2023). However, because of the limited amount of input DNA, base-pair resolution single-cell techniques provide highly sparse data and need tailored models to deal with this sparsity.

Fundamentally, at a specific cell and CpG site, DNA methylation can exist in one of four states. Either both alleles at the position are methylated, both alleles are unmethylated, or only the paternal or maternal allele is methylated. Note that, in theory, the cytosines on opposing strands of the same DNA molecule could also carry different methylation signals, but this is exceedingly rare in mammalian cells, since the dedicated DNA methyltransferase DNMT1 recognizes hemimethylated CpG sites (Goll and Bestor, 2005; Klose and Bird, 2006). Allele-specific methylation (ASM) is a

more common occurrence and is often studied in combination with QTLs (Abante *et al.*, 2020). From a genome-wide perspective, the frequency of ASM is still rather low. For example Do *et al.*, 2016 found that 2% of all regions that could be annotated with haplotype information had significant differential methylation rates in multiple human tissues. The confident detection of ASM from bisulfite sequencing data requires sufficient read depth for both alleles and is thus often prohibitive with sparse single-cell data. Therefore, methylation rate of CpG site i in an individual cell j is often modeled as a Bernoulli distributed variable

$$y_{ij} = Bern(\rho_{ij}) \tag{1.1}$$

where ρ_{ij} is the unknown true methylation rate and y is the observation in the data. ρ_{ij} can be described with a Bernoulli distribution when ignoring hemimethylation.

1.7.2.1 Genomic covariances

As mentioned previously, the cell-by-CpG site matrix produced by these techniques will be sparse, with 0.1-10% of entries covered. This makes the direct analysis of individual features challenging. On the one hand, this can be overcome by using the fact that cells are not independent measurements but can be linked by proximity on a low dimensional manifold, pseudotemporal ordering or neighborhood graph. This is the reason that continuous models have increased statistical power to detect changes. The other important consideration is that the signal of proximal CpG or GpC sites on a chromosome is not independent. The footprint of a nucleosome spans 147bp, which, with an average occurrence of GpC sites every 16bp, will affect 9 GpC sites at once. TF footprints are smaller but still can affect multiple GpC sites. The co-variance of endogenous CpG methylation will depend on multiple factors. First, the average distance of neighboring CpG sites will vary widely throughout the genome. Genome-wide, there is a depletion of CpG sites, with an average distance of 100bp between neighboring CpGs (Saxonov *et al.*, 2006), whereas in CpG islands the average distance will only be ∼10bp (Gardiner-Garden and Frommer, 1987). The spatial correlation of methylation rate between CpG sites also varies based on genomic context. Zhang *et al.*, 2015 found that there is a general decay of correlation to genome-wide background levels at a 400bp distance. It is not entirely clear what the biological underpinnings of this observed local correlation are, but it is plausible that when a methyltransferase binds to a CpG site, it is more likely to bind to neighboring CpG sites afterwards.

In summary, there are two key insights that motivate the formulation of the GPmeth model in the next section

Key insight 1 Epigenetic features measured in individual cells are not independent measurements but vary smoothly with biological processes

Key insight 2 Methylation rate and chromatin accessibility measured at base-pair resolution are not independent measurements but vary smoothly across the genome

One of the aims of this work is to provide a model that relaxes the requirement for the *a-priori* knowledge of exact region annotations by allowing for smoothly variable methylation rates within regions of interest while being able to make use of single-cell multimodal assays to model continuous temporal changes in methylation rate (see Chapter 2). The basis for this model are GPs (Rasmussen and Williams, 2006). The next Section will introduce the methodology behind GPs and showcase some of their applications.

1.8 Introduction to Gaussian Process models

In this thesis, I will propose a method to detect epigenetic changes based on single-cell assays. The core of this method is a non-linear regression of the methylation/accessibility rate over time and across the genome. While linear regression is standard in many statistical methods, including detecting epigenetic changes, non-linear regression typically comes with additional challenges of potential overfitting of the data. GPs have been used in the past to model non-linear changes in gene expression for both bulk and single-cell data (Stegle *et al.*, 2010; BinTayyash *et al.*, 2021) and have many desirable properties for this application. This Section is meant to provide a brief technical introduction to GPs and their advantages and disadvantages for the models discussed in this Thesis. For a thorough introduction, see Rasmussen and Williams, 2006.

1.8.1 Introduction to Gaussian Processes

The formal definition of a Gaussian process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution (Rasmussen and Williams, 2006). Informally, one can think of GPs as a distribution over functions that, when evaluated at any point, will have a joint Gaussian distribution.

GPs are fully specified by a mean function $\mu(\mathbf{x})$:

$$\mu(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{1.2}$$

and a covariance function or kernel $k(\mathbf{x}, \mathbf{x}')$:

$$k(\mathbf{x}, \mathbf{x}') = Cov[f(\mathbf{x}), f(\mathbf{x}')] \tag{1.3}$$

Often, it can be assumed that the mean is zero after simply subtracting the data mean from the input. This leaves the GP fully specified by its kernel, which is where the model can be constrained with a prior of choice. The choice of kernel represents assumptions about the data that is modeled, as will be shown in Section 1.8.2. The

comparison of the data fit of different covariance functions can then be used to validate those assumptions (see Section 1.8.6).

The full GP will be denoted as:

$$f(\mathbf{x}) \sim GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{1.4}$$

This means that any set of random variables f(x) will be jointly Gaussian with mean and covariance specified by the mean function and kernel, respectively. Any finite realization of this process at specified points $\mathbf{x}$ is a multivariate normal distribution.

1.8.1.1 Marginal likelihood

GPs allow us to compute key quantities, such as the marginal likelihood of input data, analytically given the above-specified model. This is an important property because it allows the optimization of the hyperparameters θ of the model with respect to a set of input data. It also allows the comparison of the likelihood of different models to determine the most appropriate model structure (see Section 1.8.6). The marginal likelihood of a GP for a set of input data $[f(\mathbf{x_1}), f(\mathbf{x_2}), ..., f(\mathbf{x_n})] = [y_1, y_2, ..., y_n] = \mathbf{y}$ is the integral of the likelihood times the prior

$$p(\mathbf{y}|\mathbf{X}, \theta) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{x}, \theta) p(\mathbf{f}|\mathbf{x}, \theta) d\mathbf{f} \tag{1.5}$$

where $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]$ are the locations of the input data. Note that we are marginalizing over the possible function values $\mathbf{f}$, hence the term marginal likelihood. I listed the hyperparameters θ explicitly here to show that they need to be given to compute the marginal likelihood. In the case of Gaussian likelihood, this integral can be evaluated explicitly, and the result is often given in logarithmic form as:

$$logp(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) - \frac{n}{2} \log(2\pi) \tag{1.6}$$

where $\mathbf{X}$ are the locations of the inputs, θ are the hyperparameters of the model (see Section 1.8.2), and Σ is the covariance matrix specified by the kernel and evaluated at the inputs such that $\Sigma_{i,j} = k(x_i, x_j)$. The marginal likelihood consists of three terms that can give the optimization process some desired properties. Intuitively, the first term $-\frac{1}{2} \log |\Sigma|$ penalizes model complexity, where low covariances between the inputs lead to a larger penalty. The second term is dependent on the data and encourages the model fit. The third term does not depend on the data or hyperparameters and is just a constant normalization term. GPs thus have a natural property of balancing model complexity against fit to the data.

1.8.1.2 Hyperparameter optimization

By optimizing the marginal likelihood with respect to θ, we can find optimal hyperparameters of the model. This can be done with gradient-based optimizers (see Rasmussen and Williams, 2006, Chapter 5).

1.8.1.3 Model predictions

After optimizing the hyperparameters θ of the GP model, we can ask it to predict values at unseen locations. This prediction will not be a point estimate but a Gaussian distribution, which we can use to compute confidence intervals for each output. The predicted distribution at an unseen input location x^* is given by:

$$\begin{aligned} p(f(x^*)|\mathbf{X},\mathbf{y},\theta) =& N(\mu(x^*) + k(x^*,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}(f(\mathbf{X}) - \mu(\mathbf{X})), \\ & k(x^*,x^*) - k(x^*,\mathbf{X})k(\mathbf{X},\mathbf{X})^{-1}k(\mathbf{X},x^*)) \end{aligned} \tag{1.7}$$

Where $\mathbf{X}$ are the locations of the training data. This is the posterior distribution of the model, which can be evaluated at any location. If we look closer at the variance term, we can see that it is equal to the prior variance at location x^* minus a positive term that shrinks the variance depending on the training data.

This will give the posterior conditioned on noise-free function values $\mathbf{y}$. In the real world, input data typically does not correspond to the function values themselves but noisy realizations of them. If we assume that the measurement errors are independent and identically distributed (i.i.d.) Gaussian, we can write

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon \tag{1.8}$$

where ε is a i.i.d Gaussian with variance σ_n^2 . The conditional distribution then becomes

$$\begin{aligned} p(f(x^*)|f(\mathbf{X})) =& N(\mu(x^*)+ \\ & k(x^*,\mathbf{X})[k(\mathbf{X},\mathbf{X}) + \sigma_n^2 I]^{-1}(f(\mathbf{X}) - \mu(\mathbf{X})), \\ & k(x^*,x^*) - k(x^*,\mathbf{X})[k(\mathbf{X},\mathbf{X}) + \sigma_n^2 I]^{-1}k(\mathbf{X},x^*)) \end{aligned} \tag{1.9}$$

This is the key predictive equation for GP prediction. From this, one can sample from the posterior, calculate the mean and variance at any location x^*, and calculate confidence intervals for the predictions.

1.8.2 Encoding assumptions about the data using covariance functions

Kernels are the positive definite functions that define the covariance between two inputs $\mathbf{x}, \mathbf{x}'$. Using different functions for the kernel is a way to impose restrictions on the prior GP that correspond to assumptions about the data.

1.8.2.1 Base Kernels

Some of the most commonly used kernels are listed below and depicted in Figure 1.12.

Linear Kernels. The linear kernel constrains the GP to produce linear functions. GPs with a linear kernel are equivalent to Bayesian linear regression (Rasmussen and Williams, 2006). The kernel function is given by:

$$k(x, x') = \sigma_f^2 (x - c)\left(x' - c\right) \tag{1.10}$$

A model with this kernel has two so-called *hyperparameters*. σ_f^2 is the kernel variance that determines the amplitude of change of the function. In the linear case, this corresponds to the slope. The second hyperparameter c determines the intercept of the function.

Constant Kernels. The constant or bias kernel is the simplest kernel that produces constant outputs.

$$k(x, x') = \sigma_f^2 \tag{1.11}$$

The output of this kernel does not depend on the input and will simply be

$$f(x) = c; c \sim N(\mu(x), \sigma_f^2) \tag{1.12}$$

Matérn Kernels. The Matérn family of kernels produces functions with different degrees of local smoothness. Their covariance function is specified by:

$$k(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{(x - x')}{\ell}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{(x - x')}{\ell}\right) \tag{1.13}$$

where Γ is the gamma function, and K_ν is the modified Bessel function of the second kind. ν determines the smoothness of the function produced and is usually fixed to one of $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$. The higher ν, the smoother the resulting function will get, and the resulting function will be $\nu - 1$ times differentiable. The hyperparameters of this kernel are σ^2, which is called the kernel variance, and ℓ, which is the kernel lengthscale. The lengthscale of a kernel is important in specifying the spatial scale

of variation that a resulting function will have. Large lengthscale kernels produce functions that vary slowly over time or space.

Squared-Exponential Kernels. The squared-exponential (SE) kernel is a special case of the Matérn family of kernels where $\nu \to \infty$. The covariance function then simplifies to:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right) \tag{1.14}$$

with the two hyperparameters σ and ℓ as before specifying the amplitude and width of the kernel.

Periodic Kernels. Periodic kernels can be used to express prior expectations of repeating the behavior of functions. The covariance function can be derived by mapping inputs to a base kernel through the transformation $u = (\cos x, \sin x)$. The result for a base SE kernel (MacKay, 1998) is given by:

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{2}{\ell^2} \sin^2\left(\pi \frac{x - x'}{p}\right)\right) \tag{1.15}$$

again with σ^2 and ℓ as variance and lengthscale hyperparameters, plus a third hyperparameter p that determines the period of the function.

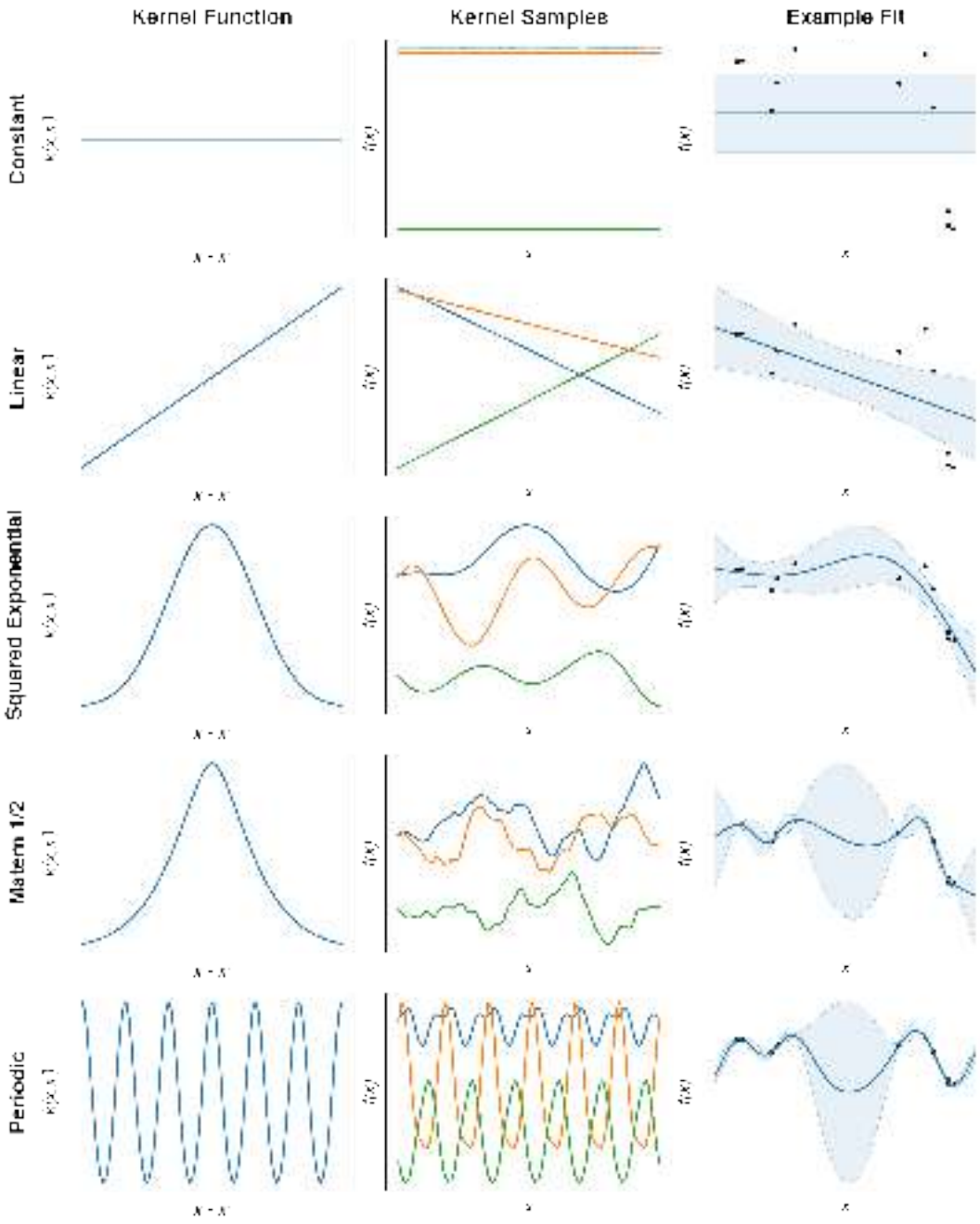


Figure 1.12 | Examples of kernels for Gaussian processes. Shown are the kernel functions (*left column*), functions sampled from the GP prior (*middle column*) and examples of the GP posterior (*right column*) where the blue line represents the predictive mean of the model and the shaded area is the 95% confidence interval. *Figure generated by Max Frank.*

These base kernels allow us to put different constraints on the predictions that a GP can make. This is useful if there is prior knowledge about the data-generating process. For example, when modeling DNA methylation rate over time, an assumption of smoothness is reasonable, but a linear change of methylation rate over time is probably too restrictive. Therefore, a kernel from the Matérn family should be a good choice.

1.8.2.2 Combining kernels

Kernels can also be combined to express expectations about functions that mix multiple elements of these base kernels. Two fundamental ways of combining kernels are multiplication and addition. This can be done for kernels that operate on the same input dimension or kernels that operate on different input dimensions.

First, I will discuss combining kernels on the same input dimension. The addition of kernels expresses the expectation that the function we want to model is a sum of functions. For example, global temperature could be expressed as a sum of a long-term SE trend that reflects the impact of climate change and a periodic trend that reflects seasonal changes. An example of such a model is depicted in Figure 1.13 (bottom row).

Multiplying kernels produces functions that can be thought of as an AND combination of the base kernels. If one of the functions is 0 the result will always be 0, which means that one kernel can be a gating function for the other kernel or control its amplitude. For example, the combination of a linear and a squared exponential kernel will produce a function with increasing variance (Fig 1.13, *second row*). Combining linear kernels through multiplication yields polynomial kernels (for example a quadratic kernel in Figure 1.13, *top row*).

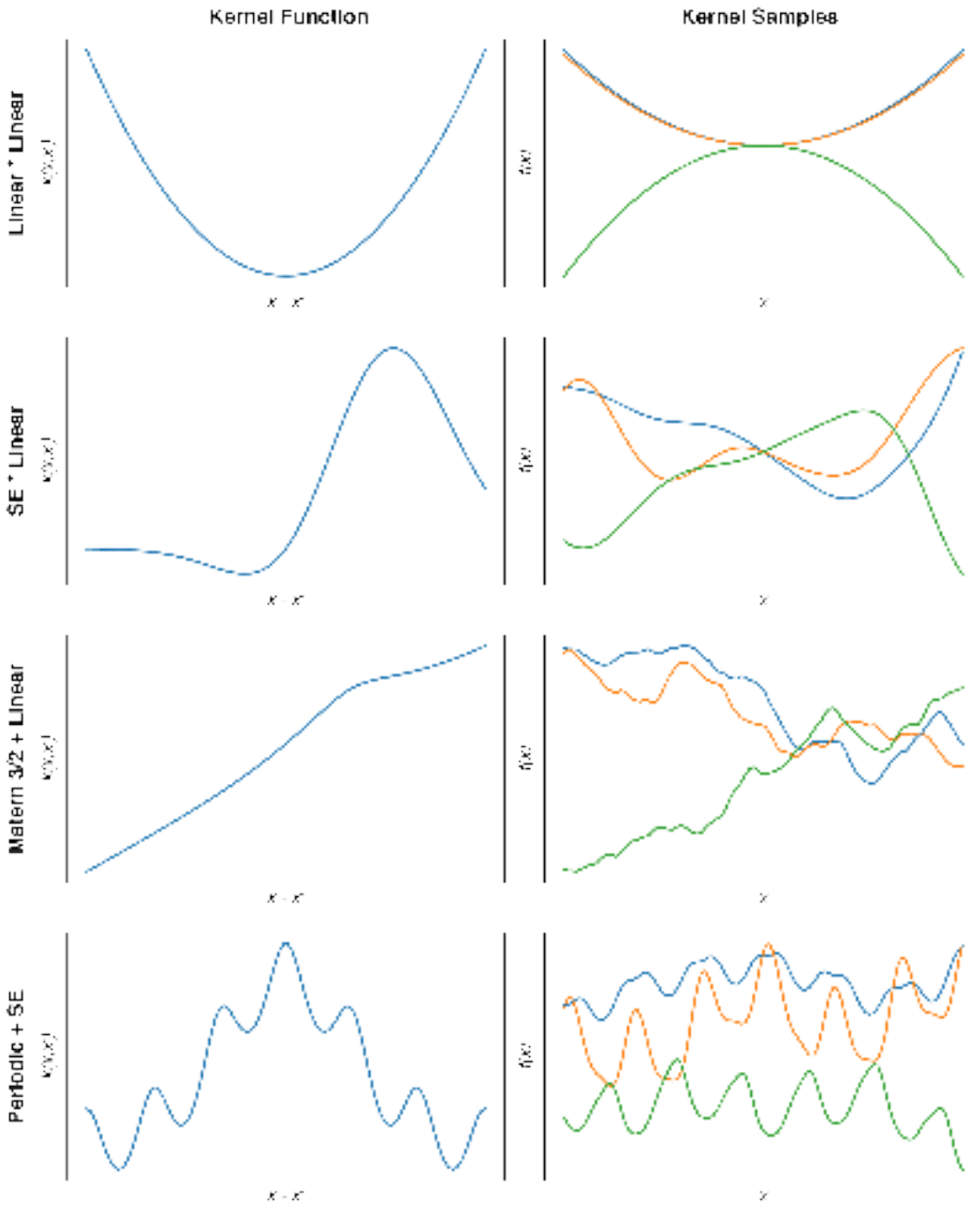


Figure 1.13 | Shown are the kernel functions (*left column*) and functions sampled from the GP prior (*right column*). The top two rows are examples of multiplicative combination of kernels, the bottom two rows are examples of additive combinations. *Figure generated by Max Frank.*

Multidimensional Kernels GPs are not limited to a single input dimension but can model functions that predict multiple dimensions. This can be done by combining kernels that are defined on different dimensions either additively or by multiplication (Fig 1.14). For example for a two-dimensional process:

$$k(\mathbf{x}, \mathbf{x}') = k_1(x_1, x_1') + k_2(x_2, x_2') \tag{1.16}$$

where the subscript indicates the active input dimension of the respective kernel. For multiplicative combination:

$$k(\mathbf{x}, \mathbf{x}') = k_1(x_1, x_1') \times k_2(x_2, x_2') \tag{1.17}$$

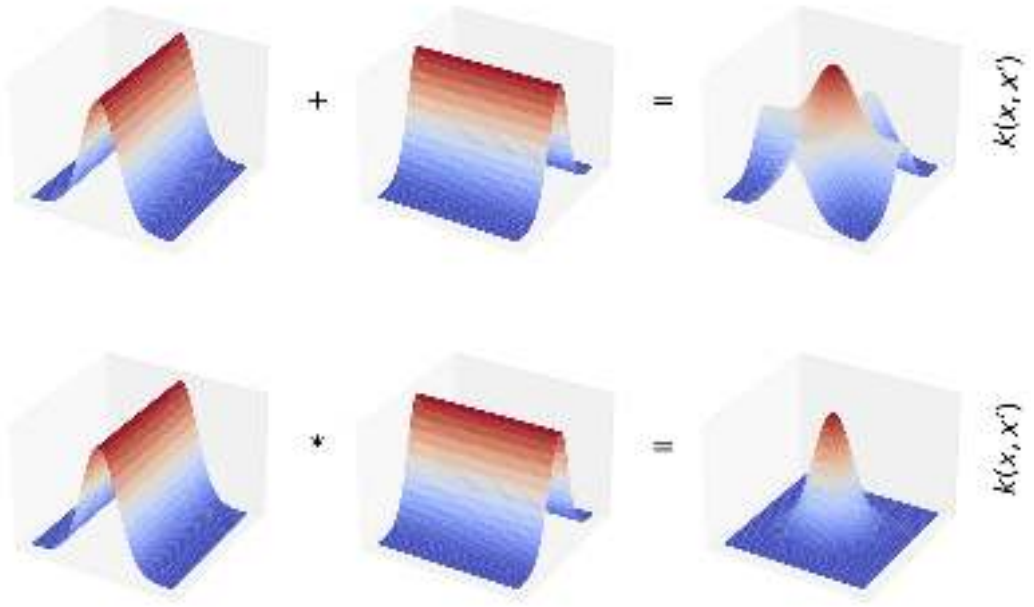


Figure 1.14 | Two one-dimensional kernels can be combined to form a two-dimensional kernel by addition (*top row*) or multiplication (*bottom row*). *Figure generated by Max Frank.*

When combining kernels, the hyperparameters of the unidimensional kernels can either be constrained to be shared across kernels or free to be trained individually. A multiplicative kernel with individually varying hyperparameters is also sometimes referred to as an automatic relevance determination (ARD) kernel. This kernel is capable of assigning very long lengthscales to input dimensions that do not have any structured variability, labeling it irrelevant.

With additive kernels, one expresses the concrete assumption that the function to be modeled is the sum of functions of the individual kernels. This makes models less flexible, but more capable of extrapolation in high dimensions if the assumption is correct. Figure 1.14 shows examples of SE kernels that are combined multiplicatively or additively. Note that the additive kernel has higher covariance further away from the center, although both kernels have the same hyperparameters.

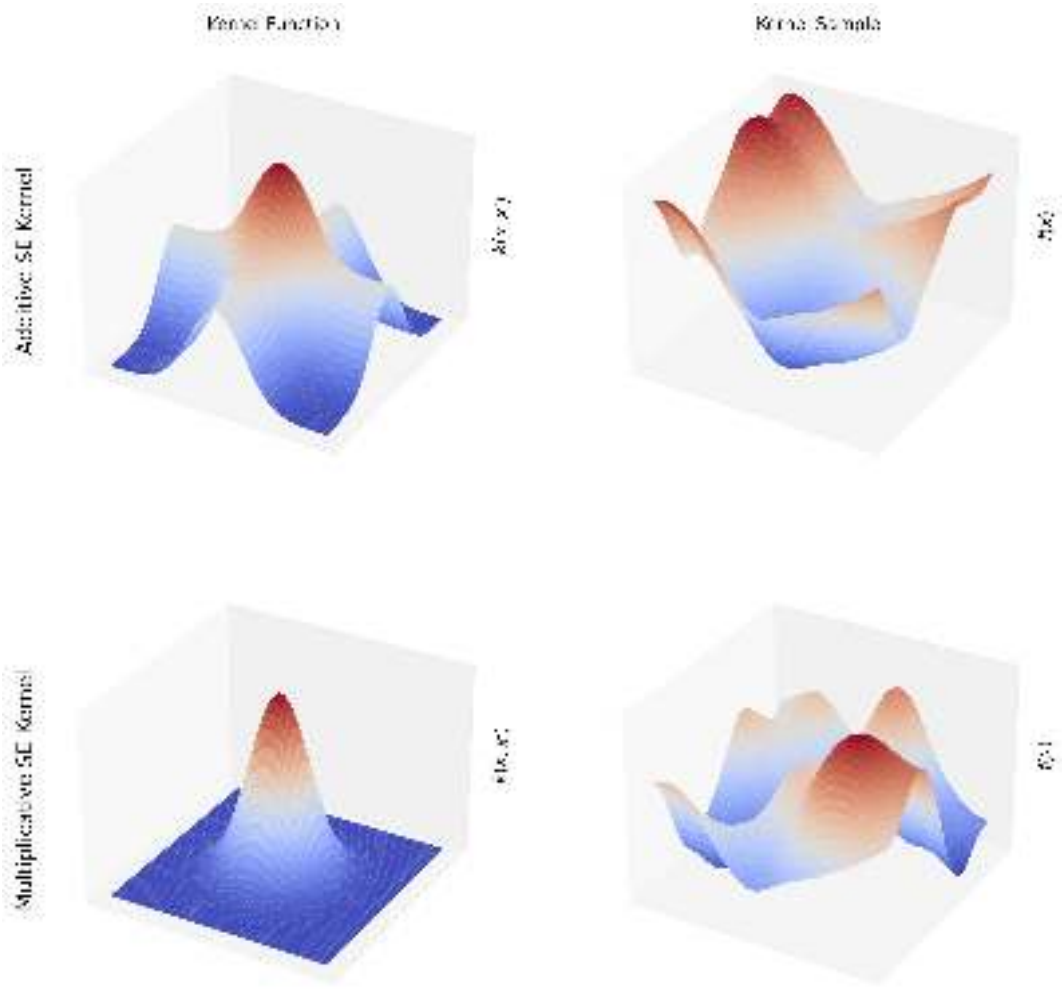


Figure 1.15 | The (*top row*) shows the kernel function (*left*) and a sample of the kernel (*right*) for an additive SE kernel. The (*bottom row*) shows the same for a multiplicative kernel. *Figure generated by Max Frank.*

1.8.3 Non-Gaussian likelihoods

So far, we have assumed that the input to a GP are noisy observations with Gaussian i.i.d. noise of a real-valued underlying function. However, GPs can also be used to model data that does not have a Gaussian likelihood assumption. For example, GP regression can be made more robust to outlier observations by using a Student' t-distributed noise model (Neal, 1997; Stegle *et al.*, 2010). When models have a non-Gaussian likelihood, their posterior is no longer analytically tractable. This requires the use of approximate inference techniques to calculate the posterior process and to train the GP. There are different inference techniques available such as the Laplace Approximation (Rasmussen and Williams, 2006), Expectation Propagation (Minka, 2001), Markov Chain Monte Carlo (Neal, 1997), and more recently, Variational Inference (Titsias, 2009; Hensman *et al.*, 2013). These methods compute an approximation to the marginal likelihood of the marginal likelihood of the model.

1.8.3.1 Classification

Another data type where Gaussian likelihoods are not appropriate is categorical data. Here, inputs are part of one of C classes. In the case of $C = 2$, we have the case of binary classification. GPs can be used to predict class probabilities for unknown input locations. To turn GP regression into a binary classifier, the idea is to use a GP prior on a latent function that is defined in the domain $(-\infty, \infty)$, and "squash" it through a mapping function $\pi(x)$ (Fig 1.16). The result of the resulting function is the probability of the class label being in class 1:

$$\pi(x) = p(y = 1|x) \tag{1.18}$$

There are two main mapping functions π that are used are $\pi(x) = \Theta(x)$, which is the cumulative function of a standard Normal distribution, and $\pi(x) = 1/(1 + e^{-z})$, the logistic function.

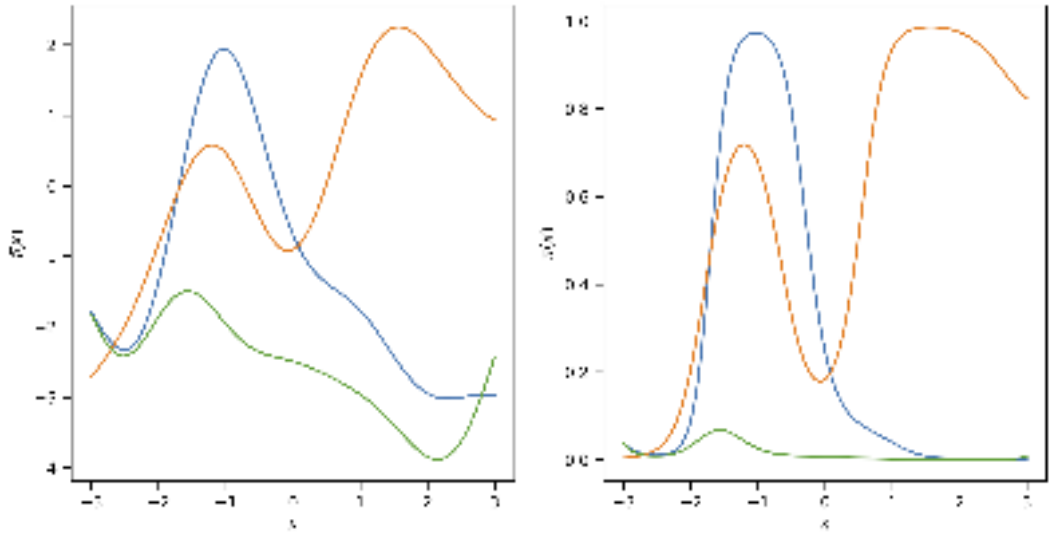


Figure 1.16 | The *left* panel shows three samples drawn from a GP with a squared-exponential kernel. The *right* panel shows the corresponding "squashed" functions that are obtained by mapping $\pi(x) = \Theta(f(x))$. *Figure generated by Max Frank.*

The inference of this model can be divided into two steps. First, the distribution of the latent variable can be computed for arbitrary time points $\mathbf{x}_*$

$$p\left(f_* \mid X, \mathbf{y}, \mathbf{x}_*\right) = \int p\left(f_* \mid X, \mathbf{x}_*, \mathbf{f}\right) p(\mathbf{f} \mid X, \mathbf{y}) d\mathbf{f} \tag{1.19}$$

Then, to produce predictions, the distribution over the latent f_* is mapped to the output space of y with the probit link function

$$\bar{\pi}_* \triangleq p\left(y_* = 1 \mid X, \mathbf{y}, \mathbf{x}_*\right) = \int \Phi\left(f_*\right) p\left(f_* \mid X, \mathbf{y}, \mathbf{x}_*\right) df_* \tag{1.20}$$

For a derivation, see Rasmussen and Williams, 2006. As mentioned above, the integral for the latent distribution is no longer analytically tractable because of the non-Gaussian likelihood and has to be approximated. I will give a brief overview of the Variational Inference approach to this approximation here. The idea of variational inference is to approximate the non-Gaussian term $p(\mathbf{f} \mid X, \mathbf{y})$ with a variational distribution $q_\psi(\mathbf{f})$ that is Gaussian and parametrized by a set of variational parameters ψ. The variational parameters are optimized so that the variational distribution is close to the original distribution in terms of the Kullback-Leibler divergence.

$$\mathrm{KL}\left(q_\psi(\mathbf{f}) || p(\mathbf{f} \mid y, X)\right) = \int q_\psi(\mathbf{f}) \ln \frac{q_\psi(\mathbf{f})}{p(\mathbf{f} \mid y, X)} \tag{1.21}$$

If we apply bayes rule to $p(\mathbf{f}|y, X)$ and rearrange the equation we get:

$$\ln p(y \mid X) - \mathrm{KL}\left(q_\psi(\mathbf{f}) || p(\mathbf{f} \mid y, X)\right) = \int q_\psi(\mathbf{f}) \ln p(y \mid \mathbf{f}, X) \mathrm{d}\mathbf{f} - \mathrm{KL}\left(q_\psi(\mathbf{f}) \| p(\mathbf{f})\right) \tag{1.22}$$

The right-hand side of this equation is referred to as the Evidence lower bound or ELBO. The ELBO is guaranteed to be smaller than or equal to the likelihood since the KL term on the left-hand side is positive or zero by definition. Thus, maximizing the right-hand side of the equation with respect to the variational parameters ψ will minimize the KL divergence. In practice, the hyperparameters of the model can also be optimized in conjunction with the variational parameters.

We can now substitute the variational approximation to the posterior at unseen prediction points $\mathbf{x}_*$ to make predictions:

$$\bar{\pi}_* \triangleq p\left(y_* = +1 \mid X, \mathbf{y}, \mathbf{x}_*\right) = \int \Phi\left(f_*\right) q_\psi(f_*) df_* \tag{1.23}$$

Fortunately, the mapping of the latent function can be computed explicitly when using a probit link since $q_\psi(f_*)$ is a Gaussian. The integral of a product of a standard cumulative Gaussian Φ and a Gaussian evaluates to

$$\int_{-\infty}^{\infty} \Phi(x) N\left(x \mid \mu, \sigma^2\right) dx = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \tag{1.24}$$

Where we can set μ to the mean and σ^2 to the variance of the latent posterior process. This gives the predictive rate parameter at test points. Notice that this is not the same as just taking the link of the mean of the latent posterior since the distribution in output space is not symmetric around its mode anymore.

1.8.4 Limitations of Gaussian Processes

While GPs offer a convenient way to express assumptions about data, there are also some drawbacks that need to be considered when working with this class of model.

Firstly, the computational complexity of computing the posterior of a GP is $O(N^3)$ in time and $O(N^2)$ in memory, which prohibits the direct application of GPs to datasets larger than 10.000 data points without specialized hardware. Fortunately, there has been progress to speed up inference with sparse versions of GPs (Hensman *et al.*, 2013). These inference schemes have been implemented in a number of libraries (GPy, 2012; Matthews *et al.*, 2017; Gardner *et al.*, 2021).

Secondly, using non-Gaussian priors for GPs necessitates the approximation of the posterior process. One efficient way of approximating the posterior is with Variational Inference (VI). Instead of calculating the posterior directly, VI makes use of an approximate Gaussian distribution that is optimized to be close to the intractable posterior in terms of its Kullback-Leibler divergence (KL). Thus, instead of calculating the likelihood of the model, one calculates an approximate posterior with an evidence lower-bound (ELBO) with respect to the hyperparameters. This ELBO is the lower bound of the log marginal likelihood of the true posterior process. There is no guarantee that the ELBO is close to the theoretical log-likelihood that could be reached with more expensive methods such as Monte-Carlo sampling, but in practice, it is often accurate enough to work with these models.

1.8.5 Overfitting

As discussed in Section 1.8.1.1, GPs have an inbuilt penalty term for more complicated model structures. This is because we integrate over the complete hypothesis space to calculate the likelihood. More complex models will have a wider hypothesis space than simple models. Therefore, complex models tend to have a better fit to the data regardless of whether they capture the data-generating process. In Fig 1.17, I illustrate the posterior function of three GPs with a squared exponential kernel with fixed lengthscales. Visually, it is readily apparent that small lengthscales will lead to overfitting. This is because small lengthscales increase the flexibility of the model. When calculating the marginal likelihood, this flexibility is penalized. The same principle applies to the variance parameter, where higher variances are penalized to prevent overly complex models.

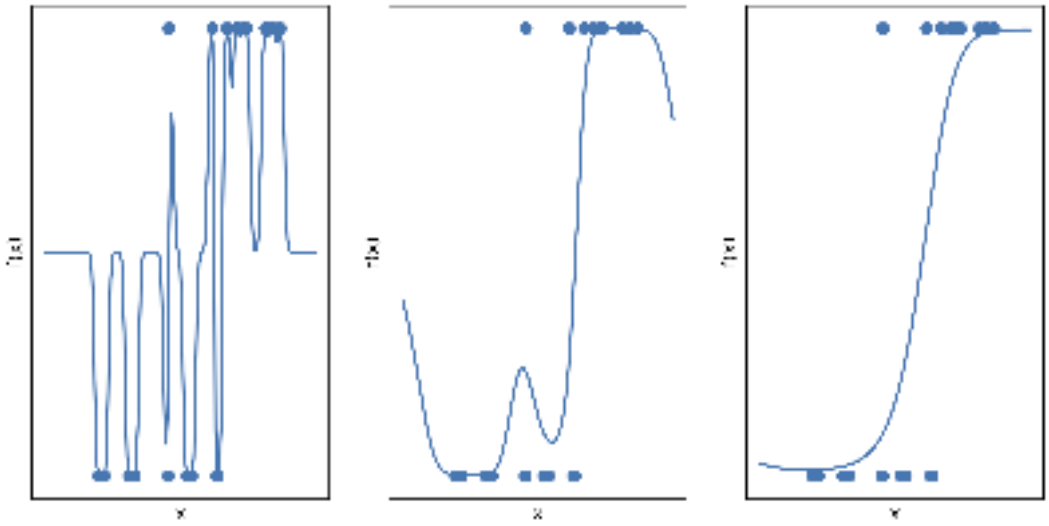


Figure 1.17 | Gaussian Process data fit with different lengthscale hyperparameters. The blue line indicates the posterior estimate of the rate parameter of a GP with Bernoulli likelihood and a squared exponential kernel with a fixed variance of 4.18 and varying lengtscales (*left*: 0.01, *middle*: 0.1, *right*: 0.3). The GP was trained with variational inference. The blue points are the training data. . *Figure generated by Max Frank.*

This principle is illustrated in Figure 1.18, which shows the ELBO estimates for a grid of hyperparameter combinations. There is an optimum in the ELBO at intermediate variance and lengthscale, where there is an optimal tradeoff of data fit and model complexity.

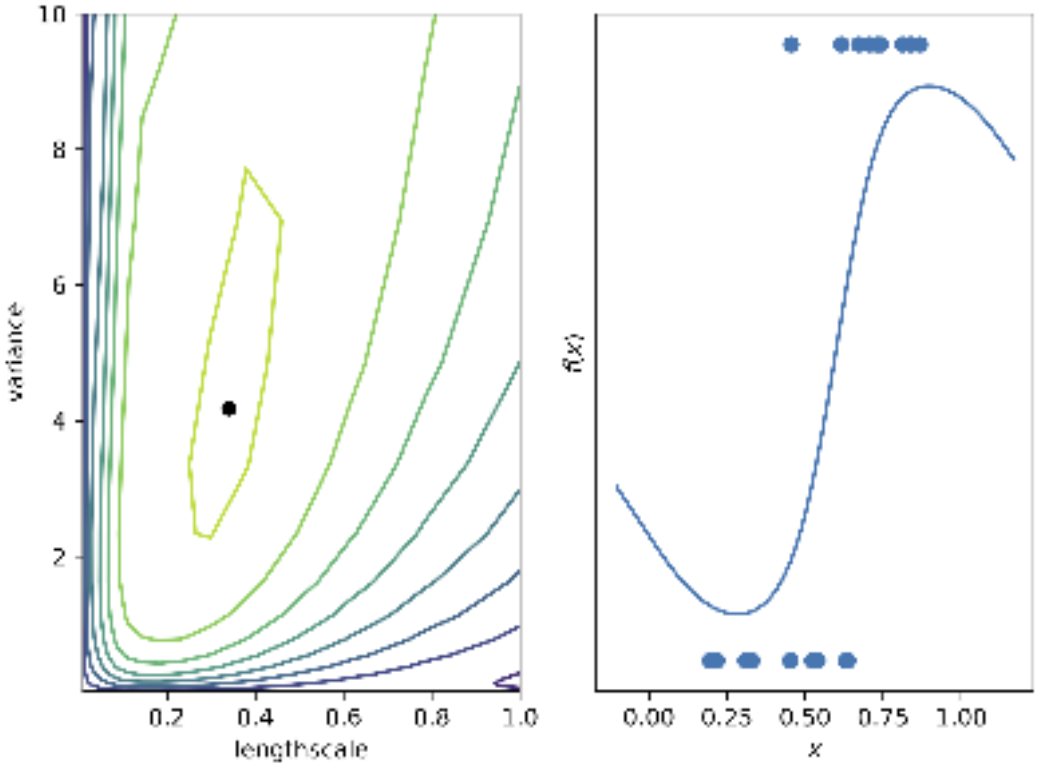


Figure 1.18 | Marginal likelihood estimate of GP for different hyperparameter settings. The *left* panel shows the contours of the ELBO estimate of the log marginal likelihood of a GP with Bernoulli likelihood and a squared exponential kernel. There is a maximum of the ELBO surface at variance=4.18 and lengthscale=0.34, as indicated by the black dot. The *right* panel shows the posterior rate estimate (blue line) of the GP with the highest ELBO. The blue points are the training data. *Figure generated by Max Frank.*

1.8.6 Hypothesis tests using Gaussian processes

The ability to encode assumptions about the data-generating processes and the robustness to overfitting makes GPs an attractive tool to model many types of data. In research applications, fitting these models is often a means to test different hypotheses. In this Section, I will describe how GP models can be used fto decide between a null hypothesis and an alternative hypothesis based on observed data points.

There are different approaches that can be used once a GP is conditioned on a set of input data.

1.8.6.1 Hypothesis testing based on hyperparameter estimates

One approach is to evaluate the posterior estimate of the hyperparameters of the GP. Depending on the setup of the GP model, the hyperparameters will have interpretable meanings for the underlying function.

For example, in the case of GP regression, the lengthscale parameter of a squared exponential kernel is inversely proportional to the influence of the input dimension

on the posterior process. As the lengthscale gets larger, the covariance between input points decreases. In the case of multivariate regression, this can be used to determine which inputs are important to determine the output values. If we extend the squared exponential kernel from Section 1.8.2 to multiple input dimensions, we get

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T M (\mathbf{x}-\mathbf{x}')}{2}\right); M = diag(\boldsymbol{\ell})^{-2} \tag{1.25}$$

where $\boldsymbol{\ell}$ is a vector of lengthscales of length equal to the number of input dimensions D. The relevance of the dth input dimension is inversely proportional to ℓ_d. This is called Automatic relevance determination (Neal, 1996) and can be used to remove irrelevant input dimensions.

1.8.6.2 Hypothesis testing based on the marginal likelihood

Similarly, whether an output is linearly dependent on an input by looking at the lengthscale hyperparameters of a regression model where the slope is modeled by a GP with a squared exponential kernel (Mulder, 2023).

$$\mathbf{y} = \boldsymbol{\beta}(\boldsymbol{x})\mathbf{x} + \epsilon \tag{1.26}$$

$$\boldsymbol{\beta}(\mathbf{x}) \sim \mathscr{GP}\left(\mathbf{0}, k\left(\mathbf{x}, \mathbf{x}'\right)\right) \tag{1.27}$$

where $k\left(\mathbf{x}, \mathbf{x}'\right)$ is a squared exponential kernel. As the inverse of the lengthscale parameter of the kernel approaches zero, the slope function $\beta(\mathbf{x})$ will become constant, and thus

$$\mathbf{y} = \beta\mathbf{x} + \epsilon \tag{1.28}$$

These two examples illustrate the interpretability of the hyperparameters of GP models. However, this does not directly provide a statistical estimate that allows the quantification of the confidence level of rejecting the null hypothesis.

A common test statistic in Bayesian modeling is the Bayes factor (BF). The Bayes factor is the ratio of the marginal likelihoods of two models. The marginal likelihood of a model is the evidence or likelihood of the model after seeing data integrated over the priors of the parameters of the model. In the case of GP regression, two models that correspond to the null and the alternative hypothesis can be used to compute a BF of the hypotheses. The marginal likelihood for a GP model M (including the hyperparameters of the model) is

$$p(D \mid M) = \int p(\theta \mid M) p(D \mid \theta, M) d\theta \tag{1.29}$$

where D is the observed data and θ is a vector of the hyperparameters of the model. This integral might be analytically intractable, depending on the prior over the

hyperparameters $p(\theta \mid M)$. Therefore, in practice, the hyperparameters are often optimized for both models, and the marginal likelihood is approximated by the maximum a posteriori likelihood with

$$p(D \mid M) = p(D \mid \hat{\theta}, M)p(\theta \mid M) \tag{1.30}$$

where $\hat{\theta}$ is the optimized set of hyperparameters. In practice, the prior over the hyperparameters is often chosen to be uninformative, which means the prior term can be excluded. Thus, the approximation of the BF for two models M_0, M_1 corresponding to the null and the alternative hypothesis becomes

$$\frac{p(D \mid M_1)}{p(D \mid M_0)} = \frac{p(D \mid \hat{\theta_1}, M_1)}{p(D \mid \hat{\theta_0}, M_0)} \tag{1.31}$$

This is sometimes referred to as the likelihood ratio. However, this approximation has to be handled with care if the models vary in their number of hyperparameters. Since the full BF integrates over all hyperparameters, models with different numbers of parameters can be compared. However, with the maximum likelihood approximation, this robustness is lost.

Using the likelihood ratio as a test statistic, one can make use of Wilks theorem (Wilks, 1938), stating that if the null hypothesis is true and the number of observed data points approaches infinity, the likelihood ratio statistic will approach a chi-squared distribution with degrees of freedom equal to the difference in number of parameters between the two models. This result has been used for calculating p-values for hypothesis tests with GPs applied to genomics (Svensson *et al.*, 2018; BinTayyash *et al.*, 2021). However, Wilks' theorem only holds true if the null model and the full model are strictly nested, meaning that the null model's parameters lie strictly within the parameter space of the full model. This assumption can often be violated with GP hypothesis tests, as will be demonstrated in the next Chapter.

1.8.7 Applications in genomics

The ability to model smooth nonlinear functions without the need to explicitly know the parametric function of the data-generating process makes GPs an attractive way to model gene expression over time courses or in spatial contexts. In this Section, I will give an overview of studies that use GPs to model genomic data.

1.8.7.1 Time course data

An early application of GPs to microarray time-course data was in the modeling of TF regulation networks. Lawrence *et al.*, 2006 studied the response of five target genes of the p53 tumor suppressor gene. They used linear and nonlinear modes of transcriptional regulation to calculate p53 expression levels only from observations of its target genes. The time course data of the target genes was modeled as a GP with a

squared exponential covariance function. They could then compare their estimations of p53 expression levels to experimentally measured levels.

In a similar fashion, Kirk and Stumpf (Kirk and Stumpf, 2009) fit GP models to gene-expression data of 800 *Arabidopsis thaliana* genes, measured in duplicates at 11 time points. They were interested in obtaining error distributions for parameter estimates of different models of gene regulation. Their approach involved sampling the GP posterior to obtain a bootstrapped dataset of gene expression time courses. They then applied the models of gene regulation to their samples to estimate the distribution of model parameters.

Stegle et al. introduced a GP model called GPTwoSample that allowed for the decision of whether two-time courses of a gene were different between two experimental conditions (Stegle *et al.*, 2010). The idea was that if the null hypothesis (no differential expression between conditions) was true, then both sets of observations could be explained by draws from a single distribution and, therefore, modeled with a single GP. In the case of differential expression, the data would need to be modeled by two independent GP models for the two conditions. The ratio of the likelihood of the two models could then be used to rank the genes according to the likelihood of differential expression. The GP models used a squared exponential kernel and also introduced the idea of using a heavy-tailed non-Gaussian likelihood to be more robust to outlier observations that are frequently observed in genomics data. Importantly, the likelihood ratio of this model can be evaluated not only for all observations but also for subsets of observations at a particular time point, allowing the authors to derive a mixture of expert models that decide for each time point if a gene is differentially expressed. This allowed the question to be asked not only if a gene is differentially expressed but also when differential expression occurs. Note that there is an important difference between performing independent tests for differential expression at each time point and this model since the covariances between measurements closely together in time are considered.

The two-sample hypothesis test was subsequently expanded to allow a continuous measure of differential expression that was not confined to time points with experimental data (Heinonen *et al.*, 2015). This was achieved by comparing the posterior concentration of the null model and the alternative model as an estimate of the confidence of the models. Both of these models need to use heuristics in order to determine the time of divergence of gene expression between conditions. Yang et al. developed a GP model with the divergence time point as an explicit parameter, allowing for a fully Bayesian estimate of when two-time series first diverge (Yang *et al.*, 2016).

Another type of hypothesis test with GPs, termed the one-sample test, was implemented by Lawrence et al. (Kalaitzis and Lawrence, 2011) shortly after the two-sample test by Stegle et al. (Stegle *et al.*, 2010). Here, the question is whether gene expression

of a gene varies over time given a single time-course measurement. Again, the test is based on computing the likelihood ratio of two models. A null model that does not allow for variation in gene expression over time and an alternative model that does. They use a GP with a squared exponential kernel and a robust Student-t likelihood to model the alternative hypothesis of a temporally varying gene. The null model is constructed by setting the variance parameter of the kernel to zero and the lengthscale parameter to infinity. They use the likelihood ratio of these model to rank genes according to their likelihood of being temporally variable.

GPs have also been extensively used to model gene expression in single-cell datasets. Here, time series are often not created by performing sequencing experiments at different time points but created from a single or a few sequencing runs using pseudotime methods. This makes temporal estimates prone to error. Furthermore, sequencing depth is reduced for an individual cell compared to bulk sequencing methods, putting increased importance on correct modeling of the noise with appropriate likelihood functions. The recently introduced GPcounts model implements two likelihoods suitable for single-cell measurements: the negative binomial likelihood and the zero-inflated negative binomial likelihood. They use their model to perform both two-sample and one-sample hypothesis tests and show on simulated data that hypothesis tests with the appropriate model likelihood have higher power compared to the same tests with Gaussian likelihood.

Because pseudotime inference in single-cell data can lead to the discovery of a branching structure in developmental datasets, GP models have also been developed to explicitly model time series that contain bifurcation events and to assign observations to branches. Note that compared to the previously discussed bulk sequencing problems, there are some subtle differences. In the bulk scenario, comparing two-time series of the same gene in two conditions, the assignment of each data point to each branch is known and fixed. Because pseudotime inference is based on noisy data, however, the association of cells to different branches should be taken with care. For this reason, GP models based on an overlapping mixture of Gaussians (Lázaro-Gredilla *et al.*, 2012) have been developed. These models construct branched trajectories and assign each cell to the trajectories in a probabilistic manner. These models can then be used to compute branching events either globally for all genes (Lönnberg *et al.*, 2017) or locally for a single gene (Boukouvalas *et al.*, 2018a).

1.8.7.2 Spatial data

With the advent of single-cell methods that allow the profiling of gene expression in a spatially resolved manner, there was an increased demand for methods detecting spatially variable genes. Svensson et al. developed such a method with SpatialDE (Svensson *et al.*, 2018; Kats *et al.*, 2021). SpatialDE can be considered an extension to the one-sample test introduced in the previous Section. It trains two GP models, allowing for spatial variability in the kernel or keeping expression constant in space.

The likelihood ratio of these models was then used to calculate a statistical estimate of spatial variability, using Wilks theorem to produce calibrated p-values (see Section 1.8.6.2). Furthermore, they provide an effect size estimate that measures the fraction of spatial variance, i.e., the fraction of total variance that is explained by spatial variation. The model also allows for the clustering of spatial patterns, which means that groups of genes with similar expression patterns can be automatically found.

2 | The GPmeth model for epigenetic single-cell data

2.1 Derivation of the GPmeth model

Based on the previous Chapters, I designed model that makes use of all the available information within single-cell epigenetics data to detect changes over the course of continuous trajectories. The first step in this process is to use single-cell gene expression data to assign cells to a position within pseudotemporal trajectories (see Section 1.7.1). This constitutes the global analysis part of the framework. The subsequent local analysis involves the identification of epigenetically regulated regions that change over the course of the inferred trajectories. This involves considerable challenges, as discussed in Section 1.7.2. These challenges are addressed by the GPmeth model.

In this Section, I describe the GPmeth model and the considerations that went into its formulation.

2.1.1 Model Description

The input data for GPmeth are single-cell base-resolution epigenetic data. Each data point for CpG/GpC site i, in cell j, can be described by

$$y_{ij} = Bern(\rho_{ij}) \tag{2.1}$$

Where ρ_{ij} is the unknown true (Bernoulli distributed) methylation rate and y is the observation in the data. Following the argumentation of Section 1.7.2 we assume that there are no hemimethylated sites, which is why methylation can be described as binary:

$$y = \begin{cases} 1 & N_{methylated}/N_{total} > 0.5 \\ 0 & N_{methylated}/N_{total} < 0.5 \end{cases} \tag{2.2}$$

Where $N_{methylated}$ are the number of reads indicating a positive methylation state, and N_{total} are the total number of reads at that site. Sites with $N_{methylated}/N_{total} = 0.5$ are discarded. For the majority of observations, all reads are methylated or unmethylated (Fig 2.1).

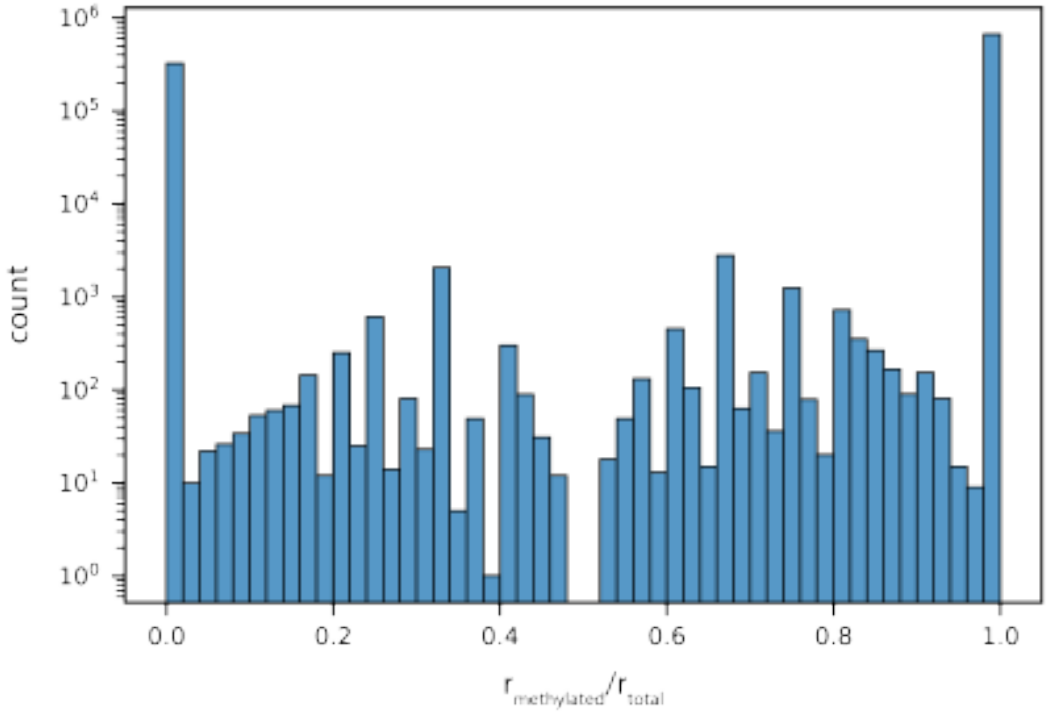


Figure 2.1 | Genome-wide scNMT methylation rate measurements. Histogram of observed CpG methylation rates in individual cells. The data shown represents a subset of one million methylation observations from the mouse gastrulation dataset. CpG sites with 0.5 methylation rate have been removed. *Figure generated by Max Frank.*

2.1.1.1 Modeling individual CpG/GpC sites

Each cell j can be associated with a position in pseudotime or cell grouping (Fig 2.2, a) using the methods described in Section 1.7.1.3. As motivated by Section 1.7.1.3 we expect the value of the underlying methylation rate to vary smoothly along time.

We start out with a model describing the underlying methylation rate ρ of an individual CpG/GpC site i along that temporal axis. The process of generating input data for an individual CpG/GpC site is outlined in Figure 2.2. First there is a global analysis producing a branch assignment as well as a pseudotime estimate t_j for each cell. Measurements of a CpG/GpC site can then be sorted along their pseudotemporal, and the methylation rate can be modeled. I choose a Gaussian Process to model the methylation rate for the reasons introduced in Chapter 1.8.5.

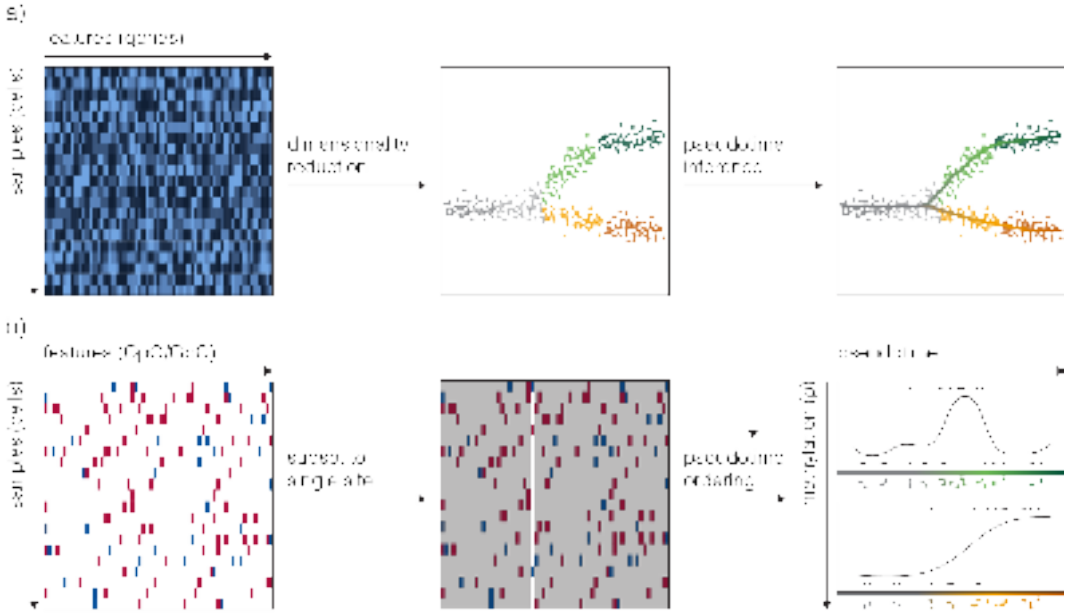


Figure 2.2 | Workflow to generate input data for GPmeth for individual sites. The top row shows how cells are assigned a pseudotemporal value and associated with a developmental branch. The bottom row shows the subsetting of the methylation data to a specific CpG/GpC site and the ordering of the cells according to pseudotime. Each regulatory region that is produced like this can then be modeled with the GPmeth model. *Figure generated by Max Frank.*

With the assumption of smoothly varying methylation, we can define a function $f_i(t)$ that describes the methylation rate trajectory of each individual CpG/GpC site i.

$$\rho_i = f_i(t_j) \tag{2.3}$$

The temporal dynamics of methylation rate are described by a Gaussian process.

$$f_i(t) \sim GP(\mu_i, k_i) \tag{2.4}$$

Because GPs produce outputs in the domain $(-\infty, \infty)$ but ρ lies in the range $[0, 1]$, the Gaussian process output pushed through an inverse probit link function (see Section 1.8.3) to guarantee a valid rate output:

$$\rho_i = \Theta(g_i(t)) \tag{2.5}$$

$$g_i(t) = GP(\mu_i, k_i) \tag{2.6}$$

Here, g_i is a nuisance function that is converted to the rate parameter with Θ, which represents the cumulative distribution function of the standard normal distribution. This ensures that the output of the GP is bounded between 0 and 1. Note that I dropped the index i of the CpG/GpC site in the second equation for brevity. From now on, it is assumed that each GP model will represent a single CpG/GpC site.

When using GPs with Gaussian likelihoods, it is common practice to scale the output data $\mathbf{y}$ to zero mean and unit variance before fitting the model. This ensures that the GP can be parametrized with a constant zero mean. In the case of binary output data, this is not possible. I, therefore, parametrize the GP with a constant mean:

$$\bar{\mu_i} = \Theta^{-1}(\bar{\mathbf{y}_\mathbf{i}}) \tag{2.7}$$

Where $\bar{\mathbf{y}}$ is the empirical mean of all observations. Since μ is the mean of the latent function f, $\bar{\mathbf{y}}$ has to be mapped to the space of the latent function via the inverse of the link function.

The GP is also parametrized by the kernel function k_i. To differentiate this kernel in later Sections and to make clear that this kernel models temporal changes I will drop the index and refer to it as k_{time}. k_{time} produces the covariance matrix of the GP. The kernel function contains the assumptions of the model. One way to encode our expectation of smoothly varying methylation rate across the genome with a radial basis function (RBF) kernel of the form

$$k_{time}(t_j, \dot{t_j}) = \sigma^2 \exp\left(-\frac{\|t_j - \dot{t_j}\|^2}{2l^2}\right) \tag{2.8}$$

Where t_j represents the pseudotime coordinate of the cell that the observation of CpG/GpC site i was made in. σ and l are hyperparameters of the model. The lengthscale parameter l controls the smoothness of the model, while σ is the kernel variance determining the amplitude of changes in the function. Other kernels can be chosen, such as a Matérn kernel or a linear kernel to encode varying assumptions about the data-generating process (see Section 1.8.2 and Fig 2.3).

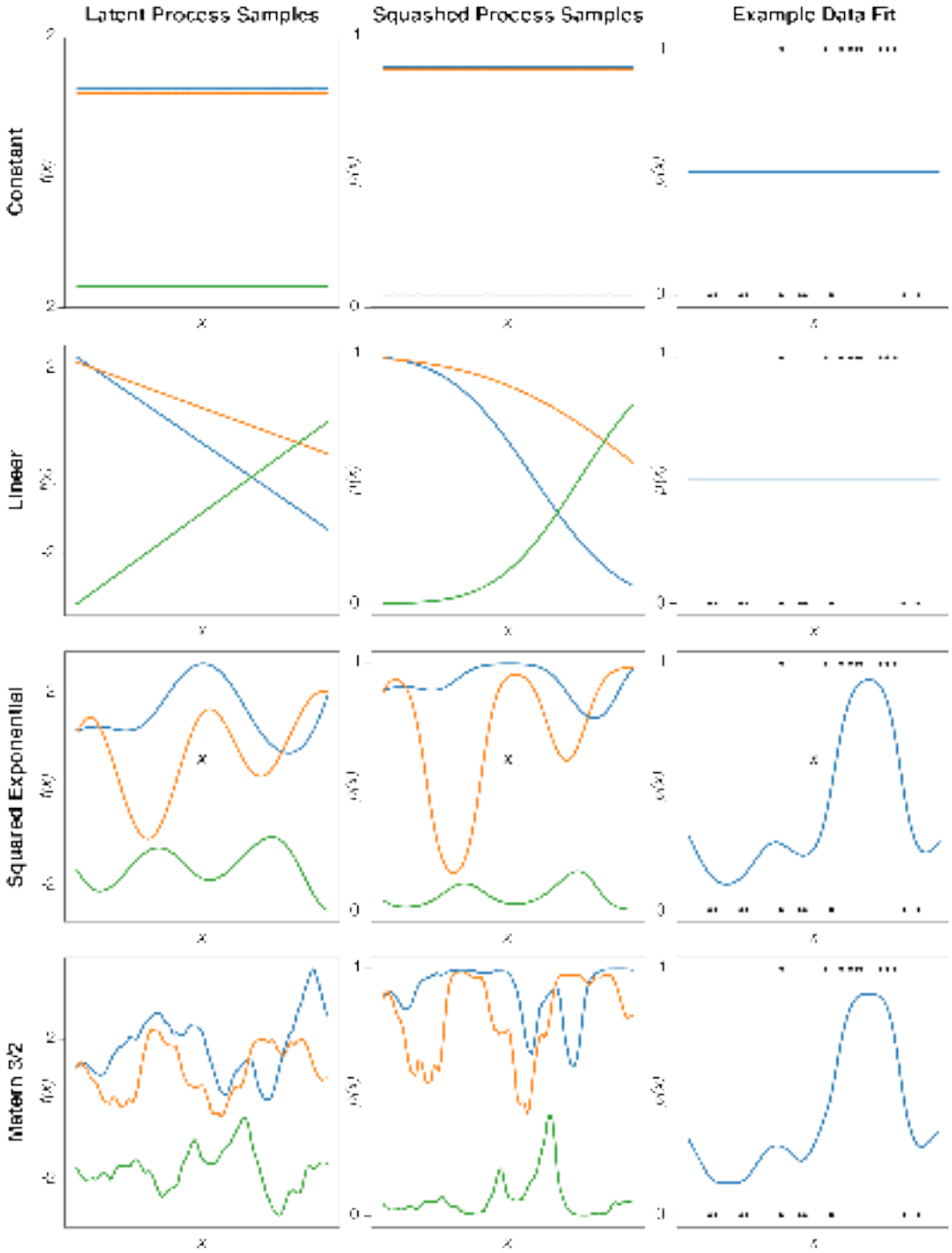


Figure 2.3 | Kernel choices for binary Gaussian Processes. Shown are samples from latent processes $f(x)$ (*left column*), their equivalents after transforming through the link function $\rho(x)$ (*middle column*), and examples of the GP posterior (*right column*) where the blue line represents the posterior estimate of $\rho(x)$ and the black dots are the training observations. Note that linear kernels become nonlinear through the transformation but remain monotonic functions. Therefore, the linear kernel is not able to properly model the example observations where the methylation rate is increasing and then decreasing, leading to a flat posterior. *Figure generated by Max Frank.*

As discussed in Section 1.8.3, the marginal likelihood of this model cannot be computed analytically. Therefore I use a variational inference approach to compute the ELBO

as an approximation. The hyperparameters of the model are optimized by standard gradient descent using the Scipy optimizer (Virtanen *et al.*, 2020).

While these kernels allow to flexibly model temporal variation of the methylation rate, we can also define a kernel that only allows constant functions over time. Then

$$k_{time}(t_j, \dot{t_j}) = \sigma^2 \tag{2.9}$$

where σ is the kernel variance. In the case of a Bernoulli likelihood with a probit link function, this variance parameter will always collapse to zero. The trained GP model then is

$$g(t) = N(\bar{\mu}, 0) \tag{2.10}$$

which leads to

$$\rho = \bar{y} \tag{2.11}$$

and the marginal likelihood is simply

$$p(\mathbf{y}|\rho) = \prod_{j=1}^{n} \rho^{y_{i,j}} (1-\rho)^{(1-y_{i,j})} \tag{2.12}$$

This models allow to perform a range of hypothesis tests on the observed data. These include:

- Does the methylation rate of the observed CpG/GpC site change over the course of a pseudotemporal trajectory? This can be achieved by computing the likelihood ratio between a model with a temporally variable kernel k_{time} and a model with a constant temporal kernel (see Section 2.1.2).
- Does the methylation rate vary smoothly over time, or can the variance be explained by grouping cells into cell types? This can be achieved by comparing a model with a smoothly varying kernel k_{time} to a model that has a constant methylation rate ρ_c for each identified group of cells c.
- Does the methylation rate vary linearly or non-linearly over time? This can be achieved by computing the likelihood ratio of a model with a non-linear kernel, such as the squared-exponential kernel or the Matérn kernel, to a model with a linear kernel.

In all these cases, the likelihood ratio of the models encoding different beliefs about the data-generating process is the test statistic that can be used to rank different CpG/GpC sites according to their likelihood of violating the null hypothesis. To get

from the likelihood ratio to a test statistic that is properly controlled at a nominal false discovery rate (FDR), the model must, however, still be calibrated (see Section 2.2.4).

I now have devised a range of tests for individual CpG/GpC sites. In practice, however, these tests would only be useful in very targeted applications where one is interested in only a handful of sites. For any genome-wide differential methylation analyses, testing individual CpG/GpC sites would run into the issue of multiple testing. Furthermore, researchers are typically not interested in individual CpG/GpC sites but want to test whether there is differential methylation in a regulatory region. Therefore, I will expand this model in the next Section.

2.1.1.2 Modeling regulatory regions

The models described in the previous Section allow for smoothly varying methylation rate over time but do not account for the co-variation of neighboring CpG/GpC sites. As discussed in Section 1.7.2.1, CpG/GpC sites that are closely positioned along the genome cannot be treated as independent observations but co-vary. Therefore, it makes sense to study CpG/GpC in the context of a genomic region as opposed to an individual site. Figure 2.4 shows the workflow of producing input data for the extended model that will be described in this Section.

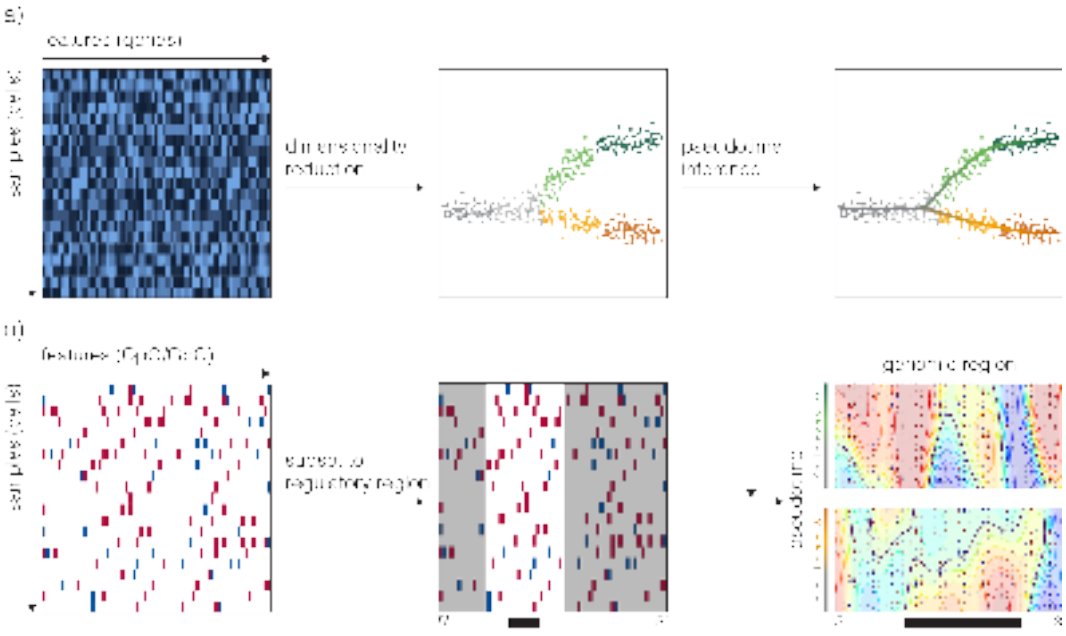


Figure 2.4 | Workflow to generate input data for GPmeth. The *top* row shows how cells are assigned a pseudotemporal value and associated with a developmental branch. The bottom row shows the subsetting of the methylation data to a specific regulatory region (e.g., enhancers, promoters, etc.) and the ordering of the cells according to pseudotime. Each regulatory region produced like this can then be modeled with the GPmeth model. The output of the GPmeth model for two trajectories is shown in the *bottom right* panel. Points represent individual CpG/GpC measurements, and the contours depict the posterior mean prediction of the model. *Figure generated by Max Frank.*

Because genomic covariance results from multiple possible regulatory processes, it is challenging to come up with an explicit model that describes the dependence of a site on its neighbors. The simplest approach would assume constant methylation for each cell across a limited genomic region. This would mean that the same model as described above can be used where each CpG/GpC of a cell that falls into the modeled region shares the same model. This assumption works well if the boundaries of regulatory regions are precisely known before the analysis. However, since these boundaries are often estimated, it would be desirable to derive a testing procedure that is somewhat robust to the choice of boundaries. I, therefore, take a non-parametric approach with a Gaussian process to model methylation rate across the genome as well. Concretely, I express the covariance between sites with a squared-exponential kernel of the form

$$k_{genome}(x_i, \dot{x_i}) = \sigma^2 \exp\left(-\frac{\|x_i - \dot{x_i}\|^2}{2l^2}\right) \tag{2.13}$$

Where x_i represents the genomic coordinate of the observation. σ and l control the smoothness and magnitude of change of the methylation rate across the genome.

As discussed in Section 1.8.2.2 there are different ways to model the co-variation of the methylation rate across the genome and across pseudotime. One choice would be to express the total covariance as an additive combination of k_{genome} and k_{time}

$$k = k_{genome} + k_{time} \tag{2.14}$$

This would mean that we assume that there are two independent processes that modulate the methylation rate in the genome dimension and in the time dimension. This assumption is likely to be too restrictive. For example, there are DNA-binding proteins that are sequence-specific and will, therefore, operate on a specific part of the genome. If these regulators are differentially expressed over time, they can influence the methylation rate, which is dependent on both genomic position and time. Another option is to multiply the genomic and temporal kernel

$$k = k_{genome} * k_{time} \tag{2.15}$$

This is a more flexible structure that will be able to model the influence of regulatory factors that depend on both time and genomic position. However, there are clear cases of factors that will not be time sensitive but dependent on position. These include nucleosomes that are not displaced during the course of time or regions that are permanently silenced by methylation. Therefore, I use a combination of the two kernels above

$$k = k_{genome} + k_{time} * k`_{genome} \tag{2.16}$$

This assumes methylation rate is an additive result of the influence of genomic factors that are constant over time (modeled with k_{genome}) and genomic factors that vary in their effect across time (modeled with $k_{time} * k`_{genome}$). This explicitly separates factors that are static over time (such as the DNA sequence, which is an important component in modeling DNA methylation (Angermueller *et al.*, 2017), from regulatory influences that vary over time but will also have a positional component, such as DNMT binding. This is the kernel of the full model for the methylation rate of a genomic region. I can now use this model to answer the question of whether there is differential methylation over the course of a developmental process. As before, this can be achieved by computing the likelihood ratio of the full model and a model that corresponds to the null hypothesis of no variation over time. This model can easily be formulated by removing the part of the kernel that takes into account temporal information

$$k_{null} = k_{genome} \tag{2.17}$$

Figure 2.5 shows draws from GP priors with the different kernels mentioned above. One can see that with the additive kernel, the temporal dynamics extend throughout the whole genomic window, inconsistent with a locally binding regulatory protein. With the multiplicative kernel, there is a dependency between the absolute value of methylation and the rate of methylation rate change at any point in the genome. This is also not a desirable property of the model. With the full kernel, the rate of change and absolute methylation level are no longer coupled. I, therefore, use the $k_{genome} + k_{time} * k`_{genome}$ kernel combination for all subsequent experiments.

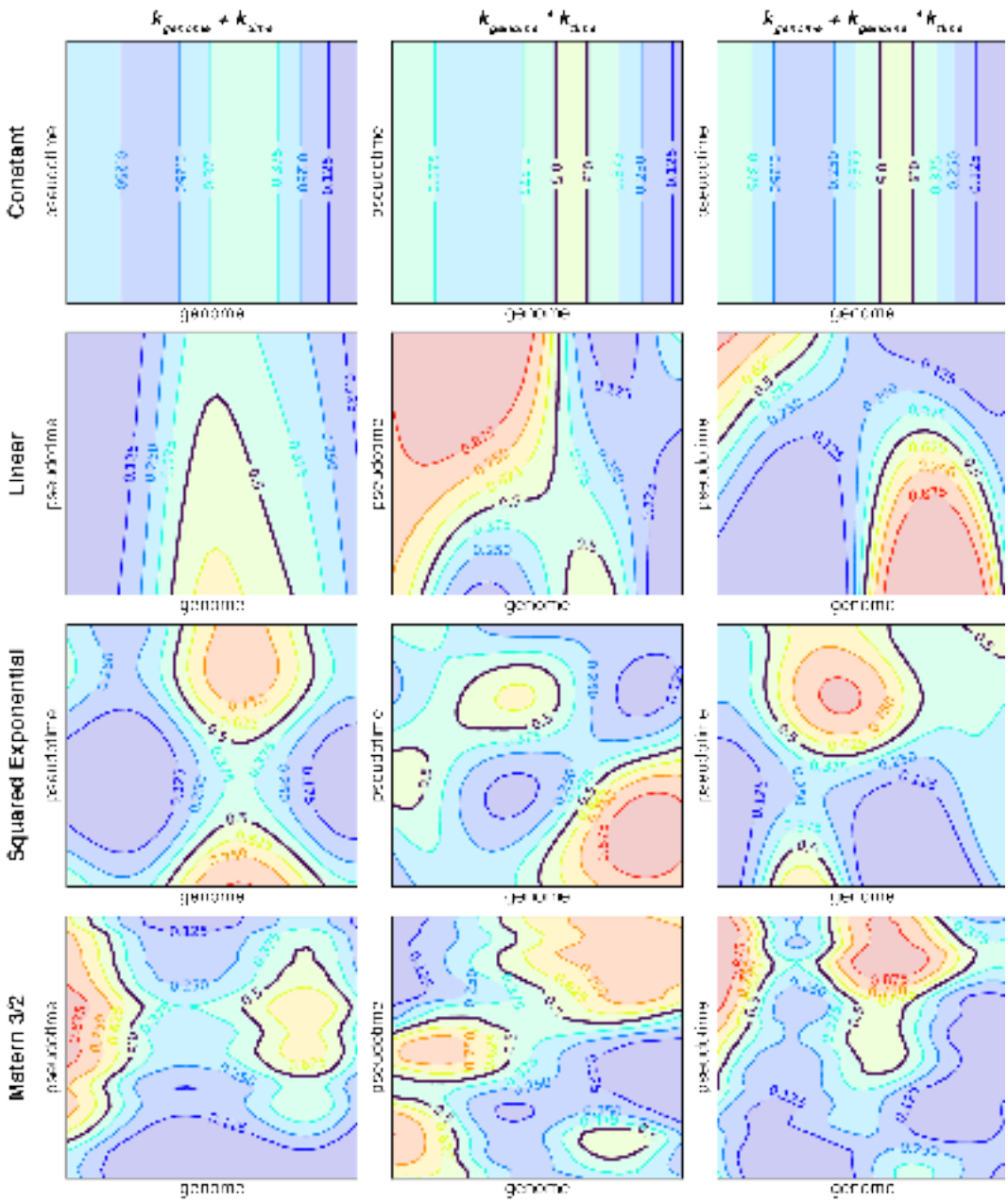


Figure 2.5 | Kernel combinations for methylation rate modeling. Shown are samples from the GP prior. Contour lines represent the methylation rate ρ(x). The column indicates the kernel construction of the GP. The row indicates the type of kernel that k_{time} is. k_{genome} is always a squared exponential kernel. All kernels have been instantiated with variance=1 and lengthscale=0.25. *Figure generated by Max Frank.*

With these kernels in hand, I can now formulate a range of models that correspond to specific assumptions about the data. The models are listed in Table 2.1.

Model name	k_{genome}	k_{time}	Assumption
Constant	0	0	No methylation changes across region or time
ConstantLinear	0	Lin	Methylation rate changes monotonically over time but not across the region
ConstantRBF	0	RBF	Methylation rate changes smoothly over time but not across the region
ConstantMatérn	0	Matérn3/2	Methylation rate changes less smoothly over time but not across the region
ConstantCategorical	0	Categorical	Methylation rate changes across cell types but not across the region
RBFConstant	RBF	0	Methylation changes only across the region
RBFLinear	RBF	Lin	Methylation rate changes monotonically over time and smoothly across the region
RBFRBF	RBF	RBF	Methylation rate changes smoothly over time and smoothly across the region
RBFMatérn	RBF	Matérn3/2	Methylation rate changes less smoothly over time and smoothly across the region
RBFCategorical	RBF	Categorical	Methylation rate changes across cell types and smoothly across the region

Table 2.1 | Models for describing methylation rate in a regulatory region.

2.1.2 Differential Testing

From the models described in 2.1, a range of statistical tests can be derived where test statistics are obtained with the comparison of models that correspond to the null and the alternative hypotheses. In the case of testing for differential methylation/accessibility over a time course, the null hypothesis is that methylation/accessibility does not change, which corresponds to a model without a temporal

kernel. The alternative hypothesis can be expressed by one of the models in Table 2.1, depending on the prior assumptions a researcher makes about the data. Figure 2.6 shows the models that can be tested against each other if the assumption is that the methylation rate is constant across the tested genomic window. In this case, the data can be visualized by aggregating the signal of all CpG/GpC sites within the input window. Here, a *Constant* model is tested against the *ConstantCategorical*, *Constant-Linear*, and *ConstantRBF* models. The *ConstantCategorical* model will describe each group (category) of cells separately. The *ConstantLinear* model allows for linearly increasing or decreasing methylation rate over time, and the *ConstantRBF* allows for nonlinear temporal trajectories. The comparison of the marginal likelihoods of the pairs of null and full models allows the calculation of test statistics that are used to reject the null hypothesis. This will be described further below.

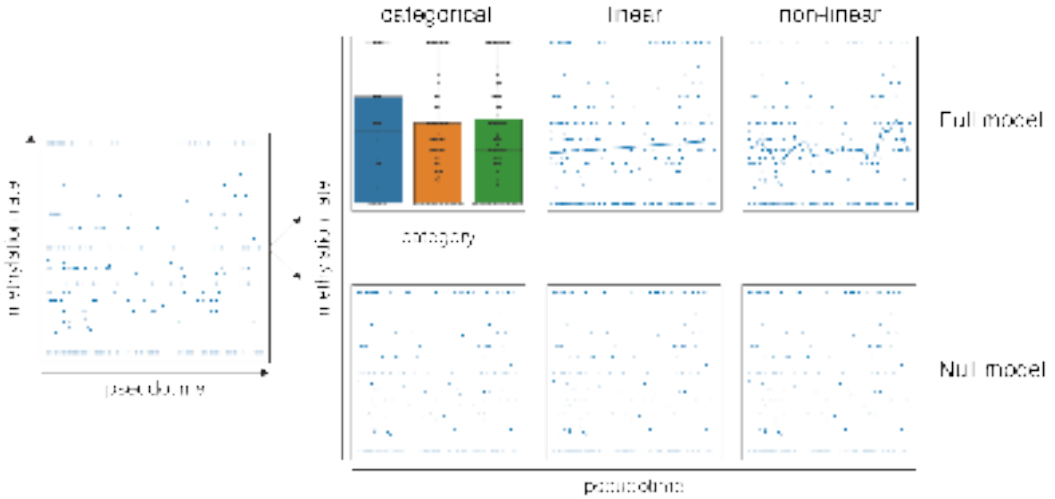


Figure 2.6 | GPmeth model comparisons without genomic variability. The *left* panel shows the input data where each point represents the average methylation rate ρ (y-axis) of all CpG/GpC sites within a predefined genomic window that was observed in a cell with associated pseudotime (x-axis). The right panel depicts three different hypothesis tests where a null model (*bottom row*) is compared to the respective full model above (*top row*). The null model for all hypothesis tests is a *Constant* model that does not allow for varying methylation rate and will regress to the mean methylation rate of all cells. The full models correspond to the *ConstantCategorical*, *ConstantLinear*, and *ConstantRBF* models. Models are described in Table 2.1. *Figure generated by Max Frank.*

As discussed in Section 1.5 and 2.1.1.2, the assumption of constant methylation rate within predefined genomic windows is often hard to justify (see Section 1.7.2, and GPmeth can also test for differential methylation/accessibility without this assumption. Figure 2.7 shows the same tests as discussed above while allowing for methylation rates to be variable within the genomic region. Note that the majority of the region is not differentially methylated/accessible. This would dilute the signal and decrease the power of the statistical test when averaging over the input window.

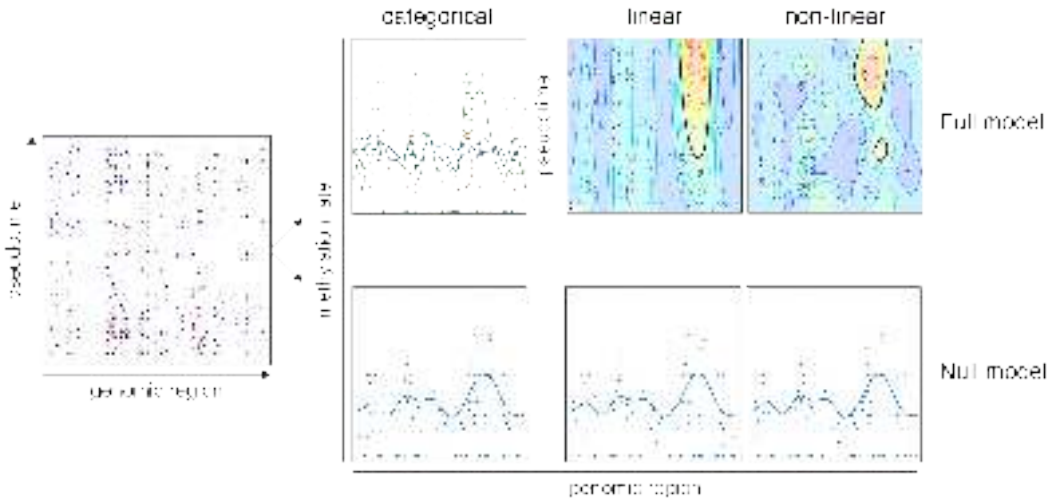


Figure 2.7 | GPmeth model comparisons with genomic variability. The *left* panel shows the input data where each point represents the genomic location (x-axis) and methylation status of a single CpG/GpC site (red=methylated, blue=unmethylated)in a cell with associated pseudotime (y-axis). The right panel depicts three different hypothesis tests where a null model (*bottom row*) is compared to the respective full model above (*top row*). The null model for all hypothesis tests is a *RBFConstant* model that allows for varying methylation rate only within the genomic window but not over time. In these plots, each point represents the average methylation rate of a single CpG/GpC site across all cells. The full models correspond to the *RBFCategorical*, *RBFLinear*, and *RBFRBF* models. Models are described in Table 2.1. *Figure generated by Max Frank.*

To obtain a statistical metric of differential methylation, we compute the likelihood ratios between the null model and the full models. Under the assumption of the null hypothesis, Wilks theorem Wilks, 1938 states that the negative log of these likelihood ratios should follow a χ^2-distribution with degrees of freedom according to the difference in hyperparameters between the full and the null model (see Section 1.8.6).

$$p(LLR \mid d) = \chi^2_d(-LLR) \tag{2.18}$$

Where L is the log-likelihood ratio of the full model with a temporal kernel and the null model without a temporal kernel.

This has been successfully employed in the context of GP hypothesis testing by Svensson et al. (Svensson *et al.*, 2018).

However, these models are not truly nested since that would require the parameters of the null model to be fixed to a value that lies strictly in the interior of the parameter space of the full model. In my case, if I set the variance parameter of k_{time} to zero, I recover the null model. Setting the variance to zero means that the parameter is fixed at the edge of the parameter space. Furthermore, due to the non-Gaussian likelihood, it is not possible to calculate the exact marginal likelihood of the model. Instead I calculate the ELBO estimate of the marginal likelihood. This turns the

inference problem into a numerical optimization problem. Therefore, the validity of the assumptions for the likelihood ratio statistic also depends on how good the approximation to the marginal likelihood is.

I, therefore, decided to validate the calibration of the model empirically by testing the model on synthetic data that represent the null hypothesis. This will be discussed further in Section 2.2.4.

2.1.3 Refinement of Differential Regions

In the above Sections, I formulated a model that describes the change in methylation rate over time for a regulatory region. In an ideal scenario, these regulatory regions should be chosen so that exactly one differential methylation 'event' takes up most or all of the region. In practice, researchers often do not have good knowledge about where in the genome differential methylation will occur. As mentioned before, a good initial guess for these regions is known regulatory elements such as enhancers and promoters. However, there is no guarantee that the boundary of these regions is chosen such that only one differential methylation 'event' happens in the chosen window.

Thus, if we get a significant test for a tested window, the next question arising is at which genomic coordinates the methylation rate change actually occurred. Fortunately, this can be readily answered since the model gives predictions of the methylation rates that can be evaluated with arbitrary precision throughout the region. A good measure of whether there is a biologically relevant change in methylation rate at a specific position in the genome is the effect size of the change or methylation rate difference. For example, many bulk studies use a threshold of a methylation rate difference of 0.3 between samples. With the model, I can ask for any point in the tested region what the maximum and minimum predictions for methylation rate are. The difference between those I termed Maximum Methylation Rate Change (MMRC), which can then be thresholded to produce refined regions of differential methylation (Fig 2.8). Note that the model is able to capture and distinguish two regions within the window that follow opposite trends of methylation rate change. This highlights the importance of a flexible nonlinear model.

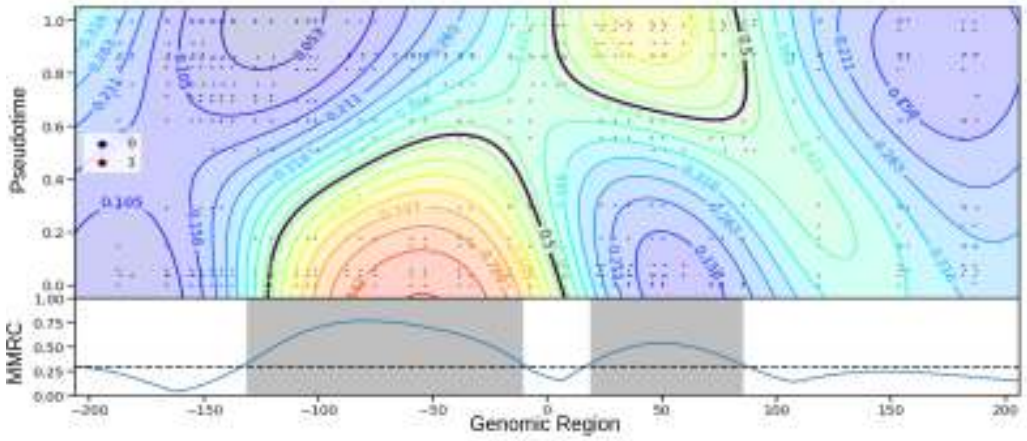


Figure 2.8 | Refinement of differential regions. The *upper* panel shows the posterior mean predictions of methylation rate of the full model (a GP with a squared-exponential kernel k_{time}) as contours. The training data (synthetically generated) are shown as points, with blue indicating an unmethylated site and red indicating a methylated site. The *lower* panel shows the maximum methylation rate change (MMRC) at each position in the genomic region as a blue line. The horizontal dotted line represents a threshold for MMRC of 0.3. The grey-shaded areas indicate refined regions where the MMRC is consistently above the threshold value. In this example, there are two differential events happening in close vicinity to each other, where one refined region is demethylated over time while the other becomes methylated. *Figure generated by Max Frank.*

2.2 Validation of the GPmeth Model on Synthetic Data

To validate the ability of GPmeth to model regulatory regions and test for differential expression, I created a synthetic dataset of regions with varying degrees of differential methylation.

2.2.1 Data generation

The goal of generating a synthetic dataset of regulatory regions was to be able to control the underlying methylation rate while mimicking the data generation and noise process of a scNMT-seq experiment as closely as possible. As described in Section 1.3.1, the DNA methylation and chromatin accessibility readouts in scNMT-seq experiments are based on bisulfite conversion of unmethylated CpG/GpC sites that are then sequenced with high throughput sequencing. To generate realistic scNMT-seq data, I, therefore, had to simulate sequencing reads. The coverage of the genome by reads is limited by the low amounts of input material in a single nucleus, rather than the sequencing depth. Based on previous scNMT-seq experiments (Angermueller *et al.*, 2016; Clark *et al.*, 2018), I assumed a fixed coverage *cov* of 9% of the genome. Next, I assumed that the number of reads that would cover a given genomic region in an individual cell would follow a binomial distribution

$$n_{reads} = Bi(floor(l_{region}/l_{read}), cov) \tag{2.19}$$

where the number of draws equals the ratio of the length of the genomic region l_{region} and the length of the reads l_{read}. The length of the reads can vary depending on the sequencing technique. Here, I used a read length of 75bp. Reads are then randomly placed in the genomic region. Each read will provide a binary readout of all CpG/GpC sites that are covered by it. To generate a simplified distribution of these sites, I assumed that they are uniformly distributed across the region with a density equal to the average genome-wide density. For CpG sites, this is roughly 0.5% outside of CpG islands and 5% within CpG islands. For GpC sites, I assumed 5% genome-wide density. Furthermore, the data depends on the number of cells assayed and their distribution across developmental trajectories. For this simulation, I assumed that 300 cells were sequenced with uniform distribution along a single temporal axis.

Figure 2.9 shows the resulting positioning of CpG/GpC sites for the experimental parameters outlined above.

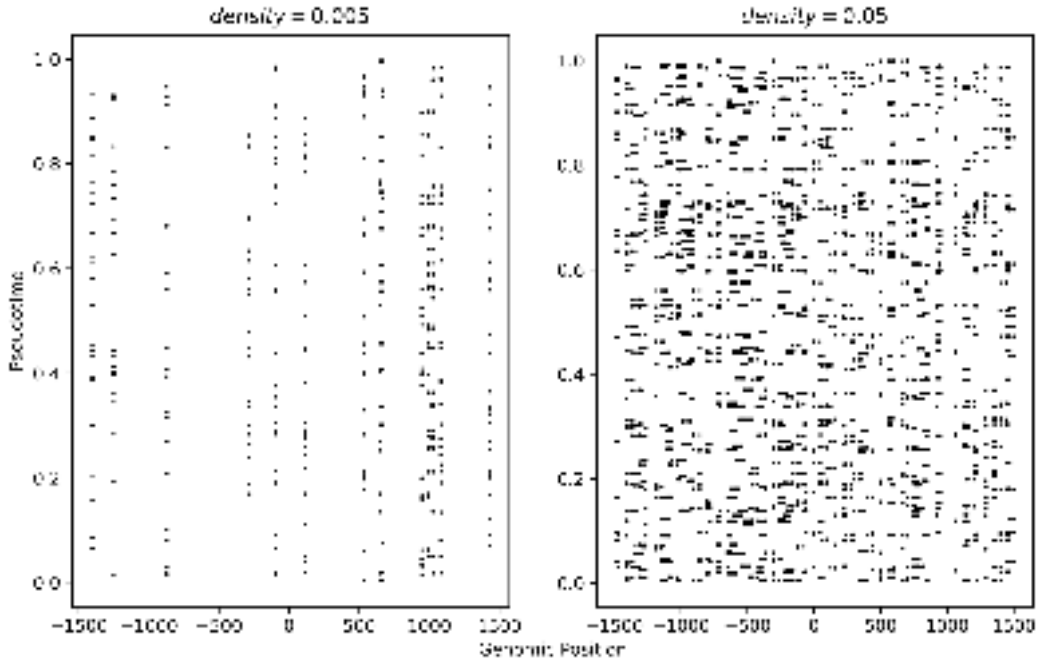


Figure 2.9 | Example simulation of CpG/GpC locations. The scatterplots show the positions of assayed CpG/GpC sites in a simulated scNMT experiment with 300 cells uniformly distributed across pseudotime. I assumed a read coverage of 0.1 and a site density of 0.005 *left* or 0.05. *Figure generated by Max Frank.*

Next, I created a generative model for the methylation rate of a simulated region. For simulating regions that correspond to the null hypothesis of no methylation change over time, the generative function is simply sampled from a GP with a genomic squared exponential kernel.

$$\rho_{null} \sim \Phi(GP(0, k_{genome}) \tag{2.20}$$

Figure 2.10 shows 50 draws of this model. The sampled methylation rate is then used to generate realistic experimental data by performing Bernoulli draws at simulated positions as described above (Fig 2.10, *right panel*).

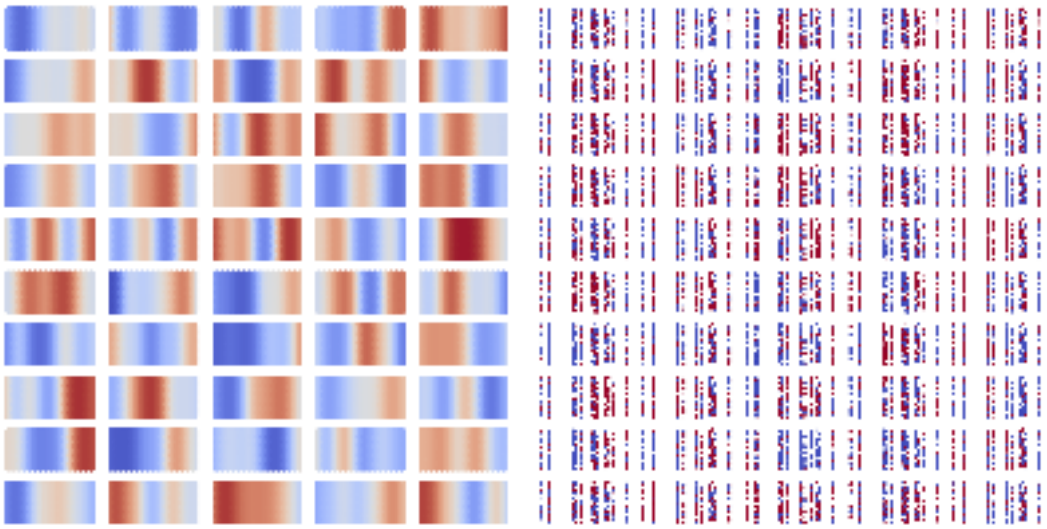


Figure 2.10 | Simulated scNMT data with no methylation change over time. The *left* panel shows 50 samples of the methylation rate ρ from a GP with a genome kernel only. The x-axis of each plot corresponds to the genome position and the y-axis corresponds to pseudotime. The *right* panel shows the Bernoulli draws at simulated positions of CpG sites. *Figure generated by Max Frank.*

For regions that correspond to the alternative hypothesis of differential methylation, I assumed that there is a single differential methylation event in the center of the regulatory region. As shown in the previous Section GPmeth is also capable of identifying multiple differential methylation events or events that are not in the center of the region. The choice for the simulation was made to simplify the analysis. For this, I specified a GP with a differential methylation kernel that is an additive kernel between k_{genome} as described above and a change window kernel that models methylation across time on a subset of the region:

$$
\begin{aligned}
\rho_{alt} &\sim \Phi(GP(0, k_{alt}) \\
k_{alt} &= k_{genome} + k_{CW} \\
k_{CW} &= k_1(x, x') * (1 - \sigma(x)) * (1 - \sigma(x')) + k_2(x, x') * \sigma(x) * \sigma(x') \\
\sigma_{x_0, x_1}(x) &= \frac{1}{e^{-s(x - x_0)}} * \frac{1}{e^{-s(x - x_1)}}
\end{aligned}
$$

The change window kernel allows the transition from a kernel k_1 outside of a window $[x_0, x_1]$ to a kernel k_2 within the window. The parameter s specifies the steepness of the transition. In this case, I chose to use a constant kernel k_1 with low variance

outside of the window and a squared exponential kernel k_2 inside of the window. The lengthscale parameter of k_2 was set to half the length of the total pseudotime, and the variance was adjusted so that a desired level of maximum methylation rate change (see Section 2.1.3 was produced. For details of the simulation process, see Section 6.1.3.

Figure 2.11 shows examples of draws from the model with varying window sizes and differential methylation rates.

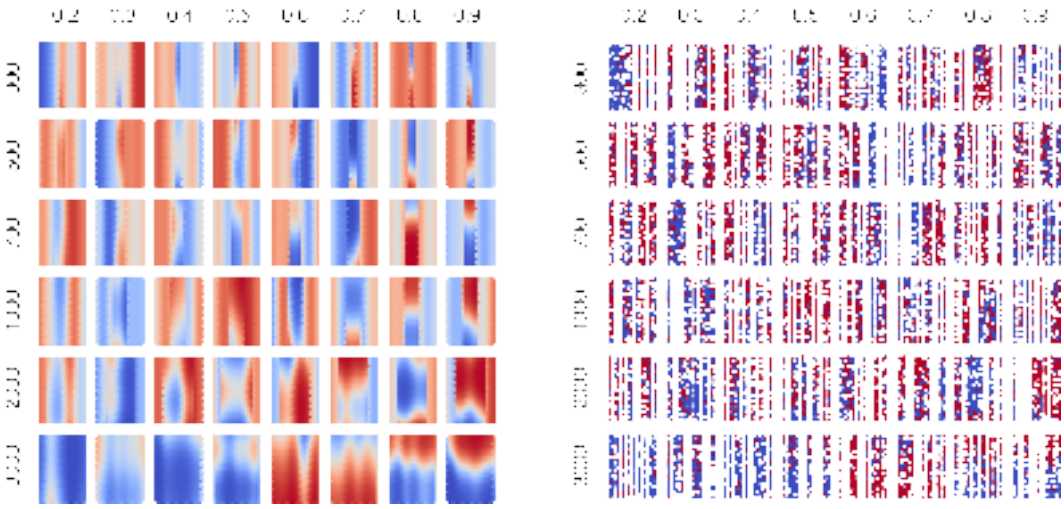


Figure 2.11 | Simulated scNMT data with methylation change over time. The *left* panel shows 50 samples of the methylation rate ρ from a GP that produces methylation change over time. The x-axis of each plot corresponds to the genome position and the y-axis corresponds to pseudotime. The *right* panel shows the Bernoulli draws at simulated positions of CpG sites. *Figure generated by Max Frank.*

2.2.2 Model Evaluation

The generative model described above was used to create a set of synthetic regions from the null hypothesis and the alternative hypothesis with different differential window sizes and methylation rate changes.

First, I tested the sensitivity of the differential methylation test with the *RBFRBF* model (Tab 2.1) at different settings for the simulation. To this end, I simulated 1000 regions with the generative model corresponding to the null hypothesis. Then I simulated 100 regions for each combination of the following scenarios:

Parameter	Values
CpG/GpC coverage	0.05, 0.005
Differential methylation window size	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000
Maximum methylation rate change	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

Table 2.2 | Simulation parameter settings

The simulated regions had a total width of 3000 bp, which means that the window that is affected by differential methylation ranges from 3% of the window to 100% of the region. Figure 2.12 shows the receiver operating characteristic curves for the different parameter settings. As expected, performance consistently improves with increasing width of the differential methylation window, as well as with increasing MMRC parameter. This result also illustrates the drastic effect of the sparsity of the CpG/GpC sites on the performance of the model. With a CpG/GpC frequency of 0.5% (Fig 2.12, *top row*), the model is not meaningfully different from a random classifier at MMRCs below 0.5. Above 0.5, the test becomes more powerful, but only if stretches of the genome larger than 500bp are affected. In contrast, with a CpG/GpC density of 5% (Fig 2.12, *bottom row*), the test has enough power to detect methylation rate changes as low as 0.2 if stretches in the genome larger than 500bp are affected or even smaller stretches if the MMRC is 0.5 or higher.

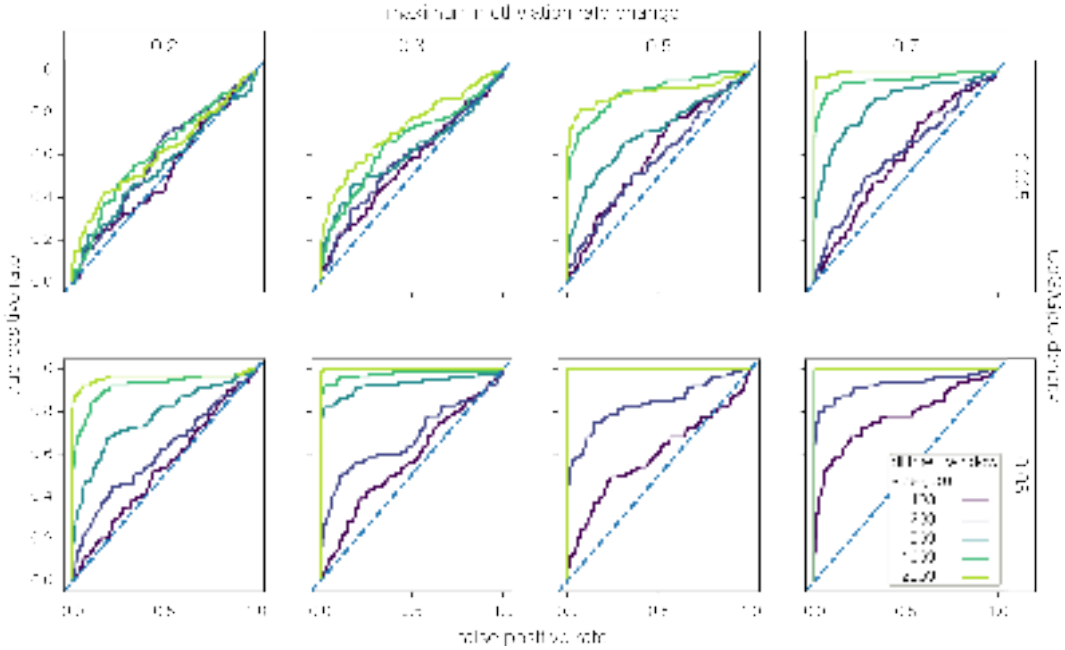


Figure 2.12 | Performance of the *RBFRBF* differential methylation test on simulated regions. Receiver operating characteristic (ROC) curve plots show the power of the *RBFRBF* model to identify differentially methylated/accessible regions in different settings. The *top row* shows ROC curves with a CpG/GpC density of 0.005, typical for endogenous methylation outside of CpG islands. The *bottom row* shows ROC curves with a CpG/GpC density of 0.05, which is typical for GpC sites and CpGs in CpG islands. Maximum methylation rate change increases from left to right. Note that for the lower CpG/GpC density scenario, the model only becomes powerful with larger methylation rate changes, while with the higher density, the model is sensitive enough to identify changes in methylation rate as low as 0.2 given a large enough differential methylation window. *Figure generated by Max Frank.*

Next, I investigated how accurately the model estimates the MMRC. The MMRC value is an important metric to evaluate the magnitude of the methylation rate change. This can be thought of as analogous to a fold-change estimate in differetial gene expression testing. Figure 2.13 shows the MMRC estimation by the *RBFRBF*

model for different simulation criteria. In general, the estimate of the MMRC is conservative, which is to be expected and indicates that the model does not overfit the data. Furthermore, the estimation becomes more accurate with larger stretches of the genome being affected and higher MMRC.

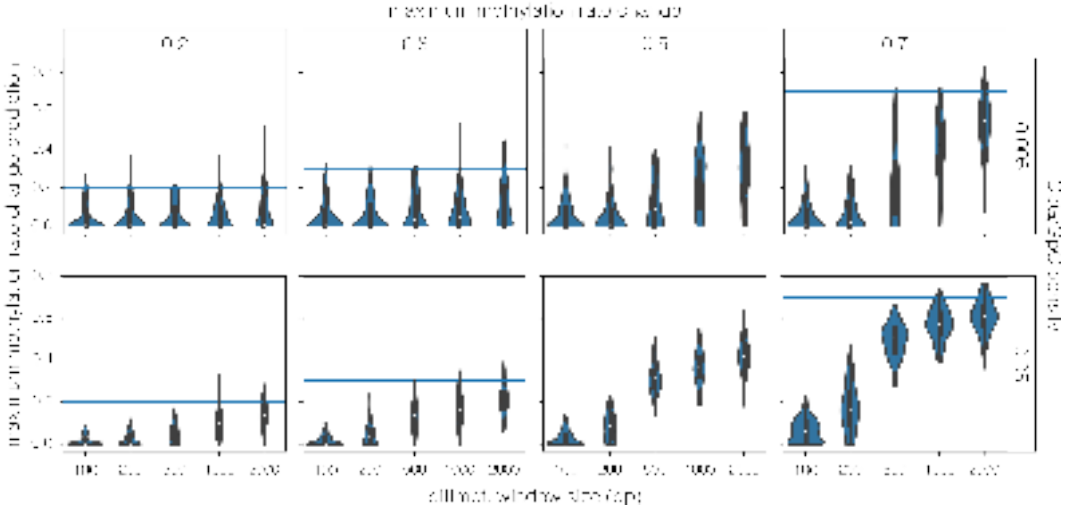


Figure 2.13 | Accuracy of the *RBFRBF* model to estimate the maximum methylation rate change (MMRC). Violin plots indicate the MMRC predictions of the *RBFRBF* model for different simulation scenarios. The blue horizontal line indicates the ground truth of the data-generating process. The *top row* shows MMRC predictions with a CpG/GpC density of 0.005, which is typical for endogenous methylation outside of CpG islands. The *bottom row* shows MMRC predictions with a CpG/GpC density of 0.05, which is typical for GpC sites and CpGs in CpG islands. The MMRC ground truth increases from left to right. The model provides a systematically conservative estimate of MMRC and gets more accurate the larger the differential methylation window size, as expected. For the higher CpG/GpC density scenario, it seems that a differential methylation window of 500 base pairs (bp) is sufficient for a good MMRC estimation in most cases. In the lower density scenario, the estimation only becomes more accurate at 1000 bp. *Figure generated by Max Frank.*

This simulation illustrates how difficult it is to detect differential methylation events in single-cell data. It is likely that many real-world regulatory events that are important for cellular decision-making just barely fulfill the criteria that this test requires to capture them. This highlights the importance of using the most powerful test possible if working with single-cell methylation or accessibility data. Therefore I also compared my model with alternative models and previously used methods.

First, compared the *RBFRBF* model to its counterpart without a genome kernel (i.e., where I assume a constant methylation rate over the genome dimension) to see if the addition of genome covariance results in increased power of the test. Then, I also wanted to test what the theoretical power of an optimal performance test for these simulated regions would be. To this end, I modeled the same regions with the generative model that originally produced the data. I fixed the hyperparameters of the generative models at the settings that produced the simulated region and only trained the variational parameters of the model. The idea here was to see how far off the *RBFRBF* model is from what is theoretically achievable with a perfect model and

in what cases a lack of power simply is due to data noise. To compare these models, I only performed simulations for the case that CpG/GpC density is 0.005, which is the more challenging scenario. Figure 2.14 shows the ROC AUC score for the *RBFRBF*, *ConstantRBF*, and generative models. When MMRC is 0.2, none of the models has large power to detect differential methylation. With MMRC of 0.3 and above, the *RBFRBF* model and generative model outperform the *ConstantRBF* model, especially for smaller diffmet window sizes. This is reassuring since it indicates that modeling genome covariance leads to more accurate models when the boundaries of the region cannot be chosen accurately. Furthermore, the *RBFRBF* model has comparable AUC scores to the generative models for all settings, meaning the performance is close to the theoretically optimal performance.

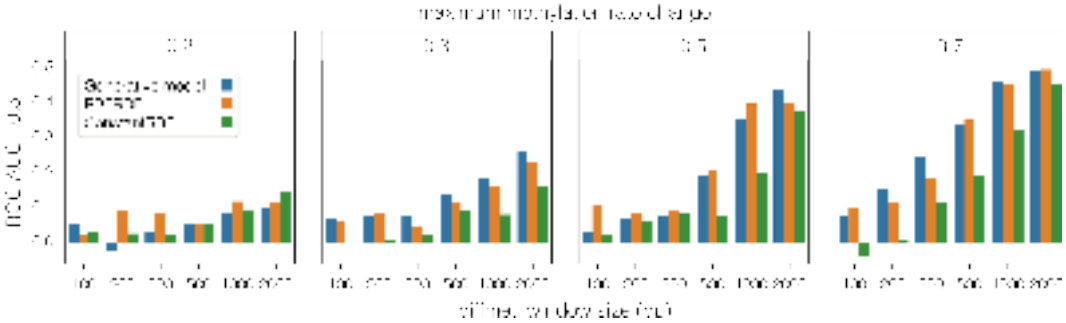


Figure 2.14 | Performance comparison of *RBFRBF*, *ConstantRBF*, and the Generative Model. Bar plots indicate the performance of different models to identify differentially methylated regions in terms of the area under the receiver operating characteristic curve (ROC AUC), where 0.5 corresponds to a model that is no better than a random classifier, and 1 corresponds to a perfect classifier. The generative model (*blue*) is the model used to simulate the data and to give a theoretical upper bound on the performance since it should be the most powerful model. The *RBFRBF* model (*orange*)is consistently performing close to the generative model. As expected, the *ConstantRBF* (*green*) model is only powerful for large differential methylation windows since it lacks the genomic kernel. *Figure generated by Max Frank.*

Next, I assessed the accuracy of the MMRC estimates of the three models. Figure 2.15 shows the MMRC estimates of the models at different settings. All three models consistently underestimate the MMRC compared to the ground truth indicated by the dashed line. The *RBFRBF* model has an intermediate performance between the *ConstantRBF* model and the generative model.

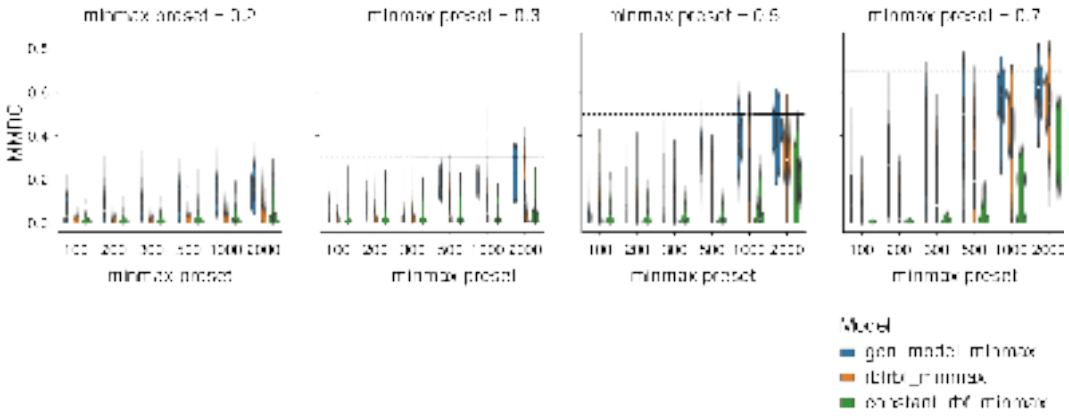


Figure 2.15 | Maximum methylation rate change (MMRC) estimation accuracy comparison of the *RBFRBF*, *ConstantRBF*, and the Generative Model. Violin plots show the estimation of MMRC by different models for various simulation scenarios. The blue horizontal line indicates the ground truth MMRC value that was used for the simulation. The generative model (*blue*), which is the model used to simulate the data, should be the optimal model. While it gives the least conservative estimate of the true MMRC, it still consistently underestimates the true rate change. This illustrates the difficulty of estimating true underlying rate parameters from sparse Bernuoulli-distributed data. The *RBFRBF* model (*orange*) yields a more conservative estimate of the true MMRC. The *ConstantRBF* (*green*) model underestimates the true MMRC drastically, especially in the case of small differential methylation windows. This is expected since it will average out the methylation rate for the whole region that is modeled. *Figure generated by Max Frank.*

2.2.3 Benchmarking GPmeth against other methods

I also used the simulated regions to compare the GPmeth method against other methods that could be applied to single-cell methylation/accessibility measurements. To my knowledge, GPmeth is the only test that models both the temporal changes and the genomic changes of methylation rate as a continuous variable (see Section 1.5.2.2). Here, I compare the GPmeth model against Fisher's exact test, which has been previously used to test for differential methylation between cell types in scNMT experiments (Argelaguet *et al.*, 2019b) and scMET (Kapourani *et al.*, 2021), which was developed specifically to model single-cell methylation data. Since both of these tests require cell types as input, I produced two artificial groups by defining an *early* and a *late* cell type by thresholding the pseudotime coordinate of each cell at half the total pseudotime. Since the generative model produces continuous change over time, the GPmeth model is expected to be more powerful. Furthermore, the two tests do not explicitly model genomic covariance but aggregate the methylation signal for each cell. For this comparison, I chose a challenging scenario for both CpG/GpC densities. With a CpG/GpC density of 0.05, I set the MMRC to 0.3 and the diffmet window width to 600bp. For the CpG/GpC density of 0.005, I set the MMRC to 0.5 and chose the same diffmet window size of 600bp. Figure 2.16 shows the ROC curves for GPmeth models with different kernels compared to scMET and Fisher's exact test. Both CpG/GpC density scenarios show the most power for GPmeth models with a kernel that models genome covariance and pseudotime covariance with a nonlinear

function. There is little difference between the performance of the *RBFRBF* model and the *RBFMatérn* models, which is not surprising since they have similar properties, and the data was generated from a model with an RBF kernel.

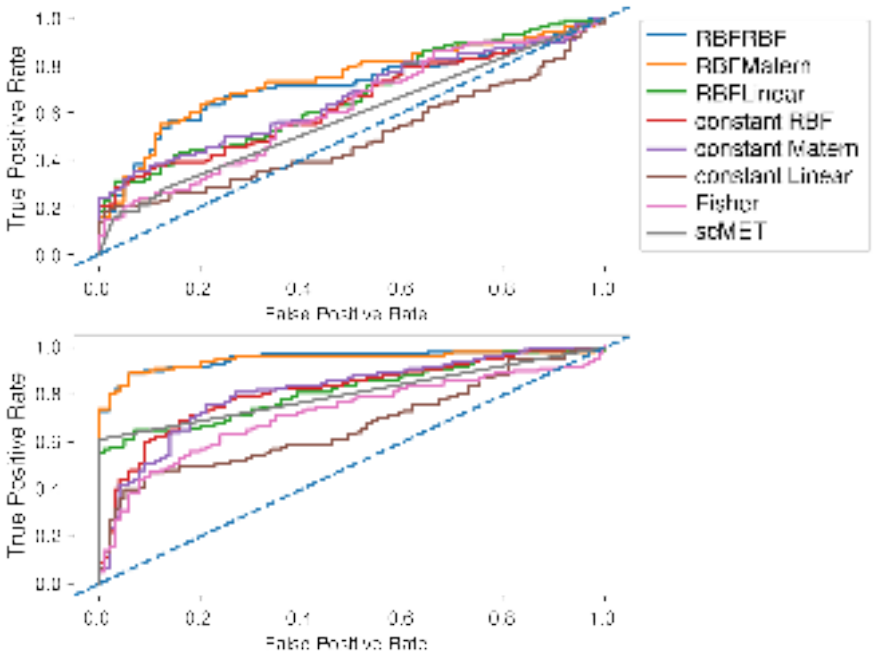


Figure 2.16 | Power comparison of GPmeth, scMET, and Fisher's exact test. Receiver operating characteristic (ROC) curve plots show the statistical power of different models to identify differentially methylated/accessible regions. The *top row* shows ROC curves with a CpG/GpC density of 0.005, an MMRC of 0.5, and a diffmet window size of 600 base pairs. The *bottom row* shows ROC curves with a CpG/GpC density of 0.05, an MMRC of 0.3, and a diffmet window size of 600 base pairs. The GP models that have both a genome and a pseudotime kernel (*RBFRBF* (*blue*), *RBFMatérn* (*orange*)) perform the most powerful in both scenarios. The GP models without a genome kernel perform more similar to the Fisher's exact test or the scMET model. *Figure generated by Max Frank.*

2.2.4 Model Calibration

As discussed above, to obtain a test statistic for differential methylation/accessibility, I compare the marginal likelihood of a full model that allows for changes in methylation rates over time against the likelihood of a null model that is constrained to a constant temporal methylation rate. The obtained test statistic is the log of the ratio of likelihoods of the full and the null model: the log-likelihood ratio (LLR). The purpose of model calibration is to determine the distribution of likelihood ratios when the null hypothesis is true. This allows to specify significance cutoffs that will have a known proportion of false positive results, or false discovery rate (FDR). In other words, this allows the computation of p-values from LLR estimates.

According to Wilks theorem (Wilks, 1938), the LLR of two nested models will approach a chi-squared (χ^2) distribution under the null as the number of observations approaches infinity. The degrees of freedom of the χ^2 distribution are determined by the difference in the number of free parameters of the two models. This was used for

model calibration in spatialDE (Svensson *et al.*, 2018) and GPcounts (BinTayyash *et al.*, 2021).

However, this is only true in the limit of infinite data points (Greven *et al.*, 2008) and for strictly nested models (Dominicus *et al.*, 2006; Self and Liang, 1987). Both of these criteria are not fulfilled here. The models are not strictly nested since the null model fixes the pseudotime kernel variance parameter at zero, which is at the edge of the full model parameter space. Furthermore, the models are optimized on sparse data, which does not guarantee that parameters are estimated perfectly. Thus, in practice, the calibration of the model with the assumption of a χ^2 null-distribution yields conservative p-values.

Therefore, I approximated an empirical null distribution by randomly permuting real input data. One way of obtaining this would be to use parametric bootstrapping, i.e., obtaining an empirical null distribution for every region tested by permuting it randomly enough times to obtain significant results. This would increase the runtime of the model by at least 1000x to get enough samples for significant p-values when considering multiple testing corrections.

With the assumption that the null distributions are similar for all regions, likelihood ratios from simulated null regions can be pooled (Listgarten *et al.*, 2013). To generate data from the null hypothesis (i.e. no differential methylation/accessibility over time), I permuted the calculated pseudotime values of cells and trained the null and the full model in the same way as for differential testing (see Section 6.1.1). I found that the pooled log-likelihood ratios can be described well with a mixture of χ^2 distributions:

$$p(LLR \mid \pi, a, d) = \pi\chi_0^2(-LLR) + (1-\pi)a\chi_d^2(-LLR) \tag{2.21}$$

Where LLR is the log-likelihood ratio of the models, π is the fraction of a χ^2 distribution with zero degrees of freedom to a χ^2 distribution with d degrees of freedom and a scale parameter a.

I used scNMT-seq data of mouse-embryonic stem cells undergoing gastrulation to calculate empirical null distributions (see Chapter 3). Specifically, I used cells during Mesoderm development. To obtain a large enough sample to fit the χ^2-mixture distribution, I produced five random pseudotime permutations of cells for each available enhancer and promoter region (for details on region definitions, see Section 3.2.1). I then trained all models in Table 2.1 on the permuted regions and calculated the LLR of each full model and the respective null model. A closer inspection of the LLR estimates revealed a clear dependency on the number of observations (Fig 2.17). This effect was stronger for permuted DNA methylation profiles than for DNA accessibility. For a discussion of possible reasons for this, see Chapter 4.

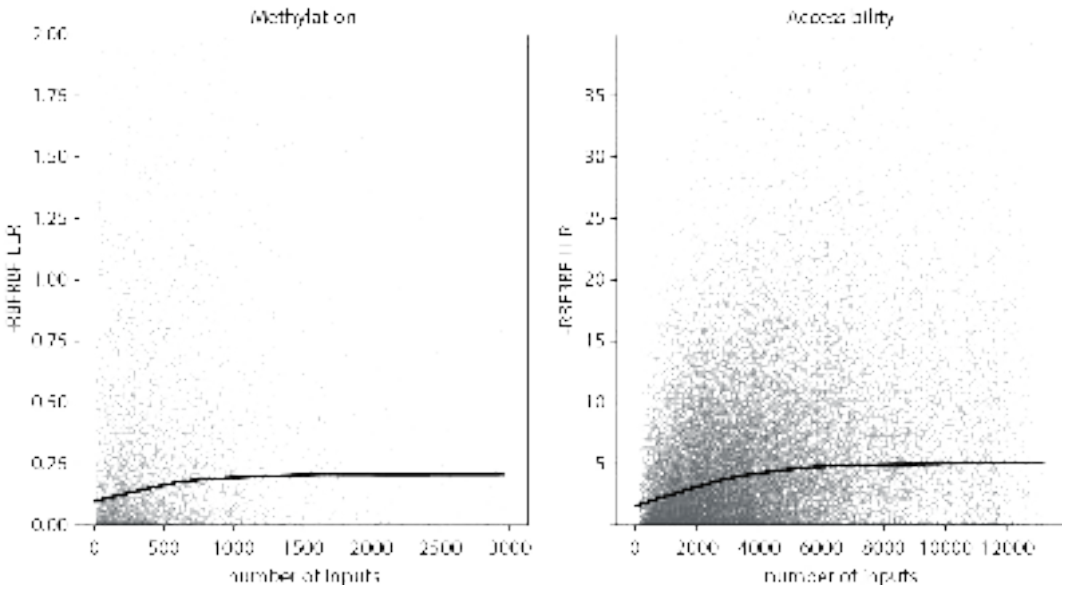


Figure 2.17 | Dependency of permuted LLR distribution on the number of observations. Every point represents the median negative log-likelihood ratio of the *RBFRBF* model for 5 permutations of an enhancer region for methylation (*left panel*) and accessibility (*right panel*). The x-axis is the number of observed methylation events in that enhancer region. The black lines represent a locally weighted scatterplot smoothing (LOWESS) fit of the data points. *Figure generated by Max Frank.*

Because of this dependency, I decided to separate promoter and enhancer regions into bins according to the number of input points and fit a χ^2-mixture distribution to each of those bins separately. Furthermore, there was a clear difference in the LLR distribution between methylation and accessibility profiles, which was partially but not completely explained by the higher density of GpC sites compared to CpG sites and the resulting higher number of inputs. Therefore, I also fit separate null distributions for methylation and accessibility profiles. To fit this null distribution, the free parameters of the χ^2-mixture, π, a, d, were estimated maximum likelihood estimation with a grid-search over the parameters. To increase the robustness of the fit, the lowest 5% and the highest 5% quantile of the LLRs were excluded. This was done to remove outliers in the distribution where random shuffling leads to genuinely differentially methylated regions by chance. Figure 2.18 shows an example of the fit of the null distribution for the *RBFRBF* model on enhancer accessibility.

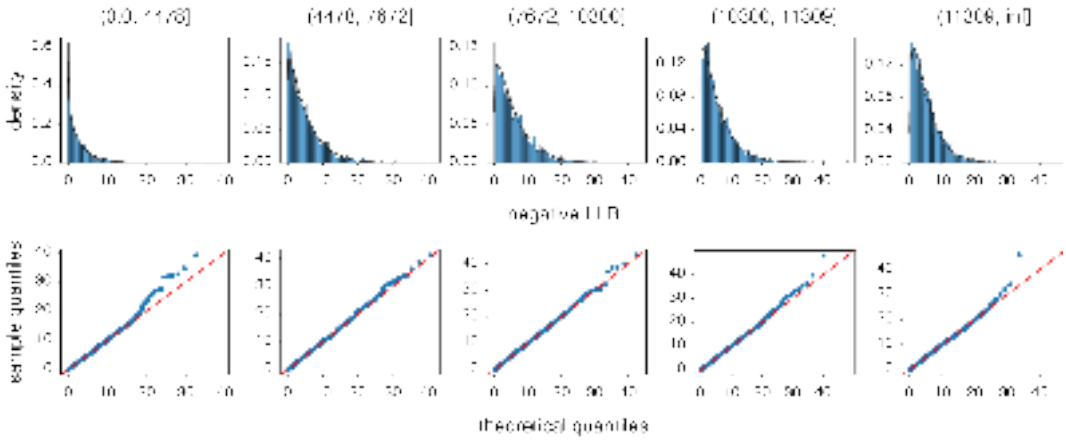


Figure 2.18 | Example fit of null distribution for the *RBFRBF* model. The *top row* shows the distribution of negative log-likelihood ratios (LLR) of the *RBFRBF* model on enhancer accessibility as a histogram plot. The blue line corresponds to the fitted parametric χ^2-distribution. The *bottom row* shows quantile-quantile plots of the expected LLR values of the parametric distribution versus the observed LLR values from shuffled data. Every *column* corresponds to a bin of input points (see column title). Note that the empirical and theoretical distribution match closely for every bin, indicating successful calibration. *Figure generated by Max Frank.*

With this approach I was able to successfully fit null distributions for every set of regions.The parameters obtained by this fit were then used to estimate p-values of non-shuffled regions in the actual gastrulation dataset which will be discussed in the next Chapter.

3 | Application of GPmeth to scNMT data of Mouse Gastrulation

This Chapter will discuss the application of the GPmeth model, described in Chapter 2, to real scNMT-seq data. The data used were published in Argelaguet *et al.*, 2019b and consists of over 1000 sequenced mouse embryonic stem cells at the gastrulation stage of development. During this process, multipotent stem cells differentiate into the three main germ layers of the embryo.

Section 3.1 will give an overview of previous work on this dataset.

In Section 3.2, I describe a processing strategy to infer pseudotime coordinates and lineage association for each cell based on RNA expression. This trajectory inference forms the basis for the subsequent analysis of the epigenome.

In Section 3.3, I will then describe the application of the GPmeth model to find differentially methylated and differentially accessible regions during Mesoderm formation (see Section 3.3.2). I benchmark the GPmeth approach in comparison with existing tools in Section 3.3.3, describing the incremental improvements of different components of the model. I then use the ability of the model to refine differentially methylated regions to perform improved TF-binding motif analysis and identify the activation timings of Mesoderm-specific TFs in Section 3.3.4. In Section 3.3.5, I investigate the temporal changes of lineage-defining enhancers, showing that for Mesoderm development pluripotency and Ectoderm-specific enhancers get inactivated before Mesoderm-specific enhancers are activated. Finally, I compare DNA methylation, chromatin accessibility and gene expression time-courses to find links between those modalities in Section 3.3.6.

3.1 Previous work

Argelaguet et al. provided the first comprehensive dataset that profiled three omics modalities in the same cell during the pluripotency exit of mouse embryonic stem cells.

scNMT-seq profiles RNA expression, DNA methylation, and chromatin accessibility in the same cell. They were able to profile 1105 cells isolated from mouse embryos at embryonic days (E) 4.5, 5.5, 6.5, and 7.5.

As expected, they find that global methylation levels of regulatory elements such as promoters and enhancers increased from an average of 25% at E4.5 to 80% at E5.5, while accessibility of these regions only dropped minimally during the same time period.

They then used multi-omics factor analysis (MOFA, Argelaguet *et al.*, 2018a) to perform dimensionality reduction with all available modalities to find shared modality-specific factors that drive the gastrulation process. They found that methylation and accessibility of enhancer elements have a stronger influence on germ layer formation than methylation and accessibility of promoter elements. They furthermore defined lineage-specific regulatory genomic regions by performing chromatin immunoprecipitation with DNA sequencing (ChIP–seq) on differentiated tissues. They defined peaks for distal H3K27ac (enhancers) and H3K4me3 (transcription start sites) that are accessible only in Ectoderm, Endoderm, and Mesoderm, respectively. One notable finding was that Ectoderm-specific enhancer elements become accessible and demethylated as early as E4.5 while Endoderm and Mesoderm enhancers only become demethylated and accessible after E5.5. Generally, they found that differentiated Ectoderm cells retain most of the regulatory signatures from pluripotent stem cells.

To track the temporal trajectories of chromatin accessibility and DNA methylation, they produced a pseudotime ordering for Endoderm and Mesoderm cells and plotted the average trajectories of both modalities for lineage-specific enhancers and promoters. They found that there is a genome-wide inverse correlation between methylation and accessibility indicating that these two modalities are tightly linked.

They also performed tests to identify differentially methylated and differentially accessible regions. For this, they aggregated the CpG/GpC methylation signal in individual cells for genomic regions of interest and performed a Fishers-exact test between groups of cells of different embryonic days and lineages. This approach implicitly assumes that methylation/accessibility is constant within the aggregated regions. Furthermore, this test requires the embryonic day covariate to faithfully capture differences in methylation over time and the annotations of regulatory regions to be precise in order to retain statistical power. These assumptions make it difficult to find smaller changes in methylation/accessibility that might happen gradually over the process of gastrulation.

The goal of applying GPmeth to this dataset was to perform a more powerful test that could identify subtle changes in DNA methylation or chromatin accessibility over time. Furthermore, the model output can be used to refine the annotation of important regulatory elements by identifying the specific boundaries of the differential methylation signal. Finally, the model can be used to compare the temporal dynamic

across modalities, opening up possibilities to speculate on regulatory mechanisms that take place.

3.2 Lineage reconstruction and pseudotime inference

The GPmeth model requires a pseudotime coordinate associated with each cell as an input. Therefore the first step was to do dimensionality reduction and pseudotime inference based on the RNA expression modality of the data. Cells were already labeled by cell type by the authors of the original publication. While the original data included cells from E4.5 to E7.5 sequenced at daily intervals, I excluded cells from E4.5 in this analysis, since these cells were too distinct from the rest of the cells, which meant that no continuous pseudotime could be established.

Because the data consisted of multiple sequencing runs performed on different embryos, there was potential for substantial batch effects that could influence the analysis. I assessed the extent of batch effects by performing dimensionality reduction by PCA and UMAP and plotting cells colored by cell type and embryo of origin. If there are no substantial batch effects, one would expect cells from a single embryo to be homogeneously distributed across all cell types of the stage at which the embryo was sequenced. Figure 3.1, *a, top row* shows that the distribution is non-homogeneous, indicating the need for batch correction. I used the bbknn (Polański *et al.*, 2020) method of the scanpy (Wolf *et al.*, 2018) library to perform batch correction. Figure 3.1 *a, bottom row* shows the UMAP after batch correction. For a detailed description of the RNA seq preprocessing see Section 6.2.2. The batch-corrected neighborhood graph was used for dimensionality reduction based on diffusion components (Fig 3.1, *b, d*). The first five diffusion components show the differentiation from pluripotent Epiblast cells to the three germ layers. Interestingly there was also a clear separation of cells within the Endoderm lineage, consisting of Gut and Notochord cells (Fig 3.1, *upper right panel*).

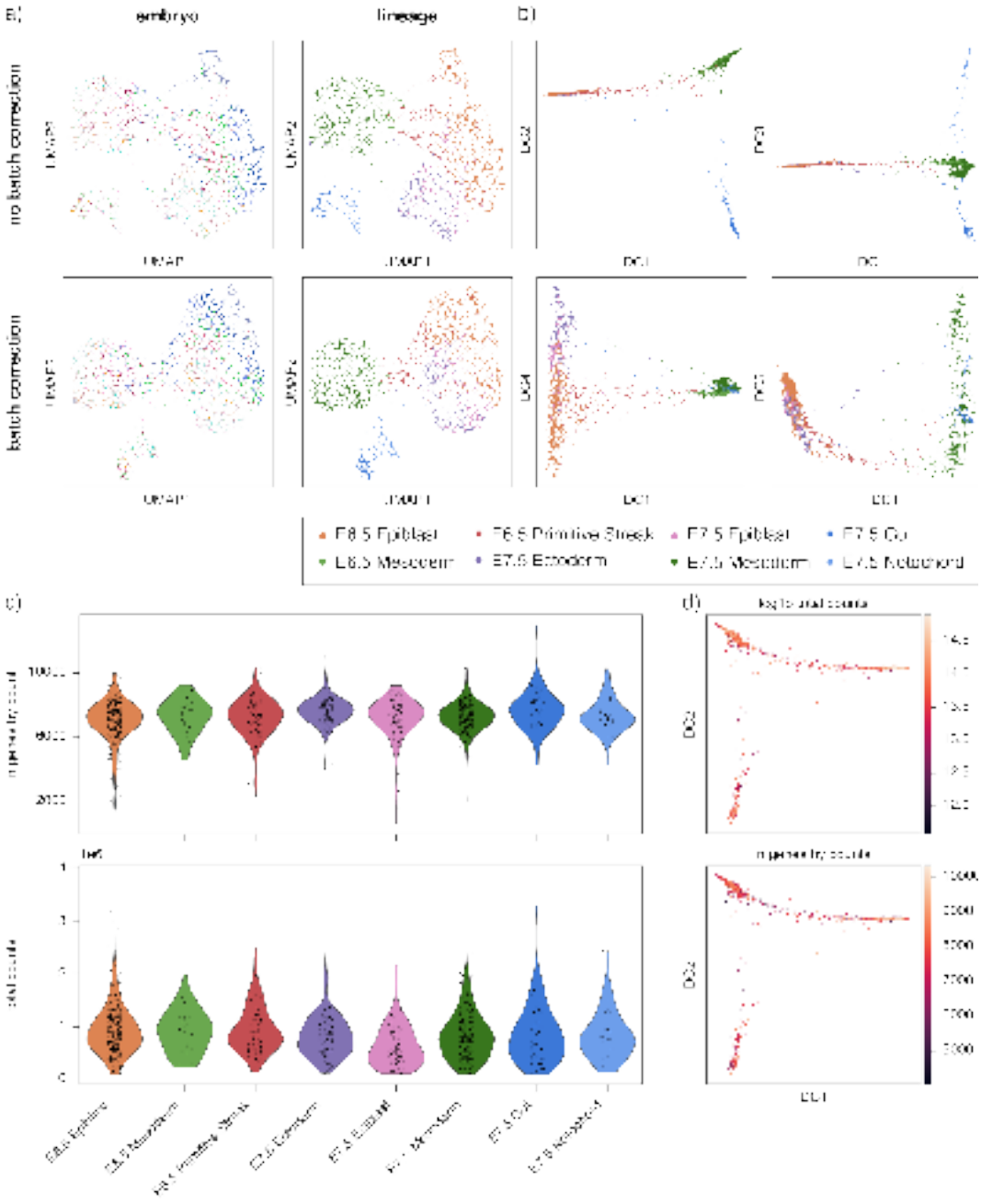


Figure 3.1 | Processing of single-cell RNA seq data and pseudotime estimation. *(a)* Effects of batch correction with bbknn on distribution of embryos. Shown are UMAPs based on the first 15 principal components that were mapped to a larger reference atlas (Pijuan-Sala *et al.*, 2019). The *left* column of UMAPs shows cells colored by the embryo of origin (colors not annotated in the legend). The *right* column shows UMAPS annotated by stage and lineage (based on annotations from the reference atals). Note that cells cluster by embryo in the *top right* UMAP embedding. After batch correcting neighbor graph calculation with bbknn embryos are more uniformly distributed within their lineages (*bottom right*). *(b)* Diffusion maps of the first five diffusion components based on the batch corrected neighborhood graph. Diffusion component one separates Mesendoderm cells from Epiblast and Ectoderm cells. Diffusion component 2 separates Endoderm cells. Diffusion component 3 distinguishes between Gut and Notochord cells. Diffusion component 4 separates Ectoderm from Epiblast cells. Diffusion component 5 differentiates between early and late Mesoderm. *(c)* Diffusion maps with the first two diffusion components. Cells are colored by technical covariates. There is no clear impact of the technical covariates on the positions of cells in the diffusion map. *Figure generated by Max Frank.*

Based on the first five diffusion components, I established a hierarchical branching structure as shown in Figure 3.2, *a*, consisting of four lineages: Ectoderm, Mesoderm, Gut, and Notochord. Panels *b, and c* show a UMAP based on the diffusion components with cell-type annotations at different levels of granularity. I then used destiny (Haghverdi *et al.*, 2016), to assign a pseudotime value to each cell (Fig 3.2, *d*). A sanity check for the lineage assignment was performed by checking the expression profiles of known marker genes for each of the inferred lineages (Fig 3.2, *e*).

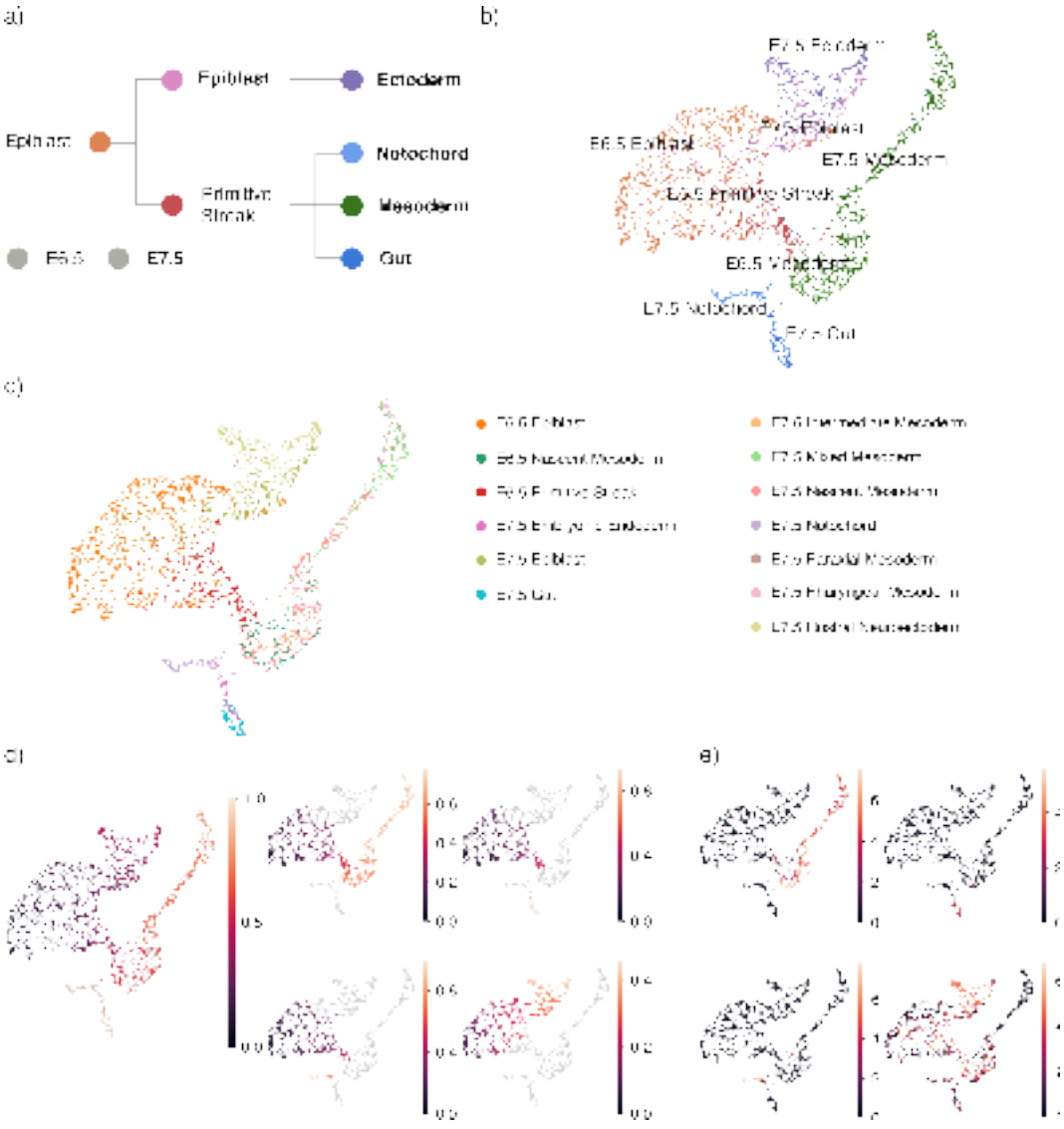


Figure 3.2 | *(a)* Schematic representation of the hierarchical model of cell states during mouse gastrulation. *(b)* Umap based on the first five diffusion components colored by embryonic day and lineage. *(c)* Umap colored by fine-grained lineage annotations. Fine-grained cell-type annotations were obtained by transferring cell-type labels from a significantly larger single-cell atlas with over 100,000 cells (Pijuan-Sala *et al.*, 2019). Briefly, matching by mutual nearest neighbors (Haghverdi *et al.*, 2018) was used on jointly normalized expression matrices to transfer labels from the atlas to scNMT cells. *(d)* Umaps colored by inferred pseudotime. On the left cells that do not have DNA accessibility or DNA methylation measurements are colored in grey. On the right only cells belonging to the respective lineage indicated are colored by pseudotime. *(e)* Umap colored by log1p transformed expression of an example marker gene for the four lineages. *Figure generated by Max Frank.*

The pseudotime coordinates were then used as input to the GPmeth model to test for differential methylation/accessibility in all four lineages.

3.2.1 NOMe-seq data preprocessing

Contrary to most single-cell RNA sequencing experiments, the raw data of the methylation and accessibility modalities, produced by the single-cell NoMe-seq assay of scNMT-seq cannot fit into the working memory of most modern computers if the goal is to do base-level modeling. This is because there are millions of CpG/CpG sites in the mammalian genome while there are only around 20,000 genes. Thus I store methylation data on individual CpG/GpC sites on disk in an indexed format that allows random access to a genomic region of interest (Li, 2011). This allows me to efficiently load all CpG/GpC sites within a genomic window into working memory and train the model, before loading the next window. For details of the implementation, see Section 6.2.6. For each cell and CpG/GpC site, the stored information consists of the number of methylated reads and the number of unmethylated reads. For reasons discussed in Chapter 2, I transform this information into a binary signal.

I then apply filtering steps to exclude ambiguous signals both on the site level and at the level of cells. For endogenous CpG methylation only sites in the genomic context of WCG (W = A or T) were retained. For GpC methylation only GCH (H = A, C, T) sites were retained. This step avoids interfering signals of the two methyltransferases. Next, I excluded cells that had low coverage (fewer than 50.000 observed CpG sites or fewer than 500.000 GpC sites). I also excluded cells with very low or very high methylation rates which indicates issues in the library preparation or bisulfite conversion steps (Fig 3.3).

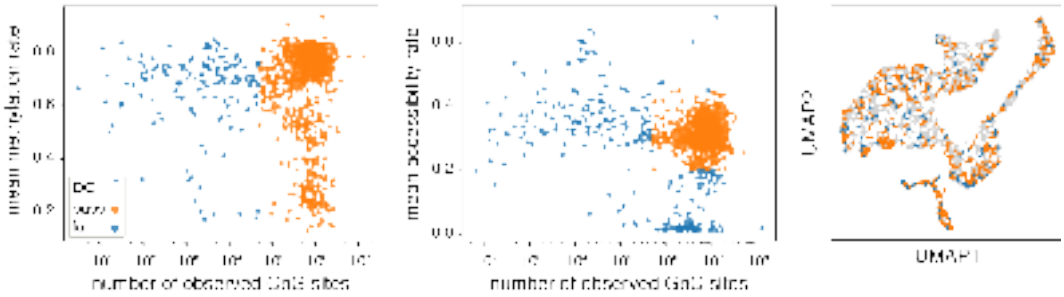


Figure 3.3 | Quality control metrics for scBS-seq. The *left* and *middle* panel show the number of observed methylation events versus the global mean methylation rate for CpG and GpC methylation respectively. Cells that were excluded in the quality control are shown in blue and cells that passed are shown in orange. The *right* UMAP depicts cells that pass the quality control criteria of all three modalities. Note that cells failing quality control are evenly distributed across the UMAP suggesting that there is no direct biological influence on sequencing quality. *Figure generated by Max Frank.*

3.2.2 Definition of regulatory regions

To generate the input for GPmeth, I used methylation/accessibility signals in two sets of regulatory regions, putative promoters and putative enhancer regions. Promoters were simply defined as a genomic window of ± 2kb around the transcription start site (TSS) of all protein-coding genes. As in Argelaguet *et al.*, 2019b, enhancers were defined as genomic windows marked by the histone mark H3K27ac detected by chromatin immunoprecipitation with DNA sequencing (ChIP–seq) peaks. ChIP seq was performed on isolated germ layers at E7.5 (Xiang *et al.*, 2020), resulting in a separate set of peaks for Ectoderm, Endoderm, and Mesoderm cells. From this lineage-specific enhancers could be defined as peaks exclusively present in one of the germ layers. Meanwhile, a comprehensive set of enhancer peaks was obtained by taking the union of the peak annotations. The ChIP-seq peaks were then extended by 500bp in either direction, resulting in an average window size of 2kb. In total, this resulted in 18,347 promoter regions and 17,386 enhancer regions.

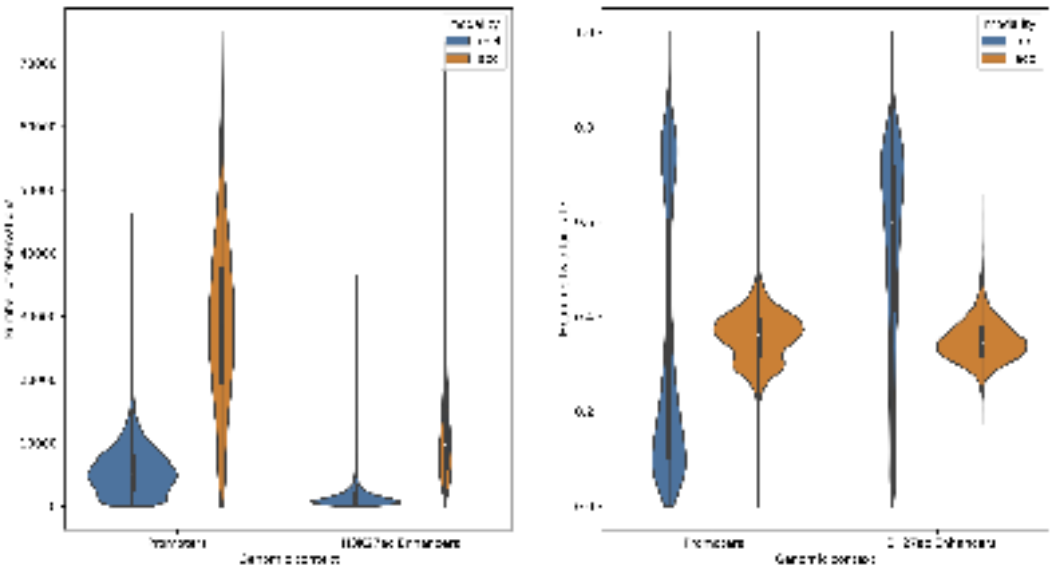


Figure 3.4 | Summary statistics of DNA methylation and accessibility in Enhancers and Promoters. The *left* panel shows the number of methylation events captured by scBS-seq for each of the assayed promoters and enhancers. As expected more methylation events are observed for DNA accessibility due to the higher frequency of GpC sites in the genome compared to CpG sites. Promoter regions also have more observations on average in line with their larger window size of 4kb compared to the average 2kb size of enhancers. The *right* panel shows the average methylation rate of the regulatory regions across all cells. Note that the average CpG methylation rate, corresponding to chromatin accessibility, is often high or low with few regions showing intermediate methylation rates. In contrast, the average GpC methylation rate is centered around 0.4. This is likely due to the fact that each region contains multiple nucleosome-occupied and nucleosome-free regions, which leads to an averaging of the signal. *Figure generated by Max Frank.*

Figure 3.4 shows the summary statistics of the number of observed CpG and GpC sites as well as the global observed methylation rates. As expected, the DNA methylation signal is sparser than the accessibility signal. Furthermore, the average DNA

methylation rate is more divergent than DNA accessibility. This is especially true for promoter regions that are either mostly methylated or mostly unmethylated, while the average accessibility rate is centered around 0.4.

GPmeth was then used to model DNA methylation and chromatin accessibility of regulatory regions for the four lineages that were identified with pseudotime analysis. I will first describe the results of this in the case of the Mesoderm lineage.

3.3 Epigenomic regulation during Mesoderm development

The Mesoderm is the middle of the three germ layers formed during gastrulation. Its cells give rise to several tissues, including many organs, parts of the circulatory system, and muscles. In this Section, I will discuss the findings of applying GPmeth to this lineage. Note that I am discussing the Mesodermal lineage first, as an illustrative example, but the GPmeth methodology will remain the same for the other three lineages that will be described in Section 3.4.

In total, 415 cells that were sequenced with scNMT-seq mapped to the Mesoderm lineage (Fig 3.2, *d,e*). Of those, 189 cells were classified as epiblast, 48 as primitive streak, and 178 as Mesoderm cells by mapping RNA expression profiles to a large single-cell atlas (Pijuan-Sala *et al.*, 2019). These cells stem from embryos at E6.5 and E7.5, during which pluripotent epiblast cells undergo gastrulation. Note that at this stage, the global wave of demethylation and remethylation is completed, and changes in regulatory activity are now cell-type specific.

I applied GPmeth to all 18,347 promoter regions and 17,386 enhancer regions defined above to test for differential DNA methylation and DNA accessibility during Mesoderm formation.

3.3.1 Model output

For each regulatory region, GPmeth fits a model of methylation rate based on the binarized CpG/GpC methylation observations. For details of the fitting process, see Chapter 2. While GPmeth has multiple options for kernel parametrization, in this case, I used the *RBFRBF* kernel that allows for nonlinear change of methylation rate across the genome and pseudotime. This model can be used to test if the methylation rate within that region changes over pseudotime. It can also locate the exact subregion where that change in methylation rate occurs.

Figure 3.5 shows an example output of GPmeth for the promoter and proximal enhancer element of a well-known Mesoderm-specific transcription factor *Mesp2*. During Mesoderm formation, the promoter of *Mesp2* is demethylated in a narrow region around the transcription start site (TSS) that GPmeth identifies to be around 400bp wide. Concurrently, the accessibility of that same region increases during Mesoderm formation before decreasing again in late-stage Mesoderm cells. 12kb

upstream of the *Mesp2* TSS, there is a putative enhancer region marked by a H3K27ac ChIP-seq peak. Interestingly, this enhancer region does not change in the center of the genomic window. Instead, the DNA methylation rate decreases, and DNA accessibility increases at the flanks of this region. GPmeth identifies this as two separate subregions.

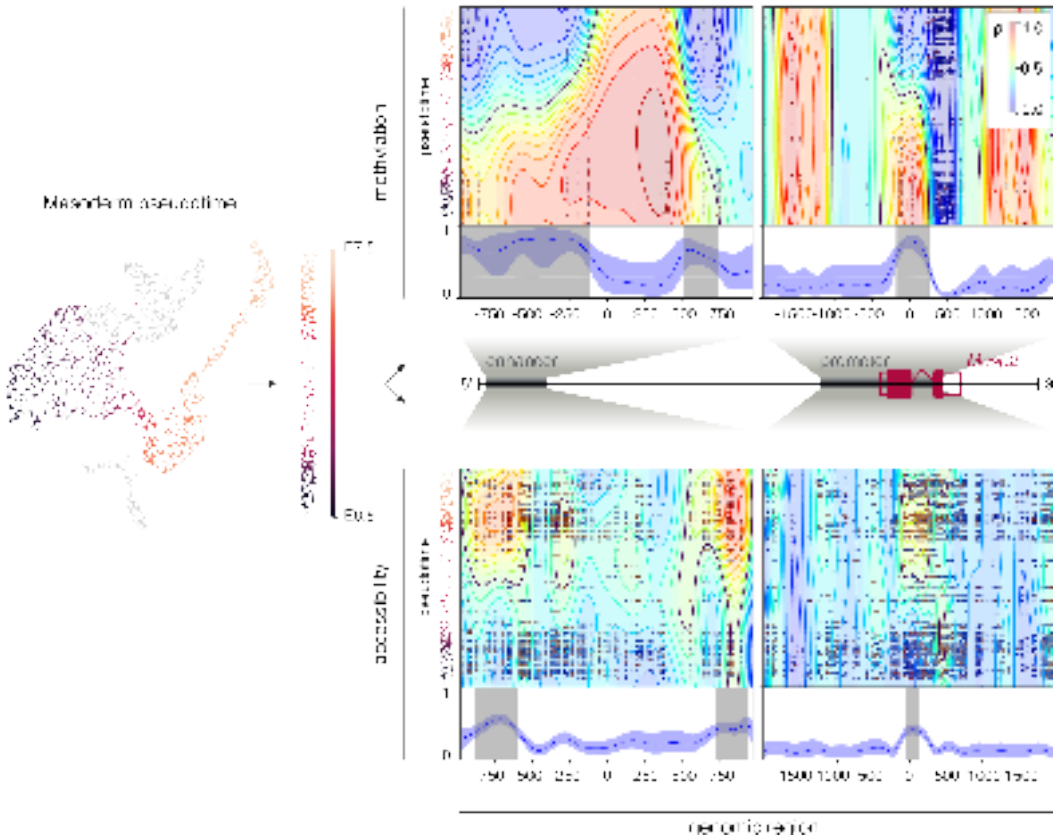


Figure 3.5 | Example output of the GPmeth model for Mesp2. The UMAP plot on the *left* depicts the pseudotime estimates for each cell of the Mesoderm lineage. These pseudotime estimates correspond to the position of each cell on the y-axis of the model output plots on the *right*. The *right* panel consists of GPmeth predictions for DNA methylation (*top row*) and accessibility (*bottom row*) for the gene promoter (*right column*) and an enhancer element (*left column*). The x-axis of the plots depicts the genomic position, with 0 corresponding to the center of the region. The scatterplot depicts the input data to the model measured by scBS-seq, where blue indicates unmethylated sites and red indicates methylated sites. The contours correspond to the posterior mean prediction of the methylation rate by the GPmeth model. Underneath the scatterplot, the blue line indicates the methylation rate change over the pseudotime of every genomic location predicted by the model. The blue-shaded regions indicate the 95% confidence interval around that prediction. Grey-shaded areas span genomic regions, where the 95% confidence interval of methylation rate change is above a threshold of 0.3. *Figure generated by Max Frank.*

This example shows how GPmeth can aid in generating hypotheses about gene regulation. The identified regions of differential methylation/accessibility present possible targets for follow-up experiments that could establish causal relationships between the different modalities. In the next Sections, I will discuss the results of running GPmeth on the complete set of promoters and enhancers.

3.3.2 Detecting DNA methylation/accessibility changes during Mesoderm development

To detect differential methylation in promoters and enhancers during Mesoderm development, I compared the likelihoods of a full model with an *RBFRBF* kernel against a null model with a *ConstantRBF* kernel (see Section 2.1.2). The *RBFRBF* kernel allows for methylation rate changes across the genomic region and pseudotime, while the *ConstantRBF* model only allows for methylation rate changes across the genomic region. The logarithm of the likelihood ratio (LLR) of these models is then used as a test statistic for the hypothesis that the methylation rate is changing over pseudotime. This statistic is then compared to a set of permuted regions, as described in Section 2.2.4 to ensure accurate p-values for each tested region. P-values were then corrected for multiple hypothesis testing according to the procedure of Benjamini-Hochberg (Benjamini and Hochberg, 1995). While the p-value determines the confidence of the model that there is a differential methylation/accessibility event within the region, it does not directly tell the user about the effect size of that differential event, i.e., the methylation rate change. In Section 2.1.3, I discussed how GPmeth estimates the maximum methylation rate change (MMRC) for every position in the tested region. The highest MMRC for every region is a good estimate of the magnitude of differential methylation and can be used as an additional cutoff to filter the results. This can be thought of analogously to defining a fold-change cutoff when testing for differential gene expression. The MMRC estimate can be seen in Figure 3.5 as the blue line underneath the model output plots. Because the output of GPmeth includes uncertainty estimates, I also obtained conservative estimates of the MMRC by taking the lowest estimation of the 95% confidence interval (CI) for this measure.

I define differentially methylated regions as regions that satisfy the following two criteria:

1. The LLR test results in a BH-adjusted p-value smaller than 0.1
2. The MMRC is larger than 0.3. In other words, the model is confident in detecting a methylation rate change of more than 30%. This can be thought of, as the equivalent of a fold-change cutoff in differential gene expression testing (Love *et al.*, 2014).

3.3.2.1 Differentially methylated regions

Running GPmeth on the methylation profiles of 18,347 promoter regions resulted in 507 significantly differentially methylated regions, of which 190 regions had a maximum methylation rate change (MMRC) of 0.3 or greater (Fig 3.6, *left*). In contrast, out of 17,385 enhancer methylation profiles, GPmeth found 2,958 regions to be significantly differentially methylated, of which 2478 had an MMRC value of more than 0.3 (Fig 3.6, *right*). This indicates that promoter methylation may not be the main driver of gene regulation during gastrulation. Enhancer methylation seems

to play a larger role in comparison. This observation was also made by Argelaguet *et al.*, 2019b, with a global analysis using MOFA (Argelaguet *et al.*, 2019a).

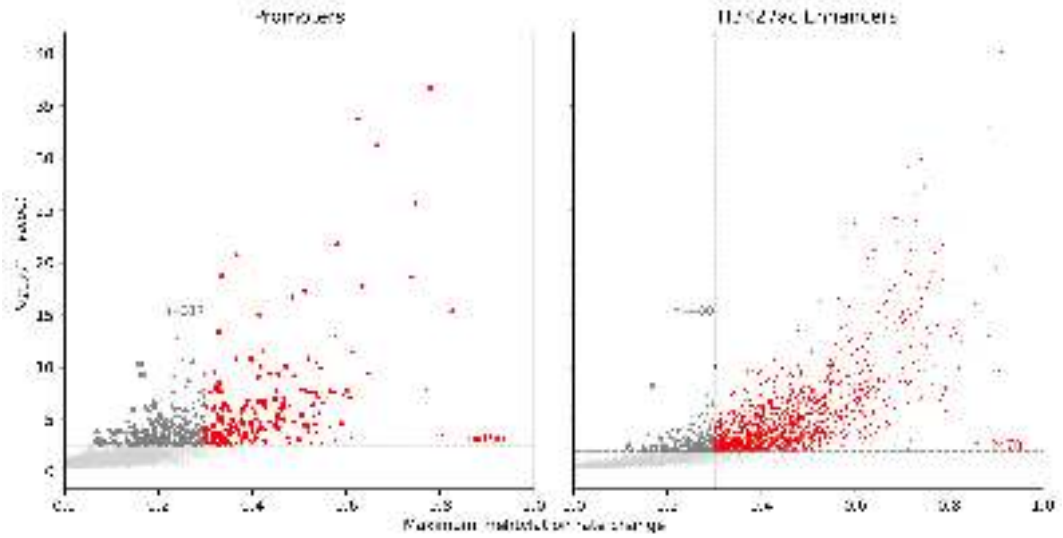


Figure 3.6 | Differentially methylated promoters and enhancers during Mesoderm development. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) during Mesoderm development. The maximum methylation rate change depicted corresponds to the maximal difference between the maximum and minimum model predictions at every point in the genomic window. The *left* panel are promoter regions, the *right* panel are enhancer regions. The horizontal dashed line corresponds to a significance cutoff of FDR<0.1 after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. *Figure generated by Max Frank.*

3.3.2.2 Differentially accessible regions

GPmeth can be applied to chromatin accessibility data from scBS-seq (GpC methylation) without modifications from the way it is applied to endogenous methylation data. However, since there are roughly ten times more GpC sites in the genome, the density of observations is higher. The presence of nucleosomes and their interplay with regulatory elements results in methylation rate trajectories that are fundamentally different from endogenous methylation. For these reasons, I used a separate calibration for the differential accessibility tests that is based on permutations of GpC methylation data (see Section 2.2.4).

With this calibration, running GPmeth on the chromatin accessibility profiles of 18,347 promoter regions resulted in 875 differentially accessible regions during Mesoderm development, of which 380 regions had an MMRC of 0.3 or greater (Fig 3.7, *left*). In contrast, out of 17,385 enhancer accessibility profiles, GPmeth found 2700 regions to be significantly differentially accessible, of which 2327 had an MMRC value of more than 0.3 (Fig 3.7, *right*). As with endogenous methylation, promoters seem to be less regulated by chromatin accessibility compared to enhancer elements during the process of Mesoderm formation.

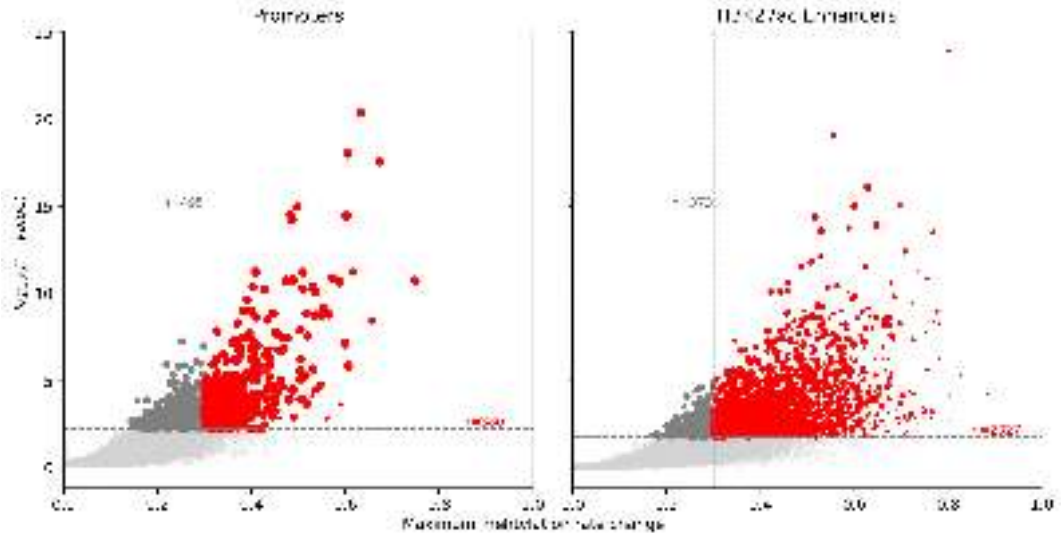


Figure 3.7 | Differentially accessible promoters and enhancers during Mesoderm development. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) during Mesoderm development. Mesoderm development. The maximum methylation rate change depicted corresponds to the maximal difference between the maximum and minimum model predictions at every point in the genomic window. The *left* panel are promoter regions, the *right* panel are enhancer regions. The horizontal dashed line corresponds to a significance cutoff of 0.05 FDR after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. *Figure generated by Max Frank.*

3.3.3 Model benchmark and comparison to other methods

In Chapter 2, I showed with synthetic data the power benefits of the GPmeth model, which should result in a larger fraction of statistically significant regions on real data. In this Section, I will show that GPmeth identifies more differentially methylated and accessible regions during Mesoderm development compared to other methods. I will highlight the two main advantages over other models, which are:

- Using a flexible genome kernel to model methylation rate with base-pair resolution
- Using a flexible nonlinear temporal kernel to model methylation rate over time

3.3.3.1 Benefits of adding a genome kernel

In this Section, I will demonstrate the benefits of the inclusion of a genome kernel into the GPmeth model. As discussed above, the inclusion of a genome kernel allows GPmeth to refine broader genomic input windows into subregions where differential methylation/accessibility occurs. Furthermore, a genome kernel should provide increased statistical power for detecting differential methylation when the subregion is small compared to the size of the full input window. I have shown this on simulated data in Section 2.2.2. To demonstrate the benefits on real data, I am comparing the full GPmeth model with a nonlinear kernel in genome and pseudotime

space (the *RBFRBF* model) against a model that still has a nonlinear kernel to model temporal changes but is constant in genome space (*ConstantRBF*). This comparison allows for the isolation of the benefits of including the genome kernel alone. Figure 3.8 shows example model fits of these four models for the promoter accessibility trajectory of Mesp2. Since the differentially accessible subregion in the center of the promoter is relatively small ($\tilde{1}$50bp), only the full GPmeth (*left panel*) model predicts an accessibility rate change larger than 0.3 for this region, since for the other models the signal is diluted by non-changing flanking GpC sites. Therefore, the models without genome kernels systematically underestimate the true rate of change in accessibility.

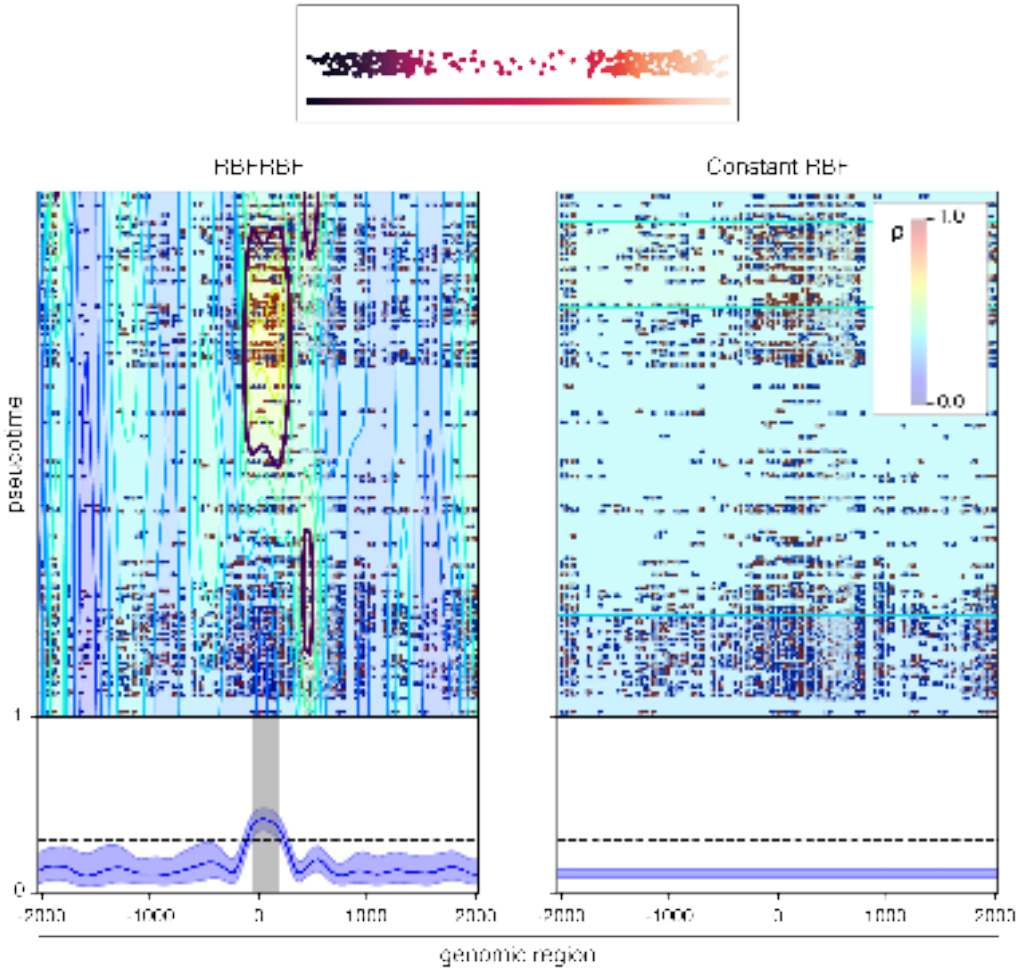


Figure 3.8 | Example model fit of Mesp2 promoter accessibility with and without genome kernel. Contour plots of GPmeth predictions for chromatin accessibility rate ρ in the promoter of the Mesp2 gene. The *left* panel depicts the output of GPmeth parametrized with an *RBFRBF* kernel, and the *right* panel with a *ConstantRBF* kernel that does not allow for variability across the genomic axis. The x-axis of the GPmeth plots depicts the genomic position, with 0 corresponding to the transcription start site. The scatterplot depicts the input data to the model measured by scNMT-seq, where blue indicates unmethylated sites and red indicates methylated sites. The contours correspond to the posterior mean prediction of the methylation rate ρ by the GPmeth model. Underneath the scatterplot, the blue line indicates the maximum methylation rate change over pseudotime (MMRC) of every genomic location predicted by the model. The blue-shaded regions indicate the 95% confidence interval around that prediction. Grey-shaded areas span genomic regions where the predicted MMRC is greater than 0.3. The panel on top represents the pseudotime trajectory of Mesoderm formation. *Figure generated by Max Frank.*

Next, these models were compared with their respective null models (see Chapter 2) to calculate the LLR and obtain significance estimates for differential accessibility. Figure 3.9 shows the p-value versus MMRC scatterplots of the four models for promoter and enhancer accessibility during Mesoderm formation.

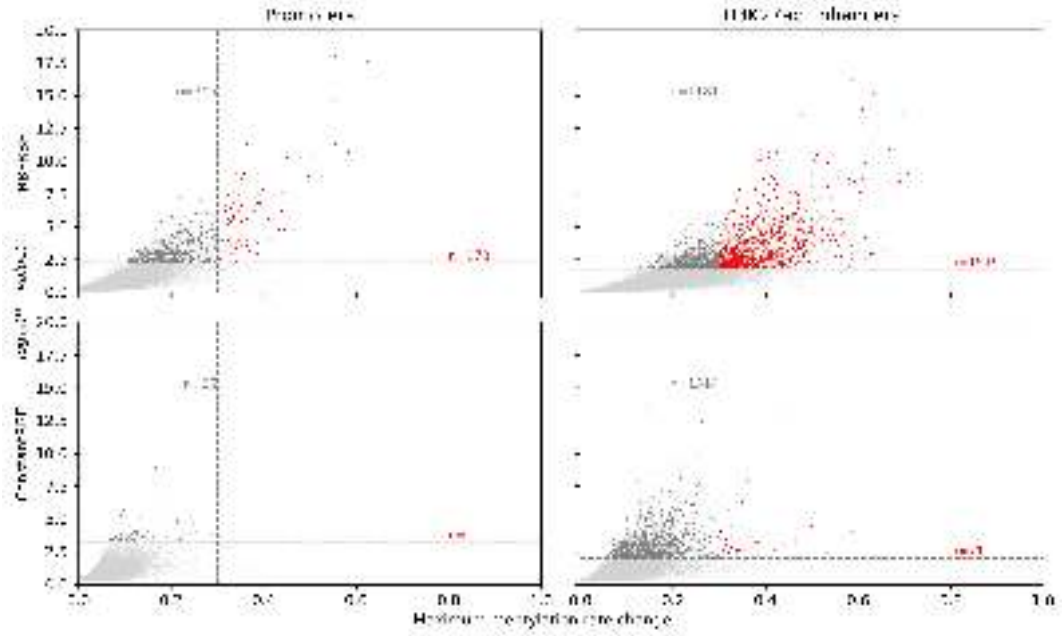


Figure 3.9 | Differential accessibility testing with and without genome kernels. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) of promoter (*left column*) and enhancer (*right column*) accessibility during Mesoderm development. Rows correspond to different models. The horizontal dashed line corresponds to a significance cutoff of 0.1 FDR after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. Dark grey dots are above the FDR threshold and red dots mark differentially accessible regions with MMRC larger than 0.3. *Figure generated by Max Frank.*

The *RBFRBF* kernel finds the highest number of significantly differential enhancers and promoters compared to all other models. When adding a minimum MMRC of 0.3, the other models find almost no differential regions. The exact numbers of regions found are listed in Table 3.1.

	Promoters		H3K27ac Enhancers	
	FDR	FDR+MMRC	FDR	FDR+MMRC
RBFRBF	875	380	2700	2327
ConstantRBF	86	1	1585	71

Table 3.1 | Number of promoter and enhancer regions found by models with and without a genome kernel. Numbers in the Significant columns have a BH-adjusted p-value smaller than 0.1. Number in the Significant + MMRC column have a BH adjusted p-value smaller than 0.1 and a MMRC larger than 0.3

To compare the models directly, Figure 3.10 shows p-values and MMRC estimates of the different models for enhancer accessibility. Assuming successful calibration and equal performance of all models, one would expect points-region estimates to be scattered around the diagonal identity line. In all comparisons, the full GPmeth models identify more regions to be significantly differentially accessible and many

more regions to have MMRCs over 0.3. Notably, the regions that were only identified by the comparison models (orange points) are mostly close to significance for the full GPmeth model. Conversely, there are regions that are highly significant for the full GPmeth model but have very high p-values in the comparison models (blue points).

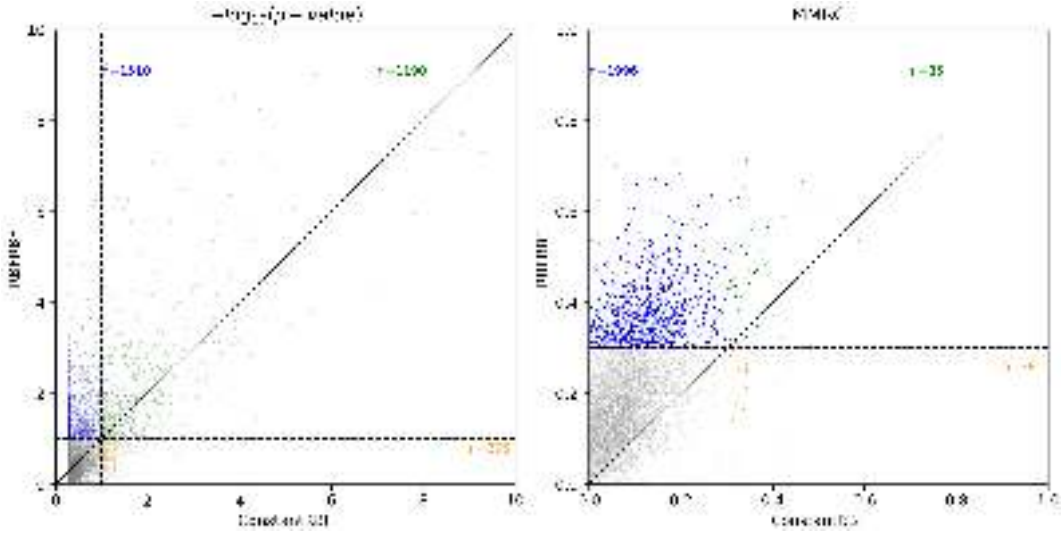


Figure 3.10 | Comparison of models with and without genome kernels for finding differentially accessible enhancers. The *top* row depicts the $-log_{10}$ p-values (BH adjusted) of the *RBFRBF* model (*y-axis*) vs. a comparison model on the *x-axis*. Horizontal and vertical dashed lines indicate p-value cutoffs at 0.1 FDR. The diagonal line is the identity line and means equal significance. Each dot represents the accessibility trajectory of an enhancer region during Mesoderm development. Blue dots are only found to be differential by the *RBFRBF* model, orange dots only by the comparison model, and green dots by both models. The *bottom row* depicts the same comparison in terms of MMRC. *Figure generated by Max Frank.*

If adding a genome kernel to the model increases statistical power by enabling the model to find smaller differential subregions, one would expect that the regions only found by the full GPmeth model to be smaller on average compared to regions that are found by the comparison models as well. To investigate this I visualized the difference in p-values of these models for different sizes of subregions found by the full GPmeth model (Fig 3.11). While subregions larger than 200bp tend to have similar p-values in the full GPmeth model and the models without a genome kernel, the smaller regions are often not detected by the comparison models, i.e. there is a large negative difference in p-values.

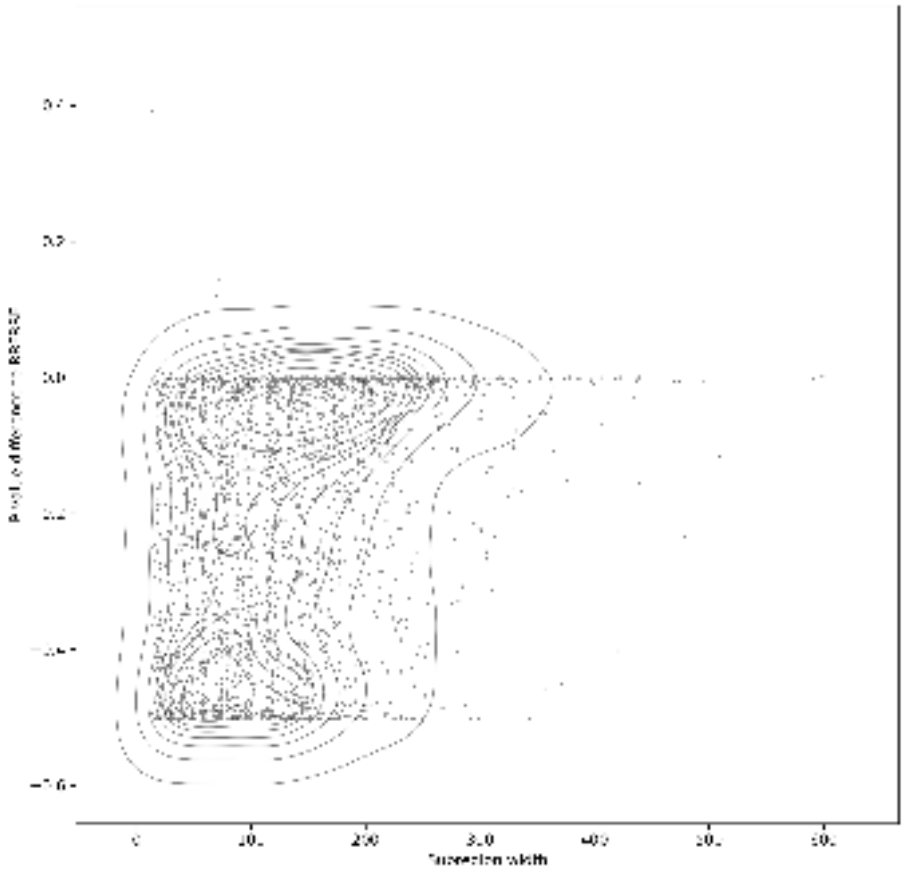


Figure 3.11 | Smaller regions of differential accessibility less likely to be detected without a genome kernel. Each plot depicts the difference in p-values (BH adjusted) between the GPmeth *RBFRBF* model and a comparison model (in the plot titles) without a genome kernel on the *y-axis* and the size of the refined subregion predicted by the *RBFRBF* model on the *x-axis*. *Figure generated by Max Frank.*

3.3.3.2 Benefits of a nonlinear temporal model

The second point that differentiates GPmeth from standard statistical models to test for differential methylation is the use of a nonlinear kernel to describe smooth methylation rate changes over time. To investigate whether a nonlinear pseudotime kernel is beneficial, I compared the *RBFRBF* (Fig 3.12, *left*) model against three models with a different temporal kernel. The first model, *RBFLinear* (Fig 3.12, *center left*), has a linear kernel to model temporal changes. The second and third models have a categorical kernel that models temporal changes and an RBF kernel in the genome dimension. The *RBFEarlyLate* model averages the methylation signal of early cells at the beginning of the pseudotime trajectory and late cells at the end (Fig 3.12, *center right*). The *RBFFirstLast* model does the same but takes only the first and last 10% of cells along pseudotime into account to model the beginning and

end stages of the trajectory (Fig 3.12, *right*). These two models were included as an analog to more traditional differential testing ideas that compare groups of cells. An example fit of these four models to the Mesp2 promoter can be seen in Figure 3.12. This time *RBFRBF* and *RBFLinear* predict accessibility rate changes above 0.3, and *RBFEarlyLate* just falls shy of the 0.3 threshold. One detail to note is that the *RBFRBF* model predicts accessibility to first increase but then decrease slightly again in mature Mesoderm cells. By contrast, the *RBFLiner* model predicts a strict increase which is due to the fact that the linear kernel does not allow for up- and down-shifts. Finally, the *RBFFirstLast* does not detect major changes in accessibility. This is due to the fact that it only includes the earliest cells with lower accessibility and the last cells, in which accessibility decreased again.

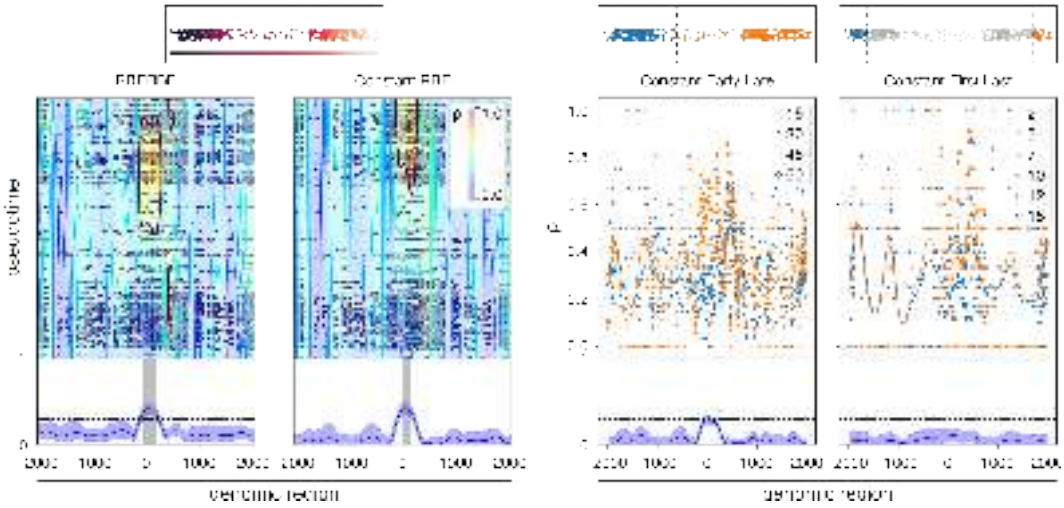


Figure 3.12 | Example model fit of Mesp2 promoter accessibility with different pseudotime kernels. Contour plots of GPmeth predictions for chromatin accessibility rate ρ in the promoter of the Mesp2 gene. The *left* two panels depict the output of GPmeth parametrized with an *RBFRBF* and an *RBFLinear* kernel, respectively. The scatterplot depicts the input data to the model measured by scNMT-seq, where blue indicates unmethylated sites and red indicates methylated sites. The x-axis depicts the genomic position, with 0 corresponding to the transcription start site. The contours correspond to the posterior mean prediction of the methylation rate ρ by the GPmeth model. Underneath the scatterplot, the blue line indicates the maximum methylation rate change over pseudotime (MMRC) of every genomic location predicted by the model. The blue-shaded regions indicate the 95% confidence interval around that prediction. Grey-shaded areas span genomic regions where the predicted MMRC is greater than 0.3. The panel on top represents the pseudotime trajectory of Mesoderm formation. The *RBFLinear* model restricts ρ to change linearly across pseudotime. The *right* two panels show the outputs of the *ConstantEarlyLate* and the *ConstantFirstLast* models, respectively. These models group cells according to a pseudotime cutoff. The pseudotime grouping of cells is depicted in the panel above the model output plots. In the model output plots, the x-axis represents the genomic dimension, and the y-axis represents the methylation rate ρ. Points are the means of measurements of individual GpC sites across cells in a pseudotime group, and their size indicates the number of cells that the mean is based on. *Figure generated by Max Frank.*

The nonlinear kernel allows the most flexibility in fitting methylation/accessibility rates, but this also comes with the necessity of strict calibration to avoid producing

false positives in differential testing due to overfitting. Therefore, simpler temporal kernels, such as the ones in the comparison models, are not necessarily less powerful if the actual methylation/accessibility rate changes fit their structure. In the case of the mouse gastrulation dataset, the temporal sampling of cells was likely not uniform, leading to an undersampling in the intermediary stages between Epiblast and differentiated germline cells (see Section 3.2). This means that most of the signal is concentrated in the early and late parts of the pseudotime trajectories, which makes it hard to find regions that have methylation/accessibility rate changes that increase and then decrease or vice versa. Therefore, a linear or categorical temporal kernel should, in fact, be a good option for this dataset. I therefore wanted to investigate if the nonlinear kernel is still competitive with the simpler alternatives. Figure 3.13 shows the results of differential testing of enhancer accessibility during Mesoderm development for the four models described above.

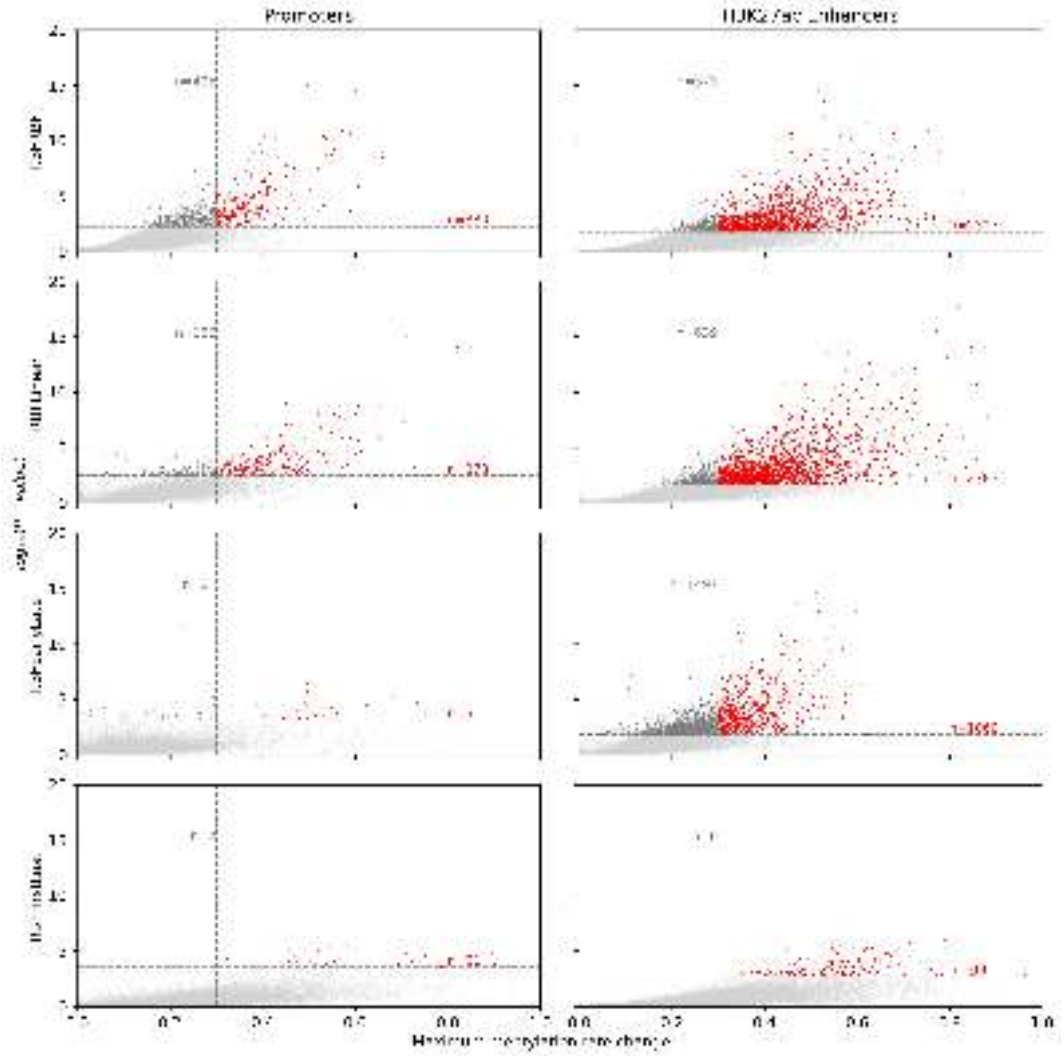


Figure 3.13 | Differential accessibility testing with different pseudotime kernels. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) of promoter (*left column*) and enhancer (*right column*) accessibility during Mesoderm development. Rows correspond to different models. The horizontal dashed line corresponds to a significance cutoff of FDR<0.1 after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. Red dots mark differentially accessible regions. *Figure generated by Max Frank.*

In this example, the model with the linear temporal kernel finds the highest number of differentially accessible enhancers (2783) followed by the *RBFRBF* model with 2327, the *RBFEarlyLate* model with 1060, and the *RBFFirstLast* model with 306. For promoter regions, the *RBFRBF* model finds 380, which is slightly more regions than the 373, the *RBFLinear* model found. The *RBFEarlyLate* and *RBFFirstLast* models both found 51 differentially accessible promoters. The detailed numbers of regions found with and without MMRC cutoffs can be found in Table 3.2.

	Promoters		H3K27ac Enhancers	
	FDR	FDR+MMRC	FDR	FDR+MMRC
RBFRBF	875	380	2700	2327
RBFLinear	559	373	3422	2783
RBFEarlyLate	112	51	2557	1060
RBFFirstLast	53	51	306	306

Table 3.2 | Number of promoter and enhancer regions found by models different pseudotime kernels. Numbers in the Significant columns have a BH-adjusted p-value smaller than 0.1. Number in the Significant + MMRC column have a BH adjusted p-value smaller than 0.1 and an MMRC larger than 0.3

To investigate the differences between the models, I compared their significance and MMRC estimates on the same enhancer regions (Fig 3.14). This comparison revealed that the *RBFLinear* kernel seems to have a slightly more favorable calibration compared to the *RBFRBF* model. This can be seen in the *top left* scatterplot of Figure 3.14 where the p-values of the two models are compared. The points show a consistent deviation from the diagonal unity line, which means that there is a consistent bias towards lower p-values for the *RBFLinear* model. In fact, looking at regions that only the *RBFLinear* model identified as significant (*orange points*), they are all close to the unity line, meaning that they fell short of significance in the *RBFRBF* model by only a small margin. Conversely, the regions only identified by the *RBFRBF* model (*blue points*), albeit fewer, are spread further from the unity line, meaning that they would not have been able to be identified by the *RBFLinear* model even with a looser significance cutoff. The same phenomenon is true for the MMRC estimates of the same models (Fig 3.14, *bottom left*). Visual investigation of the regions only found by the *RBFRBF* model revealed that the accessibility dynamics in these enhancers mostly followed an "up-down" dynamic where the accessibility rate first rises and later decreases again (data not shown). This was in line with expectations since the linear pseudotime kernel is not able to model these dynamics accurately.

The comparison of the *RBFRBF* model with the *RBFEarlyLate* model (Fig 3.14, *middle column*) revealed good agreement between the models in terms of significance estimation, but showed that the nonlinear pseudotime kernel has larger MMRC estimates for some enhancers. This is expected for those regions where the averaging of early and late cells results in a dilution of signal, i.e. where temporal dynamics cannot be captured accurately with these categories. The comparison of the *RBFRBF* model with the *RBFFirstLast* model (Fig 3.14, *right column*) showed that the *RBFFirstLast* model is systematically underpowered, due to the lower number of data points included in the model.

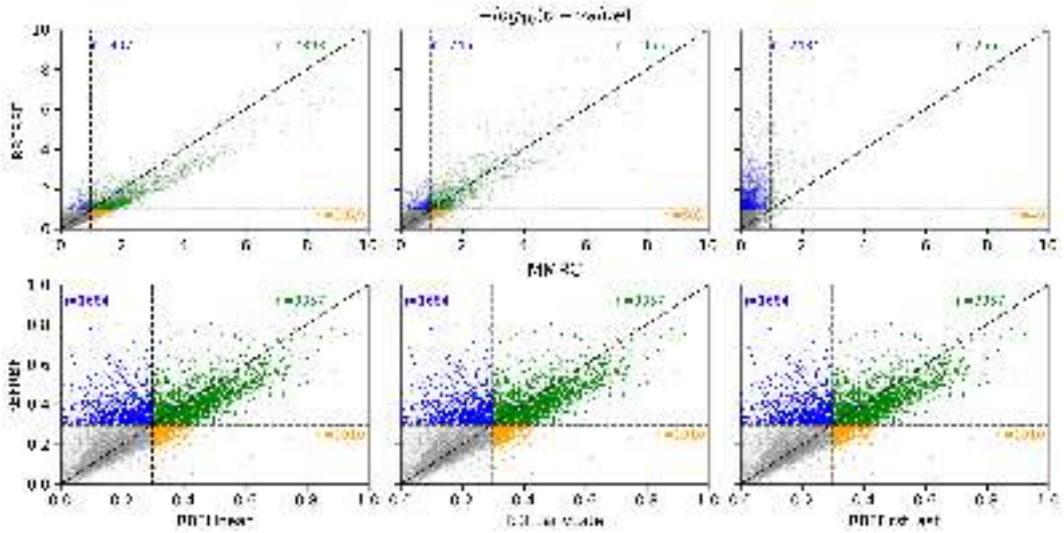


Figure 3.14 | Comparison of models with different pseudotime kernels for finding differentially accessible enhancers. The *top* row depicts the $-log_{10}$ p-values (BH adjusted) of the *RBFRBF* model (*y-axis*) vs. a comparison model on the *x-axis*. Horizontal and vertical dashed lines indicate p-value cutoffs at 0.1 FDR. The diagonal line is the identity line and means equal significance. Each dot represents the accessibility trajectory of an enhancer region during Mesoderm development. Blue dots are only found to be differential by the *RBFRBF* model, orange dots only by the comparison model, and green dots by both models. The *bottom row* depicts the same comparison in terms of MMRC. *Figure generated by Max Frank.*

These comparisons showed that the GPmeth model with an *RBFRBF* kernel is applicable, even in situations where the temporal changes of methylation/accessibility are mostly captured well by more simple models, while having the advantage of detecting more complex temporal changes, such as the ones shown in Figure 3.12. Next I compared the GPmeth model to scMet, an existing tool to test for differential methylation in single-cell data.

3.3.3.3 Comparison to scMet

As discussed in Section 1.5, there are currently few tools that are explicitly designed to detect differential methylation in single-cell data, and none that are designed to model these changes across a continuous time variable. Therefore the closest comparison is with scMet (Kapourani *et al.*, 2021), which models methylation/accessibility rate and variance in single-cell epigenomic data for predefined genomic windows. In Section 2.2.2, I showed that GPmeth was more powerful than scMet on simulated data. To compare the performance on real scNMT-seq data, I applied scMet to the same promoter and enhancer regions that GPmeth was applied to. Because scMet tests between predefined groups of cells, I defined an "early" and a "late" group, analogously to the EarlyLate pseudotime kernel discussed above, by placing a cutoff in pseudotime at 0.3. This results in the "early" group consisting mainly of E6.5 Epiblast and E6.5 Primitive Streak cells and the late group consisting of E6.5 Mesoderm and E7.5 Primitive Streak and Mesoderm cells.

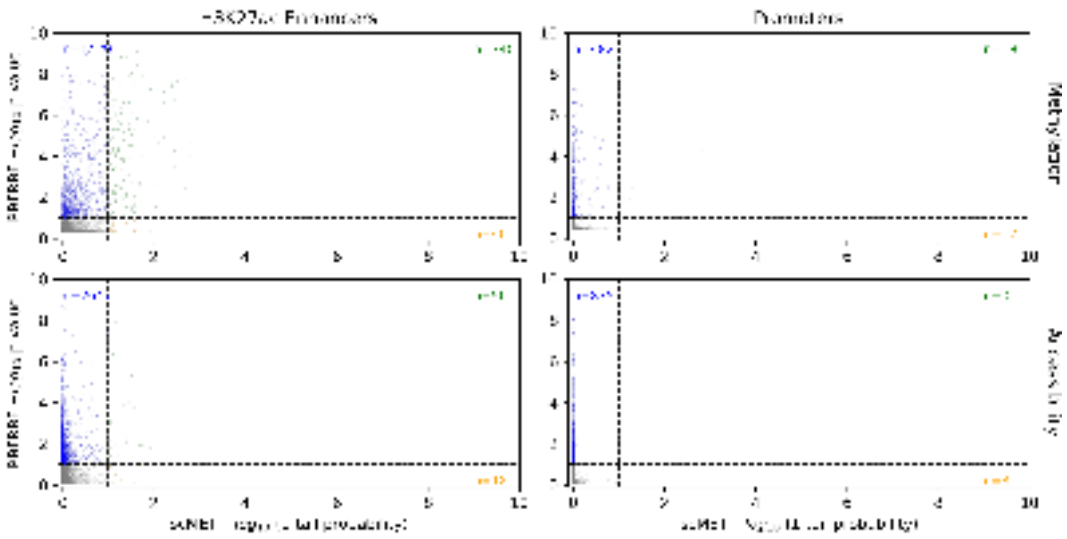


Figure 3.15 | Comparison of significance estimates of GPmeth and scMet. Scatterplots depict the $-log_{10}$ p-values (BH adjusted) of the *RBFRBF* model (*y-axis*) vs. $-log_{10}$ (1-tail probability) of scMet on the *x-axis*. Horizontal and vertical dashed lines indicate cutoffs at FDR < 0.1. The diagonal line is the identity line and represents equal significance. Each dot represents the methylation (*top row*) or accessibility (*bottom row*) trajectory of enhancer (*left column*) or promoter (*right column*) regions during Mesoderm development. Blue dots are only found to be differential by the *RBFRBF* model, orange dots only by the scMet model, and green dots by both models. *Figure generated by Max Frank.*

The scatterplots in Figure 3.15 show the comparison of the significance estimates between GPmeth on the *y-axis* and scMet on the *x-axis*. In this setting, scMet finds almost no significantly differentially methylated or accessible promoters or enhancers. This highlights the power of the GPmeth model in this setting. Importantly, this test gives an advantage to GPmeth in some ways. Firstly, the widths of the enhancer and promoter regions were chosen quite large (average 2kb and 4kb respectively), so that any methylation/accessibility changes in close proximity can be captured. Because scMet averages over the genome dimension, this dilutes the signal if the actual genomic window of differential methylation/accessibility is small. Thus, one could choose smaller regions to improve the performance of scMet, but this comes at the risk of missing signal at the edges of the windows. Secondly, the temporal cutoff to separate "early" and "late" cells could have been sub-optimal and result in the averaging of cells with different methylation/accessibility profiles. However, it should be noted that any choice of cutoff will be sub-optimal if changes over pseudotime are truly continuous in nature.

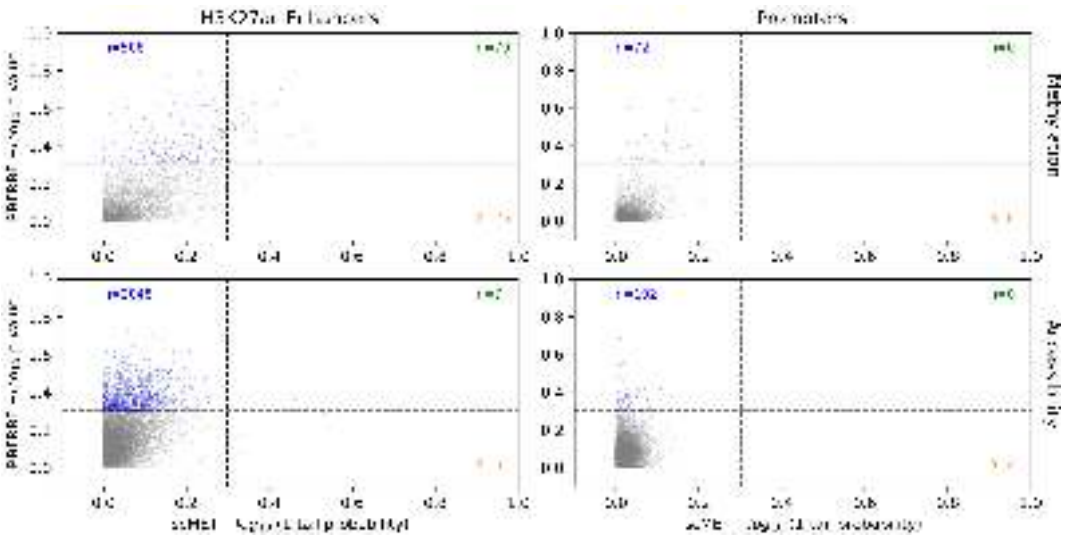


Figure 3.16 | Comparison of MMRC estimates of GPmeth and scMet. Scatterplots depict the MMRC estimate of the *RBFRBF* model (*y-axis*) vs. scMet on the *x-axis*. Horizontal and vertical dashed lines indicate an MMRC of 0.3. The diagonal line is the identity line. Each dot represents the methylation (*top row*) or accessibility (*bottom row*) trajectory of enhancer (*left column*) or promoter (*right column*) regions during Mesoderm development. Blue dots have a MMRC > 0.3 only with the *RBFRBF* model, orange dots only with the scMet model, and green by both models. *Figure generated by Max Frank.*

Figure 3.16 shows the MMRC estimates of GPmeth on the *y-axis* and scMet on the *x-axis*. scMet consistenly estimates lower changes in methylation/accessibility rates compared to GPmeth for all scenarios. This can also be expected for the same reasons as for the significance estimates. Note that this comparison of models does not directly show increases of statistical power or accuracy of the GPmeth model compared to scMet, since there is no ground truth available for these data. However, in the rest of this Chapter I will show the biological validity of the differentially methylated and differentially accessible regions found by GPmeth.

3.3.4 Analysis of refined subregions found by GPmeth

A major benefit of GPmeth is the ability to refine the originally provided genomic windows to get insights into where precisely the changes in methylation rate are occurring. For this purpose, I evaluated the posterior predictions of GPmeth for every region by taking 2000 samples from the trained models. This allows the quantification of the mean and uncertainty of the model for any point in the genome and along pseudotime. I define differentially methylated subregions as widows in the genome axis where the 5% CI of all predictions has a methylation rate change over pseudotime larger than 0.3 (Fig 3.5, grey shaded areas). Note that this threshold can be set by the user and is arbitrary. In this case, 0.3 seems to provide a sensible threshold for filtering out biologically relevant signals.

Figure 3.17 shows an overview of the refined subregions identified this way. Most differentially methylated enhancers and promoters contain a single subregion (note

the log-scale in Figure 3.17, *top left*). The identified subregions have an average width of 163 bp (Fig 3.17, *top right*). Note that this value is sensitive to the specification of the model parameters, especially to the setting of the genome kernel length scale. Therefore, the subregion width should mostly be used for comparisons of different regions that were tested with the same model specifications. When comparing the widths of promoter and enhancer subregions, no significant differences were observed (promoter mean width: 135bp, enhancers mean width: 165bp, p-value=0.11, *t-test on log-transformed widths*). When comparing the relative positioning of the identified subregions (Fig 3.17, *bottom left*), enhancer subregions are distributed around the center of the window, whereas promoter subregions are close to uniformly distributed across the window. This is surprising since one would expect differential methylation that directly influences gene expression via the promoter to be close to the TSS of the gene and, therefore, close to the 0 position. This could indicate that some of the identified subregions are, in fact, different genomic elements that happen to be in close proximity to the genes TSS. This would be another indicator that promoter methylation is not a main driver in gastrulation. Finally, we can inspect the methylation rate change prediction of the model within the found subregions (Fig 3.17, *bottom right*). This is taken as an average of all predictions of the model that fall into the subregion. Promoters have significantly lower differential methylation rates than enhancers (mean MMRC of promoters: 0.22, mean MMRC of enhancers: 0.31, p-value=$3.1e^{-31}$, *t-test on log-transformed rates*).

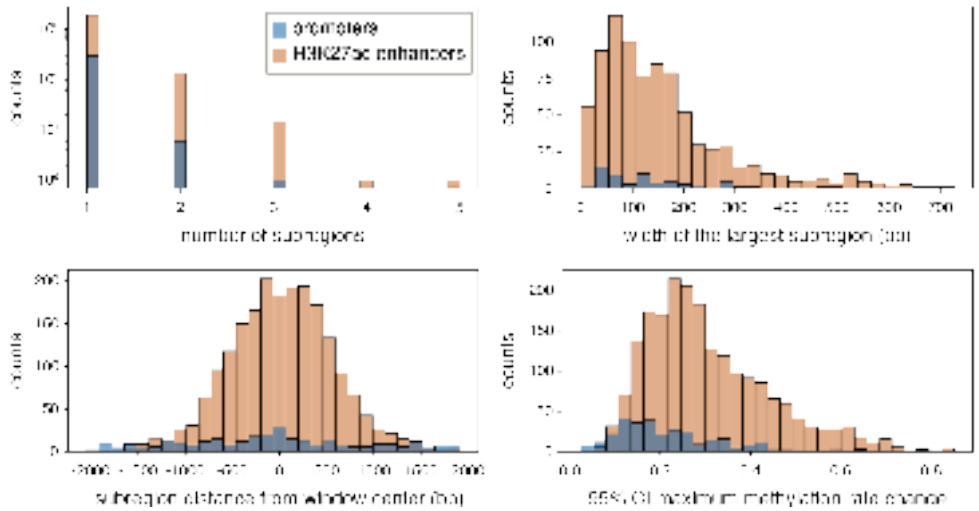


Figure 3.17 | Summary statistics of differentially methylated refined regions. The *top left* panel shows the number of subregions that are found by GPmeth for every genomic window with significant differential methylation. Note the log scale on the y-axis. The *top right* panel shows the distribution of subregion widths, i.e., the width of the genomic interval where the 95% CI MMRC is higher than a specified threshold (in this case 0.3). Most regions are smaller than 300bp (average 163bp), indicating that averaging methylation signal over larger genomic windows can dilute the signal. The *bottom left* panel shows the positioning of the center of the identified subregions relative to the center of the input genomic window. With perfect region annotations, this histogram should be a narrow Gaussian distribution around zero. The distribution of promoter subregions shows that there is no preference for differential methylation close to the TSS of the gene. Enhancer subregions are distributed around the center of the window but show a substantial spread, highlighting the need for flexible models that can tolerate inaccurate region inputs. The *bottom right* panel shows the average 95% CI MMRC of each identified subregion. Note that this is a conservative estimate of the true change in methylation rate since it includes the uncertainty of the GPmeth model. It can serve as a convenient filtering criterion to exclude regions with small or uncertain effect sizes. *Figure generated by Max Frank.*

Figure 3.18 shows an overview of the refined subregions found by GPmeth. Again, most differentially accessible enhancers and promoters contain a single subregion (note the log-scale in Figure 3.18, *top left*). The identified subregions have an average width of 120 bp (Fig3.18, *top right*). This is shorter than the 163 bp average window size that was identified for differentially methylated regions. However, this could be due to the higher density of GpC sites. When comparing the widths of promoter and enhancer subregions, there were slight, but not significant differences between the two (promoter mean width: 97bp, enhancers mean width: 122bp, p-value=0.06, *t-test on log-transformed widths*). The positioning of the identified enhancer subregions (Fig 3.18, *bottom left*) is similar to the positioning of differentially methylated subregions. However, for promoters, there is now also an enrichment in the center of the genomic window around the TSS of the gene. This was not the case for endogenous methylation. This could mean that there is a small subset of genes that will be regulated by promoter accessibility during Mesoderm formation. In fact, one of those genes is be Mesp2, which was shown in Figure 3.5. Furthermore, if we inspect the GpC methylation rate change prediction of the model within the found subregions (Fig 3.18, *bottom right*),

promoters have roughly equal differential methylation rates than enhancers (mean MMRC of promoters: 0.34, mean MMRC of enhancers: 0.33, p-value=0.24, *t-test on log-transformed rates*).

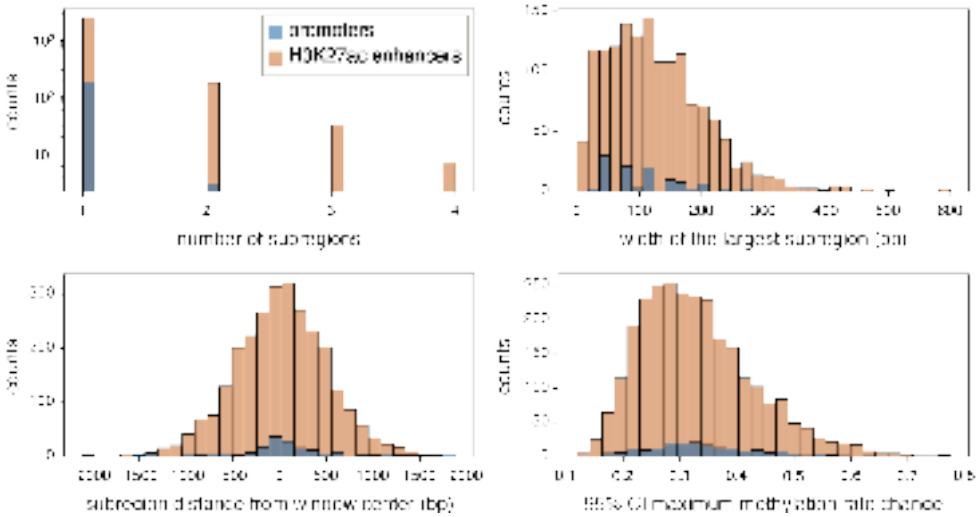


Figure 3.18 | Summary statistics of differentially accessible refined regions. The *top left* panel shows the number of subregions that are found by GPmeth for every genomic window with significant differential accessibility. Note the log scale on the y-axis. The *top right* panel shows the distribution of subregion widths, i.e., the width of the genomic interval where the 95% CI MMRC is higher than a specified threshold (in this case, 0.3). Most regions are smaller than 300bp (average 120bp), indicating that averaging methylation signals over larger genomic windows can dilute the signal. The *bottom left* panel shows the positioning of the center of the identified subregions relative to the center of the input genomic window. The *bottom right* panel shows the average 95% CI MMRC of each identified subregion. *Figure generated by Max Frank.*

The capability of GPmeth to refine input regions opens up interesting avenues for experimental follow-ups that need precise boundaries of where methylation/accessibility changes happen. To verify, whether the refined regions are biologically relevant, I performed a transcription factor binding site analysis.

3.3.4.1 Transcription factor binding site enrichment

Transcription factors (TFs) are important drivers of gastrulation and embryonic development(Meissner, 2010). They act by recognizing nucleotide patterns and binding DNA, inducing the transcription or repression of nearby genes. Transcription factors also have a complex link to DNA methylation and chromatin accessibility (Hemberger *et al.*, 2009). Because many transcription factors have both well-established roles during embryo development and known preferences for nucleotide sequences (binding motifs), I wanted to investigate if the differentially methylated/accessible regions found by GPmeth have an enrichment of relevant TF binding motifs.

To this end, I first extracted all subregions that GPmeth identified as differentially accessible during Mesdoerm development and clustered their pseudotemporal trajectories. There were four main trajectories that clustered together, two groups of

enhancers where chromatin accessibility increased with Mesoderm development and two groups where chromatin accessibility decreased (Fig 3.19, *a*). Each of these groups could further be subdivided into trajectories that change accessibility early in the pseudotime (at the primitive streak to early Mesoderm transition) and those that change later on (at the mature Mesoderm stage). I termed these four clusters "early up", "late up", "early down", and "late down" respectively. I then performed an enrichment analysis for TF binding motifs found in the JASPAR CORE vertebrate database (Rauluseviciute *et al.*, 2024) for each cluster. Each trajectory cluster showed enrichment of distinct motifs (Fig 3.19,*b,c*). Notably, enhancer subregions where GPmeth identified a decrease in accessibility are enriched in known Ectoderm and pluripotency TFs, such as POU5F1, SOX2, and SP8. This is in line with the observation in Section 3.3.2.1 that Ectoderm-defining enhancers are decreasing in accessibility during Mesoderm development. Conversely, subregions with increasing chromatin accessibility are enriched in known Mesoderm-specific TF binding sites, such as GATA4, FOXP2, EOMES, and TWIST1. Interestingly, while both downregulated trajectory clusters are enriched in similar TF binding sites (Fig 3.19,*c, green lines*), the two upregulated clusters have distinct enrichment patterns (Fig 3.19,*c, blue and purple lines*). Enhancer subregions that increase early are more enriched for the members of the T-box family of transcription factors, which are essential for the migration of nascent Mesoderm cells in the primitive streak (Costello *et al.*, 2011; Papaioannou, 2014). In particular, EOMES was shown to control the expression of Mesp1. Subregions with a late accessibility increase are mainly enriched in GATA transcription factors. The GATA family of transcription factors is known to be involved in the formation of the endocrine system (Viger *et al.*, 2008), which is one of the tissues that emerges from Mesoderm cells. This points to the intricate temporal control of gene expression that is modulated by a network of lineage-defining transcription factors and the epigenetic modifications of DNA regions they bind to.

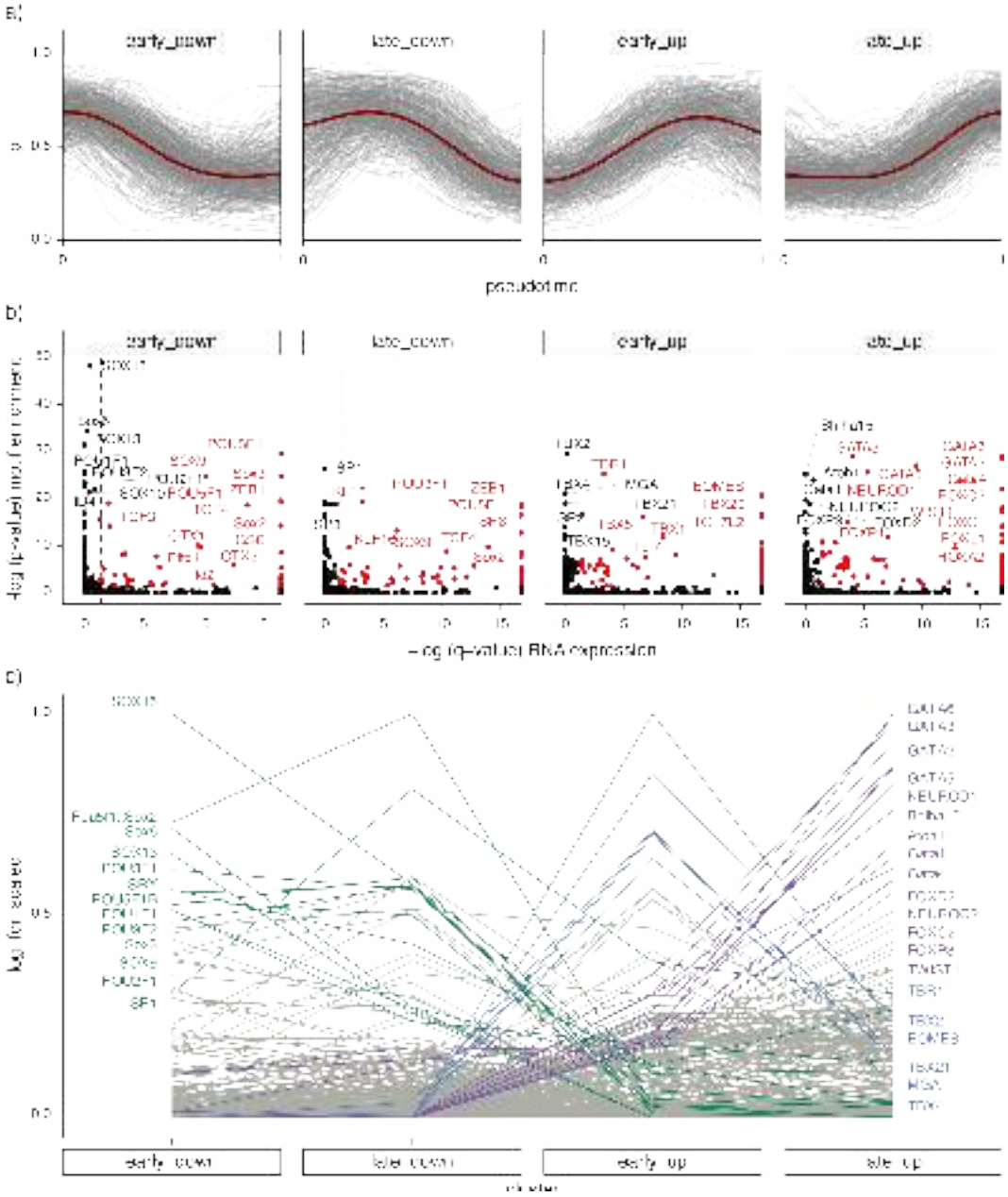


Figure 3.19 | Transcription factor enrichment of differentially accessible enhancers. *a)* k-means clustering of pseudotime trajectories. Each grey line represents the average accessibility rate ρ within a significant subregion identified by GPmeth. Dark red lines are the mean trajectory of each cluster. *b)* Transcription factor motif enrichment per cluster(Fisher's exact test, -log(p value) BH-adjusted, y-axis) versus differential RNA expression (GPcounts, -log(q-value) , x-axis) of the transcription factor. Red points are significant at FDR < 0.05 in bot enrichment and differential RNA expression. *c)* Scaled view of Transcription factor motif enrichment per cluster. *Green lines* are specifically enriched in the early down and late down clusters (FDR $< 10^{-17}$). *Blue lines* and *purple lines* are specifically enriched in the early up and late up clusters (FDR $< 10^{-23}$) respectively. *Figure generated by Max Frank.*

Since these results were in line with known gastrulation biology, I wanted to investigate if the same results could have been obtained without GPmeth's ability to identify refined subregions. To this end, I repeated the analysis with the same clustering of enhancer regions but instead provided the DNA sequences of the whole input genomic

windows to the enrichment instead of the sequences of the subregions. Figure 3.20 shows the enrichment for the whole enhancer regions by cluster. Apart from some more spurious enrichments, with this analysis, the effect of specific enrichment depending on the trajectory was lost. This is a good indicator that the refined subregions identified by GPmeth are highly likely to have functionally relevant roles in gene regulation.

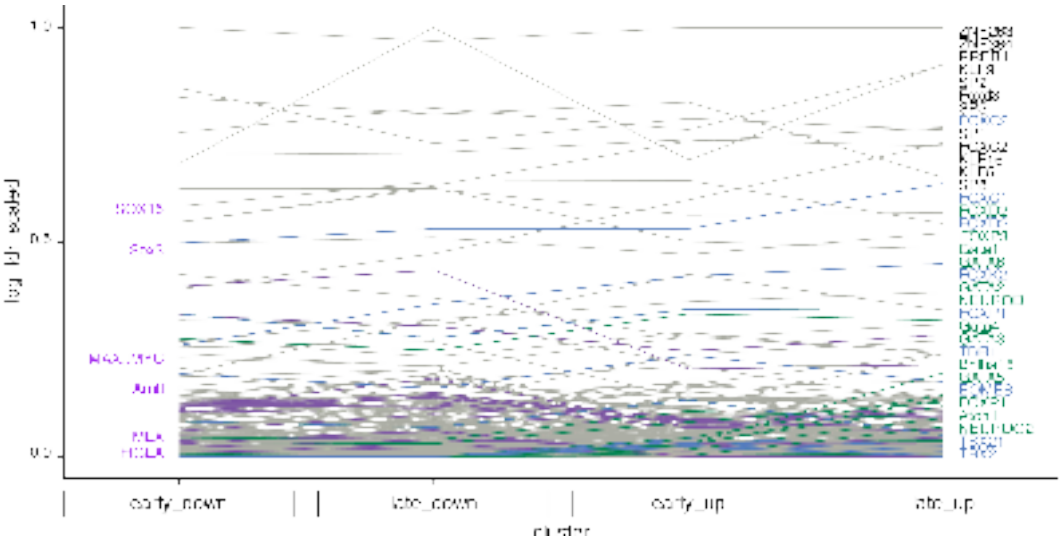


Figure 3.20 | Transcription factor enrichment of differentially accessible full enhancer regions. Scaled view of transcription factor motif enrichment per accessibility pseudotime trajectory cluster (Fig 3.19). Lines are colored the same as in the Figure above and represent cluster-specific enrichment found with the GPmeth 0.02855 workflow. Note that all lines are close to parallel, indicating that the enrichment is identical to the same for all four clusters. *Figure generated by Max Frank.*

Next I investigated the TF binding site enrichment in enhancers that are differentially methylated during Mesoderm formation. Again, 4 clusters were identified following the same temporal patterns as differentially accessible enhancers. Figure 3.21 shows the results of the enrichment analysis. Here, the trajectories where the methylation rate increased over time showed enrichment for Ectoderm and pluripotency enhancers. Interestingly binding motifs of regions in the early up trajectory include MAX and MYC, TFs present in pluripotent stem cells, whose binding is known to be impacted by DNA methylation (Cusack *et al.*, 2020; Domcke *et al.*, 2015). In enhancers that get methylated later, SOX3, SOX6, and SOX15 of the SOX family of TFs are enriched, which are implicated in brain development (Bylund *et al.*, 2003), which is in line with their later repression compared to the aforementioned pluripotency TFs. In the enhancers that get demethylated early, we again see enrichment of binding sites for members of the T-Box TF family TBX and EOMES, and enhancers with late demethylation show enrichment in GATA TFs, which is in line with the enrichment results of differentially accessible regions.

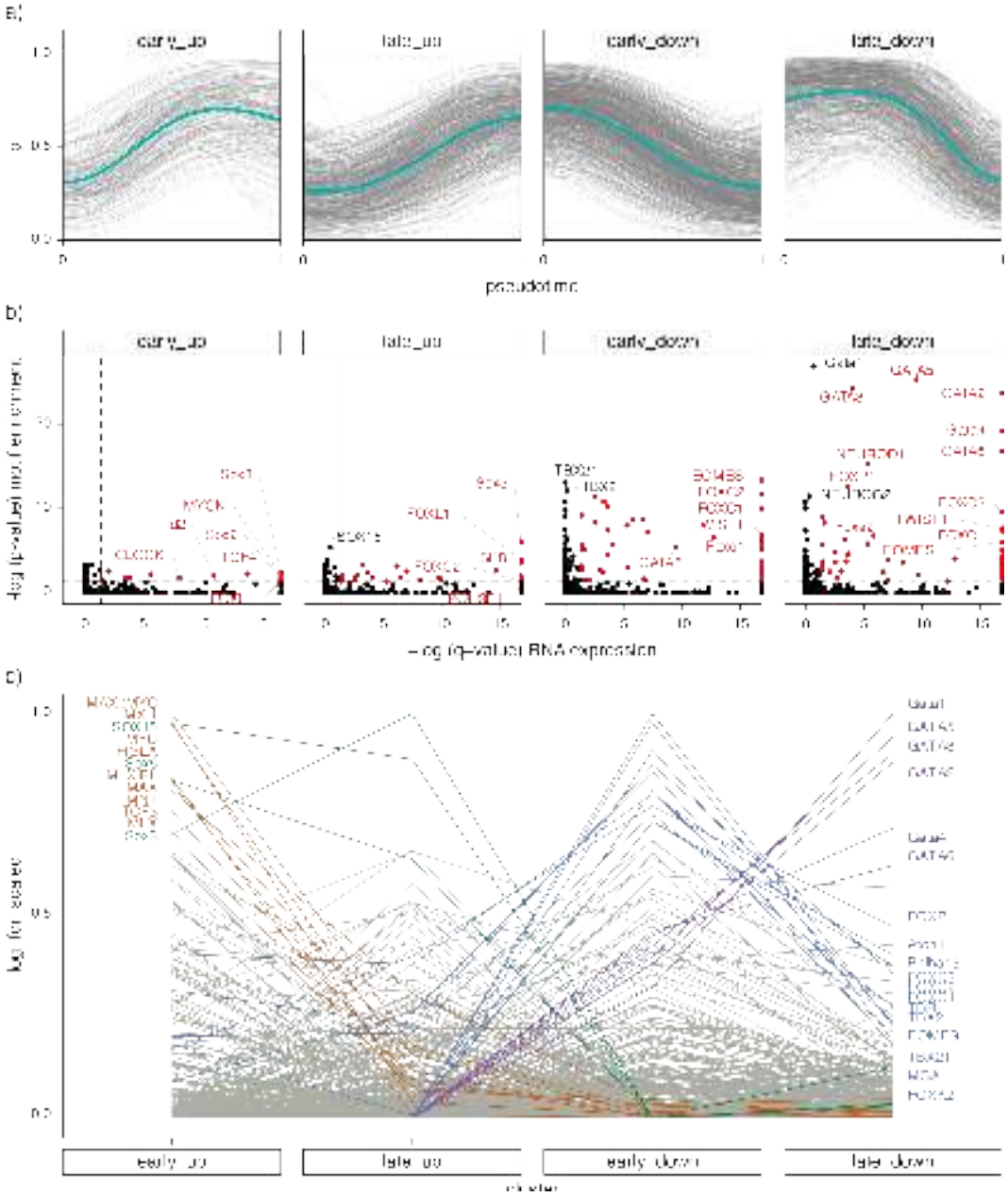


Figure 3.21 | Transcription factor enrichment of differentially methylated enhancers. *a)* k-means clustering of pseudotime trajectories. Each grey line represents the average methylation rate ρ within a significant subregion identified by GPmeth. Dark red lines are the mean trajectory of each cluster. *b)* Transcription factor motif enrichment per cluster(Fisher's exact test, -log(p-value) BH-adjusted, y-axis) versus differential RNA expression (GPcounts, -log(q-value) , x-axis) of the transcription factor. Red points are significant at FDR < 0.05 in bot enrichment and differential RNA expression. *c)* Scaled view of Transcription factor motif enrichment per cluster. *Green lines* are specifically enriched in the early down and late down clusters (FDR$< 10^{-17}$). *Blue lines* and *purple lines* are specifically enriched in the early up and late up clusters (FDR$<10^{-23}$) respectively. *Figure generated by Max Frank.*

I also investigated again if the same enrichments can be found without using the refined subregions identified by GPmeth and found the same result of nonspecific enrichment (Fig 3.22).

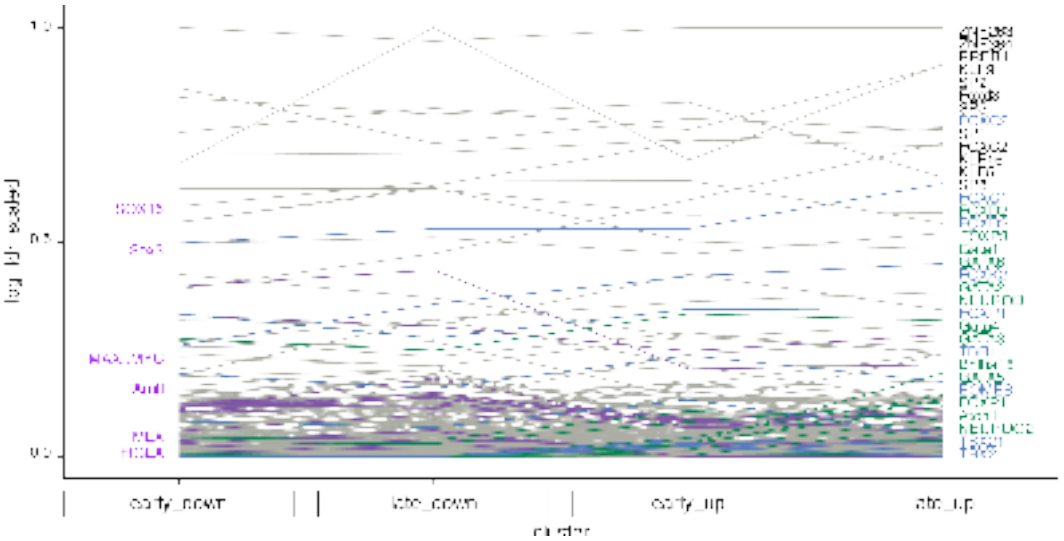


Figure 3.22 | Transcription factor enrichment of differentially methylated full enhancer regions. Scaled view of transcription factor motif enrichment per methylation pseudotime trajectory cluster (Fig 3.21). Lines are colored the same as in the above figure and represent cluster-specific enrichment found with the GPmeth workflow. Note that all lines are close to parallel, indicating that the enrichment is identical the same for all four clusters. *Figure generated by Max Frank.*

3.3.5 Analysis of lineage-defining enhancer regions

Next, I investigated whether the GPmeth results for Mesoderm enhancers are in agreement with biological expectations. Since the enhancer regions are based on the combined ChIP-seq signal of differentiated Mesoderm, Endoderm, and Ectoderm tissues, we can define lineage-specific enhancers by overlapping ChIP-seq peaks for the individual lineages with the combined signal and filtering peaks that are exclusively present in one of the tissues. This was done analogously to Argelaguet *et al.*, 2019b. Here I only considered lineage-specific enhancer regions that perfectly overlapped the regions of the combined signal. This resulted in 2122 Ectoderm-specific enhancers, 1036 Endoderm-specific enhancers and 895 Mesoderm-specific enhancers. Note that this list will not be exhaustive for all lineage-specific enhancers.

With these lineage-specific enhancers in hand, I first assessed how many of these regions show significant differential methylation/accessibility during Mesoderm formation (Fig 3.23). As expected Mesoderm-specific enhancers are most likely to be differentially methylated or accessible, with 119 (differentially methylated), 100 (differentially accessible) and 129 (both) regions out of 895 regions tested. Conversely, there are almost no Endoderm enhancers that show differential methylation or differential accessibility. Interestingly, however, there are 214 differentially methylated, 247 differentially accessible and 79 differentially methylated and accessible Ectoderm enhancers.

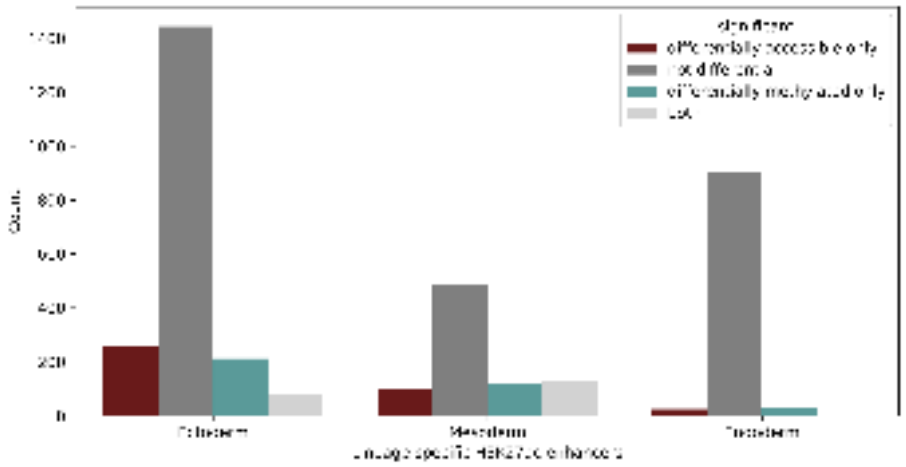


Figure 3.23 | Number of differentially methylated/accessible lineage-specific enhancers. Bar heights indicate the number of lineage-specific enhancer regions that were identified by GPmeth to be significantly differentially accessible or methylated (FDR=0.1). GPmeth found 335 out of 2122 (16%) Ectoderm-specific enhancers, 229 out of 895 (26%) Mesoderm-specific enhancers and 28 out of 1036 (3%) Endoderm-specific enhancers to be differentially accessible. Of those, 79 (Ectoderm), 129 (Mesoderm), and 3 (Endoderm) were also differentially methylated. *Figure generated by Max Frank.*

While this may be surprising at first glance, a closer inspection reveals that many Ectoderm-specific enhancers are already demethylated and highly accessible in E5.5 Epiblast cells and stay demethylated and accessible during Ectoderm formation but get methylated and closed in cells of the other two germ layers. This was also found in Argelaguet *et al.*, 2019b. Using the GPmeth output, this phenomenon can be shown by calculating the average methylation/accessibility rate across pseudotime and genomic for all lineage-specific enhancers. Figure 3.24 shows the averaged methylation predictions of the model along Mesoderm development. Here, it becomes clear that Mesoderm-specific enhancers start out highly methylated at E5.5 and become demethylated over time, while Ectoderm-specific enhancers start out with low methylation and methylation slightly increases over time. Endoderm-specific enhancers mostly stay methylated throughout Mesoderm lineage formation. Note that the *bottom right* panel of Figure 3.24 depicts the average of only 37 significantly differentially methylated regions.

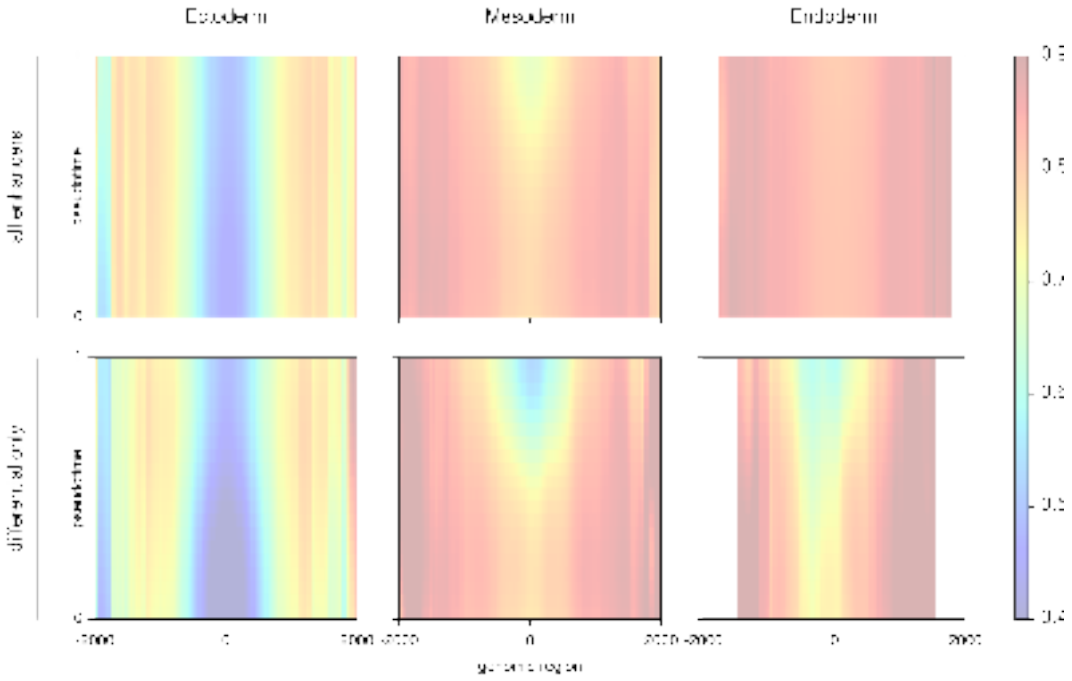


Figure 3.24 | Averaged methylation rate profiles for lineage-specific enhancers. The heatmaps represent the GPmeth posterior mean predictions, averaged across lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model for each region in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime and taking the averages of the aligned grids. The *top row* averages all lineage-specific enhancers for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR < 0.1). *Figure generated by Max Frank.*

Compared to endogenous methylation, chromatin accessibility follows the opposite trend, Figure 3.25 shows the averaged predictions of the model, showing that Mesoderm-specific enhancers start out mostly inaccessible at E5.5 and become increasingly accessible over time while Ectoderm-specific enhancers start out at intermediate accessibility rates and decrease over time. Endoderm-specific enhancers mostly stay at intermediate accessibility throughout Mesoderm lineage formation.

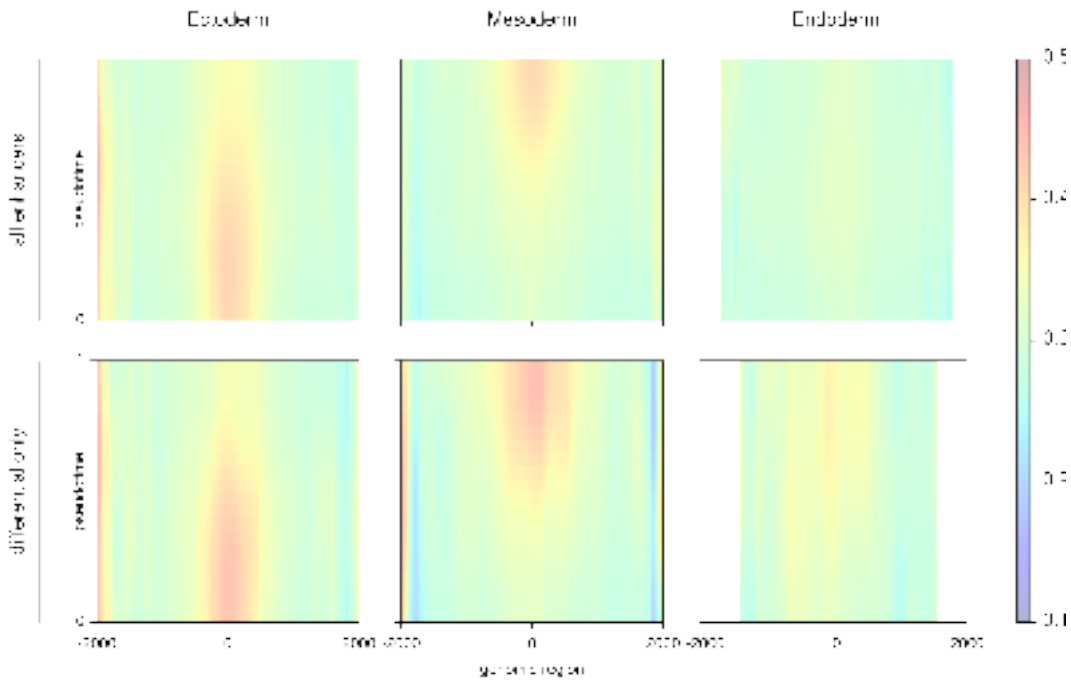


Figure 3.25 | Averaged accessibility rate profiles for lineage-specific enhancers. The heatmaps represent the GPmeth posterior mean predictions, averaged for lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime. The *top row* averages all lineage-specific enhancers for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR 0.1). *Figure generated by Max Frank.*

While these averages give insights into spatiotemporal changes in methylation rate, similar plots could be produced by simply binning the scBS-seq data across time and smaller genomic windows directly. GPmeth, however, also offers insights into the distribution of methylation rate changes based on individual regions. To investigate this, I extracted the averaged predictions of GPmeth for all significant regions within the refined subregions. For each significantly differentially methylated lineage-specific enhancer, I then visualize the subregion with a 95% CI MMRC > 0.3(Fig 3.26). Differentially methylated Mesoderm enhancers almost exclusively decrease in methylation rate over time, while the majority of differential Ectoderm enhancers increase in methylation rate over time. In comparison, the few significant Endoderm enhancers show a more mixed signal.

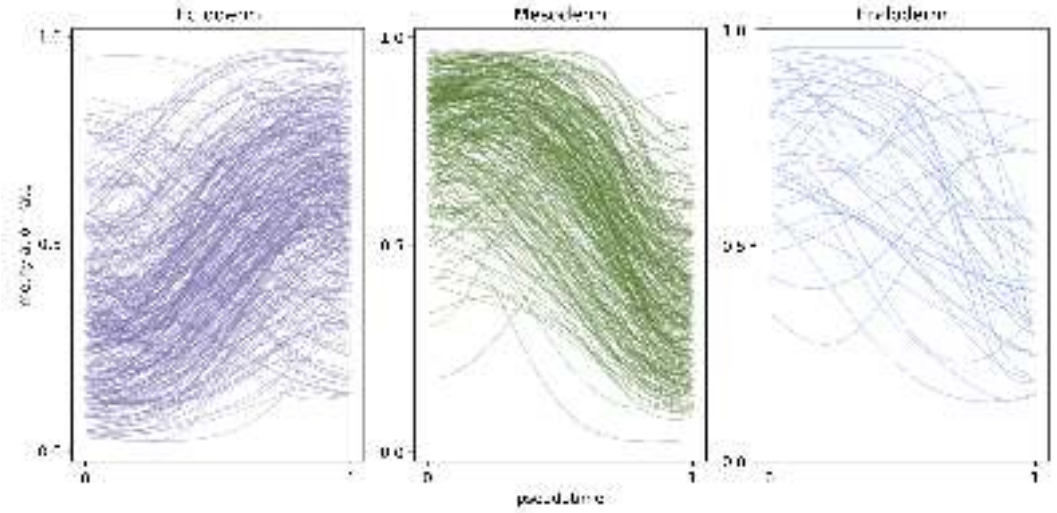


Figure 3.26 | GPmeth refined pseudotemporal methylation trajectories of lineage-specific enhancer regions. Lines represent the GPmeth posterior methylation rate averages of the refined subregions found within differentially methylated enhancers by the model. Ectoderm-specific enhancers consistently increase in methylation rate over time, while Mesoderm-specific enhancers decrease in methylation rate. Ectoderm-specific enhancers (*left*) increase methylation rate from 0.34 to 0.61 on average across the pseudotime range. Mesoderm-specific enhancers (*center*) decrease methylation rate from 0.78 to 0.36, and Endoderm-specific enhancers decrease from 0.74 to 0.43 on average. *Figure generated by Max Frank.*

To get a more quantitative measure of the pseudotemporal trends for each enhancer class, I used k-means clustering to extract patterns in methylation rate from this data. Figure 3.27 shows the extracted trends for three clusters. Note that methylation rates were scaled for each time series to make the clustering invariant to absolute methylation rate effects. For Ectoderm-specific enhancers, there is one cluster for the majority of regions that increase in methylation over time, and two clusters capture outlier patterns with decreasing methylation rate. For Mesoderm, k-means found one cluster for outlier patterns with increasing methylation rate and two distinct patterns for decreasing methylation rates, where one group of enhancers decreases in methylation rate consistently over time while the other group appears to stay highly methylated until half the pseudotime has passed and then experience rapid demethylation. Similar groups are observed in Endoderm, which could be due to some mislabeled regions or off-target effects.

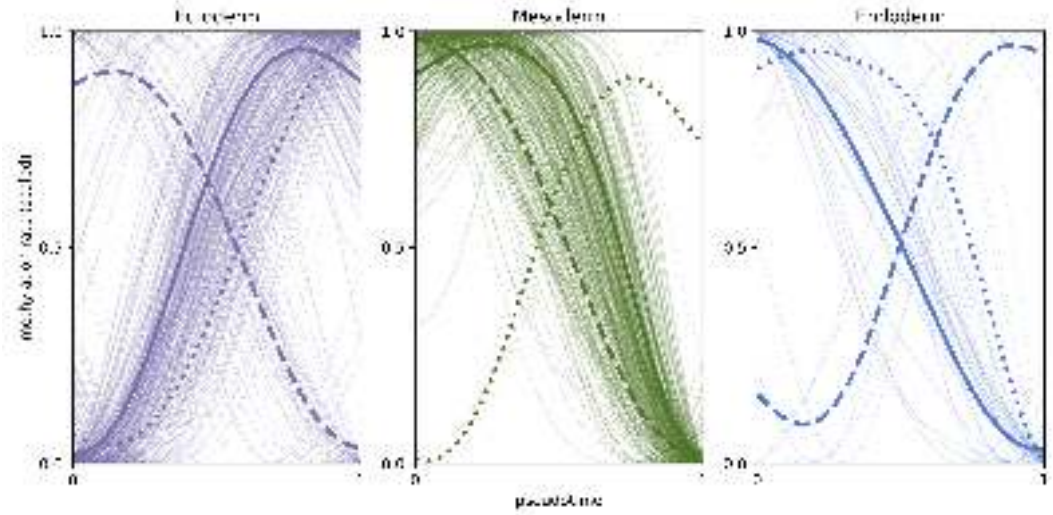


Figure 3.27 | Clustered pseudotemporal trajectories of lineage-specific enhancer regions. Shown are the same pseudotemporal trajectories of methylation rate predictions as in Figure 3.26, but every trajectory was scaled to the range of [0,1] for scale-invariant clustering. These trajectories were then clustered with k-means based on Euclidean distance. A fixed number of three clusters was specified to capture up to two different trends in the trajectories and outlier trajectories. *Figure generated by Max Frank.*

One interesting question that can be addressed by the high spatiotemporal resolution of GPmeth models is the temporal order in which certain epigenetic regulatory events happen. For example, does the increase in Ectoderm enhancer methylation precede or succeed the decrease in methylation of Mesoderm enhancers during Mesoderm formation. For this purpose, I plotted the extracted patterns against each other while inverting the trend for Ectoderm enhancers for better comparability (Fig 3.28). While there is substantial overlap between the distributions of time series, there is a clear shift where Mesoderm enhancers are demethylated after Ectoderm enhancers are methylated. This is not unexpected since there is substantial overlap between Ectoderm enhancers and pluripotency enhancers (Argelaguet *et al.*, 2019b), which should be expected to get inactivated relatively early.

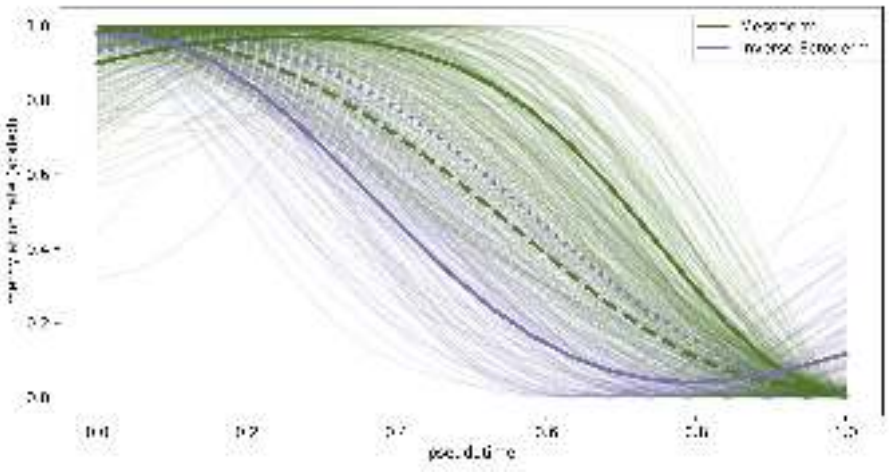


Figure 3.28 | Temporal comparison of Ectoderm and Mesoderm-specific enhancer methylation. Lines are pseudotemporal trajectories of lineage-specific enhancers as in Figure 3.27. Green lines correspond to the two major clusters of Mesoderm enhancers that are demethylated over time. Purple lines correspond to the inverse profiles (i.e., 1-methylation rate) of Ectoderm enhancers that are methylated over time. Both enhancer classes show similar temporal patterns, but Ectoderm-specific enhancer methylation tends to precede Mesoderm-specific enhancer demethylation. *Figure generated by Max Frank.*

I performed the same analysis for accessibility rate trajectories of lineage-specific enhancers, with similar results to methylation rates. Details of this analysis can be found in the Appendix (A.1).

3.3.6 Integration of molecular modalities

One of the benefits of the GPmeth model is that it allows the investigation of the relationships between the modalities that are measured in scNMT-seq in great detail. In this Section, I will describe the analyses that I performed to integrate RNA expression, DNA methylation, and chromatin accessibility over the course of Mesoderm development. First, I will discuss the analysis of promoter regions, followed by H3K27ac enhancers. GPmeth provided models for methylation and accessibility in these regions. To detect differential RNA expression, I applied another Gaussian process-based tool called GPcounts (BinTayyash *et al.*, 2021) to the expression profiles of 1171 cells measured by scNMT-seq along the same Mesoderm pseudotime course. Briefly, GPcounts fits a dynamic and a static model to the pseudotime expression profiles of each gene and performs a likelihood-ratio test to find variable genes. This means I could perform an analysis of the GP model outputs of RNA expression (GPcounts) promoter methylation and promoter accessibility (GPmeth).

3.3.6.1 Integrative analysis of Promoter regions

As described in Section 3.3.2, most promoter regions are not differentially methylated or accessible during Mesoderm development. However, there are more than 5000 genes

that change in expression during this process (Fig 3.29). Only a very small subset of 20 promoter-gene pairs is significantly changing in all three modalities. Furthermore only 26 promoters are changing both their methylation rate and their chromatin accessibility. It is tempting to conclude that the three modalities are, therefore, not linked at all in this scenario. However note that this overlap is dependent on the significance and effect size cutoffs of tests. This will be explored in more detail below.

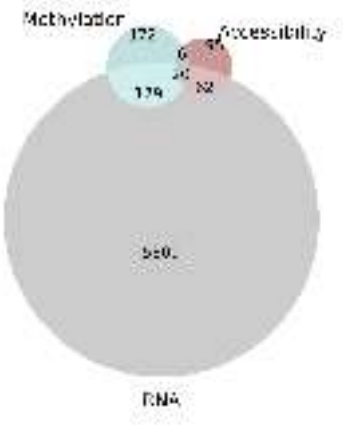


Figure 3.29 | Venn Diagram of differentially regulated genes and promoters. Number of significant promoters/genes found by GPmeth/GPcounts, respectively. The cutoff for significance with GPmeth used here was FDR<0.1 and MMRC > 0.3. For GPcounts, significant differential gene expression was defined q-value < 0.1. *Figure generated by Max Frank.*

Next, I compared the model estimates of the magnitudes of change for each modality. Figure 3.30 compares the fold-change estimates of RNA expression with the MMRC output of GPmeth for all promoters. As expected there is only a small number of promoters that change in RNA expression congruently with promoter methylation or accessibility. Furthermore, the promoter MMRC does not seem to be correlated to the fold-change of the gene even for promoter-gene pairs that are significantly differentially accessible/methylated and differentially expressed (Fig 3.30 *left and middle panel, green points*). In contrast, there seems to be a moderate link between the accessibility MMRC estimate and the methylation MMRC estimate (Fig 3.30 *right panel, green points*). This link will be explored further below.

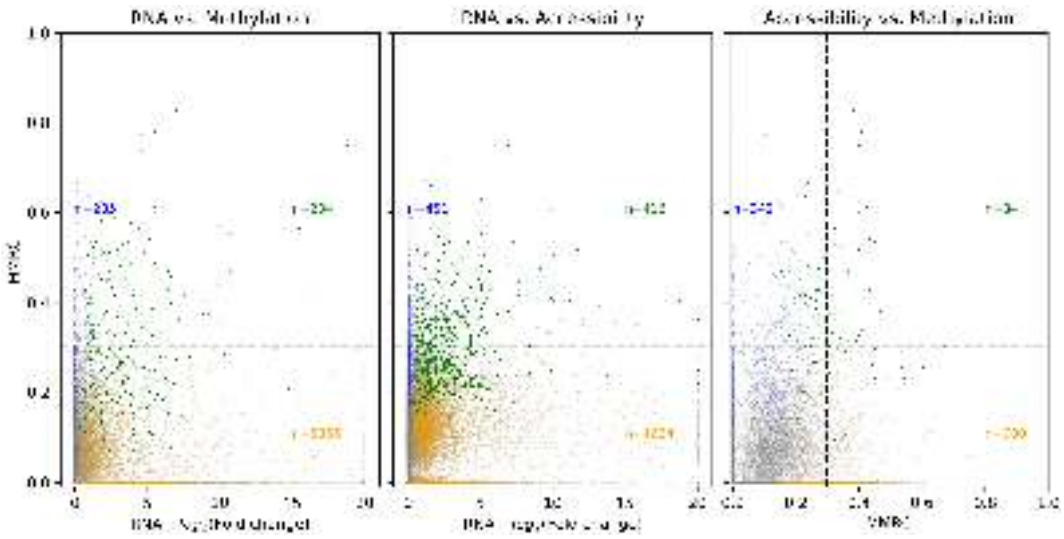


Figure 3.30 | Effect sizes of differential regulation of gene-promoter pairs. Scatterplots show the pairwise comparison of promoter methylation/accessibility and gene expression change magnitudes during Mesoderm development. The *left panel* shows -log_{10} RNA expression fold change (*x-axis*) versus the MMRC estimate of promoter methylation (*y-axis*). The *center panel* shows the -log_{10} RNA expression fold change (*x-axis*) and MMRC estimate of promoter accessibility on the *y-axis*. The *right panel* plots accessibility MMRC (*x-axis*) versus methylation MMRC (*y-axis*). *Yellow* dots points significant changes (FDR<0.1) in the modality displayed on the x-axis, *blue* points indicate significant changes of the modality on the y-axis, and *green* points indicate significant changes in both. *Figure generated by Max Frank.*

The comparison of the magnitude of change in methylation and accessibility above does not directly give any information about the link between the modalities. Therefore I extracted the posterior methylation rate estimates within refined promoter subregions at 20 equally spaced points across pseudotime, similar to Section 3.3.4. I also extracted these estimates for promoters that were not significantly differentially methylated/accessible. Then I calculated the correlation between the accessibility and methylation time series. Figure 3.31 shows the correlation estimates for four different classes of promoters: differentially methylated, differentially accessible, neither differentially methylated nor accessible (not differential), as well as differentially methylated and differentially accessible (both). The majority of promoters that GPmeth detected as both were highly inversely correlated, which would be expected if a promoter gets activated or inactivated over time. Interestingly, of the subset of regions that were only differentially methylated, most are still highly inversely correlated. This could hint at the fact that these regions are still co-regulated, but there was not enough data to detect differential accessibility. In promoters that were only differentially accessible, most regions are not more inversely correlated than the background distribution of non-differential regions. Overall, this hints at the fact that methylation changes in promoters are mostly accompanied by opposite accessibility changes, whereas accessibility changes do not necessarily result in changes in methylation rate.

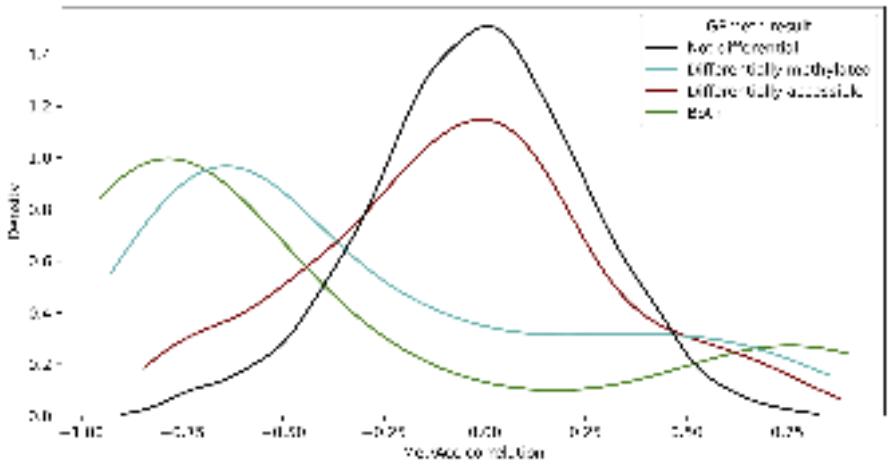


Figure 3.31 | Correlation of promoter methylation and accessibility during Mesoderm development. Kernel density estimates of the distributions of Pearson-correlation between promoter methylation and promoter accessibility time series extracted from GPmeth. The color indicates if GPmeth identified the promoter as significantly methylated/accessible at FDR<0.1 and MMRC>0.3. *Figure generated by Max Frank.*

To investigate if there is a small population of genes where methylation accessibility and RNA expression are directly linked, I calculated the correlations of promoter methylation/accessibility time series to RNA time series extracted from the GPcounts models. Figure 3.32 shows the correlations for the four classes of promoters. The expectation for classical gene regulation would be that promoter methylation is repressing the expression of a gene. Therefore, we expect a negative correlation. Promoter accessibility should induce gene expression, resulting in a positive correlation. There does not seem to be a strong enrichment for negative correlations in differentially methylated promoters. There is a small enrichment for positive correlations for differentially accessible regions, meaning that for a small subset of genes, promoter accessibility could play a role in gene regulation during Mesoderm formation, but this regulation seems independent of methylation. If we filter for correlation coefficients of >0.7 for methylation/RNA and <-0.7 for accessibility/RNA, we are left with only 6 promoter-gene pairs: Cldn4, Helb, Arg1, Sec16b, Ap1m2, Slc40a1. These genes do not seem to play special roles in Mesoderm development.

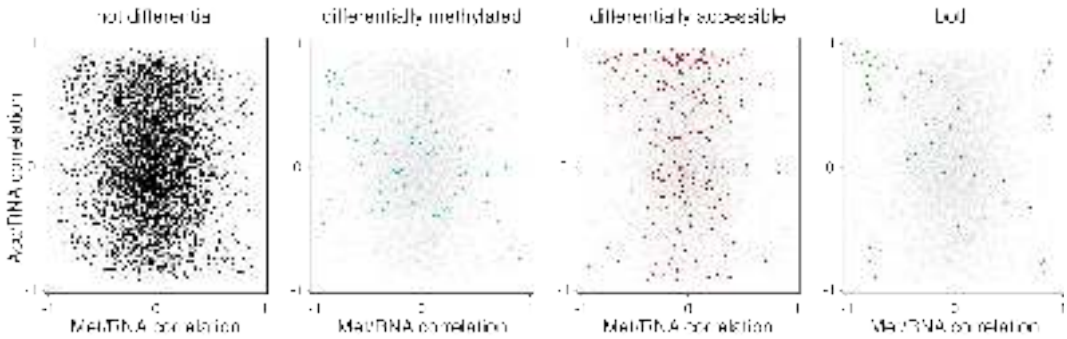


Figure 3.32 | Correlation of promoter methylation/accessibility and gene expression during Mesoderm formation. Scatterplot of Pearson-correlation between promoter methylation and gene expression time series (*x-axis*) and Pearson-correlation of promoter accessibility and gene expression time series (*y-axis*). The columns correspond to classes of promoters that GPmeth identified as differentially methylated/accessible. *Figure generated by Max Frank.*

To investigate these regions in more detail, I visualized the model outputs of all three modalities in Figure 3.33. A common theme of these regions is that the methylation change subregions identified by GPmeth (shaded grey regions) are only close to the center of the regions that harbor the TSS in one case. Conversely, the accessibility changes are close to the TSS in four out of six cases. Another interesting observation is that the subregions of differential methylation and differential accessibility do not always overlap. All this could suggest that it is not the concerted regulation of the TSS by methylation and accessibility that produces changes in gene expression. Rather, there might be regulatory regions such as enhancers that are in close proximity to the genes TSS that regulate its expression. Notably, this observation was only possible with the high genomic resolution of GPmeth and would have been missed by techniques that average signal across genomic windows.

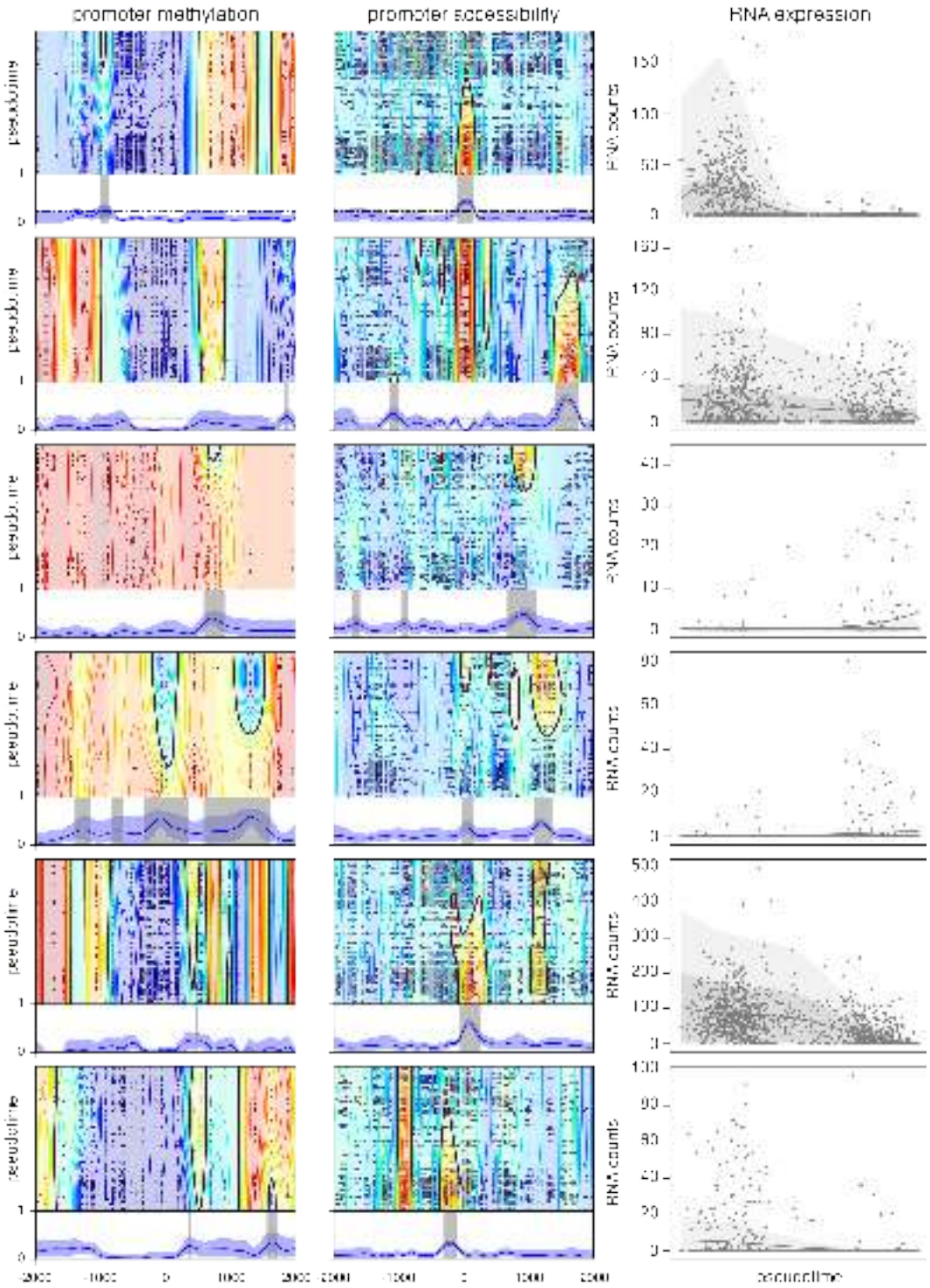


Figure 3.33 | Model predictions of promoters with potential gene regulation capabilities. The *left* and *center* columns depicts GPmeth predictions for DNA methylation and accessibility of promoters, respectively. The x-axis of the GPmeth plots depicts the genomic position, with 0 corresponding to the transcription start site. The scatterplot depicts the input data to the model measured by scNMT-seq, where blue indicates unmethylated sites and red indicates methylated sites. The contours correspond to the posterior mean prediction of the methylation rate ρ by the GPmeth model. Underneath the scatterplot, the blue line indicates the maximum methylation rate change over pseudotime (MMRC) of every genomic location predicted by the model. The blue-shaded regions indicate the 95% confidence interval around that prediction. Grey-shaded areas span genomic regions where the predicted MMRC is 0.3. The *right* column depicts the GPcounts model of RNA expression of the corresponding gene. The x-axis represents pseudotime, and the y-axis are log-scaled counts of RNA expression. Every grey point is a measurement in a cell. The grey line is the mean posterior prediction of the GPcounts model, and the dark and light grey shaded areas represent the 68% and 95% confidence interval, respectively. *Figure generated by Max Frank.*

However, when looking at just the connection between promoter accessibility and gene expression, there is a larger number of positively correlated gene-promoter pairs. Furthermore, the subregions identified by GPmeth are tightly distributed around the TSS of the gene. This distribution gets even narrower when filtering promoter-gene pairs where accessibility and RNA expression are highly correlated (Fig 3.34). In total, 94 gene-promoter pairs were significantly differentially expressed and accessible and had a Pearson-correlation coefficient of >0.7. Of these, 64 gene-promoter pairs had refined accessibility subregions within a 500bp window around the TSS. Therefore, there is a small subset of genes where promoter accessibility likely influences gene expression.

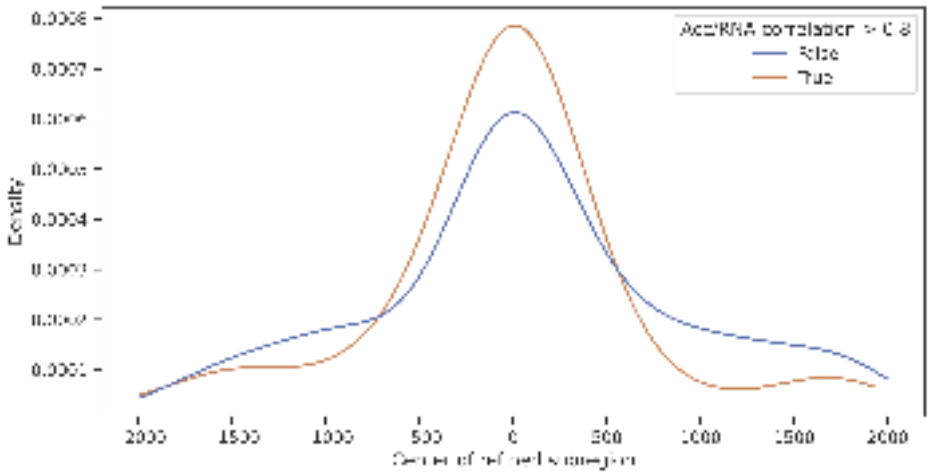


Figure 3.34 | Distance of differentially accessible subregions to transcription start sites. Kernel density estimate plots show the distribution of distances of the center of subregions that GPmeth identified as differentially accessible. The color indicates whether the temporal accessibility rate change has a Pearson-correlation coefficient with RNA expression of 0.7 or greater. *Figure generated by Max Frank.*

3.3.6.2 Integrative analysis of Enhancer regions

As discussed in Section 3.3.2, Enhancers are the main epigenetic drivers during gastrulation. I, therefore, wanted to investigate the relationships between enhancer methylation, enhancer accessibility, and gene expression. This analysis is somewhat complicated by the fact that there is no simple mapping between enhancers and genes. In this analysis, I used the simple approach to map enhancers to genes based on genomic distance. For this, I calculated the distance of the center of each H3K27ac enhancer window to the TSS of each protein-coding gene and paired each enhancer with the gene that has the smallest distance. Note that this approach produces a one-to-many mapping where one gene can be connected to multiple enhancers.

I then compared the three different modalities for each gene-enhancer pair. For enhancers, 663 regions were both significantly differentially methylated and significantly differentially accessible. This is about a third of the 2242 differentially methylated and

1980 differentially accessible enhancers in total. The overlap between both methylation/accessibility and RNA expression was similarly low, with 751 and 675 enhancers, respectively, and only 214 enhancers being significantly differential in all three modalities. For these overlaps, one has to keep in mind that they are dependent on the established gene-enhancer links, which likely contain false positives. Furthermore, as in the above Section, these overlaps are sensitive to significance cutoffs and are investigated in more detail below.

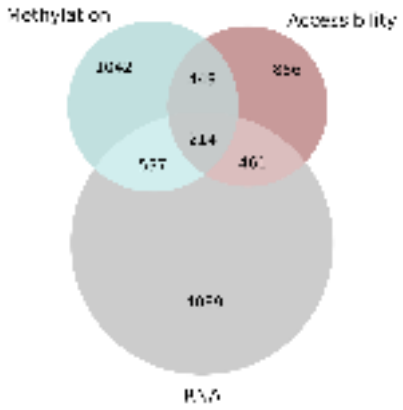


Figure 3.35 | Venn Diagram of differentially regulated genes and nearby enhancers. Number of significant enhancers/genes found by GPmeth/GPcounts, respectively. The cutoff for significance with GPmeth used here was FDR<0.1 and MMRC > 0.3. For GPcounts, significant differential gene expression was defined q-value < 0.1. *Figure generated by Max Frank.*

Next I looked at the magnitude of change in each pair of modalities (Fig 3.36. There was no visible correlation between the MMRC estimates of enhancer methylation/accessibility and RNA expression of the closest gene (Fig 3.36, *left and center panel*). Conversely, enhancer methylation MMRC and accessibility MMRC seemed to be linked (Fig 3.36, *right panel*). Interestingly, there are very few differentially methylated enhancers (*blue points*) where the accessibility MMRC estimate is close to zero, but there are quite some differentially accessible enhancers (*yellow points*) with zero methylation change.

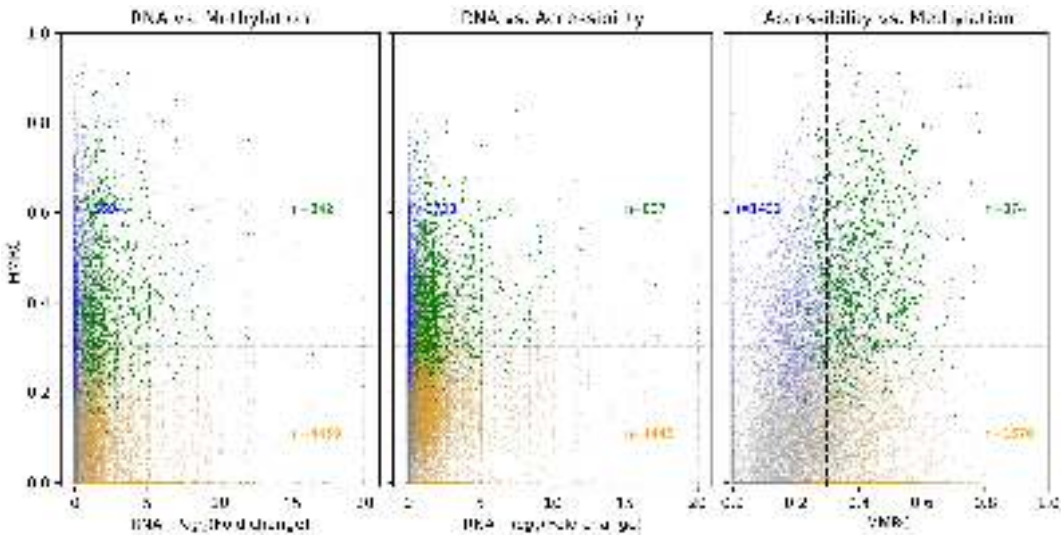


Figure 3.36 | Effect sizes of differential regulation of gene-promoter pairs. Scatterplots show the pairwise comparison of promoter methylation/accessibility and gene expression change magnitudes during Mesoderm development. The *left panel* shows -log_{10} RNA expression fold change (*x-axis*) versus the MMRC estimate of promoter methylation (*y-axis*). The *center panel* shows the -log_{10} RNA expression fold change (*x-axis*) and MMRC estimate of promoter accessibility on the *y-axis*. The *right panel* plots accessibility MMRC (*x-axis*) versus methylation MMRC (*y-axis*). *Yellow* dots points significant changes (FDR<0.1) in the modality displayed on the x-axis, *blue* points indicate significant changes of the modality on the y-axis, and *green* points indicate significant changes in both. *Figure generated by Max Frank.*

To see if the link connection between MMRC changes is due to the correlation of methylation and accessibility over the course of Mesoderm development, I extracted the time series of refined subregions identified by GPmeth as in the previous Section. Similar to promoter dynamics, methylation and accessibility are highly inversely correlated when they are labeled significant by GPmeth (Fig 3.37, *green line*). Furthermore, enhancers that are significantly differentially methylated but not significantly differentially accessible show a similar inverse correlation (*turquoise line*), and enhancers that are significantly differentially accessible but not significantly differentially methylated are bimodally distributed with peaks for no correlation and inverse correlation (*dark red line*). This hints at the fact that some regions that were not identified as significant by GPmeth are still jointly regulated by DNA methylation and chromatin accessibility, but there were simply too few data points to reach significance.

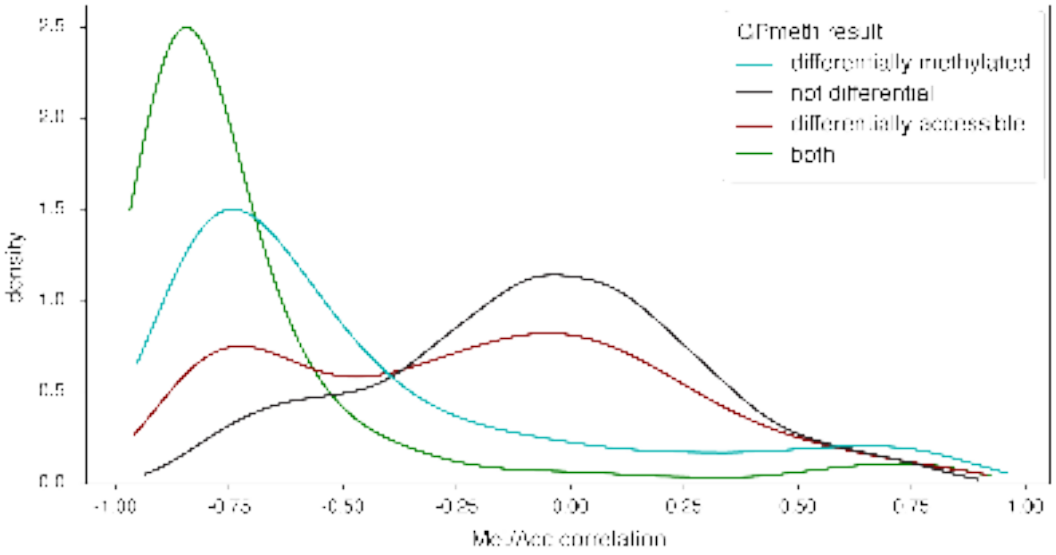


Figure 3.37 | Correlation of enhancer methylation and accessibility during Mesoderm development. Kernel density estimates of the distributions of Pearson correlation between enhancer methylation and enhancer accessibility time series extracted from GPmeth. The color indicates if GPmeth identified the enhancer as significantly methylated/accessible at FDR<0.1 and MMRC>0.3. *Figure generated by Max Frank.*

This is an interesting finding since this could provide a basis for inspecting the uncertainty estimates of the GPmeth model. One advantage of using a fully probabilistic model is that its posterior predictions are not just point estimates but are probability distributions. Therefore, we can use the posterior estimate of the model to determine how confident the model is in its prediction of accessibility rate. With this, I wanted to answer what proportions of enhancer subregions that were differentially methylated were not found as differentially accessible due to a lack of model confidence versus truly unchanging accessibility. To this end, I extracted the genomic coordinates of all 1980 differentially methylated enhancers and obtained posterior predictions for chromatin accessibility within those regions. I calculated the mean, 5%, and 95% confidence interval of those predictions and the Pearson-correlation of those time series with the methylation time series. Figure 3.38, *left panel* shows the mean MMRC estimates for methylation versus accessibility. As expected, the correlation between the modalities is more likely to be negative if both regions have large mean MMRCs. I then categorized the accessibility MMRC estimates into three groups: regions with a 95% MMRC CI smaller than 0.3 (i.e., regions where the model is confident that there are no large changes in accessibility), regions where the 5% MMRC CI is larger than 0.3 (i.e. regions where the model is confident that changes in accessibility are large) and those that fall between those criteria (i.e., regions where the model cannot confidently say whether the accessibility rate is below or above 0.3). This classification gives a lower and an upper bound for the potential number of enhancers that are co-regulated by methylation and accessibility.

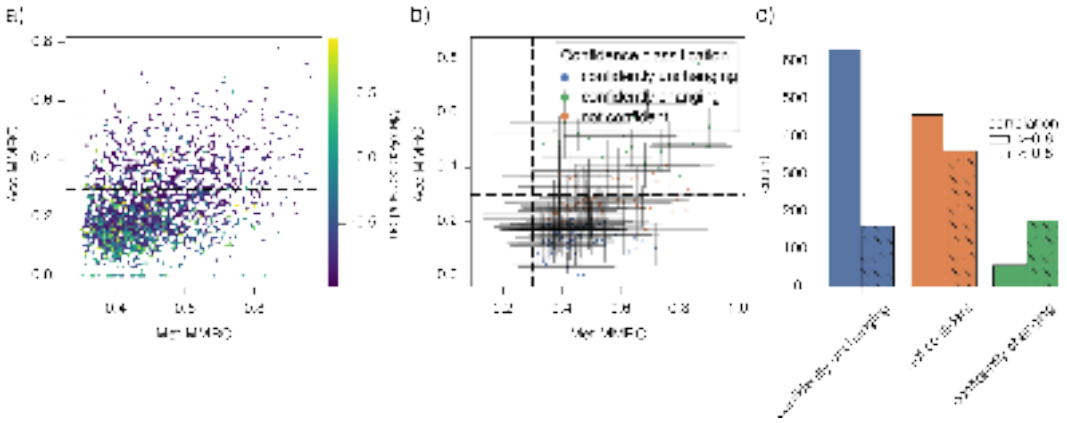


Figure 3.38 | Correlation of enhancer methylation/accessibility with uncertainty estimates. The *left and center* panels show a scatterplot of MMRC estimates of enhancer methylation (*x-axis*) and enhancer accessibility (*y-axis*). In the *left* panel, points are colored by Pearson-correlation values between methylation and accessibility time series. The *center* panel shows a random subset of 100 enhancer subregions with error bars representing the 5% and 95% quantile of MMRC predictions of the model. Points are colored by whether the 95% quantile of accessibility MMRC predictions is smaller than 0.3 ("Confidently unchanging"), or the 5% quantile is larger than 0.3 ("Confidently changing"), or if the error bars span the 0.3 cutoff ("Not confident"). The *right* panel depicts the number of enhancer subregions that fall into each confidence category. *Figure generated by Max Frank.*

Concretely, there are 178 regions where GPmeth is confident about differential accessibility, and the correlation coefficient is smaller than -0.6. This corresponds to just 9 % of all differentially methylated enhancers and is the lower bound of co-regulated regions. There are an additional 366 regions with a high inverse correlation between accessibility and methylation where GPmeths 95% CI is above 0.3. Therefore, the upper bound of co-regulated enhancers is 544, which is 27% of all differentially methylated enhancers. Note that a classical test would have resulted in point estimates for this proportion that would have likely fallen somewhere in the above-mentioned range.

Next, I wanted to investigate the relationship between enhancer epigenetics and the expression of the closest gene. I, therefore, calculated the correlations of enhancer methylation/accessibility time series to RNA time series extracted from the GP-counts model of linked genes. Figure 3.39 shows the correlations for enhancer regions annotated by GPmeth. In contrast to promoters, this analysis showed a smaller enrichment of expected correlations for enhancers that were marked as significant by GPmeth compared to the background of non-significant enhancers. Interestingly, enhancer accessibility and methylation seem to be both positively and negatively correlated with gene expression to a similar extent. This probably hints at the fact that predicting gene regulation through enhancers epigenetics requires more complex models than linking genes to their closest enhancers. This is to be expected since it is known that gene expression can be influenced through the combinatorial effects of multiple enhancers.

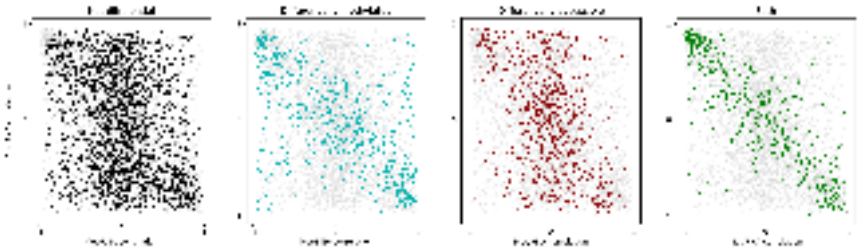


Figure 3.39 | Correlation of enhancer methylation/accessibility and gene expression during Mesoderm formation. Scatterplot of Pearson-correlation between enhancer methylation and gene expression time series (*x-axis*) and Pearson-correlation of enhancer accessibility and gene expression time series (*y-axis*). The columns correspond to classes of promoters that GPmeth identified as differentially methylated/accessible. *Figure generated by Max Frank.*

3.3.6.3 Temporal ordering of methylation and accessibility

One of the key questions in establishing models of gene regulation is whether observed correlations between different regulatory events are causally related. An important tool in establishing causality is the observation of temporal shifts between events. For example, if an enhancer is observed to be demethylated before it becomes accessible, it is impossible for the accessibility change to be the cause of the methylation rate change. Since GPmeth models should predict methylation and accessibility rate changes with high temporal resolution, I decided to investigate if there are detectable temporal shifts between the time series of the two modalities.

To this end, I started with all enhancer subregions that GPmeth identified as differentially methylated (FDR < 0.1, MMRC > 0.3). I then extracted the accessibility predictions of GPmeth at the same subregions and filtered time series that had at least a moderate inverse Pearson correlation of -0.6 or lower. This resulted in pairs of predicted time series for 429 enhancer subregions. To find temporal shifts, I used a modified measure of cross-correlation that is used to detect lag between time series in the signal processing field (Rabiner and Gold, 1975). Briefly, the idea is to slide two time series that should be aligned along each other's temporal axis while calculating the correlation between the signals at every shift position. If the correlation value peaks at a certain shift position, then it is likely that the two time series are linearly shifted with a lag corresponding to the shift position. Because the expectation for methylation and accessibility rates is inversely correlated, the lag can be determined by finding the highest inverse correlation value. Figure 3.40, (*left panel*) shows the Pearson correlation values at different temporal shifts. As expected, on average, the correlation between methylation and accessibility becomes weaker with large shifts. Interestingly, there seems to be a small but consistent positive delay in the accessibility time series. However, the distribution of calculated delays is quite broad (*right panel*), meaning that some time series are also shifted in the other direction. The mode of the delay distribution is at an accessibility delay of 0.1, which is 10% of the complete

pseudotime trajectory. Since embryos in this dataset were sequenced over the course of E6.5 to E7.5 this could roughly be translated to 5h of real time.

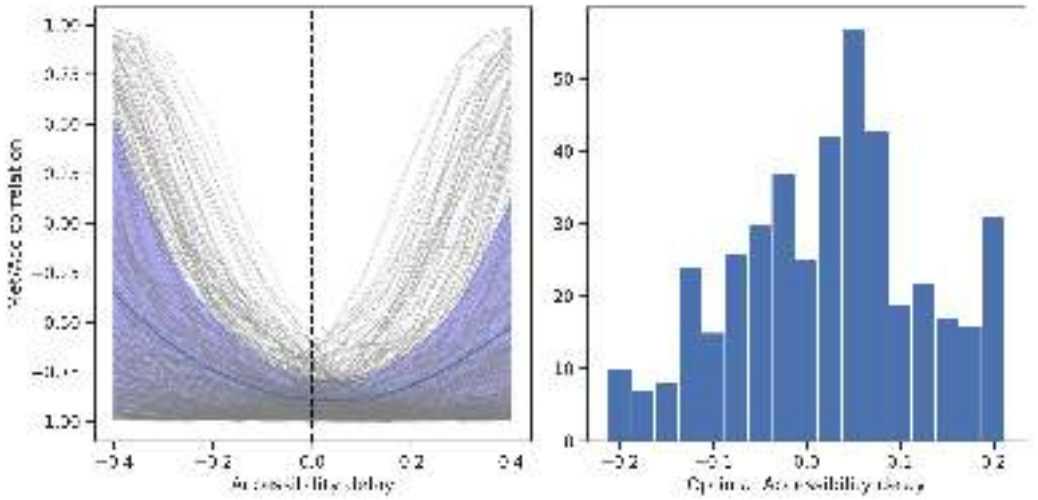


Figure 3.40 | Time delay of enhancer accessibility compared to methylation during Mesoderm formation. The *left* panel shows the Pearson-correlation between methylation and accessibility time series at different temporal shifts for differentially methylated enhancers (*grey lines*, n=429). Delay is given as a fraction of the total pseudo-timespan from E6.5 to E7.5. The blue line represents the average of all individual profiles with shaded regions indicating one standard deviation from the mean. The *right* histogram shows at which time delay the strongest inverse Pearson-correlation was observed for each enhancer. *Figure generated by Max Frank.*

Of course, this analysis is strongly dependent on the number of measured cells and the frequency at which they were sequenced. Because this dataset only consists of two sequencing time points, these results should be taken as preliminary.

3.4 Epigenomic regulation of other lineages

The majority of the results Section of this theis focussed on the Mesoderm lineage formation during mouse gastrulation. This is because this lineage had substantially more cells assigned by trajectory inference. Figure 3.41 shows the pseudotemporal assignment of each cell for the four lineages. Note that there are few cells that map to the late stages of Gut and Notochord development and the non-uniform distribution of cells. Note also that pseudotime estimation resulted in a substantially shorter total timespan for the Ectoderm lineage, a result of the high transcriptional similarity of ectoderm cells to embryonic stem cells.

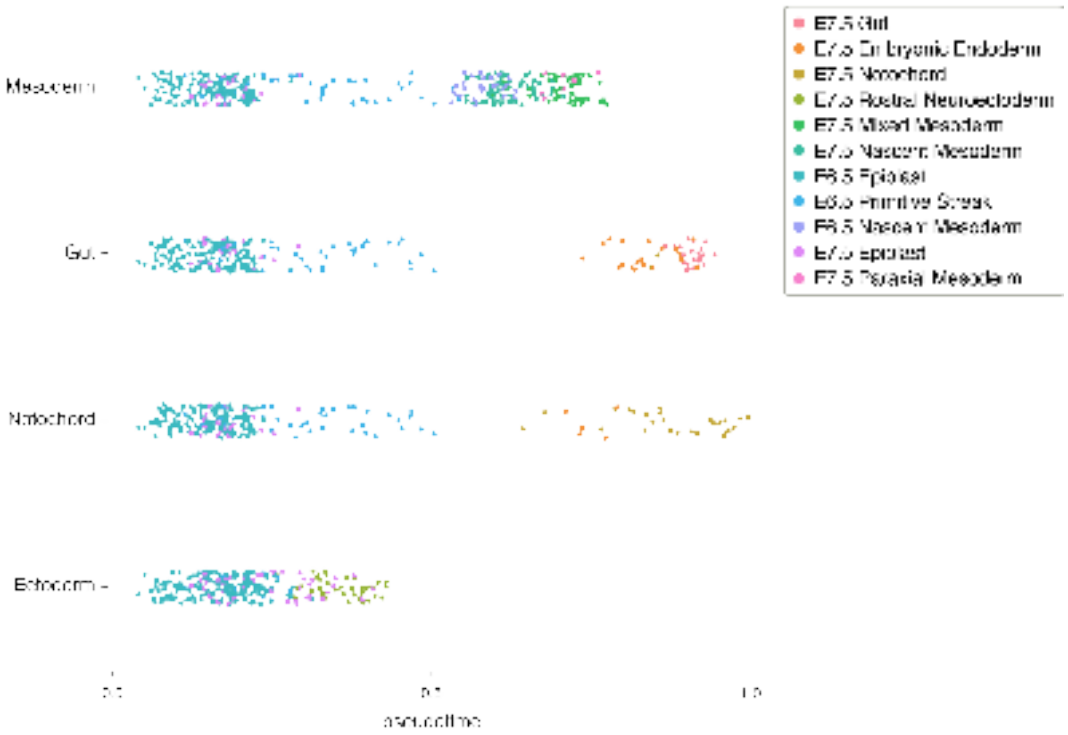


Figure 3.41 | Pseudotime estimates for all four lineages. Jitterplot shows the pseudotime assignment for each cell, in the four identified lineage trajectories. Colors indicate cell-types as assigned by the mapping to a larger single-cell transcriptomic reference atlas (Pijuan-Sala *et al.*, 2019). *Figure generated by Max Frank.*

Despite the fewer cells in lineages other than Mesoderm, GPmeth identified differentially methylated- (Fig 3.42) and accessible (Fig 3.43) enhancers and promoters. As in the Mesoderm lineage, the Notochord and Gut lineages mainly exhibit methylation and accessibility changes in enhancer regions, as evidenced by fewer differentially accessible promoters. Conversely, in the Ectoderm lineage, there are almost no differentially methylated or accessible promoters or enhancers. This is in line with the notion that cells of this lineage are epigenetically primed earlier in development (Muñoz-Sanjuán and Brivanlou, 2002; Argelaguet *et al.*, 2018a).

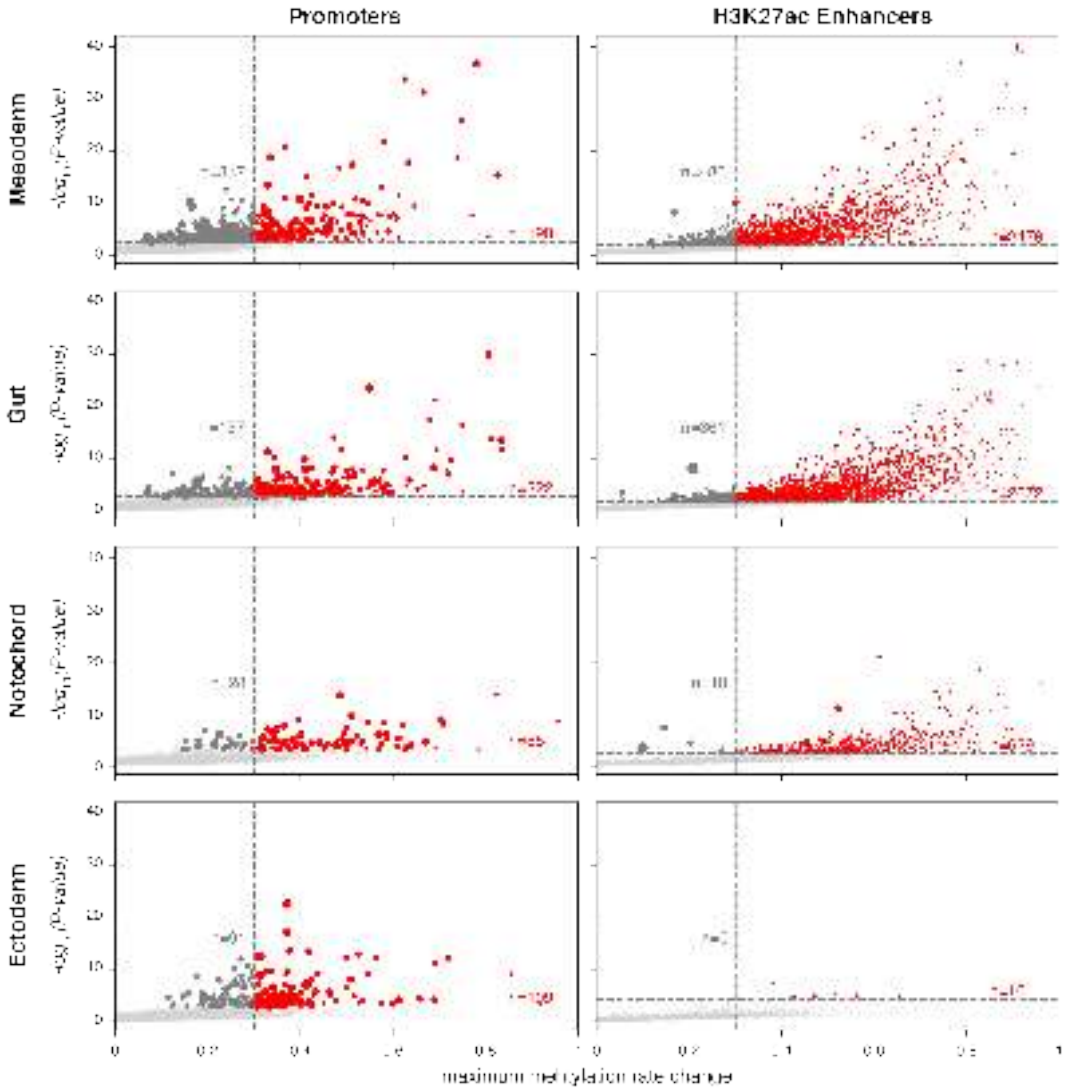


Figure 3.42 | Differential methylation of all lineages during gastrulation. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) of promoter (*left column*) and enhancer (*right column*) methylation during Mesoderm development. Rows correspond to different lineages. The horizontal dashed line corresponds to a significance cutoff of FDR<0.1 after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. Red dots mark differentially accessible regions. *Figure generated by Max Frank.*

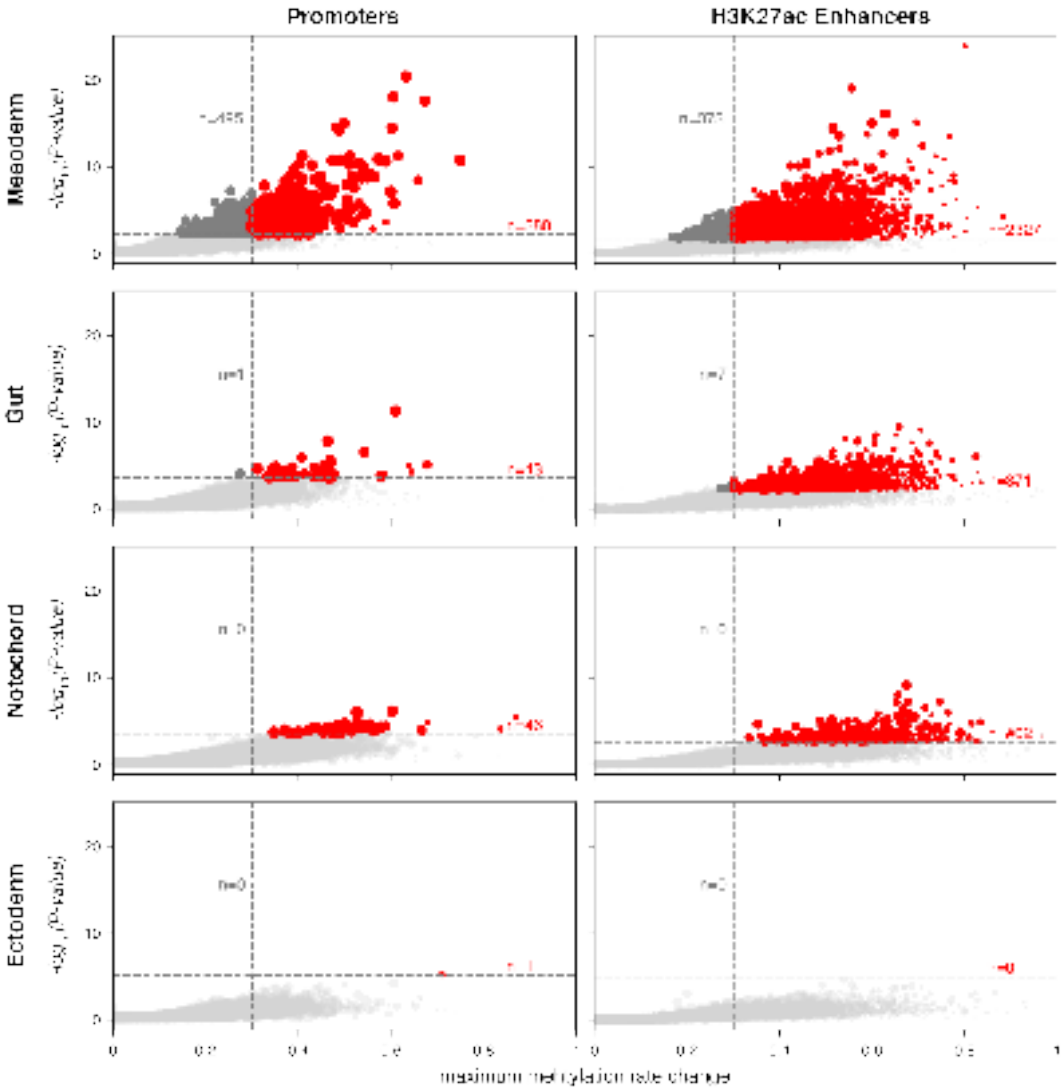


Figure 3.43 | Differential accessibility of all lineages during gastrulation. Maximum methylation rate change (MMRC) on the x-axis vs. significance on the y-axis (GPmeth -log10 p-value) of promoter (*left column*) and enhancer (*right column*) accessibility during Mesoderm development. Rows correspond to different lineages. The horizontal dashed line corresponds to a significance cutoff of FDR<0.1 after BH-adjustment for multiple testing. The vertical dashed line represents an MMRC cutoff of 0.3. Red dots mark differentially accessible regions. *Figure generated by Max Frank.*

To investigate enhancer regulation for different lineages further, I used the sets of germ-layer-specific enhancer annotations, i.e., those enhancers that are derived from ChIP-seq peaks exclusively present in one of the differentiated germ layers (see Section 3.2.2).

First, I assessed the average methylation and accessibility profiles of these enhancers over the course of Ectoderm development. As can be seen by averaged methylation and accessibility profiles in Figure 3.44, Ectoderm-specific enhancers are demethylated and highly accessible throughout Ectoderm lineage development as expected. In

contrast, lineage-specific enhancers for Mesoderm and Endoderm tissues stay highly methylated and inaccessible.

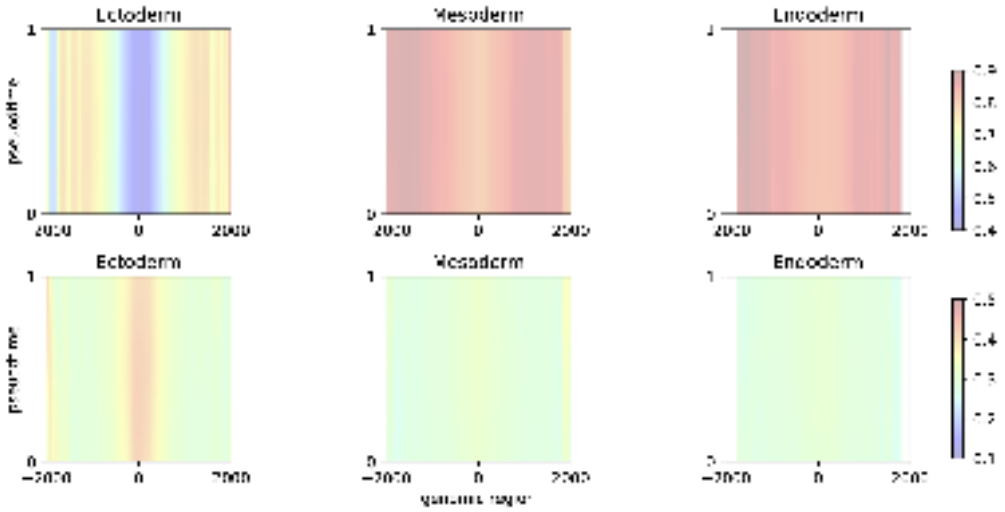


Figure 3.44 | Averaged methylation rate profiles for lineage-specific enhancers during Ectoderm development. The heatmaps represent the GPmeth posterior mean predictions, averaged across lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model for each region in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime and taking the averages of the aligned grids. The *top row* are averaged methylation profiles, the bottom row are averaged accessibility profiles. *Figure generated by Max Frank.*

Next, I investigated the enhancer dynamics of the Gut and Notochord lineage. The detailed Figures of this analysis can be found in Section A.2 and A.3. Both lineages showed increases in accessibility and decreases in methylation of Endoderm-specific enhancers (Fig A.8, A.9, A.14 and A.15). Interestingly, there were also a number of Mesoderm-specific enhancers that exhibited the same temporal dynamics. This could indicate that some Mesoderm-specific enhancers could be active in the earlier phases of Gut and Notochord development.

Next, I assessed the overlaps between all enhancers that were found to be differentially methylated by GPmeth to see if most regulatory regions are specific for only one lineage or if there are enhancers that change in methylation rate in multiple lineages. For simplicity, I excluded the Ectoderm lineage from this analysis because of the low number of differential regions. Figure 3.45) shows that many enhancers change in two or even all three of the lineages. This is not unexpected since differentiation, in general, requires the downregulation of pluripotency genes, which would be a requirement for all three lineages.

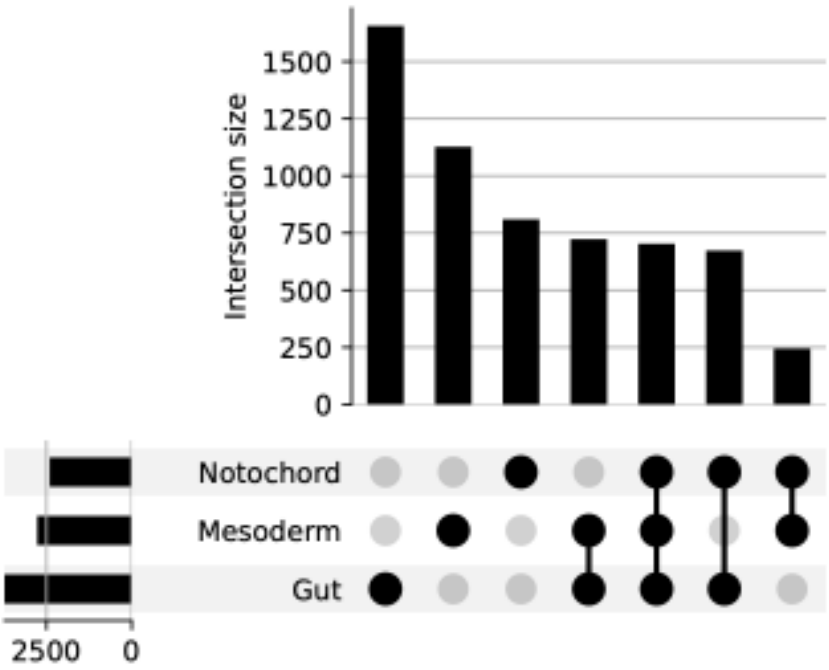


Figure 3.45 | Overlaps in differentially methylated enhancers between lineages. UpSet plot of the number of differentially methylated enhancers that are shared and exclusive to the Notochord, Mesoderm, and Gut lineage. The height of the bar indicates the number in each group of overlapping sets indicated below. The horizontal bars indicate the total number of significant regions (FDR <0.1, MMRC > 0.3) per lineage. *Figure generated by Max Frank.*

Differentially accessible enhancers exhibit similar behavior, with many regions being detected in multiple lineages (Fig 3.46).

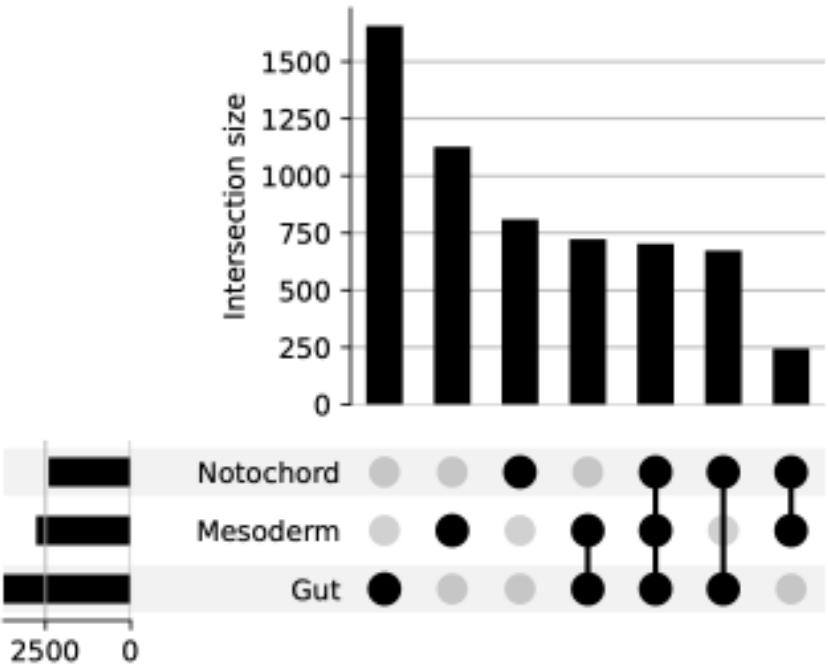


Figure 3.46 | Overlaps in differentially accessible enhancers between lineages. UpSet plot of the number of differentially accessible enhancers that are shared and exclusive to the Notochord, Mesoderm, and Gut lineage. The height of the bar indicates the number in each group of overlapping sets indicated below. The horizontal bars indicate the total number of significant regions (FDR <0.1, MMRC > 0.3) per lineage. *Figure generated by Max Frank.*

4 | Discussion

The genome, the complete set of an organism's DNA, remains largely constant across all cells of an organism and throughout its lifetime, serving as a blueprint for its biological functions. In contrast, the epigenome, which encompasses modifications such as DNA methylation and chromatin accessibility, is highly dynamic and varies significantly across different tissues and stages of development. These epigenetic marks are crucial for regulating gene expression, influencing cell type specification, and guiding cell fate decisions, underscoring the epigenome's pivotal role in organismal development and cellular differentiation.

Advancements in epigenetic profiling methods, such as ATAC-seq (Buenrostro *et al.*, 2013) for chromatin accessibility and bisulfite sequencing (Frommer *et al.*, 1992) for DNA methylation, have significantly enhanced our ability to probe the epigenetic landscape of cells. These technologies generate comprehensive data on how the epigenome is organized and how it changes in different cellular contexts, providing insights into the regulatory mechanisms that underpin gene expression and cell identity.

However, the analysis of epigenetic data, especially from single-cell assays, poses considerable challenges. Bulk analysis techniques offer a deep view of epigenetic changes across populations of cells but lack the ability to capture cell-to-cell variability and are ill suited to capture continuous changes over time. Single-cell epigenetic assays, on the other hand, reveal this heterogeneity but come with technical limitations of reduced coverage due to low genomic input material. This makes it difficult to detect subtle epigenetic changes and to study continuous biological processes, such as developmental trajectories, at reasonable costs.

Current analytical tools, including statistical tests and computational models, developed for bulk assays, are often not directly applicable or sufficiently powerful to model single-cell epigenetic data. They typically focus on identifying differences between discrete, predefined cell populations rather than capturing continuous changes across developmental pathways. This highlights the need for novel analytical approaches that can leverage the sparse and heterogeneous nature of single-cell epigenetic data.

In this Thesis, I described a strategy and modeling framework (GPmeth) to study DNA methylation and chromatin accessibility in continuous biological processes. This strategy uses the power of single-cell multimodal assays to measure transcriptomic and epigenomic modalities in the same cell. Single-cell transcriptomic analyses are well established and are able to reconstruct temporal dynamics of developmental processes from single experiments through pseudotime reconstruction. With each cell mapped onto the appropriate spot in the developmental trajectory, GPmeth can then be used to model the rate of DNA methylation or chromatin accessibility measured in parallel. Because these measurements are typically sparse, this modeling requires tailored methods. Two key features of GPmeth allow it to combat the sparsity of input data. First, it uses the pseudotemporal positions of cells as a continuous variable to share information between cells that are close in time without placing explicit assumptions on the type of temporal dynamics (such as linearity). Second, it models methylation rate at base-pair resolution while still sharing information between proximal genomic measurements. This allows the detection of differentially methylated region boundaries within a larger genomic window.

4.1 Model benchmarking and validation

I validated the GPmeth model on synthetic data that was designed to mimic scNMT-seq measurements (Section 2.2). This revealed the theoretical benefits and limitations of this model.

The GPmeth model parametrized with an *RBFRBF* kernel consistently performed close to the data-generating model in terms of correctly identifying differentially methylated regions (Section 2.2.2, Fig 2.14). Furthermore, it clearly outperformed a model that averages methylation measurements within a genomic window. As expected, the performance difference increased when the subwindow of differential methylation was decreased. When comparing GPmeth to Fisher's exact test (Fisher, 1922) and scMET (Kapourani *et al.*, 2021), there was a clear increase in statistical power that was owed to the addition of the genomic kernel and the continuous modeling of pseudotime. In practice, this means that GPmeth can detect more subtle changes in methylation/accessibility at the same FDR threshold. Notably, this could mean that GPmeth can better detect differentially methylated/accessible regions (DMRs/DMAs) that either have a smaller magnitude of methylation rate change or contain fewer CpG/GpC sites. I also used the simulation experiments to determine the theoretical limitations on what types of regions can be detected with GPmeth. GPmeth statistical power depends mainly on the following variables: the number of cells assayed, the number of CpG/GpC sites that are differentially methylated, and the magnitude of methylation rate change or maximum methylation rate change (MMRC). In my simulations, I kept the number of cells fixed to 300, which can be expected for a typical scNMT-seq experiment. I then varied the size of the subwindow with differential methylation and the MMRC to see where the detection limits lie

(Fig 2.12 and 2.13). Since CpG site density in the genome about ten times lower than that of GpC sites, the detection of endogenous methylation differences is more challenging compared to changes in chromatin accessibility. Regions with an MMRC of 0.7 or larger could be faithfully detected with affected windows of 500bp or smaller. With affected windows larger than 1000bp, GPmeth was able to detect rate changes as low as 0.5. For GpC methylation, MMRC changes as small as 0.2 were detected in affected windows larger than 500bp. With MMRC changes > 0.7, even subregions smaller than 100bp were detected. One limitation of this test is that it did not control for where along the pseudotemporal trajectory methylation rate changes occurred. This can influence the detection limit since it determines how many cells will be affected by differential methylation.

The fact that GPmeth performance was close to optimal (i.e., close to the performance of the generative model used for producing the data), yet still was not able to detect very small MMRC, or very short subregions, highlights the difficulty of this testing problem and the importance of the correct aggregation of neighboring genomic loci and cells.

I then compared GPmeth parametrized with an *RBFRBF* kernel to scMET on the real scNMT-seq dataset of mouse embryonic stem cells during Mesoderm formation (Section 3.3.3.3). It is important to note here that scMET was not designed to accommodate continuous covariates and, thus, is not inherently suited for studying developmental processes. However, I am currently unaware of any other single-cell methods designed to accommodate such an experimental design. Therefore, scMET was the closest possible comparison. Another disadvantage of scMET in this comparison was the relatively large genomic window size that I chose to use. This choice ensured that no differential subregions were missed by the GPmeth model, which is designed to handle larger input windows. scMET works with summary statistics for each cell and genomic window so that large windows will affect those summary statistics negatively. At FDR < 0.1 scMET identified 380 differentially methylated- and 68 differentially accessible enhancer regions. GPmeth identified 1769 differentially methylated and 2647 differentially accessible enhancers at the same FDR (Fig 3.15). While this does not prove that GPmeth has more statistical power, in the absence of any ground truth, many of the downstream results discussed below provide evidence that the regions identified by GPmeth are genuine.

4.2 Investigating mouse gastrulation with GPmeth

After benchmarking and validating the GPmeth model, I applied it to a scNMT-seq dataset of mouse embryonic stem-cells undergoing gastrulation (Argelaguet *et al.*, 2019b). First, I established pseudotime trajectories of lineage formation using unsupervised dimensionality reduction and pseudotime estimation techniques. This revealed four major trajectories, in which pluripotent epiblast cells differentiate

into Mesoderm, Ectoderm, Gut, and Notochord cells from embryonic day (E)6.5 to E7.5. Since the Mesoderm trajectory had the largest number of cells, many analyses in this thesis were focussing on this trajectory path. After cells were assigned pseudotime values and lineage identity, I applied GPmeth to find differentially methylated enhancers and promoters over the course of gastrulation. An immediate finding was the low number of differentially methylated or accessible promoters (190 and 380 promoters, respectively, with FDR<0.1 and MMRC>0.3 out of 18,347 tested regions). Conversely, out of 17,386 enhancer regions marked by H3K27ac, 2478 were identified as differentially methylated and 2327 as differentially accessible by GPmeth with the same criteria.

4.2.1 Promoter epigenetics

The low number of differentially methylated and accessible promoters strengthens the notion that enhancer elements are the main drivers of embryonic lineage specification, which has been observed before with alternative techniques (Cusanovich *et al.*, 2018; Zhang *et al.*, 2018) and with MOFA analysis of the same dataset (Argelaguet *et al.*, 2019b). Investigating the GPmeth model outputs of promoters in more detail revealed that the identified subregions were clustering around the TSS for accessibility but spread throughout the 4kb input window for methylation (Fig 3.18 and 3.17). This is hinting to the fact that differentially methylated promoter subregions might be different types of regulatory elements that happen to be in close proximity to the TSS of the gene. I also used the GPmeth output to investigate the relationship between promoter methylation, promoter accessibility, and gene expression. Correlation between all three modalities was only observed for a minute subset of 6 genes. Closer inspection revealed again that differentially methylated regions were not overlapping with the differential accessibility signal close to the TSS of the gene (Fig 3.33). Therefore, there seems to be no concerted regulatory mechanism that changes DNA methylation and chromatin accessibility of promoters to induce or repress gene expression. However, chromatin accessibility itself was correlated with gene expression for a small but meaningful set of 94 genes, mostly with differential accessibility in close proximity to the TSS (Fig 3.34). GO-term analysis of this gene set (data not shown) revealed expected terms such as "organism development", but also surprising enrichment for terms related to placenta formation. Furthermore, the majority of these genes decrease in gene expression and promoter accessibility during Mesoderm formation. This could be an interesting avenue to further explore if promoter accessibility could be important for the development of extraembryonic tissue arising from epiblast cells that contribute to the embryonic part of the placenta (Panja and Paria, 2021).

4.2.2 Enhancer epigenetics

Since there was a high number of H3K27ac enhancer elements identified by GPmeth both as differentially methylated (n=2478) as well as differentially accessible (n=2327),

I wanted to investigate whether methylation and accessibility are dependent on one another within these elements. This has been investigated recently in cultured mouse embryonic stem cells using single-molecule footprinting (Kreibich *et al.*, 2023; Krebs *et al.*, 2017) and produced the result that in a homogeneous population of cells only 3% of active enhancers showed a dependency of chromatin accessibility on DNA methylation. Enhancer methylation and accessibility dependency has also been investigated in a dynamic system of macrophage differentiation, using bulk technologies (Barnett *et al.*, 2020). This study found little evidence for close temporal relationship between chromatin accessibility changes and DNA methylation.

Here, I am assessing the dependency of chromatin accessibility on DNA methylation in the context of a dynamically changing system that is Mesoderm formation. 663 enhancer regions were both significantly differentially methylated and significantly differentially accessible during Mesoderm formation, which corresponds to about a third of significant enhancers. However, as seen with promoters, this overlap is not necessarily concrete evidence for co-regulation. Therefore, I assessed the correlation between the temporal change profiles of methylation and accessibility rates. Interestingly, this revealed that many differentially methylated regions are inversely correlated with differential accessibility, whereas only a subset of differentially accessible regions shows inverse temporal patterns of methylation (Fig 3.37). This could indicate that during Mesoderm development, enhancer methylation can cause accessibility changes, but the opposite is not necessarily true. To get a confident estimate of the proportion of differentially methylated enhancers that show concordant (i.e., inversely correlated) changes in accessibility, I compared GPmeth predictions at identified refined regions that were differentially methylated. Because of the probabilistic nature of GPmeth, I could separate differentially methylated enhancers into a group with a high degree of certainty of co-variability (178 or 9%) and a group with lower certainty (366 or 18%). The second group consists of enhancers where the data is too sparse to make a clear decision. While these proportions are significantly higher than described by Kreibich *et al.*, 2023, it is important to note that this analysis starts with a subset of differentially methylated enhancers, which differs from the approach by Kreibich *et al.*, 2023 who started from a larger subset of regions with intermediate chromatin accessibility. It is possible that the two sets of results cover different mechanisms of co-regulation through methylation and accessibility and that the regulation mechanism described here is not a direct result of methylation-sensitive TF binding for most enhancers but a mechanism on a longer time scale. Furthermore, these results also differ from the findings by Barnett *et al.*, 2020, where DNA methylation did not change with increases or decreases of chromatin accessibility, at least not within a time-span of hours. While the biological systems studied here and in Barnett *et al.*, 2020 are vastly different, this might be surprising. One important caveat is that the ATAC-me technology used, only captures methylation in at least partially accessible regions. Another important difference is that cells divide rapidly during embryonic development, while macrophages are terminally differentiated and do not divide.

Replication has been shown to be an important factor in DNA methylation changes (Otani *et al.*, 2013).

To gain more information about the possible mechanisms underlying this co-regulation, I investigated if there is a temporal delay in methylation and accessibility changes. Temporal delays are important in studying gene regulation because they provide additional evidence for causality relationships between linked events. I, therefore, investigated if there are linear delays between methylation and accessibility based on cross-correlation (Fig 3.40). Despite the distribution of shifts being relatively broad, there was an average delay of accessibility changes compared to methylation changes of 10% of the pseudotime range, which corresponds roughly to 5h of real-time assuming linear mapping. While these results need further validation, they are in line with methylation causing accessibility changes. However, it is important to note that this evidence is still correlative, requiring experimental follow-up.

Temporal comparison of methylation and accessibility rate trajectories also revealed that there is a temporal ordering in the activation and deactivation of lineage-specific enhancers during gastrulation (see Section 3.3.5). Mesoderm development involves the activation (i.e. demethylation and accessibility increases) of Mesoderm-specific enhancers, but also the inactivation (i.e. methylation increases and accessibility decreases) of Ectoderm-specific enhancers. This is in line with a departure from the default differentiation path of Epiblast cells to Ectoderm cells that is epigenetically programmed as early as E4.5 (Argelaguet *et al.*, 2018a). Analysis of the GPmeth trajectories revealed that Ectoderm-specific enhancers mostly get inactivated before Mesoderm-specific enhancers get activated, indicating that cells first depart from their default fate, before deciding on their Mesodermal or Endodermal trajectory.

4.2.3 GPmeth results as a basis for targeted experiments

It is important to highlight that the results described above provide testable hypotheses for follow-up experiments. For example, the ability of GPmeth to identify the precise subregions where differential methylation occurs can aid in the identification of targets for inducing methylation with genome editing techniques (Yamazaki *et al.*, 2020). This could reveal if the observed correlations between enhancer methylation and accessibility are, in fact, causal.

Furthermore, GPmeth, using scNMT-seq data with genome-wide coverage, could be used as a tool for identifying targets for subsequent targeted versions of the same experiment. Performing NOMe-seq with genome-wide coverage increases sequencing costs. Instead, subregions identified by GPmeth could be targeted by reduced representation bisulfite sequencing techniques (Meissner *et al.*, 2005; Guo *et al.*, 2015) that reduce cost and potentially allow the sequencing of more cells.

4.2.4 Limitations of this study

In this Section, I will discuss the limitations of this study and potential avenues to overcome them. While scNMT allows for unprecedented insight into gene regulatory mechanisms in individual cells, this comes with a high cost associated with this technique. Therefore, it is challenging to scale this technique to tens- or hundreds of thousands of cells. However, as the cost of sequencing decreases over time, I anticipate that more scNMT datasets of a similar scale to the one used here or larger will become available. Assaying more cells would increase the statistical power to detect differentially methylated regions. As I have shown with simulations in Section 2.2.2 the sensitivity of GPmeth is still limited when the regions of differential methylation or accessibility are small. For example, when it comes to TF footprinting, a genomic resolution of 100 bp would be ideal to faithfully detect binding events, which often only contain a handful of GpC sites and often only a single CpG site. This resolution could only be achieved by sequencing more cells or decreasing the coverage of NOMe-seq, which is difficult due to the limitation of input material. Therefore, this study likely missed some important epigenetic regulatory regions due to their size or small magnitude of change.

Another limitation is the non-uniform temporal sampling of cells. Since cells were only collected at two distinct times during the lineage-defining phase of gastrulation (E6.5 and E7.5), most cells were assigned pseudotime values at either end of the spectrum. This could be overcome by simply including one or two more sequencing runs with cells from embryos at the E7.0, E6.75, or E7.25 stages. This would be especially interesting for assessing temporal shifts between modalities in more detail.

Furthermore, most epigenetic changes in this dataset either monotonously increased or monotonously decreased during the assayed time course. One indication of this is the good performance of the GPmeth model parametrized with a linear temporal kernel. It would, therefore, be interesting to include developmental stages that go beyond E7.5. This would capture the early formation of organ structures. For example Pijuan-Sala *et al.*, 2019 included cells up to E8.5, which captured the early formation of the spinal cord, brain, heart, blood, and digestive system. Importantly, this involves the up- and subsequent down-regulation of many genes, which is likely to go hand in hand with non-linear epigenetic changes. Detecting these changes would necessitate a nonparametric model like GPmeth.

In addition, the sampling of the different germ layers was uneven in this experiment, which resulted in a lower number of cells mapping to the endoderm and ectoderm lineage. For this reason, there was an emphasis on Mesoderm development in this thesis, but a larger dataset would enable a more detailed model of the other germ layers. Especially the formation of the ectoderm layer is accompanied by only subtle changes in many regulatory elements since this trajectory is already primed in the pluripotent stage (Argelaguet *et al.*, 2019b).

5 | Future outlook

GPmeth addresses two fundamental technical problems that are still limitations in several single-cell studies. The first is the proper use of continuous covariates in experiments that study cell populations that change their characteristics smoothly across space or time. The second problem, specific to epigenomic studies, is the absence of a fixed set of features that can be measured and used for downstream analysis. The fundamental features of epigenetics are individual bases, but current limitations in experimental setup often preclude the analysis on this level of detail. Furthermore, individual bases are not independent units but are co-varying within regulatory regions. The definition of these regulatory regions is a hard problem, and there is no consensus database that works for all cell types and states. Therefore, I think GPmeth could be used to solve these problems in the following ways.

Firstly, the capability of GPmeth to identify DMRs and DMAs in a data-driven manner could be used to identify these regions genome-wide. In this study, I have started from a set of putative promoter and enhancer regions, but this could be extended to genomic windows covering whole chromosomes. However, this approach should be taken with care since this would increase the multiple testing burden. Two possible avenues could mitigate this problem. The first would be to segment the genome in a manner that separates CpGs or GpCs that are further than 150 bp apart since, beyond that distance, co-variation should be minimal. Secondly, some heuristics can be applied to exclude windows with very low coverage or low total variance in accessibility or methylation. This would allow us to find DMRs and DMAs free of bias that could be introduced by selecting putative regulatory regions based on prior knowledge.

Secondly, there are extensions that could be made to the model allowing for the explicit modeling of trajectory branching dynamics. In this study, I have determined cell-lineage associations based on pseudotime analysis alone and then separately modeled these trajectories. However, it has been shown in GP models for gene-expression data that branching dynamics can be directly included in the model structure (Yang *et al.*, 2016; Boukouvalas *et al.*, 2018b; BinTayyash *et al.*, 2021). This would also allow us to identify when differential methylation/accessibility first occurs in a probabilistic way.

Thirdly, it is possible to include spatial coordinates of cells as an additional or alternative covariate to pseudotime in the GPmeth model. With the emergence of epigenomic spatial single-cell approaches (Thornton *et al.*, 2021; Deng *et al.*, 2022), the relationships between chromatin accessibility and DNA methylation of neighboring cells can be explored. GPs have been a valuable tool for investigating the gene expression profiles that spatial single-cell RNA sequencing assays produce (Svensson *et al.*, 2018; Kats *et al.*, 2021) because of their flexibility in modeling non-linear changes. Therefore, this could be an interesting application for GPmeth.

Lastly, GPmeth was designed for NOMe-seq data which produces a base resolution output. However, the same problem of unclear definitions of regulatory features also applies to other techniques. In ATAC-seq, features are usually defined by detecting peaks of accessibility summed over all cells. Alternatively, cells are sometimes roughly clustered based on the accessibility profile of genome-wide fixed-width bins, followed by peak calling on clustered groups of cells. However, when studying a continuous developmental process, clustering could be a sub-optimal approach (Fig 1.11). Therefore, an interesting avenue would be to apply the same principles used by GPmeth to ATAC-seq datasets. For example, in a multimodal ATAC-seq dataset (Chen *et al.*, 2019; Ma *et al.*, 2020), trajectories could be established using the RNA modality. Then, a GP model could model the distribution of reads directly across pseudotime and the genome dimension, resulting in more precise regions of differential accessibility. This could improve the identification of TF binding sites or the detection of gene-enhancer pairs.

A detailed understanding of the exact genomic position of epigenetic changes, combined with high temporal resolution of these changes over biological processes has the potential to greatly enhance our understanding of how gene regulation informs cell fate. As technologies emerge with the potential to provide this information, models that can make full use of this data will be essential.

6 | Additional Methods

6.1 GPmeth

6.1.1 Model optimization

A gaussian process is trained by optimizing its marginal likelihood with respect to the hyperparameters of the model. The marginal likelihood is given by

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \mathbf{f}) \mathcal{N}(\mathbf{f} \mid \mu, K(\boldsymbol{x}, \boldsymbol{x})) \mathrm{d}\mathbf{f}, \tag{6.1}$$

where μ is the mean rate of the region and $K(\boldsymbol{x}, \boldsymbol{x})$ is the covariance matrix specified by the full kernel function. This likelihood is untraceable in the case of Bernoulli likelihood, so variational approximation is used to compute the evidence lower bound (ELBO) as an approximation.

Models are trained in a two step fashion. First the model with only the genome kernel is trained. Hyperparameters are initialized to sensible values as follows. Genome kernel lengthscales are fixed at 150bp, temporal kernel lengthscales are initialized with 0.21. The temporal kernel lengthscales are bounded between 0.2 and 100 to avoid very small lengthscales that would be biologically implausible. Kernel variances are initialized with 0.3. The model also has a fixed mean function that is set to $\Phi^{-1}(\mu_{rate})$, which is the probit function of the mean methylation/accesibility rate within the modelled region.

After the null model is trained the optimized hyperparameters of the genomic kernel are used to initialize the genomic kernels of the full models. The genomic kernel lengthscales and variances of the full model are not trained further, while the hyperparameters of the product kernel are then optimized for the full model. The model parameters of all trained models are saved in a custom Hdf5 (The HDF Group, 1997) format to be retrievable for downstream applications. The ELBO of all models is also recorded along with key model parameters in tsv format.

6.1.2 Maximum methylation rate change calculation

The maximum methylation rate change (MMRC) of each modeled region is defined as the largest change in methylation rate across pseudotime for each point in the genome. This metric is calculated by producing posterior predictions of the GPmeth model at a grid of evenly spaced location that span the modeled region in the genome and pseudotime dimension. The number of points in this grid can be varied depending on the desired resolution, with a default of 100 points across the genome dimension and 20 points across the pseudotime dimension. Then for each genomic position, the difference of the maximum and the minimum predicted posterior methylation rate is calculated. If only a point estimate is desired, this can be done with the posterior mean of the model. To calculate posterior distributions of the MMRC, the model posterior is sampled (by default n=2,000 samples). For each sample, MMRC values are calculated separately.

6.1.3 Generation of synthetic NOME-seq data

As discussed in Section 2.2, the process of generating synthetic data to assess model performance consisted of three steps:

1. Generation of realistic locations of assayed CpG/GpC sites in a typical NOME-seq experiment
2. Drawing methylation rates ρ from a generative model consisting of a GP with a changewindow kernel
3. Bernoulli sampling from the simulated methylation rate at the assayed CpG/GpC sites

This Section will describe the step of producing methylation rates with the desired properties of differential methylation over time in more detail. The generative model for ρ is

$$\begin{aligned}
&\rho_{alt} \sim \Phi(GP(0, k_{alt}) \\
&k_{alt} = k_{genome} + k_{CW} \\
&k_{CW} = k_{outside} x, x') * (1 - \sigma(x)) * (1 - \sigma(x')) + k_{inside}(x, x') * \sigma(x) * \sigma(x') \\
&\sigma_{x_0,x_1}(x) = \frac{1}{e^{-s(x-x_0)}} * \frac{1}{e^{-s(x-x_1)}}
\end{aligned} \tag{6.2}$$

as described in section 2.2. This model can be constructed with two kernels, $k_{outside}$ and k_{inside} that control the temporal change of ρ outside and inside of the change window $[x_0, x_1]$ respectively. In the simulation the change window was chosen to be placed centrally in the generated region, so that it could be specified with a single width parameter w

$$x_0 = -w/2; x_1 = w/2 \tag{6.3}$$

The choice of k_{inside} and $k_{outside}$ determine the differential methylation in the region. In this simulation I chose the outside of the window to have no differential methylation over time with a constant kernel with low variance and the inside of the window to be a squared exponential kernel with a lengthscale of 0.5, which is half the total pseudotime

$$k_{outside} = 0.000001 \tag{6.4}$$

$$k_{inside}(x, x') = \sigma_f^2 \exp\left(-\frac{(x - x')^2}{0.5}\right) \tag{6.5}$$

To test the performance of the model with different magnitudes of methylation rate change, the goal is to produce a sample that fulfills the criterion of a certain rate change at some point within the window. This corresponds to the MMRC value described in Section 2.1.3. Depending on the desired MMRC, I changed the variance parameter σ_f of k_{inside}. Because sampling from a GP will yield a stochastic result there is no direct relationship between MMRC and σ_f. Rather I chose to set

$$\sigma_f = \begin{cases} 1.3 & \text{if } MMRC > 0.7 \\ 0.7 & \text{if } 0.2 < MMRC < 0.7 \\ 0.1 & \text{if } MMRC < 0.2 \end{cases} \tag{6.6}$$

and then filtered out any samples from the model that did not fulfill the desired MMRC up to a tolerance. In practice this worked well with a tolerance parameter of 0.05. For examples of this model see Figure 2.11.

6.1.4 Model calibration

For model calibration, 18,347 promoter regions and 17,386 H3K27ac enhancer regions were selected. Pseudotimes and methylation values form from the mesoderm lineage (415 cells) scNMT-seq gastrulation dataset were used. For each region, the pseudotime values attached to each cell were randomly permuted five times resulting in a total of 91,735 shuffled promoter and 86,930 shuffled enhancer regions. After permutation these regions are not expected to show significant methylation/accessibility changes over time and thus the LLR values of the models can be used for calibration. All models in Table 2.1 were trained on the permuted regions and LLR values were calculated based on model comparisons (see Section 2.1.2). Modeled regions were grouped, based on the number of observations within the regions, into five bins. This grouping was done separately for promoter and enhancer regions. For each group a χ^2-mixture distribution was fit (see Section 2.2.4). To fit this null distribution, the free parameters of the χ^2-mixture, π, a, d, were estimated maximum likelihood

estimation with a grid-search over the parameters. To increase the robustness of the fit, the lowest 5% and the highest 5% quantile of the LLRs was excluded. The estimation of parameters was performed with a custom function adapted from the *Chi2Mixture* class of the limix package (Lippert *et al.*, 2014). To estimate significance of model comparisons on real regions, each region was matched with the appropriate bin according to it's number of input points. P-values were then calculated as the tail function value of the null distribution at the respective LLR value.

6.1.5 Software availability

GPmeth is an open-source project available at https://github.com/mffrank/gpmeth.

6.2 Additional Methods for Mouse Gastrulation

6.2.1 Definition of enhancer and promoter regions

Bed files with H3K27ac marked regions were obtained from the original publication (Argelaguet *et al.*, 2018a), and can be downloaded from ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation. For GPmeth input the union file *H3K27ac_distal_E7.5_union_intersect12_500.bed* was used. For the analysis of lineage-specific enhancer regions, *H3K27ac_distal_E7.5_Ect_intersect12_500.bed, H3K27ac_distal_E7.5_End_intersect12_500.bed* and *H3K27ac_distal_E7.5_Mes_intersect12_500.bed*, contained the regions that were detected in each of the separate germ-layers Ectoderm, Endoderm and Mesoderm respectively. To define lineage-specific enhancers, the regions were overlapped with the union file and lineage specific enhancers were defined as exact overlaps that only matched to one of the three lineage files.

6.2.2 RNA-seq preprocessing and quality control

Raw count data (see 6.2.10) was analyzed with scanpy (Wolf *et al.*, 2018). Cells with fewer than 4000 or more than 11000 genes expressed genes were removed. Then cells with more than 10% reads mapping to mitochondrial genes were removed, as well as cells with more than 3 million total reads. Then reads counts were corrected for library size and log-transformed.

6.2.3 RNA-seq dimensionality reduction and pseudotime inference

The preprocessed RNA data was mapped to a much larger single-cell atlas, as was done in Argelaguet *et al.*, 2018a, using a mutual nearest-neighbor (Haghverdi *et al.*, 2018) approach. The resulting mapped first principle components were used for further analysis. To exlude batch effects caused by differences between embyos a batch-balanced nearest-neighbor algorithm (bbknn, Polański *et al.*, 2020) was used to calculate a neighborhood graph of cells. This was then used as the input to calculate

diffusion components (Haghverdi *et al.*, 2016). Pseudotime was then calculated using Palantir (Setty *et al.*, 2019). Leiden-clustering of cells was performed based on the diffusion maps, with resolution 0.6, leading to intentional overclustering. These clusters were used to assign cells manually to the four lineages: Mesoderm, Gut, Notochord, Ectoderm. Note that some cells are multiply assigned. For example early Epiplast cells are part of all four lineages.

6.2.4 NOMe-seq data preprocessing

CpG sites (from A-C-G and T-C-G trinucleotides) and GpC sites (G-C-A, G-C-C and G-C-T trinucleotides) for every cell (see 6.2.10), were obtained in tabular Bismark (Krueger and Andrews, 2011) output format. The output for every cell was concatenated and sorted by chromosome and position. This file was then compressed and indexed with tabix (Li, 2011), enabling the fast retrieval positional subsets of the data by the GPmeth model.

6.2.5 Differential gene expression with GPcounts

For each of the four lineages established in Section 6.2.3, differential gene expression was assessed using GPcounts (BinTayyash *et al.*, 2021). Cells belonging to each lineage were filtered and their raw counts and previously calculated pseudotimes were used as an input for a one sample test with negative binomial likelihood. This test computes a test statistic based on the log-likelihood ratio between a dynamic model with an RBF temporal kernel and a static model with a constant temporal kernel. P-values were calculated assuming that the null-distribution of LLRs follows a χ^2 distribution.

6.2.6 Detailed GPmeth workflow

Figure 6.1 shows the necessary steps to train GPmeth models. After preprocessing and sorting methylation call files (Section 6.2.4), a bed formatted region file is used to generate inputs for GPmeth. For each region, methylation or accessibility values are extracted from the indexed file on disk. Cells are then matched with pseudotime values (Section 6.2.3). This results in each observation of a CpG/GpC methylation event having a genomic and peseudotemporal coordinate. This input is used to train the GPmeth models (Section 6.1.1) and calculate LLR values. LLR values are transformed into calibrated statistical estimates through the distributions of permuted null regions (Section 6.1.4).

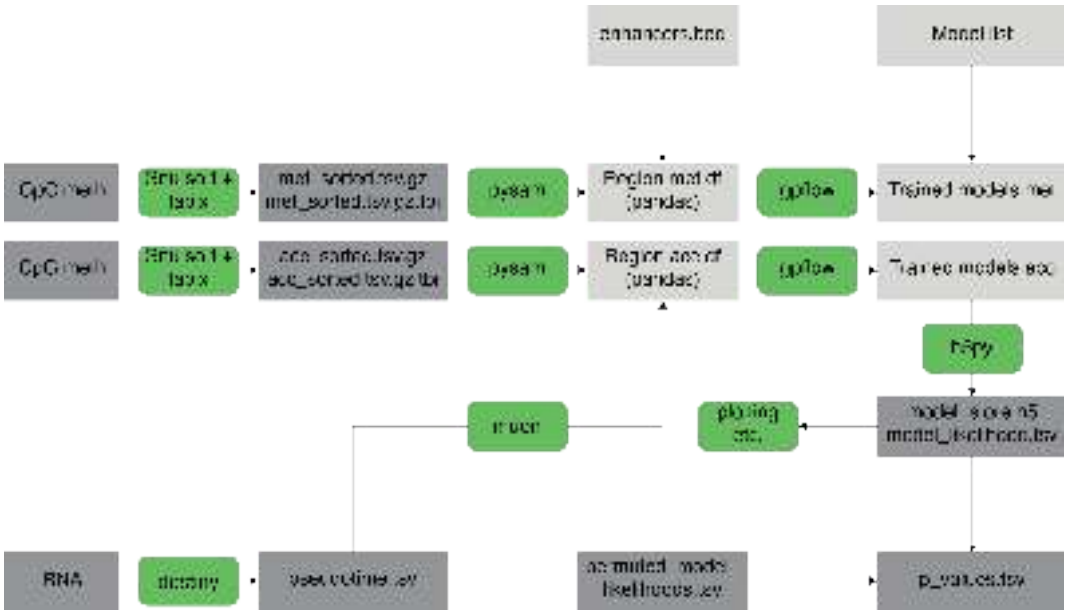


Figure 6.1 | Overview of the computational workflow of fitting the GPmeth model. *Figure generated by Max Frank.*

6.2.7 Refinement of differentially methylated regions

Refinement of differentially methylated/accessible regions was done based on thresholding the MMRC (Section 6.1.2) at each genomic position. MMRCs were calculated at 100 equally spaced points across the genomic input window. Then, neighboring points exceeding an MMRC of 0.3 were merged to define subregions. The boundaries of the subregions were defined as the leftmost and rightmost position for each region.

6.2.8 Calculation of methylation rate time-series

Time-series were based on model predictions on a grid of 100 times 20 equally spaced positions in the genome and pseudotime dimension respectively. First, points were subset based on the genomic position of the desired subregion. Then methylation rate was averaged across the genome dimension, resulting in an array of 20 equally spaced predictions of methylation rate across pseudotime.

6.2.9 Comparison to scMet

scMet takes as input the total and methylated number of observed CpG/GpC sites in each cell within a region of interest. Each cell must be associated with a group, that is compared. These quantities were calculated from the raw methylation data, and cells filtered for the mesoderm lineage and grouped according to a pseudotime threshold of 0.5 (with scaled pseudotime from zero to one). Additionally scMet can take region level statistics into account as covariates. Thus, for each region the CpG/GpC density was calculated and used in the scMet model. Cells and regions had to be additionally filterered for sparsity, to ensure robust training of the model. First cells with less than 3 CpG/GpC sites within a region were excluded. Then regions with fewer than 10

remaining cells were excluded from the analysis. Lastly only regions with a minimum total variance of 0.0001 and a total methylation rate larger than 0.05 and smaller than 0.95 were included to remove non-variable regions. Note that without these filters the optimization of scMet was not stable.

6.2.10 Data availability

The raw sequencing data can be obtained from GSE121708. Parsed data, including count matrices for RNA expression and methylation call files are available at ftp://ftp.ebi.ac.uk/pub/databases/scnmt_gastrulation. The parsed data has been used as a basis for this thesis. For details about processing of the raw sequencing data see Argelaguet *et al.*, 2019b.

A | Appendix

A.1 Analysis of accessibility of lineage-specific enhancers during Mesoderm development

For each significantly differentially accessible lineage-specific enhancer, I then visualize the subregion with the highest 95% CI MMRC (Fig A.1). Differentially accessible Mesoderm enhancers almost exclusively increase in accessibility rate over time, while the majority of differential Ectoderm enhancers decrease their accessibility rate over time. In comparison, the few significant Endoderm enhancers show a more mixed signal.

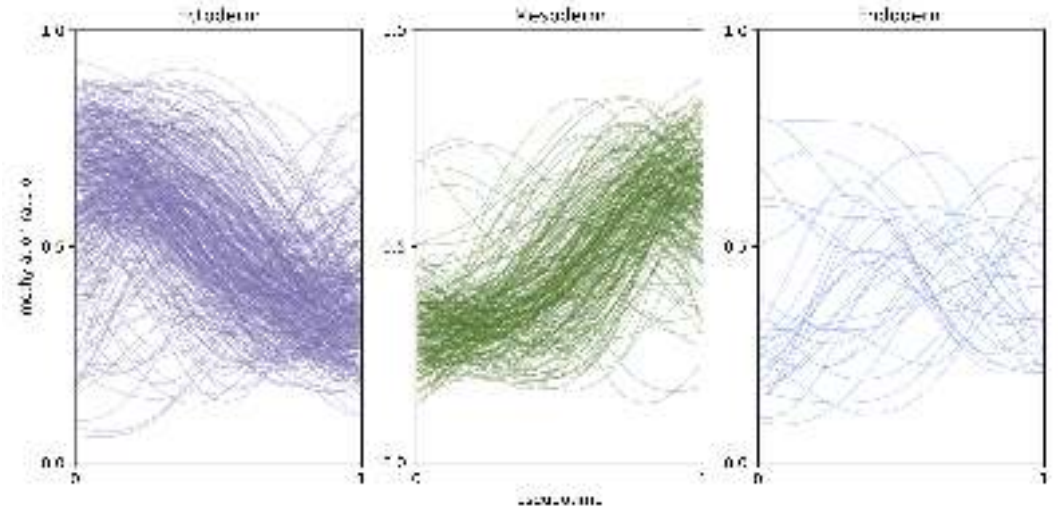


Figure A.1 | GPmeth refined pseudotemporal accessibility trajectories of lineage-specific enhancer regions. Lines represent the GPmeth posterior accessibility rate averages of the refined subregions found within differentially accessible enhancers by the model. Ectoderm-specific enhancers consistently decrease in accessibility rate over time, while Mesoderm-specific enhancers increase in accessibility rate. Ectoderm-specific enhancers (*left*) decrease the accessibility rate from 0.62 to 0.36 on average across the pseudotime range. Mesoderm-specific enhancers (*center*) increase accessibility rate from 0.31 to 0.62, and Endoderm-specific enhancers increase slightly from 0.34 to 0.40 on average. *Figure generated by Max Frank.*

I then again used k-means clustering to extract patterns in methylation rate from this data. Figure A.2 shows the extracted trends for three clusters. We again find

two major groups for Ectoderm- and Mesoderm-specific enhancers that correspond to early and late increase/decrease in accessibility.

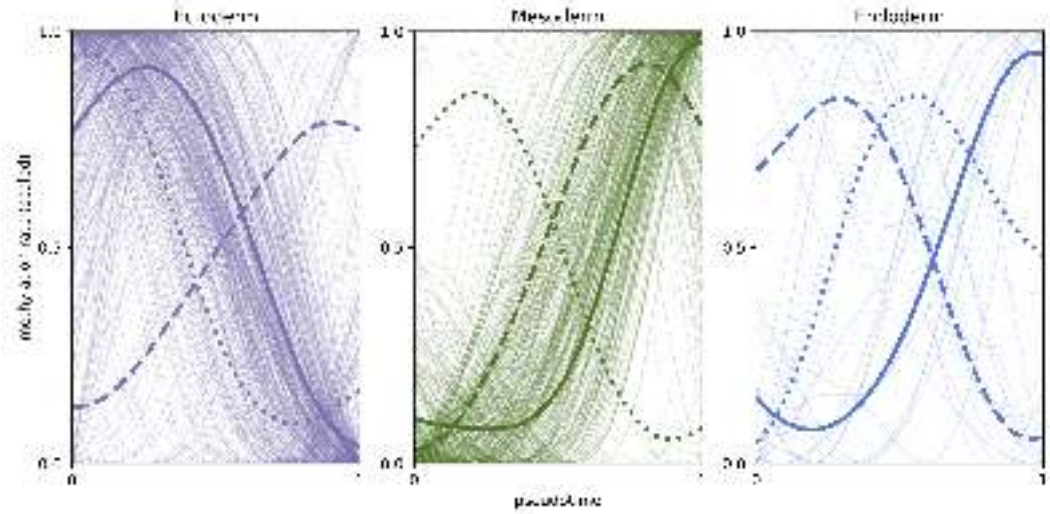


Figure A.2 | Clustered pseudotemporal accessibility trajectories of lineage-specific enhancer regions. Shown are the same pseudotemporal trajectories of accessibility rate predictions as in Figure A.1, but every trajectory was scaled to the range of [0,1] for scale-invariant clustering. These trajectories were then clustered with k-means based on Euclidean distance. A fixed number of three clusters was specified to capture up to two different trends in the trajectories and outlier trajectories. *Figure generated by Max Frank.*

I compare the pseudotemporal trajectories of these two main groups in Figure A.3). Ectoderm-specific enhancer accessibility decreases again seems to precede the increase in accessibility of Mesoderm-specific enhancers, although the effect is not quite as clear as with endogenous methylation.

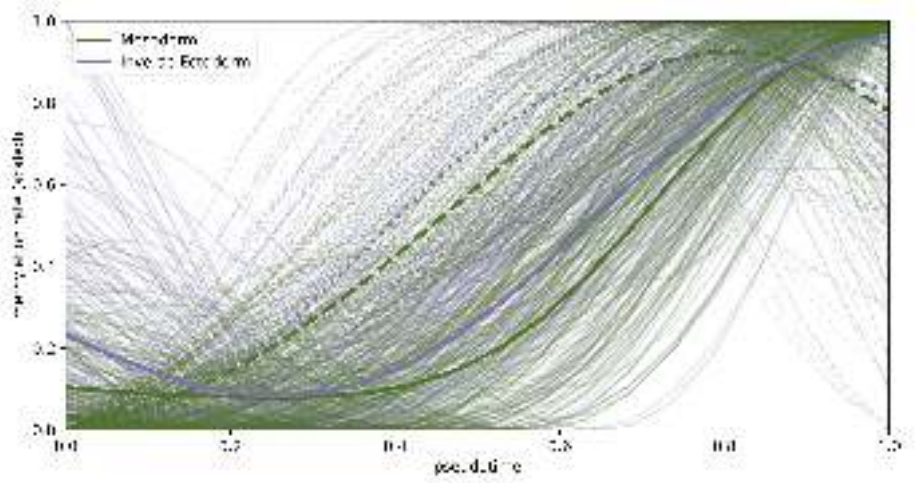


Figure A.3 | Temporal comparison of Ectoderm and Mesoderm-specific enhancer accessibility. Lines are pseudotemporal trajectories of lineage-specific enhancers as in Figure A.2. Green lines correspond to the two major clusters of Mesoderm enhancers that become more accessible over time. Purple lines correspond to the inverse profiles (i.e., 1-accessibility rate) of Ectoderm enhancers that become inaccessible over time. Both enhancer classes show similar temporal patterns, but Ectoderm-specific enhancer accessibility changes tend to precede Mesoderm-specific enhancer changes. *Figure generated by Max Frank.*

A.2 Epigenomic regulation during Gut development

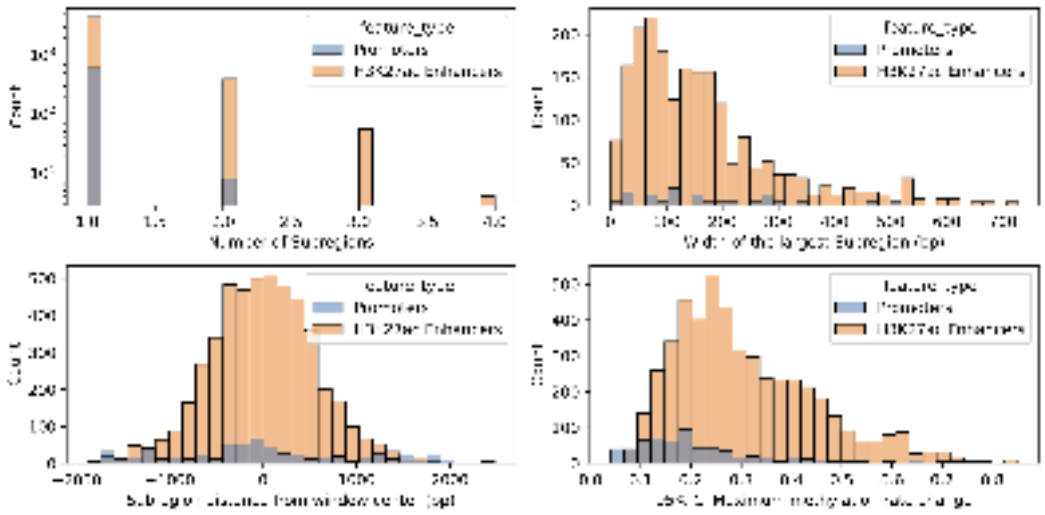


Figure A.4 | Summary statistics of differentially methylated refined regions during Gut development. The *top left* panel shows the number of subregions that are found by GPmeth for every genomic window with significant differential methylation. Note the log-scale on the y-axis. The *top right* panel shows the distribution of subregion widths, i.e. the width of the genomic interval where the 95% CI MMRC is higher than a specified threshold (in this case 0.3). The *bottom left* panel shows the positioning of the center of the identified subregions relative to the center of the input genomic window. The *bottom right* panel shows the average 95% CI MMRC of each identified subregion. *Figure generated by Max Frank.*

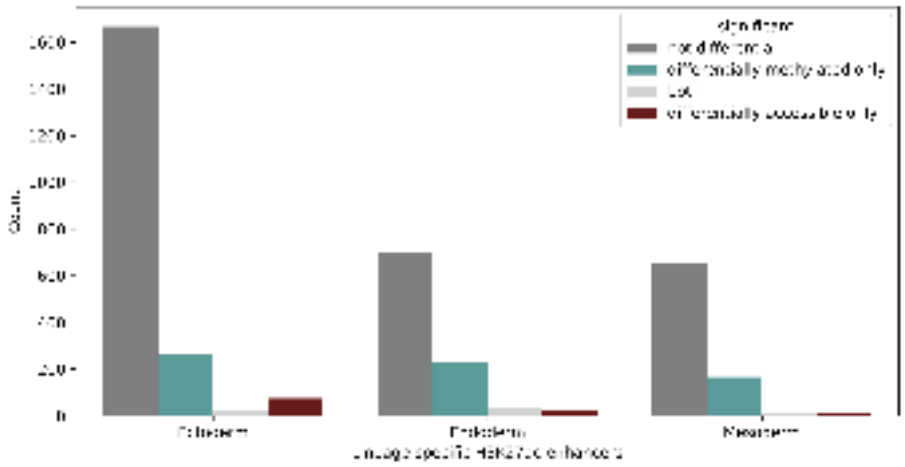


Figure A.5 | Number of differentially methylated lineage-specific enhancers during Gut development. Bar heights indicate the number of lineage-specific enhancer regions that were identified by GPmeth to be significantly differentially accessible or methylated (FDR<0.1). *Figure generated by Max Frank.*

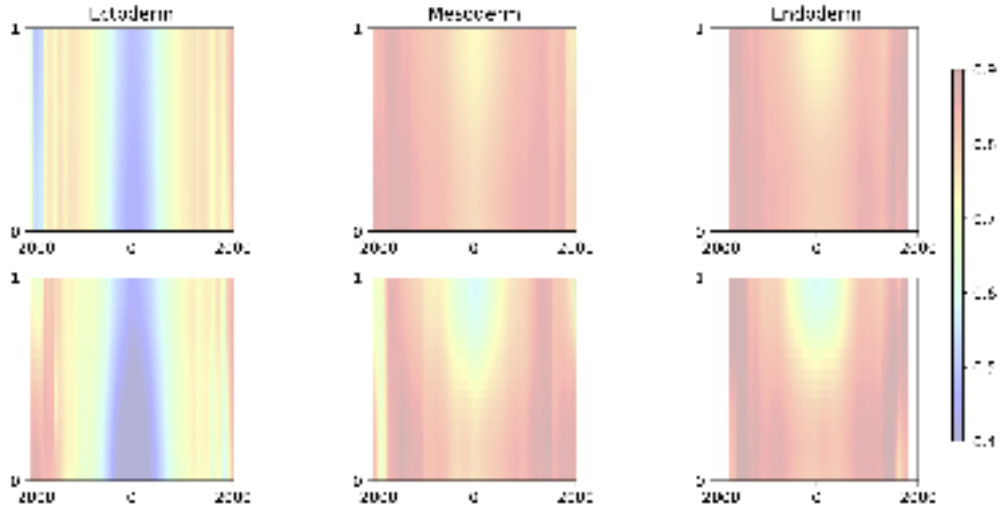


Figure A.6 | Averaged posterior methylation rate profiles for lineage-specific enhancers during Gut development. The heatmaps represent the GPmeth posterior mean predictions, averaged for lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime. The *top row* averages all lineage-specific enhancer for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR 0.1). *Figure generated by Max Frank.*

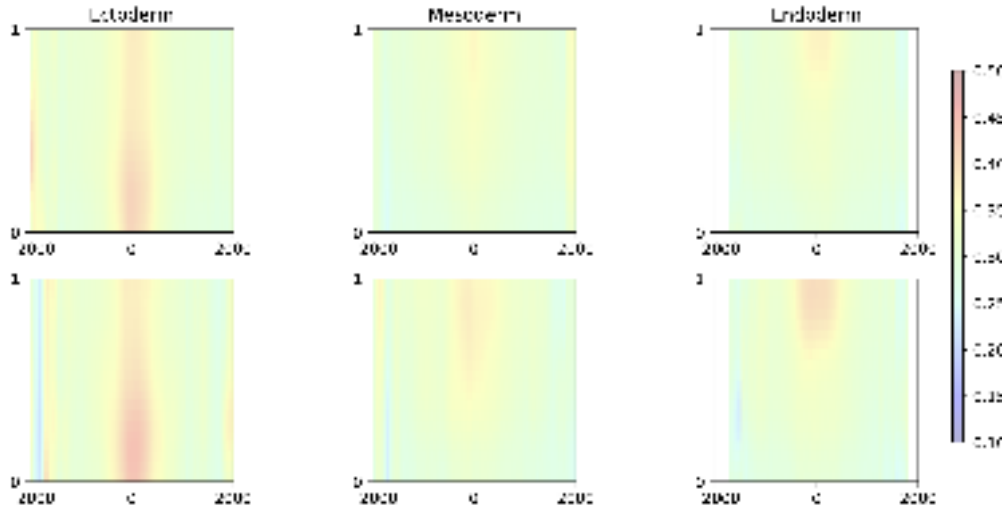


Figure A.7 | Averaged accessibility rate profiles for lineage-specific enhancers during Gut development. The heatmaps represent the GPmeth posterior mean predictions, averaged for lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime. The *top row* averages all lineage-specific enhancer for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR 0.1). *Figure generated by Max Frank.*

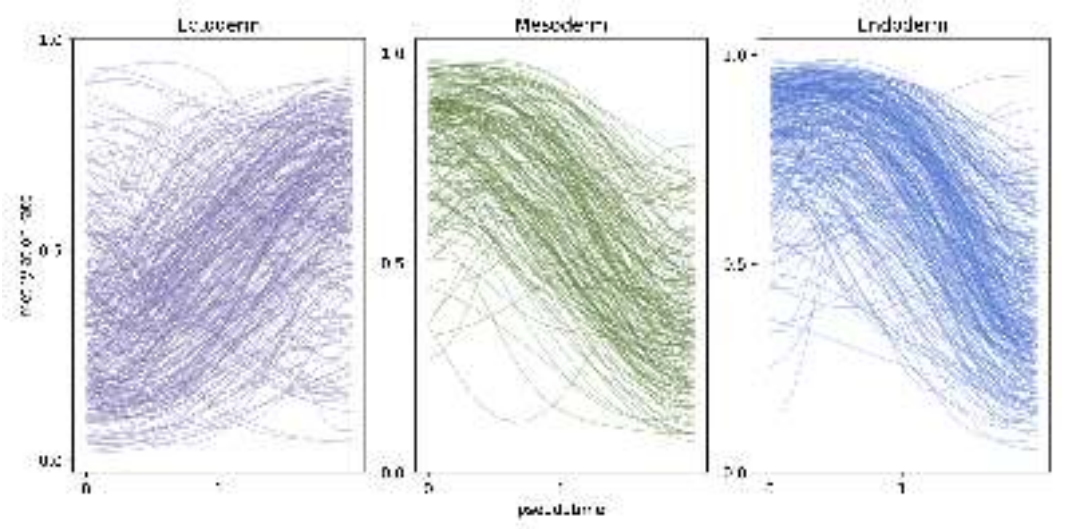


Figure A.8 | GPmeth refined pseudotemporal methylation trajectories of lineage-specific enhancer regions during Gut develompent. Lines represent the GPmeth posterior methylation rate averages of the refined subregions found within differentially methylated enhancers by the model. Shown are Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). *Figure generated by Max Frank.*

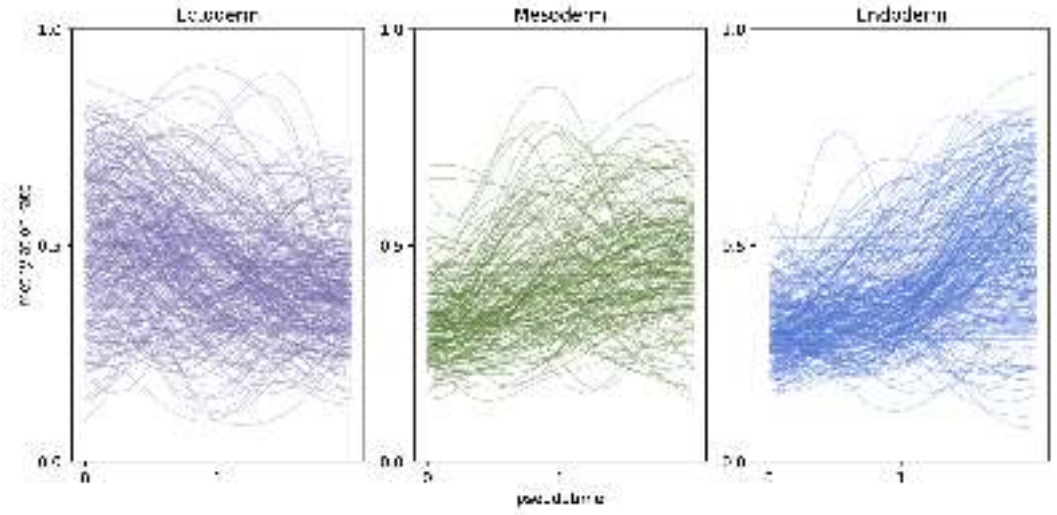


Figure A.9 | GPmeth refined pseudotemporal accessibility trajectories of lineage-specific enhancer regions during Gut develompent. Lines represent the GPmeth posterior accessibility rate averages of the refined subregions found within differentially accessible enhancers by the model. Shown are Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). *Figure generated by Max Frank.*

A.3 Epigenomic regulation during Notochord development

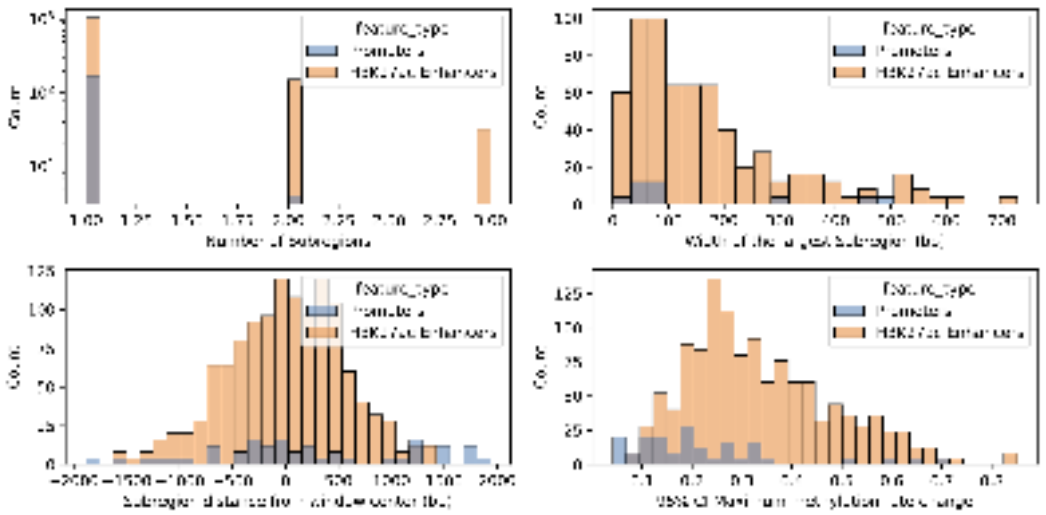


Figure A.10 | Summary statistics of differentially methylated refined regions during Notochord development. The *top left* panel shows the number of subregions that are found by GPmeth for every genomic window with significant differential methylation. Note the log-scale on the y-axis. The *top right* panel shows the distribution of subregion widths, i.e. the width of the genomic interval where the 95% CI MMRC is higher than a specified threshold (in this case 0.3). The *bottom left* panel shows the positioning of the center of the identified subregions relative to the center of the input genomic window. The *bottom right* panel shows the average 95% CI MMRC of each identified subregion. *Figure generated by Max Frank.*

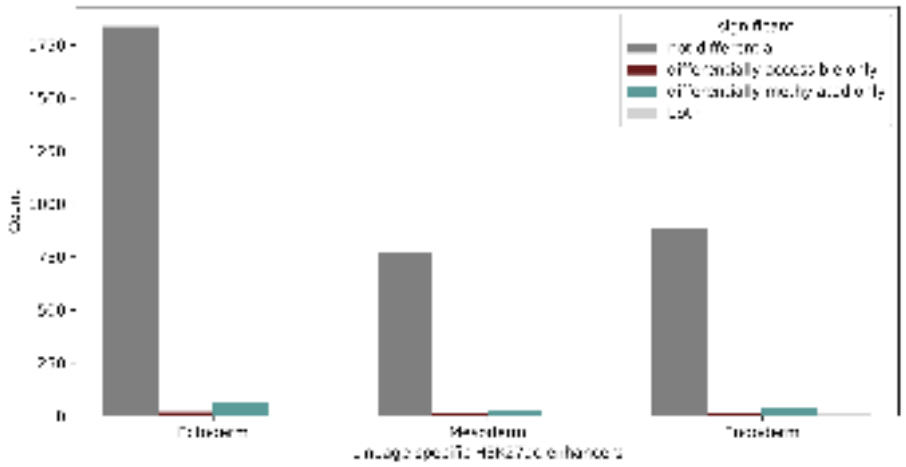


Figure A.11 | Number of differentially methylated lineage-specific enhancers during Notochord development. Bar heights indicate the number of lineage-specific enhancer regions that were identified by GPmeth to be significantly differentially accessible or methylated (FDR<0.1). *Figure generated by Max Frank.*

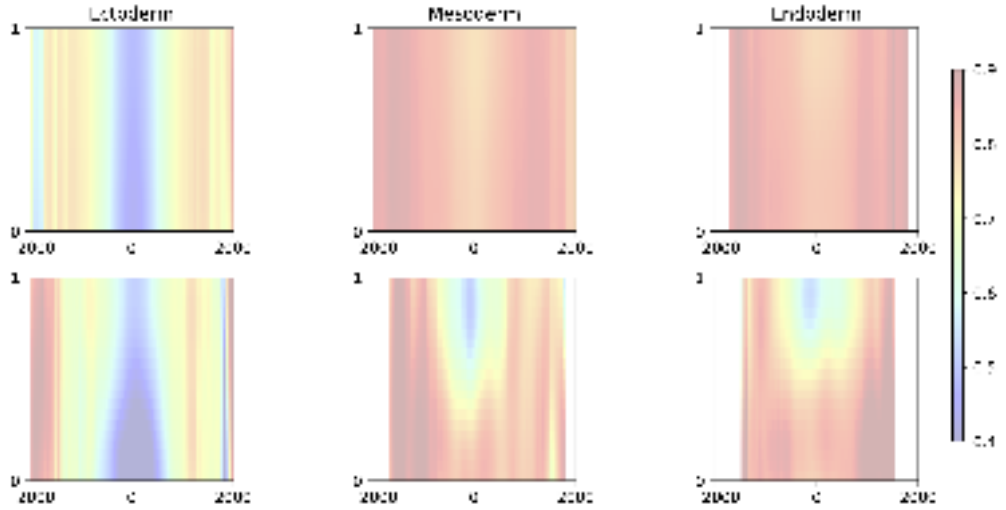


Figure A.12 | Averaged posterior methylation rate profiles for lineage-specific enhancers during Notochord development. The heatmaps represent the GPmeth posterior mean predictions, averaged for lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime. The *top row* averages all lineage-specific enhancer for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR 0.1). *Figure generated by Max Frank.*

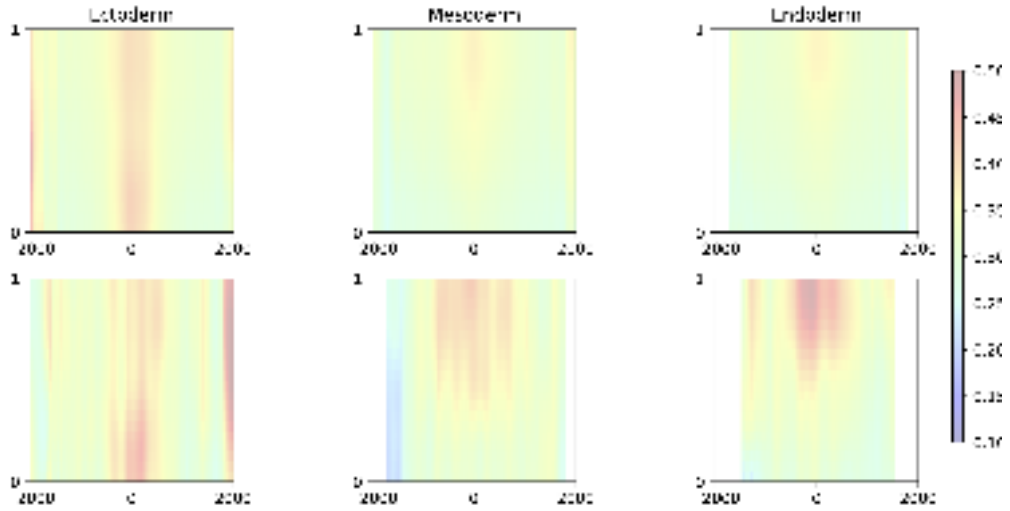


Figure A.13 | Averaged accessibility rate profiles for lineage-specific enhancers during Notochord development. The heatmaps represent the GPmeth posterior mean predictions, averaged for lineage-specific enhancer regions for Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). The averages were produced by taking predictions of the GPmeth model in a regular grid across a 4kb genomic window centered around the middle of the H3K27ac ChIP-seq peak and pseudotime. The *top row* averages all lineage-specific enhancer for the respective lineage, while the *bottom row* only averages differentially methylated enhancers (FDR 0.1). *Figure generated by Max Frank.*

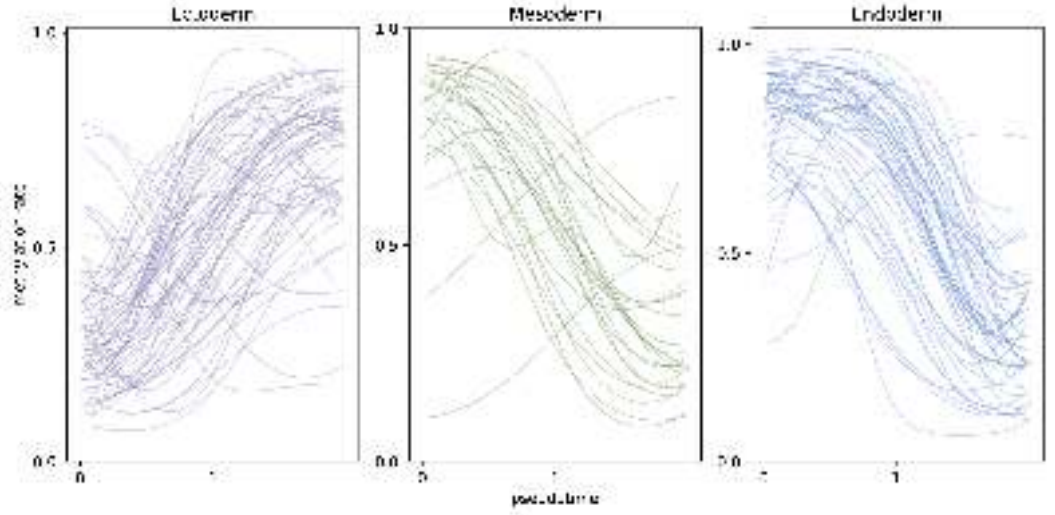


Figure A.14 | GPmeth refined pseudotemporal methylation trajectories of lineage-specific enhancer regions during Notochord developpent. Lines represent the GPmeth posterior methylation rate averages of the refined subregions found within differentially methylated enhancers by the model. Shown are Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). *Figure generated by Max Frank.*

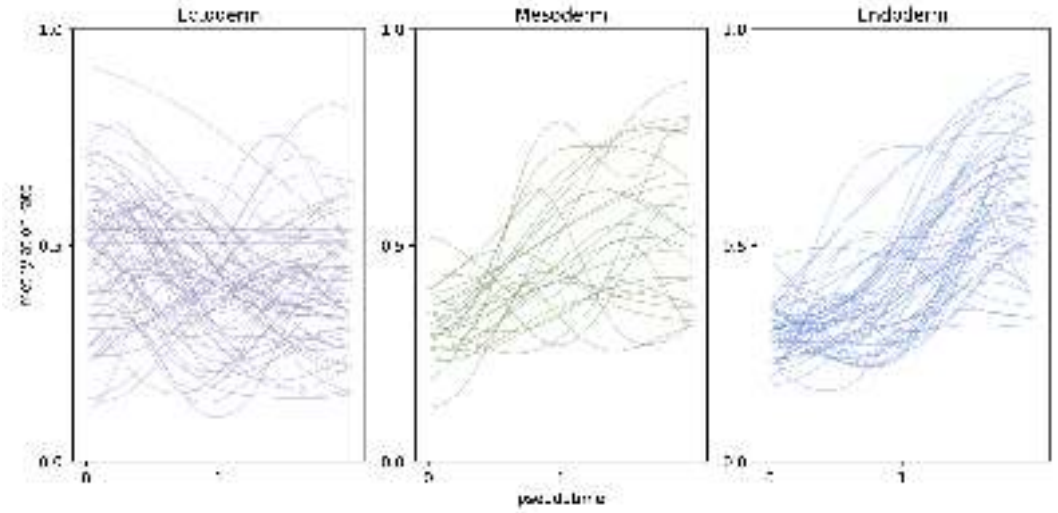


Figure A.15 | GPmeth refined pseudotemporal accessibility trajectories of lineage-specific enhancer regions during Notochord develompent. Lines represent the GPmeth posterior accessibility rate averages of the refined subregions found within differentially accessible enhancers by the model. Shown are Ectoderm enhancers (*left column*), Mesoderm enhancers (*center column*) and Endoderm enhancers (*right column*). *Figure generated by Max Frank.*

I composed this thesis with Overleaf, and used a custom document structure based on the 'Masters/Doctoral Thesis' LaTeXtemplate (www.latextemplates.com, authors Steve Gunn, Sunil Patel, vel@latextemplates), modified by Dr. Markus Mund, Dr. Jervis Vermal Thevathasan, Dr. Philipp Hoess, Dr. Yu-Le Wu, & Dr. Aline Tschanz, which is available under CC BY-NC-SA 3.0.

Licenses used in this thesis

- CC BY-NC-SA 3.0: http://creativecommons.org/licenses/by-nc-sa/3.0/
- CC BY 4.0: https://creativecommons.org/licenses/by/4.0/

Bibliography

Abante, J., Y. Fang, A. P. Feinberg, and J. Goutsias. 2020. Detection of haplotype-dependent allele-specific DNA methylation in WGBS data. *Nature Communications* **11(1)**: 5238. DOI: 10.1038/s41467-020-19077-1.

Aderem, A. 2005. Systems Biology: Its Practice and Challenges. *Cell* **121(4)**: 511–513. DOI: 10.1016/j.cell.2005.04.020.

Ahlmann-Eltze, C. and W. Huber. 2024. Analysis of multi-condition single-cell data with latent embedding multivariate regression. *bioRxiv*. DOI: 10.1101/2023.03.06.531268.

Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick, and C. E. Mason. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology* **13(10)**: R87. DOI: 10.1186/gb-2012-13-10-r87.

Allshire, R. C. and H. D. Madhani. 2018. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology* **19(4)**: 229–244. DOI: 10.1038/nrm.2017.119.

Andrews, T. S., V. Y. Kiselev, D. McCarthy, and M. Hemberg. 2021. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols* **16(1)**: 1–9. DOI: 10.1038/s41596-020-00409-w.

Angermueller, C., S. J. Clark, H. J. Lee, I. C. Macaulay, M. J. Teng, T. X. Hu, F. Krueger, S. A. Smallwood, C. P. Ponting, T. Voet, G. Kelsey, O. Stegle, and W. Reik. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods* **13(3)**: 229–232. DOI: 10.1038/nmeth.3728.

Angermueller, C., H. J. Lee, W. Reik, and O. Stegle. 2017. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology* **18(1)**: 67. DOI: 10.1186/s13059-017-1189-z.

Argelaguet, R., D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. 2019a. MOFA+: a probabilistic framework for comprehensive integration of structured single-cell data. *bioRxiv*: 837104. DOI: 10.1101/837104.

Argelaguet, R., S. J. Clark, H. Mohammed, L. C. Stapel, C. Krueger, C.-A. Kapourani, I. Imaz-Rosshandler, T. Lohoff, Y. Xiang, C. W. Hanna, S. Smallwood, X. Ibarra-Soria, F. Buettner, G. Sanguinetti, W. Xie, F. Krueger, B. Göttgens, P. J. Rugg-Gunn, G. Kelsey, W. Dean, J. Nichols, O. Stegle, J. C. Marioni, and W. Reik. 2019b. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576(7787)**: 487–491. DOI: 10.1038/s41586-019-1825-8.

Argelaguet, R., A. S. E. Cuomo, O. Stegle, and J. C. Marioni. 2021. Computational principles and challenges in single-cell data integration. *Nature Biotechnology* **39(10)**: 1202–1215. DOI: 10.1038/s41587-021-00895-7.

Argelaguet, R., T. Lohoff, J. G. Li, A. Nakhuda, D. Drage, F. Krueger, L. Velten, S. J. Clark, and W. Reik. 2022. Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv*. DOI: 10.1101/2022.06.15.496239.

Argelaguet, R., B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. 2018a. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14(6)**: e8124. DOI: 10.15252/msb.20178124.

Argelaguet, R., B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. 2018b. Multi-Omics Factor Analysis—a framework for unsupervised

integration of multi-omics data sets. *Molecular Systems Biology* **14(6)**: e8124. DOI: 10.15252/msb.20178124.

Atlasi, Y. and H. G. Stunnenberg. 2017. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews. Genetics* **18(11)**: 643–658. DOI: 10.1038/nrg.2017.57.

Auclair, G., S. Guibert, A. Bender, and M. Weber. 2014. Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. *Genome Biology* **15(12)**: 545. DOI: 10.1186/s13059-014-0545-5.

Barnett, K. R., B. E. Decato, T. J. Scott, T. J. Hansen, B. Chen, J. Attalla, A. D. Smith, and E. Hodges. 2020. ATAC-Me Captures Prolonged DNA Methylation of Dynamic Chromatin Accessibility Loci during Cell Fate Transitions. *Molecular Cell* **77(6)**: 1350–1364.e6. DOI: 10.1016/j.molcel.2020.01.004.

Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129(4)**: 823–837. DOI: 10.1016/j.cell.2007.05.009.

Benjamini, Y. and Y. Hochberg. 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. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

Bentsen, M., P. Goymann, H. Schultheis, K. Klee, A. Petrova, R. Wiegandt, A. Fust, J. Preussner, C. Kuenne, T. Braun, J. Kim, and M. Looso. 2020. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* **11(1)**: 4267. DOI: 10.1038/s41467-020-18035-1.

Bergman, Y. and H. Cedar. 2013. DNA methylation dynamics in health and disease. *Nature Structural & Molecular Biology* **20(3)**: 274–281. DOI: 10.1038/nsmb.2518.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125(2)**: 315–326. DOI: 10.1016/j.cell.2006.02.041.

BinTayyash, N., S. Georgaka, S. T. John, S. Ahmed, A. Boukouvalas, J. Hensman, and M. Rattray. 2021. Non-parametric modelling of temporal and spatial counts data from RNA-seq experiments. *Bioinformatics* **37(21)**: 3788–3795. DOI: 10.1093/bioinformatics/btab486.

Bird, A., M. Taggart, M. Frommer, O. J. Miller, and D. Macleod. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40(1)**: 91–99. DOI: 10.1016/0092-8674(85)90312-5.

Blake, J. A., R. Baldarelli, J. A. Kadin, J. E. Richardson, C. L. Smith, and C. J. Bult. 2020. Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. *Nucleic Acids Research* **49(D1)**: D981–D987. DOI: 10.1093/nar/gkaa1083.

Boukouvalas, A., J. Hensman, and M. Rattray. 2018a. BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biology* **19(1)**: 65. DOI: 10.1186/s13059-018-1440-2.

Boukouvalas, A., J. Hensman, and M. Rattray. 2018b. BGP: identifying gene-specific branching dynamics from single-cell data with a branching Gaussian process. *Genome Biology* **19(1)**: 65. DOI: 10.1186/s13059-018-1440-2.

Buenrostro, J. D., P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf. 2013. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods* **10(12)**: 1213–1218. DOI: 10.1038/nmeth.2688.

Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523(7561)**: 486–490. DOI: 10.1038/nature14590.

Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36(5)**: 411–420. DOI: 10.1038/nbt.4096.

Bylund, M., E. Andersson, B. G. Novitch, and J. Muhr. 2003. Vertebrate neurogenesis is counteracted by Sox1–3 activity. *Nature Neuroscience* **6(11)**: 1162–1168. DOI: 10.1038/nn1131.

Cao, J., M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, and J. Shendure. 2019. The single-cell

transcriptional landscape of mammalian organogenesis. *Nature* **566(7745)**: 496–502. DOI: 10.1038/s41586-019-0969-x.

Chen, J., A. Schlitzer, S. Chakarov, F. Ginhoux, and M. Poidinger. 2016. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature Communications* **7(1)**: 11988. DOI: 10.1038/ncomms11988.

Chen, S., B. B. Lake, and K. Zhang. 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology* **37(12)**: 1452–1457. DOI: 10.1038/s41587-019-0290-0.

Clark, S. J., R. Argelaguet, C.-A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, O. Stegle, and W. Reik. 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications* **9(1)**: 1–9. DOI: 10.1038/s41467-018-03149-4.

Costello, I., I.-M. Pimeisl, S. Dräger, E. K. Bikoff, E. J. Robertson, and S. J. Arnold. 2011. The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. *Nature Cell Biology* **13(9)**: 1084–1091. DOI: 10.1038/ncb2304.

Crick, F. 1970. Central Dogma of Molecular Biology. *Nature* **227(5258)**: 561–563. DOI: 10.1038/227561a0.

Cuomo, A. S. E., T. Heinen, D. Vagiaki, D. Horta, J. C. Marioni, and O. Stegle. 2022. CellRegMap: a statistical framework for mapping context-specific regulatory variants using scRNA-seq. *Molecular Systems Biology* **18(8)**: e10663. DOI: 10.15252/msb.202110663.

Cuomo, A. S. E., D. D. Seaton, D. J. McCarthy, I. Martinez, M. J. Bonder, J. Garcia-Bernardo, S. Amatya, P. Madrigal, A. Isaacson, F. Buettner, A. Knights, K. N. Natarajan, L. Vallier, J. C. Marioni, M. Chhatriwala, and O. Stegle. 2020. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications* **11(1)**: 810. DOI: 10.1038/s41467-020-14457-z.

Cusack, M., H. W. King, P. Spingardi, B. M. Kessler, R. J. Klose, and S. Kriaucionis. 2020. Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Research* **30(10)**: 1393–1406. DOI: 10.1101/gr.257576.119.

Cusanovich, D. A., J. P. Reddington, D. A. Garfield, R. M. Daza, D. Aghamirzaie, R. Marco-Ferreres, H. A. Pliner, L. Christiansen, X. Qiu, F. J. Steemers, C. Trapnell, J. Shendure, and E. E. M. Furlong. 2018. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature* **555(7697)**: 538–542. DOI: 10.1038/nature25981.

Dahlet, T., A. Argüeso Lleida, H. Al Adhami, M. Dumas, A. Bender, R. P. Ngondo, M. Tanguy, J. Vallet, G. Auclair, A. F. Bardet, and M. Weber. 2020. Genome-wide analysis in the mouse embryo reveals the importance of DNA methylation for transcription integrity. *Nature Communications* **11** 3153. DOI: 10.1038/s41467-020-16919-w.

Dann, E., N. C. Henderson, S. A. Teichmann, M. D. Morgan, and J. C. Marioni. 2022. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology* **40(2)**: 245–253. DOI: 10.1038/s41587-021-01033-z.

Deng, Y., M. Bartosovic, S. Ma, D. Zhang, P. Kukanja, Y. Xiao, G. Su, Y. Liu, X. Qin, G. B. Rosoklija, A. J. Dwork, J. J. Mann, M. L. Xu, S. Halene, J. E. Craft, K. W. Leong, M. Boldrini, G. Castelo-Branco, and R. Fan. 2022. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609(7926)**: 375–383. DOI: 10.1038/s41586-022-05094-1.

Do, C., C. F. Lang, J. Lin, H. Darbary, I. Krupska, A. Gaba, L. Petukhova, J.-P. Vonsattel, M. P. Gallagher, R. S. Goland, R. A. Clynes, A. Dwork, J. G. Kral, C. Monk, A. M. Christiano, and B. Tycko. 2016. Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. *The American Journal of Human Genetics* **98(5)**: 934–955. DOI: 10.1016/j.ajhg.2016.03.027.

Domcke, S., A. F. Bardet, P. Adrian Ginno, D. Hartl, L. Burger, and D. Schübeler. 2015. Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528(7583)**: 575–579. DOI: 10.1038/nature16462.

Dominicus, A., A. Skrondal, H. K. Gjessing, N. L. Pedersen, and J. Palmgren. 2006. Likelihood Ratio Tests in Behavioral Genetics: Problems and Solutions. *Behavior Genetics* **36(2)**: 331–340. DOI: 10.1007/s10519-005-9034-7.

Du, P., X. Zhang, C.-C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11(1)**: 587. DOI: 10.1186/1471-2105-11-587.

Du, Z., H. Zheng, B. Huang, R. Ma, J. Wu, X. Zhang, J. He, Y. Xiang, Q. Wang, Y. Li, J. Ma, X. Zhang, K. Zhang, Y. Wang, M. Q. Zhang, J. Gao, J. R. Dixon, X. Wang, J. Zeng, and W. Xie. 2017. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547(7662)**: 232–235. DOI: 10.1038/nature23263.

Fan, X., P. Lu, H. Wang, S. Bian, X. Wu, Y. Zhang, Y. Liu, D. Fu, L. Wen, J. Hao, and F. Tang. 2022. Integrated single-cell multiomics analysis reveals novel candidate markers for prognosis in human pancreatic ductal adenocarcinoma. *Cell Discovery* **8(1)**: 1–16. DOI: 10.1038/s41421-021-00366-y.

Feng, H. and H. Wu. 2019. Differential methylation analysis for bisulfite sequencing using DSS. *Quantitative biology (Beijing, China)* **7(4)**: 327–334. DOI: 10.1007/s40484-019-0183-8.

Fisher, R. A. 1922. On the Interpretation of $\chi 2$ from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society* **85(1)**: 87–94. DOI: 10.2307/2340521.

Frommer, M., L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul. 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 of the United States of America* **89(5)**: 1827–1831. DOI: 10.1073/pnas.89.5.1827.

Galupa, R. and E. Heard. 2018. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annual Review of Genetics* **52(1)**: 535–566. DOI: 10.1146/annurev-genet-120116-024611.

Gardiner-Garden, M. and M. Frommer. 1987. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196(2)**: 261–282. DOI: 10.1016/0022-2836(87)90689-9.

Gardner, J. R., G. Pleiss, D. Bindel, K. Q. Weinberger, and A. G. Wilson. 2021. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. DOI: 10.48550/arXiv.1809.11165.

Giansanti, P., P. Samaras, Y. Bian, C. Meng, A. Coluccio, M. Frejno, H. Jakubowsky, S. Dobiasch, R. R. Hazarika, J. Rechenberger, J. Calzada-Wack, J. Krumm, S. Mueller, C.-Y. Lee, N. Wimberger, L. Lautenbacher, Z. Hassan, Y.-C. Chang, C. Falcomatà, F. P. Bayer, S. Bärthel, T. Schmidt, R. Rad, S. E. Combs, M. The, F. Johannes, D. Saur, M. H. de Angelis, M. Wilhelm, G. Schneider, and B. Kuster. 2022. Mass spectrometry-based draft of the mouse proteome. *Nature Methods* **19(7)**: 803–811. DOI: 10.1038/s41592-022-01526-y.

Gilbert, S. F. 2000. Early Mammalian Development. *Developmental Biology. 6th edition.* Sinauer Associates.

Goll, M. G. and T. H. Bestor. 2005. Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry* **74(1)**: 481–514. DOI: 10.1146/annurev.biochem.74.010904.153721.

GPy. 2012. GPy: A Gaussian process framework in python.

Granja, J. M., M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* **53(3)**: 403–411. DOI: 10.1038/s41588-021-00790-6.

Greer, E. L. and Y. Shi. 2012. Histone methylation: a dynamic mark in health, disease and inheritance. *Nature Reviews Genetics* **13(5)**: 343–357. DOI: 10.1038/nrg3173.

Greven, S., C. M. Crainiceanu, H. Küchenhoff, and A. Peters. 2008. Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *Journal of Computational and Graphical Statistics* **17(4)**: 870–891. DOI: 10.1198/106186008X386599.

Grunstein, M. 1997. Histone acetylation in chromatin structure and transcription. *Nature* **389(6649)**: 349–352. DOI: 10.1038/38664.

Guo, H., P. Zhu, F. Guo, X. Li, X. Wu, X. Fan, L. Wen, and F. Tang. 2015. Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature Protocols* **10(5)**: 645–659. DOI: 10.1038/nprot.2015.039.

Haghverdi, L., M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13(10)**: 845–848. DOI: 10.1038/nmeth.3971.

Haghverdi, L., A. T. L. Lun, M. D. Morgan, and J. C. Marioni. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* **36(5)**: 421–427. DOI: 10.1038/nbt.4091.

Hanna, C. W., H. Demond, and G. Kelsey. 2018. Epigenetic regulation in development: is the mouse a good model for the human? *Human Reproduction Update* **24(5)**: 556–576. DOI: 10.1093/humupd/dmy021.

Hansen, K. D., B. Langmead, and R. A. Irizarry. 2012. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology* **13(10)**: R83. DOI: 10.1186/gb-2012-13-10-r83.

Hao, Y., S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. 2021. Integrated analysis of multimodal single-cell data. *Cell* **184(13)**: 3573–3587.e29. DOI: 10.1016/j.cell.2021.04.048.

Hebestreit, K., M. Dugas, and H.-U. Klein. 2013. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics* **29(13)**: 1647–1653. DOI: 10.1093/bioinformatics/btt263.

Heinonen, M., O. Guipaud, F. Milliat, V. Buard, B. Micheau, G. Tarlet, M. Benderitter, F. Zehraoui, and F. d'Alché-Buc. 2015. Detecting time periods of differential gene expression using Gaussian processes: an application to endothelial cells exposed to radiotherapy dose fraction. *Bioinformatics* **31(5)**: 728–735. DOI: 10.1093/bioinformatics/btu699.

Hemberger, M., W. Dean, and W. Reik. 2009. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nature Reviews Molecular Cell Biology* **10(8)**: 526–537. DOI: 10.1038/nrm2727.

Henikoff, S. and A. Shilatifard. 2011. Histone modification: cause or cog? *Trends in Genetics* **27(10)**: 389–396. DOI: 10.1016/j.tig.2011.06.006.

Hensman, J., N. Fusi, and N. D. Lawrence. 2013. Gaussian Processes for Big Data. *Uncertainty in Artificial Intelligence*. **29** AUAI Press.

Heumos, L., A. C. Schaar, C. Lance, A. Litinetskaya, F. Drost, L. Zappia, M. D. Lücken, D. C. Strobl, J. Henao, F. Curion, H. B. Schiller, and F. J. Theis. 2023. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24(8)**: 550–572. DOI: 10.1038/s41576-023-00586-w.

Iwafuchi-Doi, M. and K. S. Zaret. 2014. Pioneer transcription factors in cell reprogramming. *Genes & Development* **28(24)**: 2679–2692. DOI: 10.1101/gad.253443.114.

Jambhekar, A., A. Dhall, and Y. Shi. 2019. Roles and regulation of histone methylation in animal development. *Nature Reviews Molecular Cell Biology* **20(10)**: 625–641. DOI: 10.1038/s41580-019-0151-1.

Ji, Z. and H. Ji. 2016. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research* **44(13)**: e117. DOI: 10.1093/nar/gkw430.

Jiang, C. and B. F. Pugh. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* **10(3)**: 161–172. DOI: 10.1038/nrg2522.

Jin, W., Q. Tang, M. Wan, K. Cui, Y. Zhang, G. Ren, B. Ni, J. Sklar, T. M. Przytycka, R. Childs, D. Levens, and K. Zhao. 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528(7580)**: 142–146. DOI: 10.1038/nature15740.

Jones, P. A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13(7)**: 484–492. DOI: 10.1038/nrg3230.

Jühling, F., H. Kretzmer, S. H. Bernhart, C. Otto, P. F. Stadler, and S. Hoffmann. 2016. metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Research* **26(2)**: 256–262. DOI: 10.1101/gr.196394.115.

Kalaitzis, A. A. and N. D. Lawrence. 2011. A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression. *BMC Bioinformatics* **12(1)**: 180. DOI: 10.1186/1471-2105-12-180.

Kaluscha, S., S. Domcke, C. Wirbelauer, M. B. Stadler, S. Durdu, L. Burger, and D. Schübeler. 2022. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nature Genetics* **54(12)**: 1895–1906. DOI: 10.1038/s41588-022-01241-6.

Kapourani, C.-A., R. Argelaguet, G. Sanguinetti, and C. A. Vallejos. 2021. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. *Genome Biology* **22(1)**: 114. DOI: 10.1186/s13059-021-02329-8.

Kapourani, C.-A. and G. Sanguinetti. 2019. Melissa: Bayesian clustering and imputation of single-cell methylomes. *Genome Biology* **20(1)**: 61. DOI: 10.1186/s13059-019-1665-8.

Kats, I., R. Vento-Tormo, and O. Stegle. 2021. SpatialDE2: Fast and localized variance component analysis of spatial transcriptomics. DOI: 10.1101/2021.10.27.466045.

Ke, Y., Y. Xu, X. Chen, S. Feng, Z. Liu, Y. Sun, X. Yao, F. Li, W. Zhu, L. Gao, H. Chen, Z. Du, W. Xie, X. Xu, X. Huang, and J. Liu. 2017. 3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis. *Cell* **170(2)**: 367–381.e20. DOI: 10.1016/j.cell.2017.06.029.

Kelly, T. K., Y. Liu, F. D. Lay, G. Liang, B. P. Berman, and P. A. Jones. 2012. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research* **22(12)**: 2497–2506. DOI: 10.1101/gr.143008.112.

Kelsey, G., O. Stegle, and W. Reik. 2017. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358(6359)**: 69–75. DOI: 10.1126/science.aan6826.

Kim, S. and J. Wysocka. 2023. Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell*. Reimagining the Central Dogma **83(3)**: 373–392. DOI: 10.1016/j.molcel.2022.12.032.

Kirk, P. D. W. and M. P. H. Stumpf. 2009. Gaussian process regression bootstrapping: exploring the effects of uncertainty in time course data. *Bioinformatics* **25(10)**: 1300–1306. DOI: 10.1093/bioinformatics/btp139.

Klemm, S. L., Z. Shipony, and W. J. Greenleaf. 2019. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20(4)**: 207–220. DOI: 10.1038/s41576-018-0089-8.

Klose, R. J. and A. P. Bird. 2006. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences* **31(2)**: 89–97. DOI: 10.1016/j.tibs.2005.12.008.

Kouzarides, T. 2007. Chromatin Modifications and Their Function. *Cell* **128(4)**: 693–705. DOI: 10.1016/j.cell.2007.02.005.

Krebs, A. R., D. Imanci, L. Hoerner, D. Gaidatzis, L. Burger, and D. Schübeler. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Molecular cell* **67(3)**: 411–422.e4. DOI: 10.1016/j.molcel.2017.06.027.

Kreibich, E., R. Kleinendorst, G. Barzaghi, S. Kaspar, and A. R. Krebs. 2023. Single-molecule footprinting identifies context-dependent regulation of enhancers by DNA methylation. *Molecular cell* **83(5)**: 787–802.e9. DOI: 10.1016/j.molcel.2023.01.017.

Kribelbauer, J. F., O. Laptenko, S. Chen, G. D. Martini, W. A. Freed-Pastor, C. Prives, R. S. Mann, and H. J. Bussemaker. 2017. Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. *Cell Reports* **19(11)**: 2383–2395. DOI: 10.1016/j.celrep.2017.05.069.

Krueger, F. and S. R. Andrews. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27(11)**: 1571–1572. DOI: 10.1093/bioinformatics/btr167.

Lambert, S. A., A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. 2018. The Human Transcription Factors. *Cell* **172(4)**: 650–665. DOI: 10.1016/j.cell.2018.01.029.

Lammers, N. C., Y. J. Kim, J. Zhao, and H. G. Garcia. 2020. A matter of time: Using dynamics and theory to uncover mechanisms of transcriptional bursting. *Current Opinion in Cell Biology*. Differentiation and disease **67** 147–157. DOI: 10.1016/j.ceb.2020.08.001.

Larsen, F., G. Gundersen, R. Lopez, and H. Prydz. 1992. CpG islands as gene markers in the human genome. *Genomics* **13(4)**: 1095–1107. DOI: 10.1016/0888-7543(92)90024-M.

Lawrence, N., G. Sanguinetti, and M. Rattray. 2006. Modelling transcriptional regulation using Gaussian Processes. *Advances in Neural Information Processing Systems*. **19** MIT Press.

Lázaro-Gredilla, M., S. Van Vaerenbergh, and N. D. Lawrence. 2012. Overlapping Mixtures of Gaussian Processes for the data association problem. *Pattern Recognition* **45(4)**: 1386–1395. DOI: 10.1016/j.patcog.2011.10.004.

Lee, H. J., T. A. Hore, and W. Reik. 2014. Reprogramming the methylome: erasing memory and creating diversity. *Cell Stem Cell* **14(6)**: 710–719. DOI: 10.1016/j.stem.2014.05.008.

Lee, J., D. Y. Hyeon, and D. Hwang. 2020. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine* **52(9)**: 1428–1442. DOI: 10.1038/s12276-020-0420-2.

Lee, T. I. and R. A. Young. 2013. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152(6)**: 1237–1251. DOI: 10.1016/j.cell.2013.02.014.

Li, H. 2011. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27(5)**: 718–719. DOI: 10.1093/bioinformatics/btq671.

Lippert, C., F. P. Casale, B. Rakitsch, and O. Stegle. 2014. LIMIX: genetic analysis of multiple traits. *bioRxiv*. DOI: 10.1101/003905.

Listgarten, J., C. Lippert, E. Y. Kang, J. Xiang, C. M. Kadie, and D. Heckerman. 2013. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29(12)**: 1526–1533. DOI: 10.1093/bioinformatics/btt177.

Lönnberg, T., V. Svensson, K. R. James, D. Fernandez-Ruiz, I. Sebina, R. Montandon, M. S. F. Soon, L. G. Fogg, A. S. Nair, U. N. Liligeto, M. J. T. Stubbington, L.-H. Ly, F. O. Bagger, M. Zwiessele, N. D. Lawrence, F. Souza-Fonseca-Guimaraes, P. T. Bunn, C. R. Engwerda, W. R. Heath, O. Billker, O. Stegle, A. Haque, and S. A. Teichmann. 2017. Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves TH1/TFH fate bifurcation in malaria. *Science Immunology* **2(9)**: eaal2192. DOI: 10.1126/sciimmunol.aal2192.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15(12)**: 1–21. DOI: 10.1186/s13059-014-0550-8.

Luecken, M. D. and F. J. Theis. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology* **15(6)**: e8746. DOI: 10.15252/msb.20188746.

Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389(6648)**: 251–260. DOI: 10.1038/38444.

Ma, S., B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, T. Law, C. Lareau, Y.-C. Hsu, A. Regev, and J. D. Buenrostro. 2020. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183(4)**: 1103–1116.e20. DOI: 10.1016/j.cell.2020.09.056.

Macaulay, I. C., W. Haerty, P. Kumar, Y. I. Li, T. X. Hu, M. J. Teng, M. Goolam, N. Saurat, P. Coupland, L. M. Shirley, M. Smith, N. Van der Aa, R. Banerjee, P. D. Ellis, M. A. Quail, H. P. Swerdlow, M. Zernicka-Goetz, F. J. Livesey, C. P. Ponting, and T. Voet. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods* **12(6)**: 519–522. DOI: 10.1038/nmeth.3370.

Macaulay, I. C., C. P. Ponting, and T. Voet. 2017. Single-Cell Multiomics: Multiple Measurements from Single Cells. *Trends in Genetics* **33(2)**: 155–168. DOI: 10.1016/j.tig.2016.12.003.

MacKay, D. J. C. 1998. Introduction to Gaussian Processes. *Neural Networks and Machine Learning*. NATO ASI Series. Kluwer Academic Press: 133–166.

Mattei, A. L., N. Bailly, and A. Meissner. 2022. DNA methylation: a historical perspective. *Trends in Genetics* **38(7)**: 676–707. DOI: 10.1016/j.tig.2022.03.010.

Matthews, A. G. d. G., M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. 2017. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research* **18(40)**: 1–6.

Maurano, M. T., H. Wang, S. John, A. Shafer, T. Canfield, K. Lee, and J. A. Stamatoyannopoulos. 2015. Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell Reports* **12(7)**: 1184–1195. DOI: 10.1016/j.celrep.2015.07.024.

Mayo, T. R., G. Schweikert, and G. Sanguinetti. 2015. M3D: a kernel-based test for spatially correlated changes in methylation profiles. *Bioinformatics (Oxford, England)* **31(6)**: 809–816. DOI: 10.1093/bioinformatics/btu749.

Meissner, A. 2010. Epigenetic modifications in pluripotent and differentiated cells. *Nature Biotechnology* **28(10)**: 1079–1088. DOI: 10.1038/nbt.1684.

Meissner, A., A. Gnirke, G. W. Bell, B. Ramsahoye, E. S. Lander, and R. Jaenisch. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research* **33(18)**: 5868–5877. DOI: 10.1093/nar/gki901.

Metzker, M. L. 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics* **11(1)**: 31–46. DOI: 10.1038/nrg2626.

Millán-Zambrano, G., A. Burton, A. J. Bannister, and R. Schneider. 2022. Histone post-translational modifications — cause and consequence of genome function. *Nature Reviews Genetics* **23(9)**: 563–580. DOI: 10.1038/s41576-022-00468-7.

Minka, T. P. 2001. A family of algorithms for approximate Bayesian inference. Thesis. Massachusetts Institute of Technology.

Moore, L. D., T. Le, and G. Fan. 2013. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38(1)**: 23–38. DOI: 10.1038/npp.2012.112.

Mulder, J. 2023. Bayesian Testing of Linear Versus Nonlinear Effects Using Gaussian Process Priors. *The American Statistician* **77(1)**: 1–11. DOI: 10.1080/00031305.2022.2028675.

Muñoz-Sanjuán, I. and A. H. Brivanlou. 2002. Neural induction, the default model and embryonic stem cells. *Nature Reviews Neuroscience* **3(4)**: 271–280. DOI: 10.1038/nrn786.

Neal, R. M. 1996. Bayesian Learning for Neural Networks. **118** Lecture Notes in Statistics. New York, NY: Springer. ISBN: 978-1-4612-0745-0. DOI: 10.1007/978-1-4612-0745-0.

Neal, R. M. 1997. Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. DOI: 10.48550/arXiv.physics/9701026.

Nica, A. C. and E. T. Dermitzakis. 2013. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368(1620)**: 20120362. DOI: 10.1098/rstb.2012.0362.

Nordström, K. J. V., F. Schmidt, N. Gasparoni, A. Salhab, G. Gasparoni, K. Kattler, F. Müller, P. Ebert, I. G. Costa, DEEP consortium, N. Pfeifer, T. Lengauer, M. H. Schulz, and J. Walter. 2019. Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data. *Nucleic Acids Research* **47(20)**: 10580–10596. DOI: 10.1093/nar/gkz799.

Otani, J., H. Kimura, J. Sharif, T. A. Endo, Y. Mishima, T. Kawakami, H. Koseki, M. Shirakawa, I. Suetake, and S. Tajima. 2013. Cell Cycle-Dependent Turnover of 5-Hydroxymethyl Cytosine in Mouse Embryonic Stem Cells. *PLOS ONE* **8(12)**: e82961. DOI: 10.1371/journal.pone.0082961.

Pan, Q., O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40(12)**: 1413–1415. DOI: 10.1038/ng.259.

Panigrahi, A. and B. W. O'Malley. 2021. Mechanisms of enhancer action: the known and the unknown. *Genome Biology* **22(1)**: 108. DOI: 10.1186/s13059-021-02322-1.

Panja, S. and B. C. Paria. 2021. Development of the Mouse Placenta. *Advances in anatomy, embryology, and cell biology* **234** 205–221. DOI: 10.1007/978-3-030-77360-1_10.

Papaioannou, V. E. 2014. The T-box gene family: emerging roles in development, stem cells and cancer. *Development (Cambridge, England)* **141(20)**: 3819–3833. DOI: 10.1242/dev.104471.

Park, Y., M. E. Figueroa, L. S. Rozek, and M. A. Sartor. 2014. MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics (Oxford, England)* **30(17)**: 2414–2422. DOI: 10.1093/bioinformatics/btu339.

Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg. 2014. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols* **9(1)**: 171–181. DOI: 10.1038/nprot.2014.006.

Pijuan-Sala, B., J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. V. Tyser, D. L. L. Ho, W. Reik, S. Srinivas, B. D. Simons, J. Nichols, J. C. Marioni, and B. Göttgens. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566(7745)**: 490–495. DOI: 10.1038/s41586-019-0933-9.

Polański, K., M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park. 2020. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36(3)**: 964–965. DOI: 10.1093/bioinformatics/btz625.

Pott, S. 2017. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *eLife* **6** e23203. DOI: 10.7554/eLife.23203.

Pourrajab, F. and S. Hekmatimoghaddam. 2021. Transposable elements, contributors in the evolution of organisms (from an arms race to a source of raw materials). *Heliyon* **7(1)**: e06029. DOI: 10.1016/j.heliyon.2021.e06029.

Qiu, X., Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell. 2017. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods* **14(10)**: 979–982. DOI: 10.1038/nmeth.4402.

Rabiner, L. and B. Gold. 1975. Theory and Application of Digital Signal Processing. Prentice-Hall signal processing series. Prentice-Hall. ISBN: 978-0-13-914101-0.

Rasmussen, C. E. and C. K. I. Williams. 2006. Gaussian Processes for Machine Learning. The MIT Press. ISBN: 0-262-18253-X.

Rauluseviciute, I., R. Riudavets-Puig, R. Blanc-Mathieu, J. A. Castro-Mondragon, K. Ferenc, V. Kumar, R. B. Lemma, J. Lucas, J. Chèneby, D. Baranasic, A. Khan, O. Fornes, S. Gundersen, M. Johansen, E. Hovig, B. Lenhard, A. Sandelin, W. W. Wasserman, F. Parcy, and A. Mathelier. 2024. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52(D1)**: D174–D182. DOI: 10.1093/nar/gkad1059.

Reményi, A., H. R. Schöler, and M. Wilmanns. 2004. Combinatorial control of gene expression. *Nature Structural & Molecular Biology* **11(9)**: 812–815. DOI: 10.1038/nsmb820.

Reuter, J. A., D. V. Spacek, and M. P. Snyder. 2015. High-Throughput Sequencing Technologies. *Molecular Cell* **58(4)**: 586–597. DOI: 10.1016/j.molcel.2015.05.004.

Rice, J. C., S. D. Briggs, B. Ueberheide, C. M. Barber, J. Shabanowitz, D. F. Hunt, Y. Shinkai, and C. D. Allis. 2003. Histone Methyltransferases Direct Different Degrees of Methylation to Define Distinct Chromatin Domains. *Molecular Cell* **12(6)**: 1591–1598. DOI: 10.1016/S1097-2765(03)00479-9.

Ritchie, M. D., E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim. 2015a. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics* **16(2)**: 85–97. DOI: 10.1038/nrg3868.

Ritchie, M. E., B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. 2015b. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43(7)**: e47. DOI: 10.1093/nar/gkv007.

Rulands, S., H. J. Lee, S. J. Clark, C. Angermueller, S. A. Smallwood, F. Krueger, H. Mohammed, W. Dean, J. Nichols, P. Rugg-Gunn, G. Kelsey, O. Stegle, B. D. Simons, and W. Reik. 2018. Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. *Cell Systems* **7(1)**: 63–76.e12. DOI: 10.1016/j.cels.2018.06.012.

Saelens, W., R. Cannoodt, H. Todorov, and Y. Saeys. 2019. A comparison of single-cell trajectory inference methods. *Nature Biotechnology* **37(5)**: 547–554. DOI: 10.1038/s41587-019-0071-9.

Sainsbury, S., C. Bernecky, and P. Cramer. 2015. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* **16(3)**: 129–143. DOI: 10.1038/nrm3952.

Saxonov, S., P. Berg, and D. L. Brutlag. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences* **103(5)**: 1412–1417. DOI: 10.1073/pnas.0510310103.

Schultz, M. D., Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang, and J. R. Ecker. 2015. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523(7559)**: 212–216. DOI: 10.1038/nature14465.

Self, S. G. and K.-Y. Liang. 1987. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association* **82(398)**: 605–610. DOI: 10.1080/01621459.1987.10478472.

Setty, M., V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, and D. Pe'er. 2019. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology* **37(4)**: 451–460. DOI: 10.1038/s41587-019-0068-4.

Smallwood, S. A., H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods* **11(8)**: 817–820. DOI: 10.1038/nmeth.3035.

Smith, Z. D., M. M. Chan, T. S. Mikkelsen, H. Gu, A. Gnirke, A. Regev, and A. Meissner. 2012. A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484(7394)**: 339–344. DOI: 10.1038/nature10960.

Solnica-Krezel, L. and D. S. Sepich. 2012. Gastrulation: Making and Shaping Germ Layers. *Annual Review of Cell and Developmental Biology* **28(1)**: 687–717. DOI: 10.1146/annurev-cellbio-092910-154043.

Song, L. and G. E. Crawford. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols* **2010(2)**: pdb.prot5384. DOI: 10.1101/pdb.prot5384.

Sönmezer, C., R. Kleinendorst, D. Imanci, G. Barzaghi, L. Villacorta, D. Schübeler, V. Benes, N. Molina, and A. R. Krebs. 2021. Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Molecular cell* **81(2)**: 255–267.e6. DOI: 10.1016/j.molcel.2020.11.015.

Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Schöler, E. v. Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schübeler. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480(7378)**: 490–495. DOI: 10.1038/nature10716.

Stegle, O., K. J. Denby, E. J. Cooke, D. L. Wild, Z. Ghahramani, and K. M. Borgwardt. 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. DOI: 10.1089/cmb.2009.0175.

Stegle, O., S. A. Teichmann, and J. C. Marioni. 2015. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16(3)**: 133–145. DOI: 10.1038/nrg3833.

Stern, C. 2004. Gastrulation: From Cells to Embryo. Cold Spring Harbor Laboratory Series. Cold Spring Harbor Laboratory Press. ISBN: 978-0-87969-707-5.

Stockwell, P. A., A. Chatterjee, E. J. Rodger, and I. M. Morison. 2014. DMAP: differential methylation analysis package for RRBS and WGBS data. *Bioinformatics (Oxford, England)* **30(13)**: 1814–1822. DOI: 10.1093/bioinformatics/btu126.

Stoeckius, M., C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* **14(9)**: 865–868. DOI: 10.1038/nmeth.4380.

Stuart, T. and R. Satija. 2019. Integrative single-cell analysis. *Nature Reviews Genetics* **20(5)**: 257–272. DOI: 10.1038/s41576-019-0093-7.

Stuart, T., A. Srivastava, S. Madad, C. A. Lareau, and R. Satija. 2021. Single-cell chromatin state analysis with Signac. *Nature methods* **18(11)**: 1333–1341. DOI: 10.1038/s41592-021-01282-5.

Svensson, V., S. A. Teichmann, and O. Stegle. 2018. SpatialDE: identification of spatially variable genes. *Nature Methods* **15(5)**: 343–346. DOI: 10.1038/nmeth.4636.

Talbert, P. B., M. P. Meers, and S. Henikoff. 2019. Old cogs, new tricks: the evolution of gene expression in a chromatin context. *Nature Reviews Genetics* **20(5)**: 283–297. DOI: 10.1038/s41576-019-0105-7.

Tee, W.-W. and D. Reinberg. 2014. Chromatin features and the epigenetic regulation of pluripotency states in ESCs. *Development* **141(12)**: 2376–2390. DOI: 10.1242/dev.096982.

The HDF Group. 1997. Hierarchical Data Format, version 5.

Thornton, C. A., R. M. Mulqueen, K. A. Torkenczy, A. Nishida, E. G. Lowenstein, A. J. Fields, F. J. Steemers, W. Zhang, H. L. McConnell, R. L. Woltjer, A. Mishra, K. M. Wright, and A. C. Adey. 2021. Spatially mapped single-cell chromatin accessibility. *Nature Communications* **12(1)**: 1274. DOI: 10.1038/s41467-021-21515-7.

Titsias, M. 2009. Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. **5** Proceedings of Machine Learning Research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR: 567–574.

Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32(4)**: 381–386. DOI: 10.1038/nbt.2859.

Van den Berge, K., H. Roux de Bézieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement. 2020. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications* **11(1)**: 1201. DOI: 10.1038/s41467-020-14766-3.

Venkatesh, S. and J. L. Workman. 2015. Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology* **16(3)**: 178–189. DOI: 10.1038/nrm3941.

Vermeulen, M., K. W. Mulder, S. Denissov, W. W. M. P. Pijnappel, F. M. A. v. Schaik, R. A. Varier, M. P. A. Baltissen, H. G. Stunnenberg, M. Mann, and H. T. M. Timmers. 2007. Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* **131(1)**: 58–69. DOI: 10.1016/j.cell.2007.08.016.

Viger, R. S., S. M. Guittot, M. Anttonen, D. B. Wilson, and M. Heikinheimo. 2008. Role of the GATA Family of Transcription Factors in Endocrine Development, Function, and Disease. *Molecular Endocrinology* **22(4)**: 781–798. DOI: 10.1210/me.2007-0513.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17** 261–272. DOI: 10.1038/s41592-019-0686-2.

Wang, D., L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia, and S. Liu. 2012a. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28(5)**: 729–730. DOI: 10.1093/bioinformatics/bts013.

Wang, H., M. T. Maurano, H. Qu, K. E. Varley, J. Gertz, F. Pauli, K. Lee, T. Canfield, M. Weaver, R. Sandstrom, R. E. Thurman, R. Kaul, R. M. Myers, and J. A. Stamatoyannopoulos. 2012b. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research* **22(9)**: 1680–1688. DOI: 10.1101/gr.136101.111.

Wang, L., J. Zhang, J. Duan, X. Gao, W. Zhu, X. Lu, L. Yang, J. Zhang, G. Li, W. Ci, W. Li, Q. Zhou, N. Aluru, F. Tang, C. He, X. Huang, and J. Liu. 2014. Programming and Inheritance of Parental DNA Methylomes in Mammals. *Cell* **157(4)**: 979–991. DOI: 10.1016/j.cell.2014.04.017.

Wang, Z., C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* **40(7)**: 897–903. DOI: 10.1038/ng.154.

Wang, Z., M. Gerstein, and M. Snyder. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10(1)**: 57–63. DOI: 10.1038/nrg2484.

Warden, C. D., H. Lee, J. D. Tompkins, X. Li, C. Wang, A. D. Riggs, H. Yu, R. Jove, and Y.-C. Yuan. 2013. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic Acids Research* **41(11)**: e117. DOI: 10.1093/nar/gkt242.

Weber, M., I. Hellmann, M. B. Stadler, L. Ramos, S. Pääbo, M. Rebhan, and D. Schübeler. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics* **39(4)**: 457–466. DOI: 10.1038/ng1990.

Wilcoxon, F. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1(6)**: 80–83. DOI: 10.2307/3001968.

Wilks, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics* **9(1)**: 60–62.

Wolf, F. A., P. Angerer, and F. J. Theis. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19(1)**: 15. DOI: 10.1186/s13059-017-1382-0.

Workman, J. L. 2006. Nucleosome displacement in transcription. *Genes & Development* **20(15)**: 2009–2017. DOI: 10.1101/gad.1435706.

Wu, J., B. Huang, H. Chen, Q. Yin, Y. Liu, Y. Xiang, B. Zhang, B. Liu, Q. Wang, W. Xia, W. Li, Y. Li, J. Ma, X. Peng, H. Zheng, J. Ming, W. Zhang, J. Zhang, G. Tian, F. Xu, Z. Chang, J. Na, X. Yang, and W. Xie. 2016. The landscape of accessible

chromatin in mammalian preimplantation embryos. *Nature* **534(7609)**: 652–657. DOI: 10.1038/nature18606.

Xiang, Y., Y. Zhang, Q. Xu, C. Zhou, B. Liu, Z. Du, K. Zhang, B. Zhang, X. Wang, S. Gayen, L. Liu, Y. Wang, Y. Li, Q. Wang, S. Kalantry, L. Li, and W. Xie. 2020. Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency. *Nature Genetics* **52(1)**: 95–105. DOI: 10.1038/s41588-019-0545-1.

Yamazaki, T., Y. Hatano, R. Taniguchi, N. Kobayashi, and K. Yamagata. 2020. Editing DNA Methylation in Mammalian Embryos. *International Journal of Molecular Sciences* **21(2)**: 637. DOI: 10.3390/ijms21020637.

Yan, F., D. R. Powell, D. J. Curtis, and N. C. Wong. 2020. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21(1)**: 22. DOI: 10.1186/s13059-020-1929-3.

Yang, J., C. A. Penfold, M. R. Grant, and M. Rattray. 2016. Inferring the perturbation time from biological time course data. *Bioinformatics* **32(19)**: 2956–2964. DOI: 10.1093/bioinformatics/btw329.

Yin, Y., E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schübeler, C. Vinson, and J. Taipale. 2017. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356(6337)**: eaaj2239. DOI: 10.1126/science.aaj2239.

Yoder, J. A., C. P. Walsh, and T. H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*. Epigenetics **13(8)**: 335–340. DOI: 10.1016/S0168-9525(97)01181-5.

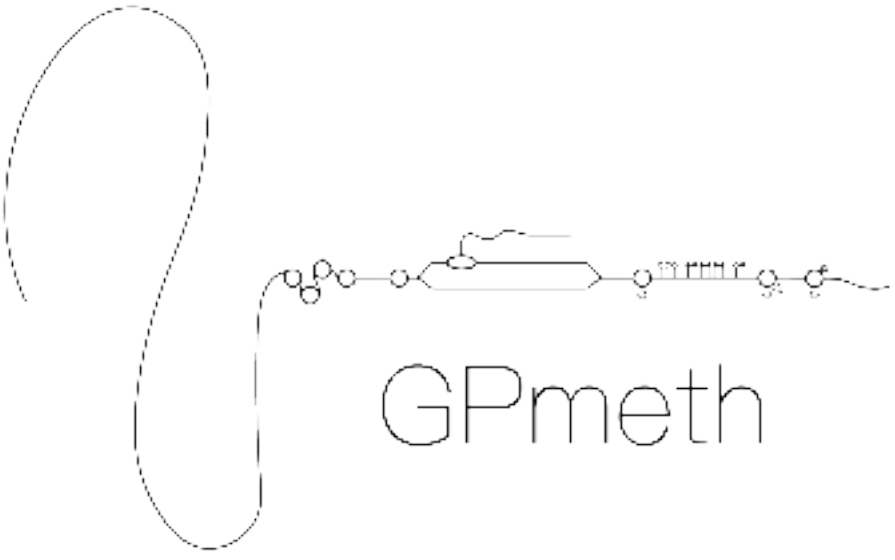
Zhang, B., H. Zheng, B. Huang, W. Li, Y. Xiang, X. Peng, J. Ming, X. Wu, Y. Zhang, Q. Xu, W. Liu, X. Kou, Y. Zhao, W. He, C. Li, B. Chen, Y. Li, Q. Wang, J. Ma, Q. Yin, K. Kee, A. Meng, S. Gao, F. Xu, J. Na, and W. Xie. 2016. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature* **537(7621)**: 553–557. DOI: 10.1038/nature19361.

Zhang, K., N. R. Zemke, E. J. Armand, and B. Ren. 2024. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nature Methods*: 1–11. DOI: 10.1038/s41592-023-02139-9.

Zhang, W., T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology* **16(1)**: 14. DOI: 10.1186/s13059-015-0581-9.

Zhang, Y., Y. Xiang, Q. Yin, Z. Du, X. Peng, Q. Wang, M. Fidalgo, W. Xia, Y. Li, Z.-a. Zhao, W. Zhang, J. Ma, F. Xu, J. Wang, L. Li, and W. Xie. 2018. Dynamic epigenomic landscapes during early lineage specification in mouse embryos. *Nature Genetics* **50(1)**: 96–105. DOI: 10.1038/s41588-017-0003-x.

Ziller, M. J., H. Gu, F. Müller, J. Donaghey, L. T.-Y. Tsai, O. Kohlbacher, P. L. De Jager, E. D. Rosen, D. A. Bennett, B. E. Bernstein, A. Gnirke, and A. Meissner. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500(7463)**: 477–481. DOI: 10.1038/nature12433.


GPmeth