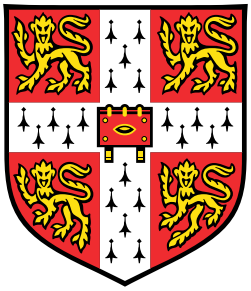
Statistical methods for the integrative analysis of single-cell multi-omics data



Ricardo Argelaguet Calado

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

September 2020

Declaration

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the individual declarations at the beginning of each chapter. To further indicate the parts of the thesis for which I used data of others or for which others were involved in interpreting the results, I use the pronoun "we". For the parts of my thesis that are purely my own work, I use the pronoun "I". The contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 60,000 words exclusive of tables, footnotes, bibliography, and appendices and contains less than 150 figures.

Ricardo Argelaguet Calado
September 2020

Abstract

Title
Statistical methods for the integrative analysis of single-cell multi-omics data

Name
Ricardo Argelaguet Calado

Summary
Single-cell profiling techniques have provided an unprecedented opportunity to study cellular heterogeneity at the molecular level. This represents a remarkable advance over traditional bulk sequencing methods, particularly to study lineage diversification and cell fate commitment events in heterogeneous biological processes. While the large majority of single-cell studies are focused on quantifying RNA expression, transcriptomic readouts provide only a single dimension of cellular heterogeneity. Recently, technological advances have enabled multiple biological layers to be probed in parallel one cell at a time, unveiling a powerful approach for investigating multiple dimensions of cellular heterogeneity. However, the increasing availability of multi-modal data sets needs to be accompanied by the development of suitable integrative strategies to fully exploit the data generated. In this thesis I worked in collaboration with different research groups to introduce innovative experimental and computational strategies for the integrative study of multi-omics at single-cell resolution.

The first contribution is the development of scNMT-seq, a protocol for the simultaneous profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. I demonstrate how this assay provides a powerful approach for investigating regulatory relationships between the epigenome and the transcriptome within individual cells.

The second contribution is Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of multi-omics data sets. MOFA is a Bayesian latent variable model that can be viewed as a statistically rigorous generalization of Principal Component Analysis to multi-omics data. The method provides a principled approach to retrieve, in an unsupervised manner, the underlying sources of sample heterogeneity while at the same time disentangling which axes of heterogeneity are shared across multiple modalities and which are specific to individual data modalities.

The third contribution is the generation of a comprehensive molecular roadmap of mouse gastrulation at single-cell resolution. We employed scNMT-seq to simultaneously profile RNA expression, DNA methylation and chromatin accessibility for hundreds of cells, spanning multiple time points from the exit from pluripotency to primary germ layer specification. Using MOFA, and other tools, I performed an integrative analysis of the multi-modal measurements, revealing novel insights into the role of the epigenome in regulating this key developmental process.

The fourth contribution is an extended formulation of the MOFA model tailored to the analysis of large-scale single-cell data with complex experimental designs. I extended the model to incorporate

a flexible regularisation that enables the joint analysis of multiple omics as well as multiple sample groups (batches and/or experimental conditions). In addition, I implemented a GPU-accelerated stochastic variational inference framework, thus enabling the scalable analysis of potentially millions of samples.

Acknowledgements

My supervisor John Marioni for his guidance and continuous support throughout my PhD.

My supervisor Oliver Stegle for hiring me as a MSc student and introducing me to the exciting field of factor analysis models and single-cell genomics.

Wolf Reik, Stephen Clark, Tim Lohoff and Carine Stapel for being the best experimental collaborators one could ask for.

Damien Arnol, for for making my PhD journey much more fun.

Andreas Kapourani, for insightful discussions on Bayesian statistics and for being an excellent host in Greece.

Marc Jan Bonder, Davis McCarthy, Daniel Seaton and Yuanhua Huang for being excellent role models.

Leah Rosen, for providing me with much needed motivation in my last year.

The entire MOFA team: Britta Velten, Damien Arnol, Danila Bredikhin and Yonatan Deloro, for making me feel so proud of the tool we developed.

Bas Haak, for introducing me to the world of the microbiome.

The Robinson College community, for being my second home.

The Cambridge Blues Basketball Team, for unforgettable trainings, games and Varsity matches.

My friends, for being there.

Last, but not least, my family and my girlfriend Mari for your patience and unconditional support, this thesis is dedicated to you.

Table of contents

Chapter 1

Introduction

1.1 Introduction to single-cell sequencing

Next-generation sequencing technologies have revolutionised the study of biological systems by enabling the genome-wide profiling of molecular layers in an unbiased manner, including the genome [92] the epigenome [93] and the transcriptome [175, 22, 215, 210]. Traditionally, bulk sequencing approaches were used to profile a large number of cells at once and report an average molecular readout. However, these methods are unable to capture differences between individual cells and are thus of limited use when studying heterogeneous biological processes [99, 223, 227]. The gradual development of low-input sequencing technologies resulted in an explosion of single-cell sequencing technologies, most of which focused on profiling the transcriptome. In contrast to bulk protocols, single-cell technologies provide an unprecedented opportunity to study the molecular variation associated with cellular heterogeneity, lineage diversification and cell fate commitment [152].

The field of single-cell sequencing has largely been driven by the quantification of the messenger RNA (mRNA). In less than a decade, the field of single-cell transcriptomics has experienced an exponential growth of scale in terms of number of cells profiled, driven by incremental optimisations of reagent volumes, decreases in consumable costs, as well as intelligent innovations in the capture, separation, and barcoding of cells [294]. The earliest high-throughput single-cell RNA sequencing (scRNA-seq) technologies were published between 2009 and 2011, yielding a handful of cells. In 2019, there are studies that have achieved the astonishing milestone of profiling the transcriptome for more than a million cells in a single experiment [49]. With the development of efficient commercial platforms, the maturation of scRNA-sequencing technologies has provided major insights on the study of lineage diversification and cell fate commitment [152, 99, 223, 227]. In 2020, we are at the stage of a major endeavour to generate transcriptomic atlases for different tissues, embryos and even entire adult organisms. The most ambitious of all is the Human Cell Atlas, aimed at building a reference map for all cells in the human body [249].

1.1.1 Single-cell RNA sequencing

scRNA-seq protocols differ extensively in terms of scalability, cost and sensitivity [294, 161]. Broadly speaking, they can be classified into plate-based and droplet-based methods. In plate-based methods such as CEL-seq [109] and Smart-seq [244, 231], cells are isolated using micropipettes or flow cytometry into individual wells of a plate, where the library preparation is performed. Although plate-based strategies have limitations in terms of throughput and scalability, their main advantage is the higher quality of libraries and the full length transcript information (in the case of Smart-seq)

which enables a more accurate quantification of splicing variants [124], allele-specific expression [73] and transcription-degradation kinetics [160].

Droplet-based methods are based on the use of droplet microfluidics technology [338]. By capturing cells in individual droplets, each containing all necessary reagents for library preparation, this protocol allows the profiling of thousands of cells in a single experiment. This class of methods include InDrop [150, 344], Drop-seq [191] and the commercial 10x Genomics Chromium platform [342]. As a trade-off, the increasing throughput of droplet-based approaches comes at the expense of reduced sensitivity [343, 322, 295].

More recently, a third type of scRNA-seq methodology emerged based on a combinatorial cellular indexing strategy [47, 258, 49], which has permitted the sequencing of more than a million cells in a single experiment for a fraction of the cost of other methods, albeit at the cost of much lower sensitivity.

1.1.2 Single-cell sequencing of the epigenome

While the vast majority of single-cell technologies are focused on quantifying RNA expression, transcriptomic readouts provide a single dimension of cellular heterogeneity and hence contain limited information on the molecular determinants of phenotypic variation [254]. Consequently, gene expression markers have been identified for a myriad of biological systems, but the role of the accompanying epigenetic changes in driving cell fate decisions remains poorly understood [99, 140, 28]. Significant effort has been placed to obtain epigenetic measurements at single-cell resolution by adapting bulk methods for small quantities of input material, a strategy thas has been successful for a variety of molecular layers, including DNA methylation [274], chromatin accessibility [69, 50, 57], histone modifications [157] and chromatin conformation [157].

DNA methylation

DNA methylation is a stable epigenetic modification that is strongly associated with transcriptional regulation and lineage diversification in both developmental and adult tissues [132, 226, 165, 275]. Its classical roles include the silencing of repetitive elements, inactivation of the X chromosome, gene imprinting, and repression of gene expression [135]. Consequently, the disruption of the DNA methylation machinery is associated with multiple dysfunctions, including cancer [26], autoimmune diseases [178] and neurological disorders [10].

Protocols for profiling DNA methylation in single cells have emerged from their bulk counterparts, most notably bisulfite sequencing (BS-seq) [274, 102, 98, 90]. The underlying principle of BS-seq is the treatment of the DNA with sodium bisulfite before DNA sequencing, which converts unmethylated cytosine (C) residues to uracil (and after PCR amplification, to thymine (T)), leaving 5-methylcytosine residues intact. The resulting C→T transitions can then be detected by DNA sequencing [93, 60, 58]. Nevertheless, the high degree of DNA degradation caused by the purification steps and the bisulfite treatment impaired the use of conventional BS-seq with low starting amounts

of DNA. To address this problem, [274] adapted the post-bisulfite adaptor tagging (PBAT) protocol by doing multiple rounds of 3' random primer amplification. In addition, when the bisulfite treatment is performed before ligation of adaptors, rather than afterwards, loss of adapter-tagged molecules is minimised, demonstrating the potential to use scBS-seq from low-input material.

Chromatin accessibility

In eukaryotes, the genome is packed into a compact complex of DNA, RNA and proteins called chromatin. Several layers of chromatin condensation have been identified, the fundamental of which being the nucleosome, which consists of a string of $\approx$ 150bp of DNA wrapped around histone proteins, with $\approx$ 80bp of DNA connecting them [151, 312]. The positioning of the nucleosomes along the DNA provide an important layer of gene regulation, mostly by exposing or sheltering transcription factor binding sites [130]. In general, active regulatory regions tend to have low occupancy of nucleosomes, whereas inactive regions show a high density of nucleosomes [289]. Thus, the profiling of DNA accessibility and transcription factor footprints represents an important dimension to understand the regulation of gene expression.

Traditionally, three main experimental approaches have been used to profile bulk chromatin accessibility in a genome-wide and high-throughput manner [219] (Figure 1.1): DNase sequencing (DNase-seq) [279], transposase-accessible chromatin followed by sequencing (ATAC-seq) [41] and Nucleosome Occupancy and Methylome-sequencing (NOMe-seq) [139].

- **DNase-seq**: cells are incubated with DNAse I, an enzyme that in low concentrations cuts nucleosome-free regions, hence releasing accessible sites that are subsequently sequenced [279]. Although this methodology became one of the gold standards to profile chromatin accessibility by the ENCODE consortium [62, 302], it has now been reported that DNase I introduces significant cleavage biases, thus affecting the reliability of transcription factor footprints inferred from the DNase-seq data [112].

- **ATAC-seq**: cells are incubated with a hyperactive mutant Tn5 transposase, an enzyme that inserts artifical sequencing adapters into nucleosome-free regions. Subsequently, the adaptors are purified, PCR-amplified and sequenced. In the recent years it has displaced DNase-seq as the *de facto* method for profiling chromatin accessibility due to its fast and sensitive protocol [39, 312].

- **NOMe-seq**: follows a very different strategy than the previous technologies. Cells are incubated with a GpC methyltransferase (M.CviPI), which labels accessible (or nucleosome depleted) GpC sites by DNA methylation. In mammalian genomes, cytosine residues in GpC dinucleotides are methylated at a very low rate [144]. Hence, after M.CviPI treatment, GpC methylation marks can be interpreted as direct read outs for chromatin accessibility. [139]. NOMe-seq has a range of appealing properties in comparison with count-based methods such as ATAC-seq or DNAseq-seq. First, one can obtain simultaneous information of CpG DNA methylation with little additional cost, permitting the user to effectively measure two molecular layers for the price of one. Second, the resolution of the method is determined by

the frequency of GpC sites within the genome (≈ 1 in 16 bp), rather than the size of a library fragment (usually >100 bp). This allows the quantification of nucleosome positioning and transcription factor footprints at high resolution [139, 237]. Third, non-sequenced fragments can be easily discriminated from inaccessible chromatin. This implies that lowly accessible sites will not suffer from increased technical variation (due to low read counts) compared to highly accessible sites. The downsides of the approach are the high sequencing depth requirements and the need to discard read outs from GCG positions (21% of all CG sites) and CGC positions (27%), as I will discuss later in this thesis.

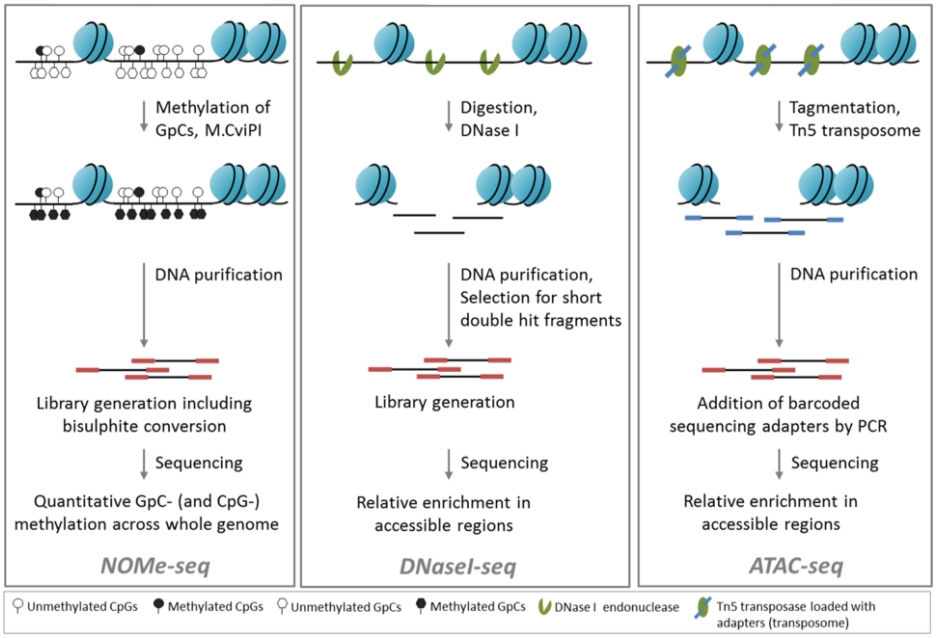


Figure 1.1: High-level overview of the workflows for the three main chromatin accessibility assays: NOMe-seq, DNase-seq and ATAC-seq. Reprinted from [219] with minor modifications.

As with DNA methylation, ATAC-seq [40], NOMe-seq [237] and DNase-seq [131] have also been adapted for single cells. Due to its cost-effective strategy, single-cell ATAC-seq (scATAC-seq) has become the most popular technique to profile open chromatin [69, 50, 57]. Compared with bulk ATAC-seq, scATAC-seq libraries are notably sparse. In a saturated library, [69] reported a range of ≈ 500 to ≈ 70,000 mapped reads per cell, with a median of ≈ 2500. As the authors report, this represents less than 25% of the molecular complexity expected from 500-cell bulk experiments. Yet, despite the low coverage, the authors showed that cell-type mixtures can be confidently deconvoluted. Later, in a pioneer effort, [68] generated an atlas of chromatin accessibility for different mouse tissues, defining the first *in vivo* landscape of the regulatory genome at single-cell resolution.

1.1.3 Multi-modal single-cell sequencing

Cellular phenotypes result from the combination of multiple levels of cellular regulation. Undoubtedly, no single "-omics" technology can capture the intricacy of complex molecular mechanisms, but the

collective information has the potential to draw a more comprehensive picture of biological processes [110, 254].

The profiling of multi-omic readouts at the bulk level is relatively simple, as the same tissue can be dissociated into different aliquots, where each assay can be performed independently [254]. This strategy is also used with single-cell assays, but it has the important downside that the different molecular layers cannot be unambiguously matched, hence limiting the insights that can be inferred from the data. Therefore, many single-cell multi-omics technologies are being developed, which seek to obtain multiple molecular readouts from the same cell. The development of these technologies will help us understand the fundamental regulatory principles that connect the different molecular layers. In addition, integrative analyses that simultaneously pool information across multiple data modalities (-omics) and across multiple studies promise to deliver more comprehensive insights into the complex variation that underlies different cellular populations.

Notably, the early success and rapid development of single-cell multi-modal methods has led to their recognition as Method of the year in 2019 by the journal *Nature Methods* [203]. However, their development is still in pilot stages and at the time of writing there is no commercial platform available, limiting its widespread use by the community. As reviewed in [290, 55], multi-modal measurements can be obtained using three broad strategies:

- **Application of a non-destructive assay before a destructive assay**: a prominent example of this is the sorting of cells based on protein surface markers using (multiparameter) fluorescence-activated cell sorting (FACS) followed by high-throughput sequencing [228]. Although simple and efficient, this approach requires prior knowledge of protein surface markers, and is limited by the spectral overlap of fluorescence reporters.

- **Physical isolation of different cellular fractions followed by high-throughput sequencing**: this technique was pioneered with the introduction of genome and transcriptome sequencing (G&T-seq) [189]. After cell lysis, the mRNA fraction is separated from the genomic DNA fraction using biotinylated or paramagnetic oligo(dT) beads, followed by the independent sequencing of the mRNA and the DNA. This strategy allows for the simultaneous profiling of transcriptomic measurements with (epi)-genomic measurements, including DNA sequence, copy number variation, DNA methylation or chromatin accessibility [189, 122, 12, 123].

- **Conversion of different molecular layers to a common format that can be measured using the same readout**: prominent examples of this are the simultaneous measurement of surface proteins and mRNA expression as in Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq [288]) and RNA expression and protein sequencing assay (REAP-seq [230]). The idea is to incubate cells with antibodies tagged with oligonucleotides that target specific protein surface proteins. This allows both protein surface markers and mRNA levels to be simultaneously measured using a single sequencing round. Notably, this strategy is significantly more powerful than FACS, as the DNA barcodes can be resolved at the sequence level with much higher sensitivity than using fluorescence markers. A second prominent example is NOMe-seq, described in Section 1.1.2. By labelling accessibile GpC sites

with DNA methylation marks, one can simultaneously measure endogenous DNA methylation and chromatin accessibility using a single bisulfite sequencing assay.

Although single-cell multi-modal technologies have proven successful, they still face numerous difficulties, both on the experimental and the computational front, including limited scalability, low coverage and high levels of technical noise. These difficulties, which are also inherent to single-cell uni-modal technologies, generally get exacerbated when doing multi-modal profiling. To quote Cole Trapnell, one of the pioneers of single-cell data analysis: *When you do a multi-omic assay, you're combining all the bad things from multiple protocols* [81]. A clear example of these challenges is sci-CAR [48], a combinatorial indexing strategy that combines scRNA-seq and scATAC-seq to profile gene expression and chromatin accessibility in the same cell. This is a promising approach that reported, for the first time, the profiling of both modalities in thousands of cells. However, the chromatin accessibility modality yielded $\approx$ 10-fold less complexity than previous (already sparse) uni-modal scATAC-seq experiments.

I envision that in the next few years significant efforts will be made to obtain more scalable and cheaper multi-modal measurements from single cells. However, as cost and scalability remain a barrier for high-resolution multi-modal technologies, the development of computational methods that are capable of uncovering biological signal across multiple data modalities while overcoming the technical biases and missing information that are inherent to single-cell experiments, will be a cornerstone of data analysis.

1.2 Single-cell transcriptomics data analysis

From a computational perspective, the rapid development of single-cell technologies has introduced unprecedented challenges for the statistical community, and novel computational methods need to be developed (or adapted) for interrogating the data generated [283]. The vast majority of methods are focused on RNA expression, spanning multiple tasks that include normalisation [159], feature selection [308], differential expression [143], clustering [146], cell type recognition [1], pseudotime inference [106], detection of gene regulatory networks [2] and batch correction [105], among others. Analysis tools have been wrapped into popular platforms such as *Seurat* [46], the *Bioconductor* class *SingleCellExperiment* [9] and *Scanpy* [328].

In this section I will provide a brief overview of a typical scRNA-seq analysis pipeline, paying particular attention to the methods that I have used throughout this thesis.

1.2.1 Read alignment and gene expression quantification

The first step in the computational pipeline is to demultiplex the DNA barcodes in order to identify reads that originate from the same cell. This is particularly important when multiple experiments are pooled into a common sequencing library. This task is significantly more complex than in bulk

data, owing to the large number of cells and the high rates of errors that can introduce nucleotide missmatches [299].

Subsequently, trimmed reads are aligned to the appropriate reference transcriptome. Gene expression is represented as an integer matrix of counts, with rows representing genomic features (typically genes) and the columns representing individual cells.

1.2.2 Quality control

Incomplete cell lysis or failures during library preparation may result in poor quality cells that need to be removed for a successful downstream analysis. Typical quality control metrics are the total number of reads detected per cell, the number of genes expressed and the fraction of mitochondrial genes. Cells that are outliers for some of these metrics are filtered out. Importantly, even though there is a generic strategy to assess the quality control for scRNA-seq samples, the specific thresholds vary between datasets and technologies, and care must be taken to always visualise the quality control metrics [184].

A common source of technical variability in single-cell experiments is the existence of doublets, which occurs when multiple cells co-locate in the same well or in the same droplet and are thus assigned the same cell barcode. This results in cells that appear as mixtures of different cellular populations and can be mistaken for non-existing intermediate populations or transitory states. Thus, it is important to remove doublets so that they do not compromise the downstream analysis. In small-scale plate-based technologies, most doublets can be excluded simply by microscope inspection, but in large-scale droplet-based technologies one needs to adopt data-driven heuristics to exclude multiplet libraries [198].

1.2.3 Normalisation

The first step of quality control is essential to remove poor quality cells, but the number of detected molecules and transcripts can vary widely even among cells that pass the quality control. This variability is not only due to biological heterogeneity but may also be technical. Any of the library preparation steps may lead to technical variability, such as PCR amplification biases, differences in RNA capture and reverse transcription efficiency. In addition, the stochasticity of the amplification process produces *dropout* events, in which no read counts are observed for genes that are expressed [315]. There has been considerable debate on how to deal with the high proportion of zero counts, and multiple statistical frameworks have been devised, including zero-inflated negative binomial models [253]. However, recent reports suggest that droplet-based scRNA-seq measurements can be explained by simple Poisson statistics [293, 266].

Regardless of the statistical model, data normalisation steps are necessary to eliminate (or at least reduce) the technical variation. Methods that were developed for bulk RNA-seq, including *TMM* [257] and *DEseq2* [181] are not successful for scRNA-seq owing to the large number of zeros that dominate the gene expression matrix.

In this thesis I used the methodology implemented in the *scran* package [159]. Briefly, this normalisation procedure divides the gene counts by a size factor per cell and subsequently applies a log transformation with pseudocount on each observation. The essential innovation for single-cell data is to pool expression values from multiple cells (resulting in fewer zeros) and to subsequently deconvolve the cell-specific size factors using a linear system of equations.

Recent work has suggested that global size factors do not effectively normalise all genes at the same time, and different groups of genes require specific size factors in order to remove technical biases [104]. In this thesis I have not explored this approach, but it showcases how data normalisation is still an open and debated topic.

1.2.4 Dimensionality reduction

A key principle of biological datasets is that covariation patterns between the features (i.e. genes) results from differences in underlying processes that can be inferred and interpreted. This key assumption sets off an entire statistical framework of exploiting the redundancy encoded in the data set to reduce the dimensionality of the data in an unsupervised fashion.

Principal Component Analysis (PCA) is the most popular technique for dimensionality reduction of scRNA-seq data [184]. A typical analysis pipeline performs clustering, graph inference and other downstream analyses on the (denoised) latent PCA space defined by the top N principal components (where components are ranked by variance explained). Importantly, by maximising the variance explained, PCA implicitly assumes a normal distribution for each feature. Therefore, it is important that the data is log transformed, which, as outlined above, converts integer counts into continuous measurements, before PCA is performed. In addition, the log transformation prevents signal being driven by a small number of extremely highly-expressed genes (because in the raw counts the variance of each gene is proportional to its mean expression).

PCA defines a linear transformation from the high-dimensional space to the low-dimensional space where each component captures an orthogonal source of variation. Capturing the biological signal in most single-cell datasets require a relatively high number of components. Unfortunately, humans do not have the ability to make visual representations of more than three dimensions at the same time, so for the purposes of visualisation further dimensionality reduction is typically applied using non-linear methods, including t-Distributed Stochastic Neighbor Embedding (t-SNE) [188] and Uniform Manifold Approximation and Projection (UMAP) [199]. Both methods have been extensively applied, although UMAP is gaining popularity for larger datasets because it is better at preserving the global structure than is t-SNE, which is aimed at solely preserving local structure.

1.2.5 Clustering

Unsupervised clustering is arguably one of the most powerful applications of single-cell genomics, as it underpins the ability to define cell types in a coherent, systematic and unbiased manner. Although clustering is still largely empirical and no strong consensus exists on the methodology

and the parameters, it is applied in virtually any single-cell data set [145]. The most popular clustering algorithm has traditionally been k-means, which iteratively identifies k cluster centroids, and assigns each cell to the nearest centroid. This method is simple, fast and efficient for medium-sized datasets. For large-scale datasets, however, the use of community-detection algorithms on coarse-grained graphs has become more popular [184]. Briefly, the first step of commnity-detection methods is to build a k-nearest-neigbourhood graph using a cell-to-cell similarity metric, where each node corresponds to one cell. Then, tighly connected communities are detected by maximising a modularity score, where the modularity quantifies the assignment of nodes to communities when contrasted to a random network.

1.2.6 Inference of developmental trajectories

In many biological systems, and particularly during embryonic development, cells display a continuous spectrum of states and so discrete clustering may be inappropriate. Due to the destructive nature of single-cell assays, experiments are not capable of measuring the changes individual cells undergo over time. However, differentiating cells are typically asynchronised and display a continuous spectrum of molecular states that reflect the underlying trajectory. Computational methods have been developed to reconstruct this continuity using latent mathematical representations, and are often termed pseudotime methods [262]. The aim of pseudotime methods is to generate an ordering of cells according to some metric, which is usually (but not necessarily) some approximation of real time that is inferred from the data. A myriad of pseudotime methods have been developed, with tailored assumptions depending on the nature of the input data and the expected topology of the trajectory (linear, bifurcating, etc.), among other variables [262].

1.3 Integrative analysis of single-cell omics

Despite the explosion of statistical methods for scRNA-seq data analysis, to date only a few methods have been published with the aim of performing data integration across experiments and data modalities. This is currently defined as one of the grand challenges in single-cell data science [162].

1.3.1 Defining the common coordinate framework

The first step when performing data integration is to consider which coordinate framework can be used to anchor the different data modalities. This is generally dictated by the experimental design itself, and leads to three broad types of data integration strategies (Figure 1.2):

- Cells as the common coordinate (*vertical* integration): when the different data modalities are derived from the same cell in *matched* multi-omic assays. Examples of technologies that generate such data are single-cell Methylome & Transcriptome (scM&T-seq) [12], Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq) [288] and Single-nucleus chromatin accessibility and RNA expression sequencing (SHARE-seq) [187].

- Genomic features as the common coordinate framework (*horizontal* integration): when multiple data modalities of the same type are profiled in different sets of cells. We call this *non-matched* multi-omics and the main advantage is that it is significantly easier and cheaper to obtain than *matched* multi-omics, and as a result most of the current datasets to date belong in this category. An example of this would be performing scRNA-seq on cells from the same tissue across different groups of donors, where the set of genes represents the anchors.
- No common coordinate framework in the high-dimensional space (*diagonal* integration): when both cells and genomic features are different between experiments. An example of *horizontal* integration would be the profiling of RNA expression in one set of cells, and chromatin accessibility in an independent set of cells.

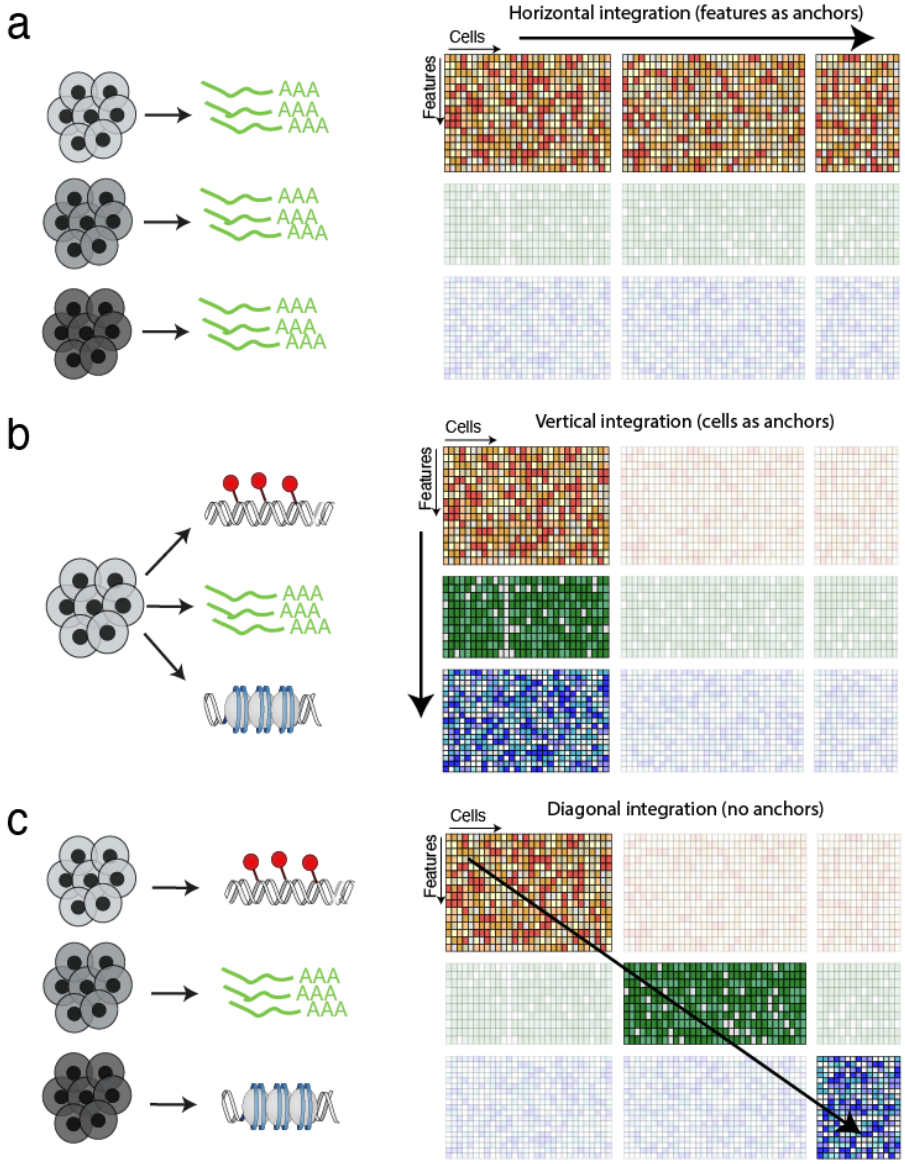


Figure 1.2: Defining the data integration strategy: choosing the common coordinate framework.
Schematic representation of (a) *Horizontal* integration, when features act as anchors (b) *Vertical* integration, when cells act as anchors (c) *Diagonal* integration, when no anchors exist.

1.3.2 Challenges of data integration

The joint analysis of multiple data sources must tackle numerous challenges, some of which are:

- **Heterogeneous data modalities**: measurements collected using different technologies generally exhibit heterogeneous statistical properties and have to be modelled under different statistical assumptions. For example, combining count (i.e. gene expression) and binary traits (i.e.somatic mutations) under the same statistical framework is not a trivial task. In the statistics community, this is commonly refered to as the multi-view learning problem [332, 170].

- **Overfitting**: as the number of molecular layers increases but the number of samples remains limited, modelling strategies need to account for the risk of overfitting, which can lead to poor generalisation. For example, scM&T-seq captures the methylation status for potentially millions of CpG sites, but the experimental designs are typically restricted to only a few hundred cells. This is a classic case of a large p and small n problem in high-dimensional statistics.

- **Discriminating biological vs technical sources of variation**: multi-omics datasets typically contain undesired sources of heterogeneity, both technical and biological [254]. Prominent examples are batch effects or cell cycle variation, respectively. If not accounted for, such strong sources of variability can hide the signal of interest [42]. Therefore, understanding and account for undesired cell-to-cell variation hsa to be performed before other computational pipelines are applied [202].

- **Missing data**: a major problem in some single-cell methodologies is the large amounts of missing information. For example, in a typical single-cell bisulfite experiment less than 10% of all CpG sites in the genome are measured [274]. This poses important challenges to some of the conventional statistical methods that do not handle missing information. Furthermore, assays differ in terms of how missing data is defined. For bisulfite sequencing methods, the missing values are distinguishable from the observed values. However, for other methods such as scRNA-seq or scATAC-seq, the absence of a sequencing read does not distinguish between the event that the genomic feature was not measured from that the readout was indeed zero [59].

- **Scalability**: as sequencing costs decrease and technologies improve, we anticipate that multi-modal datasets will follow a similar trend as scRNA-seq, where in the span of less than ten years the size of the experiments increased from the order of tens to milions of cells [294].

- **Assay noise**: because of the small amounts of starting material, single-cell technologies are inherently noisy and result in large amounts of technical noise [283]. Hence, in most cases, inspection of individual genes or cells tends to be unreliable. To overcome this challenge, computational frameworks are required to pool information across cells and/or genes to delinate the signal from noise and to obtain reliable statistical estimates [314]. Prominent examples of this are (empirical) Bayesian approaches that are able to borrow information

across cells and/or genes and propagate uncertainity when doing inference and predictions [143].

- **Principled validation of the model outputs**: The assessment of data integration outputs is one of the most challenging steps. In most cases, no ground truth is available and one has to assess the model fit by relying on quality control statistics and/or assessing the impact of the integration on downstream analysis tasks (i.e. differential expression, dimensionality reduction, clustering, etc.). The evaluation of the model fit is particularly hard for non-linear methods that can be prone to overfit.

1.3.3 Defining the methodology

Once the common coordinate framework is defined, one needs to choose the data integration strategy. These can fall into two classes: *local* and *global*, a notation inspired from integrative approaches that have been pursued at the bulk level [254].

Local analyses refer to associations between specific features across different molecular layers, with the aim of detecting putative interactions between them. Prominent examples are associations between genetic variants and gene expression (expression quantitative trait loci, eQTL) or correlations between the epigenetic status of putative regulatory elements and gene expression of nearby genes. The restriction to a local search space is often necessary, to ensure the problem is computationally tractable. For example, cis eQTL mapping is most prevalent because by testing only proximal genetic variants for each gene, the effects of the multiple testing burden is reduced [218]. Since such association analyses are typically performed per feature (across cells), *local* analyses generally require unambiguous matching between the modalities, and therefore require *matched* multi-modal assays, and thus belongs to the category of *vertical* analysis. Methodologically, most *local* analyses rely on different flavours of linear regression models, with different modelling assumptions depending on the nature of the molecular readouts. This can include non-gaussian likelihoods, sparsity assumptions to prevent overfitting or probabilistic terms to account for random effects. For example, linear mixed models (LMMs) are a popular framework for performing genetic analyses [208]. In a LMM, a random effect term is added to account for the population structure and relatedness between individuals, which may affect both the phenotype and the genotype, thus leading to spurious associations if not accounted for.

While useful for characterising genetic variants or identifying putative regulatory elements, *local* analyses have limited capacity to discover complex maps of molecular heterogeneity that result from interactions between genomic features. An alternative strategy for data integration is to exploit the full spectrum of measurements to identify cellular states defined by the coordinated action of multiple genomic elements. For example, cell cycle phase or pluripotency potential are cellular properties that are determined by gene regulatory networks and thus cannot be studied with *local* analyses. *Global* integration is typically (but not always) performed using unsupervised dimensionality reduction approaches that find common modes of variation between molecular layers. Alternatives have been proposed that perform transformations on each data type before merging them into a common similarity network, e.g. using kernel or graph-based approaches [163,

319]. Popular examples of global integration methods that have been adapted from the statistical literature are derived as different flavours of matrix factorisation: PCA, Canonical Correlation Analysis (CCA, implemented in Seurat [46]), Group Factor Analysis (MOFA [15, 14], introduced in this thesis), Projected Least Squares (PLS, implemented in DIABLO [273]) and Non-negative Matrix factorisation (NMF, implemented in LIGER [325]), among others. Although all of these methods share important similarities, the assumptions underlying each model are heavily dependent on the common coordinate framework adopted. As such, the output of each data integration method is the result of different assumptions and thus has specific challenges and diagnostics that must be addressed accordingly.

Horizontal integration

Horizontal integration strategies define features as the common anchors in *unmatched* experiments of the same type. This task is faced in most large-scale projects where sequencing data is generated across multiple batches, as uncontrollable differences in the experimental procedure result in systematic deviations in the observed RNA expression across the different batches. If left unaccounted for, these sources of technical variation can mask relevant biological variability and thus complicate the interpretation of the downstream analysis. *Horizontal* integration is currently the most common task, and it is typically regarded as a batch correction problem, where the aim is to remove undesired technical variation across batches while preserving the biological variation contained within each batch [309]. With the growing availability of reference atlases, epitomised by the Human Cell Atlas project [249], this is arguably one of the most important steps in a single-cell analysis pipeline.

Linear batch correction methods that were originally developed for bulk datasets such as *limma* [255] and *ComBat* [134]) are not successful for single-cell experiments, mainly because they assume identical (or at least, known) cell type composition across batches. In practice, however, the abundance of cellular subpopulations can vary even between biological replicates due to subtle differences in the library preparation or in the sampling procedures. As a consequence, the majority of *horizontal* integration methods developed for single-cell data rely on non-linear (or locally linear) strategies that account for differences in cell type compositions.

Several integrative methods have been developed and benchmarked. These include MNN [105], Seurat v3 [46], LIGER [325], Harmony [154], BBKNN [236], scVI [180], Conos [23], among others, which have been benchmarked in an independent study [185]. Despite sharing similar principles, these each employ different methodologies. In particular, MNN and Seurat v3 detect mutual nearest neighbors in a joint low-dimensional space, defined by either principal components (MNN) or canonical covariates (Seurat v3). LIGER, on the other hand, performs integrative NMF and disentangles dataset-specific factors versus shared factors, followed by the construction of a neighborhood graph using the shared factors. Harmony learns a cell-specific linear correction function by successive rounds of k-means clustering on a principal component space. BBKNN performs correction on a neighbourhood graph, which results in much faster computations at the expense of losing single-cell resolution. Finally, scVI is a Bayesian variational autoencoder with a probabilistic formulation which includes random variables that account for batch-specific variation.

Most of these methods also share a common set of challenges. First, a classical problem of non-linear integration methods is over-correction, which occurs when the batch correction vector is wrongly estimated and the algorithm forcibly merges non-matching subpopulations of cells [185]. This can occur for example when there are no common axes of biological variation between the datasets. Second, most methods perform data integration in a denoised latent space, typically using principal components or canonical covariates. This step undoubtedly improves most batch correction algorithms, but the high-dimensional observations (i.e. the gene expression counts) can be severely distorted as a result of the batch alignment, and other downstream gene-based analyses such as gene marker detection or differential expression analysis can be problematic [105].

Vertical integration

Vertical integration strategies take advantage of the unambiguous assignment between the molecular profiles in *matched* multi-modal experiments and thus define cells as anchors between data modalities. This facilitates the detection of co-variation patterns across features and permits two data integration strategies: gene-based *local* analysis and a cell-based *global* analysis.

Local analysis

In *local* analyses, the challenge is to distinguish true interactions between features from spurious associations that can result from global confounding effects. To correct for global confounding effects (both technical and biological) affecting the expression phenotype in eQTL mapping, methods such as Principal Component Analysis and PEER [284] are often used to identify factors that capture global expression trends, which can be added as covariates in the linear (mixed) model framework. Similarly, the use of a kinship matrix is used to account for global genotype effects that result from population substructure and individual relatedness. Mapping eQTL using single-cell genomics has led to the identification of cell type-specific eQTL in rare cell populations, which would have been masked using bulk assays [327]. Additionally, [67] combined differentiations of iPSCs across multiple donors and single cell expression profiles to show how eQTL influence expression dynamically along a differentiation trajectory. Single cell eQTL mapping is growing as a field, and it promises to provide an extra layer to our understanding of genetic regulation at the molecular level. As methods to assay various molecular traits at single cell resolution become more established, non-expression single cell QTL mapping, where genomic variants are associated with changes in DNA methylation, histone modifications or protein level at single cell resolution may also become routine.

Global analysis

When having a single data modality, PCA is the paradigmatic method for global analysis, and will be discussed in more detail in Chapter 3. Briefly, PCA infers an orthogonal projection of the data onto a low-dimensional space such that the variance explained by the projected data is maximised. The key for the popularity of PCA is its linearity assumption, which ensures that the resulting principal components are simple and interpretable. PCA has also been applied as an integrative method for multi-modal data by extracting principal components from each modality and subsequently comparing them. This approach was attempted in one of the first multi-modal datasets, where scM&T-seq was used to simultaneously profile RNA expression and DNA methylation on

61 embryonic stem cells [12]. The authors found that a small number of PCs derived from mRNA expression displayed significant correlations with PCs derived from DNA methylation, which suggests that some global sources of variation are preserved across data modalities, but a large fraction of the variation is uncorrelated. This simple analysis provides the intuition for some of the more advanced multi-omic integration methods aimed at performing variance decomposition across data modalities.

An alternative strategy has been to apply PCA after concatenation of the datasets, but this has important limitations when applied to datasets where the features are structured into non-overlapping views (referred to as multi-view data in the machine learning literature). First, PCA extracts components that maximise the variance explained, but it is difficult to quantify the contribution that each component has from each data modality. Second, in its vanilla implementation, PCA does not handle missing values and hence imputation is required when cells do not have measurements available in all data modalities. This is a frequent problem in *matched assays*, as cells might pass quality control for one data modality but not the other. Third, by maximising the variance explained, PCA implicitly assumes a normal distribution for each feature, and is not well suited for the integration of binary and count-based readouts.

Generalisations of PCA for the integration of multi-omics data have been devised by adapting multi-view learning methods from the statistics literature. Although most of these methods were originally devised for bulk data, the majority of them remain applicable to single-cell multi-modal data. This includes MOFA [15], JIVE [179], PLS [273], MCIA [201] and iNMF [325], all of which use different flavours of matrix factorisation to perform unsupervised dimensionality reduction. As I will discuss in this thesis, the matrix factorisation framework is very appealing due to its simplicity, interpretability, scalability and limited overfitting. This framework has also proven to be an excellent choice for extracting interpretable signatures from sparse and noisy observations such as single-cell measurements.

Diagonal integration

The third type of data integration problem occurs when no common coordinate framework exists in the high-dimensional space. This task is faced in *unmatched* experiments when different molecular layers are profiled in different subsets of cells, and is arguably the hardest type of integration. *Diagonal* integration methods are generally aimed at reconstructing a low-dimensional manifold that captures co-variation across data modalities. Thus, a critical assumption of this integrative strategy is the existence of a latent manifold with similar topology between the data modalities. For example, this could represent cells sampled from a common differentiation trajectory or cells sampled from a common set of discrete subpopulations.

The simplest strategy that has been employed to solve a *diagonal* integration task is to transform it into a simpler *horizontal* integration task. This can be achieved by summarising molecular measurements over genomic elements that can be unambiguously linked (i.e. RNA expression and promoter methylation). Using this strategy, methods such as LIGER [325] and Seurat [291] have been successful at integrating unmatched epigenetic and transcriptomic experiments from the same tissue, and even across different species. However, this strategy relies on fragile biological

assumptions and is prone to fail in scenarios where molecular layers are not strongly correlated. A good example is early embryonic development where promoter DNA methylation and/or chromatin accessibility are not as predictive of RNA expression [14] as in terminally differentiated cell types.

Alternatively, a few methods have attempted to solve this problem by reconstructing technology-invariant integrated latent spaces. The first method to be developed was MATCHER [324], a gaussian process latent variable model. However, this method relies upon the strong assumption that biological variation is defined by a unidimensional axis of variation. Some recent methods, including MMD-MA [176], SCIM [282] and UnionCom [51] have generalised MATCHER to account for complex multivariate trajectories. However, no independent benchmarking has yet been performed, and the biological insights extracted from these methods have been relatively limited.

1.4 Thesis overview

In this PhD thesis I sought to develop computational strategies for data integration for single-cell multi-omics. In particular my research focused on the *vertical* integration task, where cells are the common coordinate framework in *matched* assays.

In Chapter 2 I introduce single-cell nucleosome, methylation and transcription sequencing (scNMT-seq), an experimental protocol for the genome-wide profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. While some approaches have reported unbiased genome-wide measurements of up to two molecular layers, scNMT-seq allows, for the first time, the simultaneous profiling of three molecular layers at single cell resolution. We validate the assay using a simple prototype experiment, and we show how scNMT-seq can be used to study coordinated epigenetic and transcriptomic heterogeneity along a simple differentiation process.

In Chapter 3 I present Multi-Omics Factor Analysis (MOFA), a statistical framework for the integration of multi-omics datasets. MOFA is a latent variable model that offers a principled approach to explore, in a completely unsupervised manner, the underlying sources of sample heterogeneity in multi-omics data. After validating the model features using simulated data, we applied MOFA to a cohort of chronic lymphocytic leukaemia patients. In a quick unsupervised analysis, MOFA revealed the most important dimensions of disease heterogeneity, connected to clinical markers that are commonly used in practice. In a second application we show how MOFA can cope with noisy single-cell multi-modal data, identifying coordinated transcriptional and epigenetic changes along a differentiation process.

In Chapter 4 I discuss how we combined scNMT-seq and MOFA to study the role of epigenetic layers during mouse gastrulation, a critical embryonic stage that spans exit from pluripotency to primary germ layer specification. In this study we built the first triple-omics roadmap of mouse gastrulation, which enabled us to perform an integrative study that revealed novel insights on the dynamics of the epigenome. Notably, we show that cells committed to mesoderm and endoderm undergo widespread epigenetic rearrangements, driven by demethylation in enhancer marks and by concerted changes in chromatin accessibility. In contrast, the epigenetic landscape of ectodermal cells remains in a

default state, resembling earlier stage epiblast cells. This work provides a comprehensive insight into the molecular logic for a hierarchical emergence of the primary germ layers.

In Chapter 5 I propose an improved formulation of the MOFA framework (MOFA+) aimed at performing integrative analysis of large-scale (single-cell) datasets with complex experimental designs. We introduce key methodological developments, including a fast stochastic variational inference framework and multi-group generalisation in the structure of the prior distributions. All together, this allows MOFA+ to disentangle heterogeneity across sample groups (i.e. studies or experimental conditions) and data modalities (i.e. omics) in very large datasets. After benchmarking the new features using simulated data, we applied it to single-cell datasets of different scales and designs.

Finally, Chapter 6 summarises this thesis and provides an outlook of future research.

Chapter 2

Joint profiling of chromatin accessibility DNA methylation and transcription in single cells

In this Chapter I describe scNMT-seq, an experimental protocol for genome-wide profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. First, I show a validation of the quality of the molecular readouts, including a comparison with existing technologies. Subsequently, I showcase how scNMT-seq can be used to reveal coordinated epigenetic and transcriptomic heterogeneity along a differentiation process.

The work discussed in this Chapter results from a collaboration with the group of Wolf Reik (Babraham Institute, Cambridge, UK). It has been peer-reviewed and published in [59]. The methodology was conceived by Stephen Clark, who performed most of the experiments. Felix Krueger processed and managed sequencing data. I performed all the computational analysis shown in this chapter. John C. Marioni, Oliver Stegle and Wolf Reik supervised the project. The article was jointly written by Stephen Clark and me, with input from all authors.

2.1 Description of the experimental protocol

scNMT-seq builds upon two previous multi-modal protocols: single-cell Methylation and Transcriptome sequencing (scM&T-seq) [12] and Nucleosome Occupancy and Methylation sequencing (NOMe-seq) [139, 237]. An overview of the protocol is shown in Figure 2.1.

In the first step (the NOMe-seq step), cells are sorted into individual wells and incubated with a GpC methyltransferase (M.CviPI). This enzyme labels accessible (or nucleosome depleted) GpC sites via DNA methylation[144, 139]. In mammalian genomes, cytosine residues in GpC dinucleotides are methylated at a very low rate. Hence, after M.CviPI treatment, GpC methylation marks can be interpreted as direct read outs for chromatin accessibility, as opposed to the CpG methylation readouts, which can be interpreted as endogenous DNA methylation[144, 139].

In a second step (the scM&T-seq step), the DNA molecules are separated from the mRNA using oligo-dT probes pre-annealed to magnetic beads. Subsequently, the DNA fraction undergoes single-cell bisulfite conversion[274], whereas the RNA fraction undergoes Smart-seq2 [231].

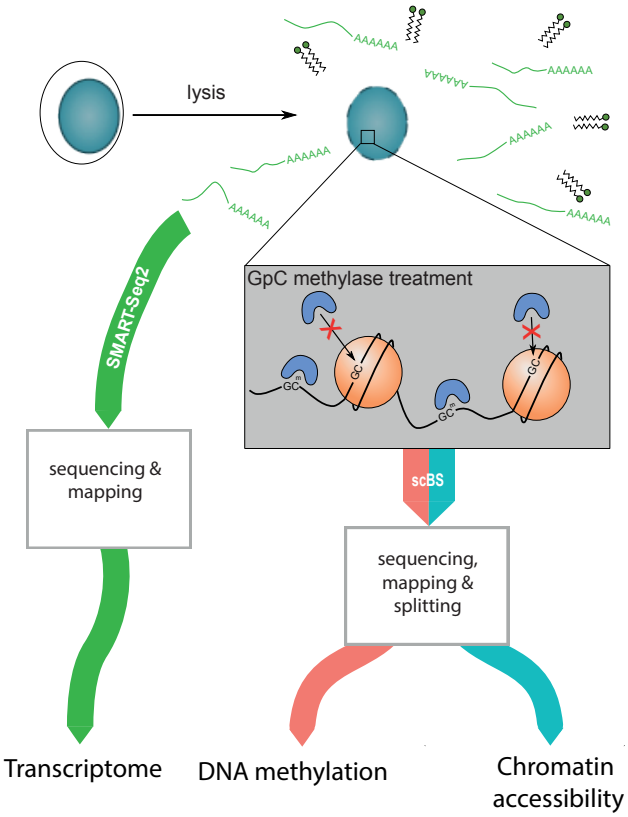


Figure 2.1: scNMT-seq protocol overview.
In the first step, cells are isolated and lysed. Second, cells are incubated with a GpC methyltransferase. Third, the RNA fraction is separated using oligo-dT probes and sequenced using Smart-seq2. The DNA fraction undergoes scBS-seq library preparation and sequencing. Finally, CpG Methylation and GpC chromatin accessibility data are separated computationally.

As discussed in Section 1.1.2, NOMe-seq has a range of appealing properties in comparison with count-based methods such as ATAC-seq or DNAseq-seq. First, the obvious gain of simultaneously measuring another epigenetic readout such as DNA methylation with little additional cost. Second, the resolution of the method is determined by the frequency of GpC sites within the genome ($\approx$ 1 in 16 bp), rather than the size of a library fragment (usually >100 bp). This allows the robust inspection of individual regulatory elements, nucleosome positioning and transcription factor footprints [139, 237, 219]. Third, missing data can be easily discriminated from inaccessible chromatin. Importantly, this implies that lowly accessible sites will not suffer from increased technical variation (due to low read counts) compared to highly accessible sites. Finally, the M.CviPI enzyme shows less sequence motif biases than the DNAse or the Tn5 transposase [219].

The downsides of the approach are the limited scalability associated with plate-based methods, and the need to discard read outs from (1) GCG positions (21% of all CpG sites), as it is intrinsically not possible to distinguish endogenous methylation from *in vitro* methylated bases, and (2) CGC positions (27%), to mitigate off-target effects of the enzyme [139]. This filtering step reduces the number of genome-wide cytosines that can be assayed from 22 million to 11 million.

2.2 Description of the data processing pipeline

After DNA sequencing, reads undergo quality control and trimming using TrimGalore to remove the flanking 6bp (the random primers), adaptor contamination and poor-quality base calls. Subsequently, trimmed reads are aligned to the corresponding genome assembly. Here we used Bismark [156] with the additional –NOMe option, which produces CpG report files containing only ACG and TCG trinucleotides and GpC report files containing only GCA, GCC and GCT positions.

Following [274], a bernoulli model is assumed for each CpG and GpC site in each cell after removal of duplicate alignments, which results in binary methylation calls. Notice that the use of a bernoulli model is an exclusive property of single-cell bisulfite sequencing data, for the vast majority of sites only one allele is observed per cell (due to data sparsity). This contrasts with bulk bisulfite sequencing data, where each dinucleotide typically contains multiple reads (originating from different cells) and thus a binomial model is more appropriate than a bernoulli estimate.

Finally, when quantifying DNA methylation and chromatin accessibility over genomic features (i.e. promoters or enhancers) a binomial model is assumed for each cell and feature, where the number of successes is the number of methylated CpGs (or GpCs) and the number of trials is the total number of CpGs (or GpCs) that are observed within the specific cell and genomic feature.

2.3 Data validation

2.3.1 Coverage

We validated scNMT-seq in 70 EL16 mouse embryonic stem cells (ESCs), together with 3 cells processed without M.CviPI enzyme treatment (i.e. using scM&T-seq). The use of this relatively simple and well-studied *in vitro* system allows us to compare our DNA methylation and chromatin accessibility statistics to published data [274, 12, 91].

First, we compared the theoretical maximum coverage that could be achieved with the empirical coverage (Figure 2.2). Despite the reduction in theoretical coverage due to the removal of ambiguous CCG and GCG sites, we observed, for DNA methylation, a median of $\approx$ 50% of promoters, $\approx$ 75% of gene bodies and $\approx$ 25% of active enhancers captured by at least 5 CpGs in each cell. Nevertheless, limited coverage is indeed observed for small genomic contexts such as p300 ChIP-seq peaks (median of $\approx$ 200bp).
For chromatin accessibility, coverage was larger than that observed for endogenous methylation due to the higher frequency of GpC dinucleotides, with a median of $\approx$ 85% of gene bodies and $\approx$ 75% of promoters measured with at least 5 GpCs.

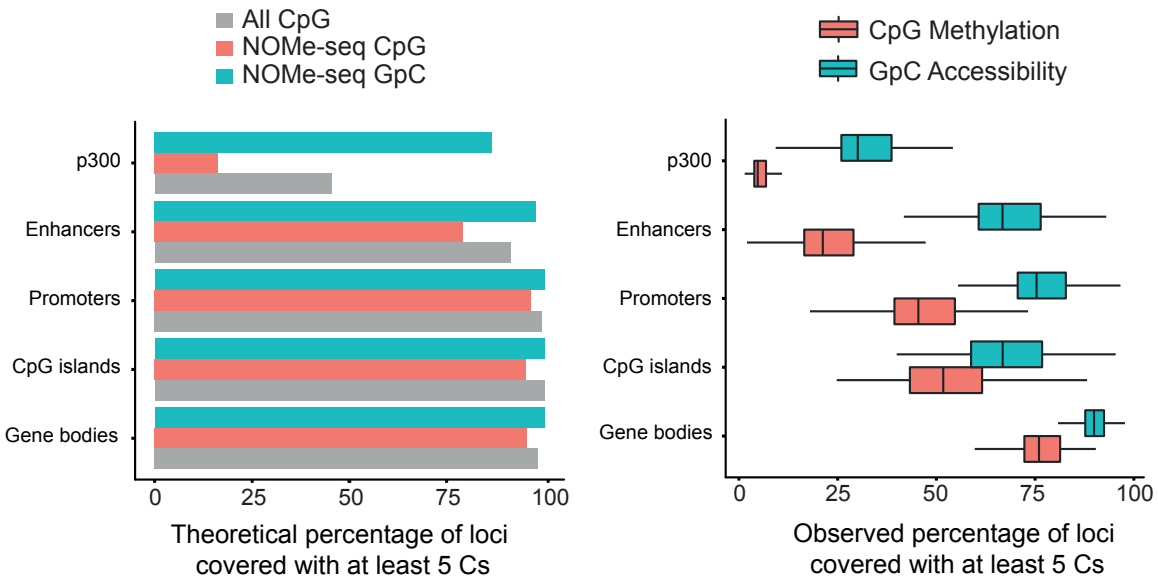


Figure 2.2: Coverage statistics for CpG DNA methylation and GpC chromatin accessibility.
(a) Fraction of loci with at least 5 CpG (red) or GpC (blue) dinucleotides per genomic context, after exclusion of the conflicting trinucleotides. The grey bar shows the total number of CpGs without exclusion of trinucleotides. (b) Empirical coverage per genomic context in a data set of 61 mouse ES cells. The empirical coverage is quantified as the fraction of loci with at least 5 CpG (red) or GpC (blue) observed. The boxplots summarise the distribution across cells, showing the median and the 1st and 3rd quartiles.

Next, we compared the DNA methylation coverage with a similar data set profiled by scM&T-seq [12] (Figure 2.3), where the the conflicting trinucleotides are not excluded.
Despite scNMT-seq yielding less CpG measurements, we find little differences in coverage when quantifying DNA methylation over genomic contexts, albeit these become evident when down-sampling the number of reads.

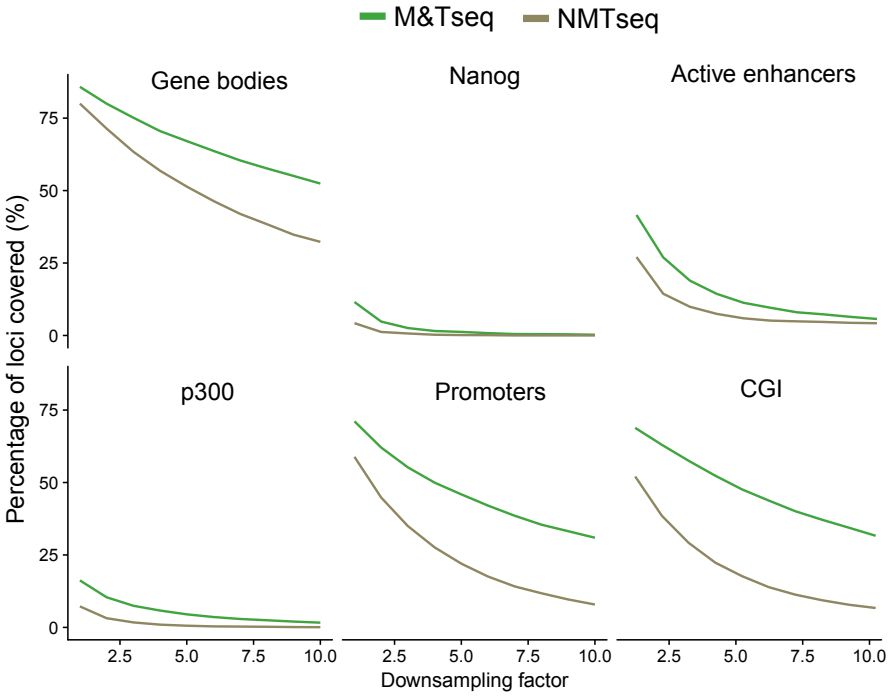


Figure 2.3: Comparison of the empirical coverage of DNA methylation with scM&T-seq [12].
The y-axis displays the fraction of loci covered with at least 5 CpG sites. The x-axis displays the downsampling factor, where the value of 1 corresponds to no downsampling (i.e. the base line). To facilitate the comparison, we selected two cells that were sequenced at equivalent depth.

2.3.2 Consistency with previous studies

To assess the consistency with previous studies we quantified DNA methylation and chromatin accessibility using a running window throughout the genome. The resulting methylomes were compared to datasets from the same cell lines profiled with similar technologies, including scM&T-seq[12], scBS-seq[274] and bulk BS-seq[91]. We find that most of the variation is not attributed to the technology but to differences in culture condition (Figure 2.4). This result is expected, as cells grown in 2i media remain in a native pluripotency state that is associated with genome-wide DNA hypomethylation [91]. Interestingly, the serum-cultured cells processed in this study overlapped with 2i-cultured cells from previous datasets. suggesting that they remained in a more pluripotent state. The most likely explanation for this variation is the differences in the cell lines (we used female EL16 versus male E14 in [12, 274, 91]). Previous studies have shown that female ESCs tend to show lower levels of mean global methylation, which is consistent with a more pluripotent phenotype [345].

In terms of accessibility, no NOMe-seq measurements were available for ESCs at the time of the study, so we compared it to bulk DNase-seq data from the same cell type, yielding good consistency between datasets ($R = 0.74$).

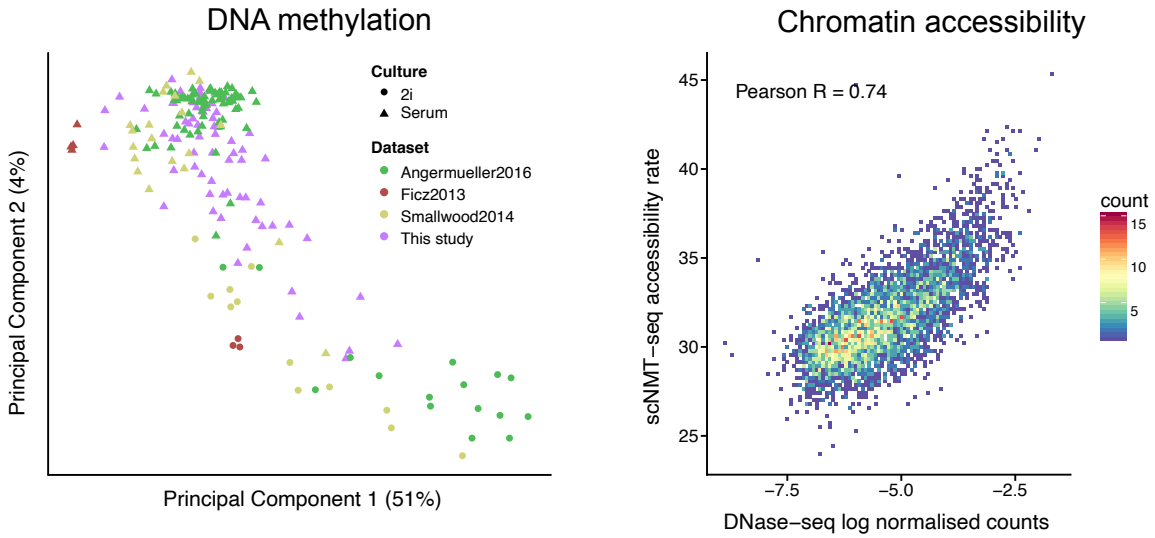


Figure 2.4: Comparison of unsupervised genome-wide quantifications to published datasets.
(a) Principal Component Analysis of 1kb running windows. Missing values were imputed using the average methylation rate per locus.
(b) Scatter plot of chromatin accessibility quantified over 10kb running windows of scNMT-seq data versus published bulk DNase-seq. For DNase-seq, accessibility is quantified as the log2 reads. The Pearson correlation was weighted by the GpC coverage in scNMT-seq data.

2.3.3 Quantification of DNA methylation and chromatin accessibility in known regulatory regions

We pseudobulked the data across all cells and examined DNA methylation and chromatin accessibility levels at loci with known regulatory roles. We found that in CTCF binding sites and DNaseI hypersensitivity sites DNA methylation was decreased while chromatin accessibility was increased, as previously reported [237]. As a control, we observe that cells which did not receive M.CviPI treatment showed globally low GpC methylation levels ($\approx 2\%$, Figure 2.5).

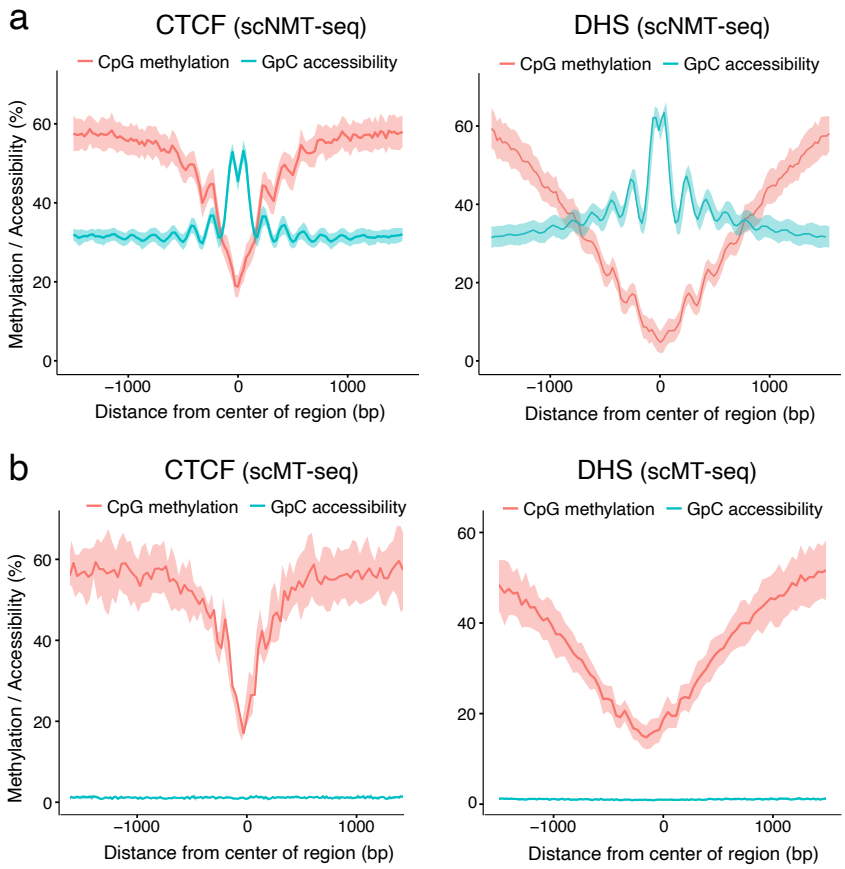


Figure 2.5: Accessibility and methylation profiles in regulatory genomic contexts. First, we pseudobulk the data set by pooling information across all cells. Next, we compute running averages of the CpG methylation (red) and the GpC accessibility (blue) in consecutive non-overlapping 50bp windows. Solid line displays the mean across all genomic elements within a given annotation and the shading displays the corresponding standard deviation.
(a) Profiles for scNMT-seq cells. (b) Profiles for scMT-seq cells

2.3.4 Quantification of the association between molecular layers.

We attempted to reconstruct the expected directional relationships between the transcriptome and the epigenome, namely the positive association between RNA expression and chromatin accessibility and the negative association between DNA methylation and RNA expression [302, 12].
To get a measure of the association (or coupling) between two molecular layers, we quantified a linear association per cell (across genes). Notice that this approach is not exclusive to single-cell data and can also be computed (more accurately) with bulk measurements. Reassuringly, this analysis confirmed, even within single cells, the expected positive correlation between chromatin accessibility and RNA expression, and the negative correlations between RNA expression and DNA methylation, and between DNA methylation and chromatin accessibility.

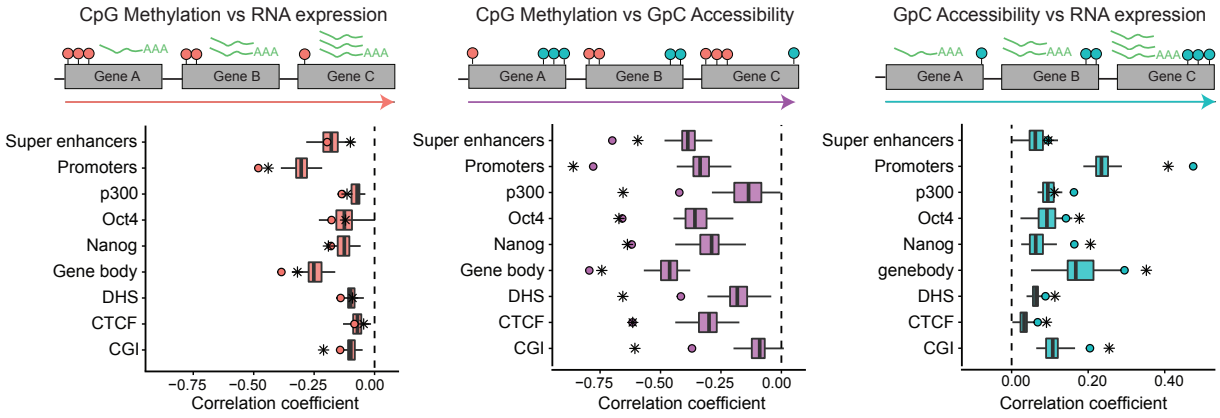


Figure 2.6: Quantification of linear associations between molecular layers.
The top diagram illustrates the computation of an association test per cell (across all loci in a given genomic context). The left panel shows DNA methylation versus RNA expression. The middle panel shows DNA methylation versus chromatin accessibility. The right panel shows RNA expression versus chromatin accessibility. The x-axis displays the Pearson correlation coefficients between two molecular layers, per genomic context (y-axis). The box plots summarise the distribution of correlation coefficients across cells. The dots and stars show the linear associations quantified in pseudo-bulked scNMT-seq data and published bulk data from the same cell types [91, 62], respectively.

2.4 Application to an embryoid body differentiation data set

2.4.1 Identification of genomic elements with coordinated variability across molecular layers

Having validated the quality of scNMT-seq data with a simple and relatively homogeneous data set, we next explored its potential to identify coordinated heterogeneity between the transcriptome and the epigenome.

We generated a second data set of 43 embryonic stem cells (after quality control), where we induced a differentiation process towards embryoid bodies by removing the LIF media for 3 days.

Dimensionality reduction on the RNA expression data reveals the existence of two subpopulations: one with high expression of pluripotency markers (*Esrrb* and *Rex1*) and the other with high expression of differentiation markers (*T* and *Prtg*).

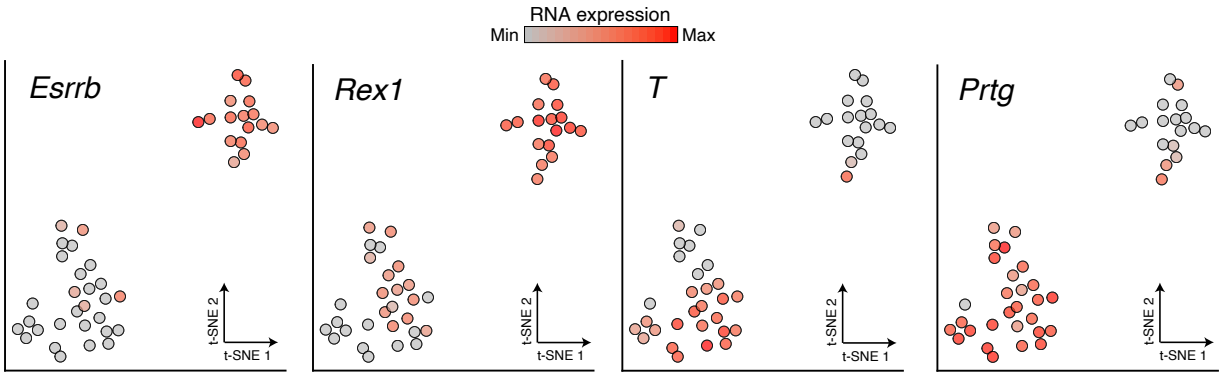


Figure 2.7: t-SNE representation of the RNA expression profiles for the embryoid body cells.
The scatter plots show a t-SNE [188] representation of the EB data. Cells are coloured based on expression of pluripotency factors (top) and differentiation markers (bottom).

Next, we tested for locus-specific associations between pairwise combinations of molecular layers (correlation across cells, Figure 2.8).
First, considering correlations between DNA methylation and RNA expression, we identified a majority of negative associations, reflecting the known relationship between these two layers. In contrast, we obtained largely positive associations between chromatin accessibility and RNA expression, mainly in promoters, p300 binding sites and super enhancer regions. Finally, we found mostly negative associations between DNA methylation and chromatin accessibility. This confirms the expected direction of association between molecular layers, as reported in bulk studies.
As an illustrative example, we display the *Esrrb* locus, a gene involved in early development and pluripotency [225]. A previous study [12], identified a super enhancer near the gene that showed a high degree of correlation between DNA methylation and RNA expression changes. In our study, we find *Esrrb* to be expressed primarily in the pluripotent cells, consistent with its role in early development. When examining the epigenetic dynamics of the corresponding super enhancers, we observe a strong negative correlation between DNA methylation and RNA expression, thus replicating previous findings. Additionally, we observe a strong negative relationship between DNA methylation and chromatin accessibility, indicating the two epigenetic layers are tightly coupled.

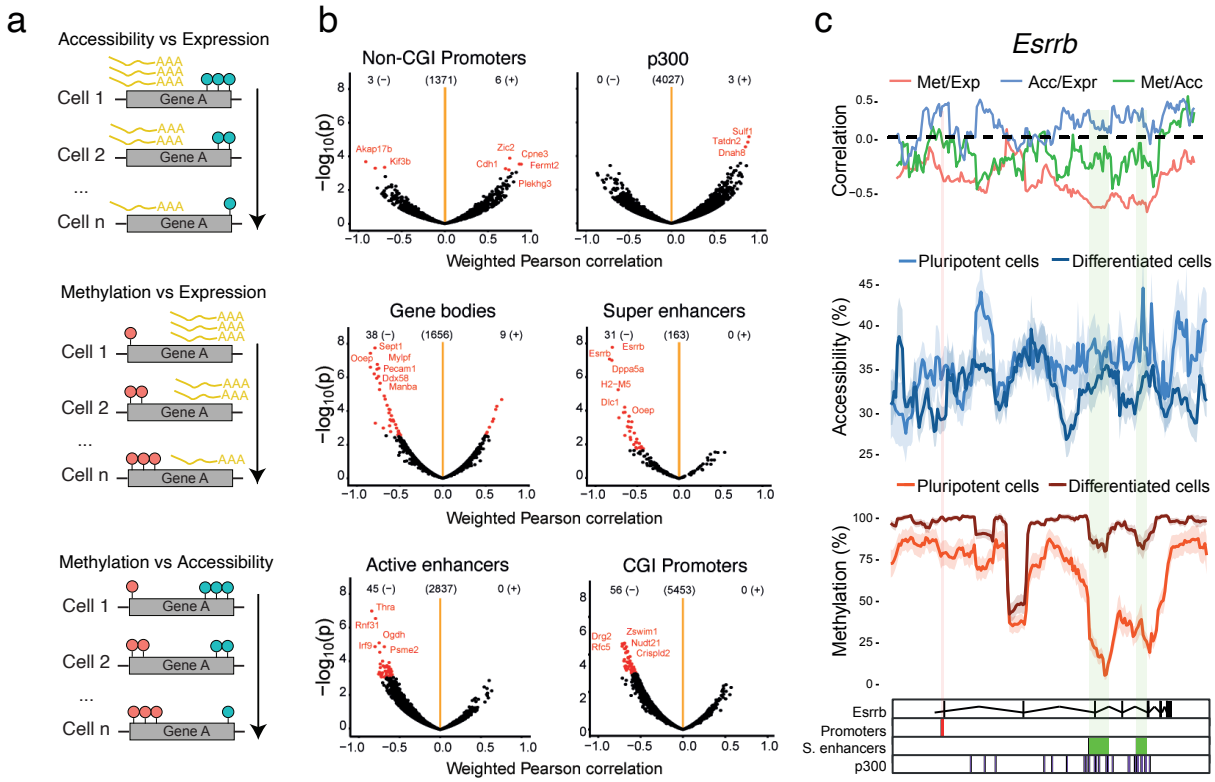


Figure 2.8: scNMT-seq enables the discovery of novel associations between transcriptomics and epigenetics at individual loci.
(a) Illustration for the correlation analysis, which results in one association test per locus (across cells).
(b) Pearson correlation coefficient (x-axis) and log10 p-value (y-axis) from association tests between different molecular layers, stratified by genomic contexts. Significant associations (FDR<0.1), are highlighted in red.
(c) Zoom-in view of the *Esrrb* gene locus. Shown from top to bottom are: Pearson correlation between each pair of molecular layers. Accessibility (blue) and methylation (red) profiles shown separately for pluripotent and differentiated sub-populations; mean rates (solid line) and standard deviation (shade) were calculated using a running window of 10kb with a step size of 1kb. Track with genomic annotations highlighting the position of regulatory elements.

2.4.2 Exploration of epigenome and transcriptome connections

The use of single-cell technologies has permitted the unbiased study of continuous trajectories by computationally reconstructing the *pseudotemporal* dynamics from the molecular profiles [310, 106, 263]. A novel opportunity unveiled by the introduction of single-cell multi-modal technologies is the study of epigenetic dynamics along trajectories inferred from the transcriptome. To explore this idea, we applied a diffusion-based pseudotime method [106] to the EB data set, using the RNA expression of the 500 genes with highest biological overdispersion [186]. The first diffusion component was used to reconstruct a pseudotemporal ordering of cells from pluripotent to differentiated states:

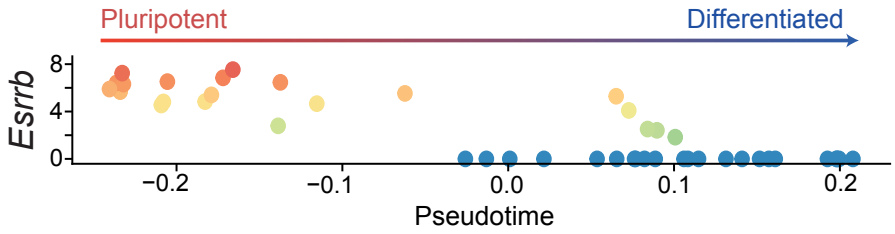


Figure 2.9: Reconstruction of developmental trajectory in embryoid body cells from the RNA expression data.
Each dot corresponds to one cell. The y-axis displays expression of *Esrrb*, a canonical pluripotency marker, and the x-axis shows the position of the cells in the first diffusion component.

Using the pseudotime reconstruction we investigated whether the strength of association between molecular layers (as calculated in Figure 2.6) are affected along the developmental trajectory. To do this, we correlated the correlation coefficient across genes between each pair of molecular layers (one value per cell) versus the pseudotime position (Figure 2.10). Importantly, this analysis is possible by the continuous nature of single-cell data and by the ability of scNMT-seq to profile three molecular layers at the same time.
We observe that for DNA methylation and chromatin accessibility, the negative correlation coefficients decreases in important regulatory genomic contexts (Figure 2.10), such that pluripotent cells have a notably weaker methylation-chromatin coupling than differentiated cells. This suggests that the strength of regulation between molecular layers can be altered during cell fate decisions.

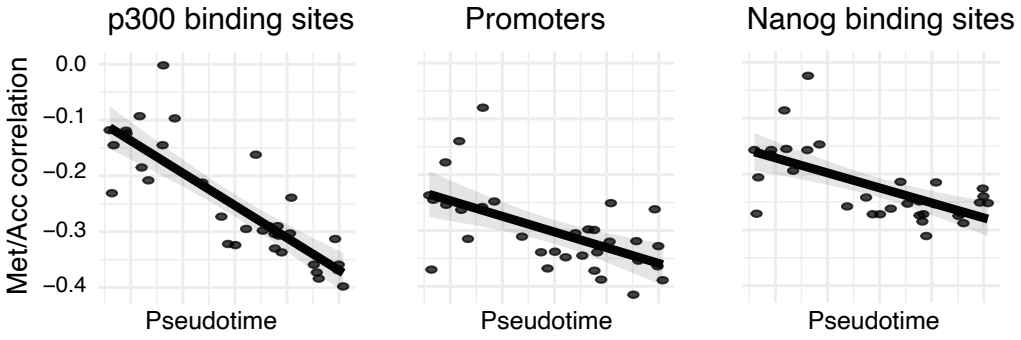


Figure 2.10: Developmental trajectory is associated changes in methylation-accessibility coupling.
Shown is the location of each cell in pseudotime (x-axis) and the corresponding Pearson correlation coefficients between methylation and accessibility (y-axis) in three different genomic contexts with regulatory roles.

2.5 Conclusions and open perspectives

In this Chapter I have introduced single-cell nucleosome, methylation and transcriptome sequencing (scNMT-seq), an experimental protocol for the genome-wide profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. This novel assay is an important step forward in the field of single-cell multi-modal sequencing. Yet, as with other protocols, the technology is still

in a very early stage and numerous developments are expected to occur in the next years. Some lines of research that I believe are important to improve scNMT-seq are the following:

- **Scalability**: scRNA-seq protocols are reaching the astonishing numbers of millions of cells per experiment. This contrasts with the limited cell numbers achieved in current multi-modal assays, including scNMT-seq [49, 50, 101]. As in scRNA-seq, the maturation of multi-modal techniques will display a trade-off between sensitivity and scalability [55]. scNMT-seq already provides high-resolution measurements, thus effort should be placed on making the protocol more scalable, which can be achieved by a series of technical improvements. First, barcodes are currently added at the end of the protocol, which limits cell numbers to the size of the plate. As in droplet-based methods or combinatorial indexing methods, adding the barcodes at the start of the protocol would enable the simultaneous processing of multiple pools of samples [74, 212]. Second, the physical separation of mRNA from genomic DNA is performed at the beginning of the protocol, one cell at a time. Given that it is a time-consuming and expensive process, this step should be performed after pooling [74]. Finally, albeit sequencing costs are decreasing [294], the sequencing of scNMT-seq libraries remains expensive due to genome-wide coverage. Hence, I anticipate that efforts to decrease the library size by a pre-selection of the genetic material will be indispensable. Examples of such strategies are the digestion by restriction enzymes as in RRBS [102], an initial round of ATAC protocol to select open chromatin [281] or the pull-down of specific genomic regions using capture probes.

- **Imputation of missing epigenetic data**: because of the low amounts of starting material, single-cell methylation protocols are limited by incomplete CpG coverage [11]. This becomes even more pronounced in scNMT-seq where almost $\approx 50\%$ of CpG dinucleotides are removed to avoid technical biases (see Section 2.3.1). Nonetheless, as discussed in Section 2.1, an important advantage of bisulfite approaches is that missing data can be discriminated from inaccessible chromatin (unlike in scATAC-seq). Therefore, the imputation of DNA methylation data will likely be a critical step to enable genome-wide analysis. Most of the imputation methods developed for bulk data are unsuccesful because they do not account for cell-to-cell variability [11]. A successful single-cell strategy based on deep learning has been proposed (DeepCpG [11]), but is a complex model that is difficult to train and does not scale to large datasets. Faster and accurate Bayesian approaches have also been considered (Melissa [136]), albeit the model is restricted to a small feature set and cannot perform genome-wide imputation.

- **Adding more molecular layers**: scNMT-seq can be adapted both experimentally and computationally to profile additional molecular layers. From the computational side, one could exploit the sequence information in the libraries to infer copy number variation or single nucleotide variants [235, 89, 197, 84]. This approach has been successful at delineating the clonal substructure of somatic tissues and at tracking mutational signatures in cancer tissues. In addition, the full length transcript information enables the quantification of splice variants[124], allele-specific fractions [73] and RNA velocity information [160]. From the experimental side, scNMT-seq could be combined with novel single-cell assays that quantify protein expression [288], transcription factor binding [211] and histone modifications [137].

- **Long reads**: the scNMT-seq libraries that were generated for this study contained short reads (75bp) that do not provide sufficient information about the regional context of the individual DNA molecule. By sequencing NOMe-seq libraries with long-read nanopore sequencing technology [167] showed that one can obtain phased methylation and chromatin accessibility measurements and structural changes from a single assay. This approach could potentially unveil a more comprehensive understanding of the epigenome dynamics and its regulatory role on RNA expression.

Chapter 3

Multi-Omics Factor Analysis (MOFA), a Bayesian model for integration of multi-omics data

The work described in this Chapter results from a collaboration with Wolfgang Huber's group at the EMBL (Heidelberg, Germany). It has been peer-reviewed and published in [15]. The method was conceived by Florian Buettner, Oliver Stegle and me. I performed most of the mathematical derivations and implementation, but with significant contributions from Damien Arnol and Britta Velten. The CLL data application was led by Britta Velten whereas the single-cell application was led by me, but with joint contributions in either cases. Florian Buettner, Wolfgang Huber and Oliver Stegle supervised the project.
The article was jointly written by Britta Velten and me, with contributions from all authors.

3.1 Theoretical foundations

Mathematical notation

- Matrices are denoted with bold capital letters: $\mathbf{W}$
- Vectors are denoted with bold non-capital letters: $\mathbf{w}$. If the vector comes from a matrix, we will use a single index to indicate the row that it comes from. If two indices are used, the first one corresponds to the row and the second one to the column. The symbol ':' denotes the entire row/column. For instance, $\mathbf{w}_i$ refers to the ith row from the $\mathbf{W}$ matrix, whereas $\mathbf{w}_{:,j}$ refers to the jth column.
- Scalars are denoted with non-bold and non-capital letters: w. If the scalar comes from a 1-dimensional array (a vector), a single subscript will indicate its position in the vector. If the scalar comes from a 2-dimensional array, two indices will be shown at the bottom: the first one corresponding to the row and the second one to the column. For instance, $w_{i,j}$ refers to the value from the ith row and the jth column of the matrix $\mathbf{W}$, and w_i to the ith value of the vector $\mathbf{w}$.
- $\mathbf{0}_k$ is a zero vector of length k.
- $\mathbf{I}_k$ is the identity matrix with rank k.

– $\mathbb{E}_q[x]$ denotes the expectation of x under the distribution q. When the expectations are taken with respect to the same distribution many times, we will avoid cluttered notation and we will instead use $\langle x \rangle$.
– $\mathcal{N}\left(x \,|\, \mu, \sigma^2\right)$: x follows a univariate normal distribution with mean μ and variance σ^2.
– $\mathcal{G}\left(x \,|\, a, b\right)$: x follows a gamma distribution with shape and rate parameters a and b.
– $\text{Beta}\left(x \,|\, a, b\right)$: x follows a beta distribution with shape and rate parameters a and b.
– $\text{Ber}(x|\theta)$: x follows a Bernoulli distribution with parameter θ.
– $\mathbb{1}_0$: Dirac delta function centered at 0.
– $Tr(\mathbf{X})$: Trace of the matrix $\mathbf{X}$

Graphical notation for probabilistic models

Probabilistic models can be represented in a diagrammatic format (i.e. a graph or a network) that offers a compact visual representation of complicated systems of probability distributions [31]. In a graphical model the relationship between the nodes becomes more explicit, namely their conditional independence properties which allow the joint distribution over all variables to be factorised into a series of simpler products involving subsets of variables [31]. The basic unit of a network is the node, which represents the different types of variables, including observed variables, unobserved probabilistic variables and unobserved parameters. The nodes are connected by unidirectional edges (arrows) which capture the conditional independence relationship between the variables.

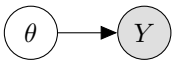
For this thesis we adapted the graphical notations from [76]:

Observed variables 

Unobserved probabilistic variables 

Unobserved parameters 

Repetition of node θ_n for $n \in [\![1; N]\!]$ 

Conditional dependency between nodes: $p(Y, \theta) = p(Y|\theta)p(\theta)$ 

3.1.1 Probabilistic modelling

A scientific model is a simple theoretical representation of a complex natural phenomenon to allow the systematic study of its behaviour. The general idea is that if a model is able to explain some observations, it might be capturing its true underlying laws and can therefore be used to make future predictions. In particular, statistical models are a powerful abstraction of nature. They consist of a set of observed variables and a set of (hidden) parameters. The procedure of fitting the parameters using a set of observations is called inference or learning.

One of the major challenges of inference when dealing with real datasets is the distinction between signal and noise. An ideal model should learn only the information relevant to gain explanatory power while disregarding the noise. However, this is a non-trivial task in most practical situations. Very complex models will tend to overfit the training data, capturing large amounts of noise and consequently leading to a bad generalisation performance to independent datasets. On the other hand, simplistic models will fit the data poorly, resulting in low explanatory power.

The ideas above can be formalised using the framework of probability and statistics.

3.1.2 Maximum likelihood inference

A common approach is to define a statistical model of the data $\mathbf{Y}$ with a set of parameters $\boldsymbol{\theta}$ that define a probability distribution $p(\mathbf{Y}|\boldsymbol{\theta})$, called the likelihood function. A simple approach to fit a model is to estimate the parameters $\hat{\boldsymbol{\theta}}$ that maximise the likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{Y}|\boldsymbol{\theta})$$

This process is called maximum likelihood learning [287, 31, 214]. However, in this setting there is no penalisation for model complexity, making maximum likelihood solutions can overfit when the data is relatively sparse. Generalisations that account for model complexity have been proposed that include regularising terms that shrink parameters to small values. However, these are often particular cases of the more general framework of Bayesian statistics [111, 31, 214].

3.1.3 Bayesian inference

In the Bayesian framework, the parameters themselves are treated as random unobserved variables and we aim to obtain probability distributions for $\boldsymbol{\theta}$, rather than a single point estimate. To do so, prior beliefs are introduced into the model by specifying a prior probability distribution $p(\boldsymbol{\theta})$. Then, using Bayes' theorem [25], the prior hypothesis is updated based on the observed data $\mathbf{Y}$ by means of the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ function, which yields a posterior distribution over the parameters:

$$p(\boldsymbol{\theta}|\mathbf{Y}) = \frac{p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{Y})}$$

where $p(\mathbf{Y})$ is a constant term called the marginal likelihood, or model evidence [31, 214].
The choice of the prior distribution is a key part of Bayesian inference and captures beliefs about the distribution of a variable before the data is taken into account. With asymptotically large sample sizes, the choice of prior has negligible effects on the posterior estimates, but it becomes critical with sparse data [31, 214, 97].

There are two common considerations when defining the prior distributions. The first relates to the incorporation of subjective information, or predefined assumptions, into the model. For example, one could adapt the prior distribution to match the results from previous experiments (i.e. an

informative prior). Alternatively, if no prior information is available one could set set uninformative priors by following maximum entropy principles [129].

The second strategy is based on convenient mathematical properties to make inference tractable. If the likelihood and the prior distributions do not belong to the same family of probability distributions (they are not conjugate) then inference becomes more problematic [242, 31, 214, 97]. The existence of conjugate priors is one of the major reasons that justify the widespread use of exponential family distributions in Bayesian models [97].

Again, the essential point of Bayesian inference is that an entire posterior probability distribution is obtained for each unobserved variable. This has the clear advantage of naturally handling uncertainity in the estimation of parameters. For instance, when making predictions, a fully Bayesian approach attempts to integrate over all possible values of all unobserved varaibles, effectively propagating uncertainity across multiple layers of the model. Nevertheless, this calculation is sometimes intractable and one has to resort to point estimates [31, 214, 97]. The simplest approximation to the posterior distribution is to use its mode, which leads to the maximum a posteriori (MAP) estimate:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(\mathbf{Y}|\boldsymbol{\theta})$$

This is similar to the maximum likelihood objective function, but with the addition of a term $p(\boldsymbol{\theta})$. When the prior distribution is well chosen, this term penalises for model complexity. Therefore, in contrast to standard (non-penalised) maximum likelihood inference, Bayesian approaches naturally handle the problem of model complexity and overfitting [31, 214, 97]. At the limit of infinite observations, the influence of the prior to the posterior is negligible and the MAP estimate converges towards the Maximum likelihood estimate, hence providing a rational link between the two inference frameworks.

Deterministic approaches for Bayesian inference

The central task in Bayesian inference is the direct evaluation of the posterior distributions and/or the computation of expectations with respect to the posterior distributions. In sufficiently complex models, closed-form solutions are not available and one has to resort to approximation schemes, which broadly fall into two classes: stochastic or deterministic [97, 33].

Stochastic approaches hinge on the generation of samples from the posterior distribution via a Markov Chain Monte Carlo (MCMC) framework. Such techniques have the appealing property of generating exact results at the asymptotic limit of infinite computational resources. However, in practice, sampling approaches are computationally demanding and suffer from limited scalability to large datasets [33].
In contrast, deterministic approaches are based on analytical approximations to the posterior distribution, which often lead to biased results. Yet, given the appropriate settings, these approaches are potentially much faster and scalable to large applications [31, 214, 33].

3.1.4 Variational inference

Variational inference is a deterministic family of methods that have been receiving widespread attention due to a positive balance between accuracy, speed, and ease of use [33, 336]. The core framework is derived below.

In variational inference the true (but intractable) posterior distribution $p(\mathbf{X}|\mathbf{Y})$ is approximated by a simpler (variational) distribution $q(\mathbf{X}|\mathbf{\Theta})$ where $\mathbf{\Theta}$ are the corresponding parameters. The parameters, which we will omit from the notation, need to be tuned to obtain the closest approximation to the true posterior.
The distance between the true distribution and the variational distribution is calculated using the KL divergence:

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = -\int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} d\mathbf{X}$$

Note that the KL divergence is not a proper distance metric, as it is not symmetric. In fact, using the reverse KL divergence $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{Y}|\mathbf{X}))$ defines a different inference framework called expectation propagation [204].

If we allow any possible choice of $q(\mathbf{X})$, then the minimum of this function occurs when $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$. Nevertheless, since the true posterior is intractable to compute, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of distributions $q(\mathbf{X})$ that are tractable to compute and subsequently seek the member of this family for which the KL divergence is minimised.

Doing some calculus it can be shown (see [31, 214]) that the KL divergence $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$ is the difference between the log of the marginal probability of the observations $\log(\mathbf{Y})$ and a term $\mathcal{L}(\mathbf{X})$ that is typically called the Evidence Lower Bound (ELBO):

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = \log p(\mathbf{Y}) - \mathcal{L}(\mathbf{X})$$

Hence, minimising the KL divergence is equivalent to maximising $\mathcal{L}(\mathbf{X})$ (Figure 3.1):

$$\begin{aligned} \mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y}) \Big) d\mathbf{X} \\ &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Y})] - \mathbb{E}_q[\log q(\mathbf{X})] \end{aligned} \tag{3.1}$$

The first term is the expectation of the log joint probability distribution with respect to the variational distribution. The second term is the entropy of the variational distribution. Importantly, given a simple parametric form of $q(\mathbf{X})$, each of the terms in Equation (3.1) can be computed in closed form. In some occasions, we will use the following form for the ELBO:

$$\mathcal{L}(\mathbf{X}) = \mathbb{E}_q[\log p(\mathbf{Y}|\mathbf{X})] + (\mathbb{E}_q[\log p(\mathbf{X})] - \mathbb{E}_q[\log q(\mathbf{X})]) \tag{3.2}$$

where the first term is the expectation of the log likelihood and the second term is the difference in the expectations of the p and q distributions of each hidden variable.

In conclusion, variational learning involves minimising the KL divergence between $q(\mathbf{X})$ and $p(\mathbf{X}|\mathbf{Y})$ by instead maximising $\mathcal{L}(\mathbf{X})$ with respect to the distribution $q(\mathbf{X})$.

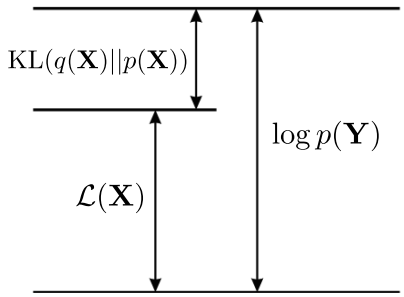


Figure 3.1: The quantity $\mathcal{L}(\mathbf{X})$ provides a lower bound on the true log marginal likelihood $\log p(\mathbf{Y})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(q||p)$ between the variational distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$

There are several approaches to define $q(\mathbf{X})$, the two most commonly used are called (unparametric) mean-field and (parametric) fixed-form [336, 33].

Mean-field variational inference

The most common type of variational Bayes, known as the mean-field approach, assumes that the variational distribution factorises over M disjoint groups of unobserved variables [267]:

$$q(\mathbf{X}) = \prod_{i=1}^{M} q(\mathbf{x}_i) \tag{3.3}$$

where typically all unobserved variables are assumed to be independent. Importantly, notice that no parametric assumptions were placed regarding the nature of $q(\mathbf{x}_i)$.

Evidently, in sufficiently complex models where the unobserved variables have dependencies this family of distributions do not contain the true posterior (Figure 3.2). Yet, this is a key assumption to obtain an analytical inference scheme that yields surprisingly accurate results [32, 88, 36].

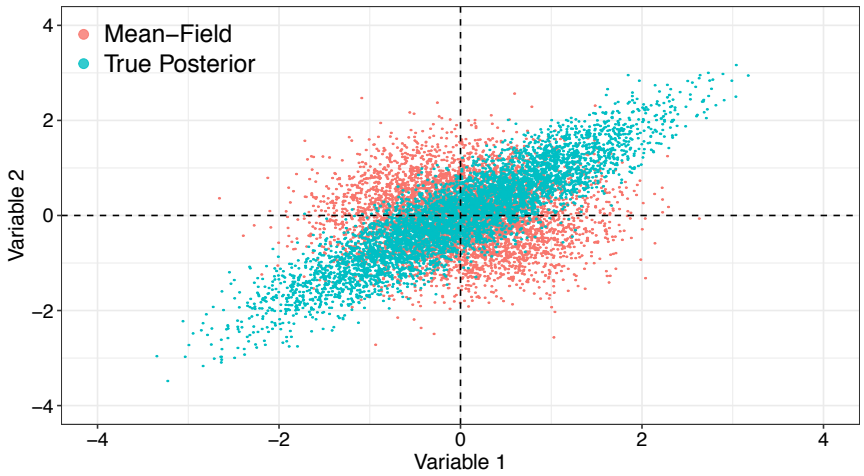


Figure 3.2: Illustrative example of sampling from a true posterior distribution (blue) versus a fitted mean-field varaitional distribution (red) in a model with two (correlated) unobserved variables. The mean-field approximation wrongly assumes that the unobserved variables are independent.

Using calculus of variations (derivations can be found in [31, 214]), it follows that the optimal distribution $q(\mathbf{X})$ that maximises the lower bound $\mathcal{L}(\mathbf{X})$ is

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const} \tag{3.4}$$

where $\mathbb{E}_{-i}$ denotes an expectation with respect to the q distributions over all variables $\mathbf{x}_j$ except for $\mathbf{x}_i$.
The additive constant is set by normalising the distribution $\hat{q}_i(\mathbf{z}_i)$:

$$\hat{q}(\mathbf{x}_i) = \frac{\exp(\mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})])}{\int \exp(\mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})]) d\mathbf{X}}$$

While the form of $\hat{q}(\mathbf{x}_i)$ is not restricted to a specific parametric form, it can be shown that when using conjugate priors, the distributions $\hat{q}_i(\mathbf{x}_i)$ have the same functional form as the priors $\hat{p}(\mathbf{x}_i)$.

Fixed-form variational inference

An alternative straightforward strategy is to directly define a parametric form for the distribution $q(\mathbf{X})$ with some parameters $\boldsymbol{\Theta}$. Once the choice of $q(\mathbf{X})$ is made, the parameters $\boldsymbol{\Theta}$ are optimised to minimise $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$ (the variational problem):

$$\hat{\boldsymbol{\Theta}} = \arg\min_{\boldsymbol{\Theta}} \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \tag{3.5}$$

$$= \mathbb{E}[\log(q(\mathbf{X})) - \log(p(\mathbf{X}, \mathbf{Y}))] \tag{3.6}$$

Numerically optimising this function requires the evaluation of expectations with respect to $q(\mathbf{X})$. In closed form, this is only feasable for a limited group of variational distributions. Alternatively, one can attempt Monte Carlo approximations, but in practice this turns to be slow and leads to high-variance estimates [36, 245, 36].

Typically, one would choose this distribution to factorise over parameters and to be of the same (exponential) family as the prior $p(\mathbf{X})$. In such case there is a closed form coordinate-ascent scheme available, and it turns out that the fixed-form formulation is equivalent to the (non-parametric) mean-field derivation when using conjugate priors.
Unfortunately, for generic models with arbitrary families of distributions, no closed-form variational distributions exist [336, 33].

However, while the parametric assumption certainly limits the flexibility of variational distributions, the advantage of this formulation is that it opens the possibility to use fast gradient-based methods for the inference procedure [119, 245].

3.1.5 Expectation Propagation

Expectation Propagation (EP) is another deterministic strategy with a similar philosophy as the Variational approach. It is also based on minimising the KL divergence between a variational distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$, but while variational inference minimises $KL(p||q)$, EP maximises the reverse KL-divergence $KL(q||p)$.

Interestingly, this simple difference leads to an inference scheme with strikingly different properties. This can be understood by inspecting the differences between the two KL divergence formulas:

Variational inference:

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = -\int q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \tag{3.7}$$

Expectation propagation:

$$\mathrm{KL}(p(\mathbf{X}|\mathbf{Y})||q(\mathbf{X})) = -\int p(\mathbf{X}|\mathbf{Y}) \log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \tag{3.8}$$

In regions of $\mathbf{X}$ where the true posterior density $p(\mathbf{X}|\mathbf{Y})$ is small, setting a large density for $q(\mathbf{X})$ has a much stronger penalisation in Equation (3.8) than in Equation (3.7). Hence, EP tends to avoid areas where the density $p(\mathbf{X}|\mathbf{Y})$ is very low, even if it does not correspond to areas of very high-density in $p(\mathbf{X}|\mathbf{Y})$. In contrast, in Equation (3.7) there is a strong penalty for having low-density $q(\mathbf{X})$ values.
As discussed in [31], the practical consequences of this duality can be observed when the posterior is multi-modal. In VI, $q(\mathbf{X})$ converges towards areas of high-density in $p(\mathbf{X}|\mathbf{Y})$, namely local optima. In contrast, EP tends to capture as much non-zero density regions from $p(\mathbf{X}|\mathbf{Y})$ as possible, thereby averaging across all optima. In the context of doing predictions, the VI solution is much more desirable than the EP solution, as the average of two good parameter values is not necessarily a good value itself.
A detailed mathematical treatment of EP, including derivations for specific examples, can be found in [31, 214, 204]

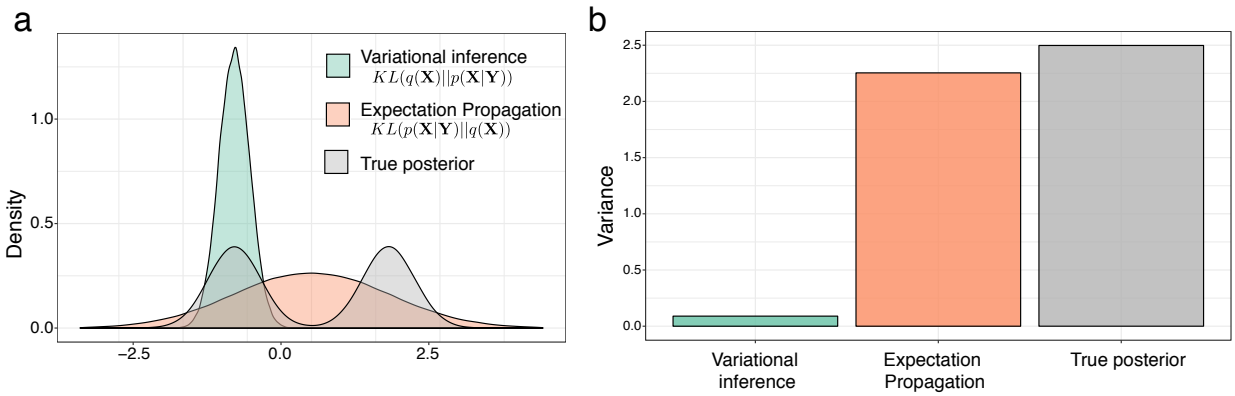


Figure 3.3: Illustrative comparison of Variational inference and Expectation Propagation. Shown is the (a) Density and (b) Variance of the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$ (grey), the variational distribution (orange) and the expectation propagation distribution (green).

3.1.6 Conclusions

In this section we have introduced Bayesian modelling and variational inference methods, which will be used later in this chapter.
More generally, variational inference is growing in popularity for the analysis of big datasets and it has been applied to a myriad of different problems, including genome-wide association studies [52], population genetics, [243], network analysis [264] and natural language processing [34].

Yet, despite its increasing success, there is significant room for improvement. First and foremost, the theoretical guarantees of variational inference are not as developed as in sampling-based MCMC schemes [33, 336, 216]. As an example, the mean-field setting makes strong independence assumptions about the parameters. Although it tends to be surprisingly effective, it is not clear in which applications the dependencies between the parameters are important enough that the mean-field approximation could potentially break.
More generally, an open research problem is understanding what are the statistical properties of the variational posterior with respect to the exact posterior [33, 336].

As we shall demonstrate later, alternative strategies have been considered to allow some dependencies between the variables, resulting in *structured* mean-field approximations [118, 304]. However, they often lead to very complex (if not intractable) inference frameworks.

Finally, another area of extensive research is how to extend the applicability of VI to non-conjugate models. As discussed in Section 3.1.3, the ELBO of non-conjugate models contains intractable integrals, and setting up an inference scheme requires the use of either stochastic Monte Carlo approximations or deterministic approximations that introduce additional lower bounds [336, 271, 82]. In this thesis we follow this rationale to derive an inference framework for a model with non-Gaussian likelihoods.

3.1.7 Latent variable models for genomics

With the exponential growth in the use of high-throughput genomics, biological datasets are increasingly high dimensional, both in terms of samples and features. A key principle of biological datasets is that variation between the features results from differences in underlying, often unobserved, processes. Such processes, whether driven by biological or technical effects, are manifested by coordinated changes in multiple features. This key assumption sets off an entire statistical framework of exploiting the redundancy encoded in the data set to learn the (latent) sources of variation in an unsupervised fashion. This is the aim of dimensionality reduction techniques, or latent variable models [153, 285, 168, 238, 164, 286, 202].

Mathematical formulation

Given a dataset $\mathbf{Y}$ of N samples and D features, latent variable models attempt to exploit the dependencies between the features by reducing the dimensionality of the data to a potentially small set of K latent variables, also called factors. The mapping between the low-dimensional space and the high-dimensional space is performed via a function $f(\mathbf{X}|\mathbf{\Theta})$ that depends on some parameters $\mathbf{\Theta}$.
The choice of $f(\mathbf{X}|\mathbf{\Theta})$ is essentially the field of dimensionality reduction. A trade-off exists between complexity and interpretation: while non-linear functions such as deep neural networks provide more explanatory power, this leads to considerable challenges in interpretation [339]. Hence, for most applications where interpretability is essential, $f(\mathbf{X}|\mathbf{\Theta})$ is assumed to be linear:

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T \tag{3.9}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times K}$ is a matrix that contains the low-dimensional representation for each sample (i.e. the factors). The matrix $\mathbf{W} \in \mathbb{R}^{D\times K}$ contains the weights, which provide the linear mapping between the features and the factors.
Note that the aim of dimensionality reduction is to exploit the coordinated heterogeneity between features, and hence features can be assumed to be centered without loss of generality.

The inference procedure consists in learning the values of all unobserved variables, including factors and weights. As we shall demonstrate, different inference schemes and assumptions on the prior distributions lead to significantly different model outputs [248].

3.1.8 Principal Component Analysis

Principal Component Analysis (PCA) is the most popular technique for dimensionality reduction [120, 252].Two formulations of PCA exist [31]: in the maximum variance formulation, the aim is to infer an orthogonal projection of the data onto a low-dimensional space such that variance explained by the projected data is maximised. Formally, the aim in PCA is to infer the matrix $\mathbf{W}$ such that the variance of $\mathbf{Z}$ (the projected data) is maximised. If we consider a single latent factor,

the variance of the projected data is:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{z}_n - \hat{\mathbf{z}})^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{w} - \hat{\mathbf{y}}^T\mathbf{w})^2$$

where $\hat{\mathbf{y}}$ is a vector with the feature-wise means. If we assumed centered data this simplifies to:

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{w})^2$$

Some algebra allows us to define this equation in terms of the (centered) data covariance matrix: $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T$:

$$\begin{aligned}\sigma^2 = &\frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n^T\mathbf{w})^T(\mathbf{y}_n^T\mathbf{w}) \\ =&(\mathbf{w}^T\mathbf{y}_n)(\mathbf{y}_n^T\mathbf{w}) \\ =&\mathbf{w}^T(\mathbf{y}_n\mathbf{y}_n^T)\mathbf{w} \\ =&\mathbf{w}^T\mathbf{S}\mathbf{w}\end{aligned}$$

Thus, for a single principal component, the optimisation problem is:

$$\arg\max_{\|\mathbf{w}\|=1} = \mathbf{w}^T\mathbf{S}\mathbf{w} \tag{3.10}$$

The k-th principal component can be found by subtracting from $\mathbf{Y}$ the reconstructed data by the previous $k-1$ principal components. If we define $\mathbf{z}_k = \mathbf{w}_k^T\mathbf{Y}$ to be the k-th principal component:

$$\hat{\mathbf{Y}} = \mathbf{Y} - \sum_{k=1}^{K}(\mathbf{z}_k\mathbf{w}_k^T)$$

Re-applying Equation (3.10) defines the new optimisation problem.

In its minimum error formulation, the aim is to find an equivalent projection that minimises the mean squared error between the observations and the data reconstructed using the principal components:

$$\arg\max_{\|\mathbf{w}\|=1} \|\mathbf{Y} - \sum_{k=1}^{K}(\mathbf{z}_k\mathbf{w}_k^T)\|^2$$

where $\|\cdot\|^2$ is the Frobenius norm.

Remarkably, in both cases, solving the optimisation problems via Lagrange multipliers leads to master eigenvalue-eigenvector equation:

$$\mathbf{S}\mathbf{w}_k = \lambda_k\mathbf{w}_k \tag{3.11}$$

where the weight vectors $\mathbf{w}_k$ can be calculated as the eigenvectors of the covariance matrix $\mathbf{S}$ [31].

Interestingly, the reason why the maximum variance solution and the minimum reconstruction error solution are the same can be understood by applying Pythagoras theorem to the right triangle defined by the projection of a sample $\mathbf{y}_n$ to a weight vector $\mathbf{w}$ (Figure 3.4). Assuming again centered data, the variance of $\mathbf{y}_n$ is $\|\mathbf{y}_n\| = \mathbf{y}_n^T\mathbf{y}_n$. This variance decomposes as the sum of the variance in the latent space $\|\mathbf{z}_n\| = \mathbf{z}_n^T\mathbf{z}_n$ and the residual variance after reconstruction $\|\mathbf{y}_n - \mathbf{z}_n\mathbf{w}^T\|$:

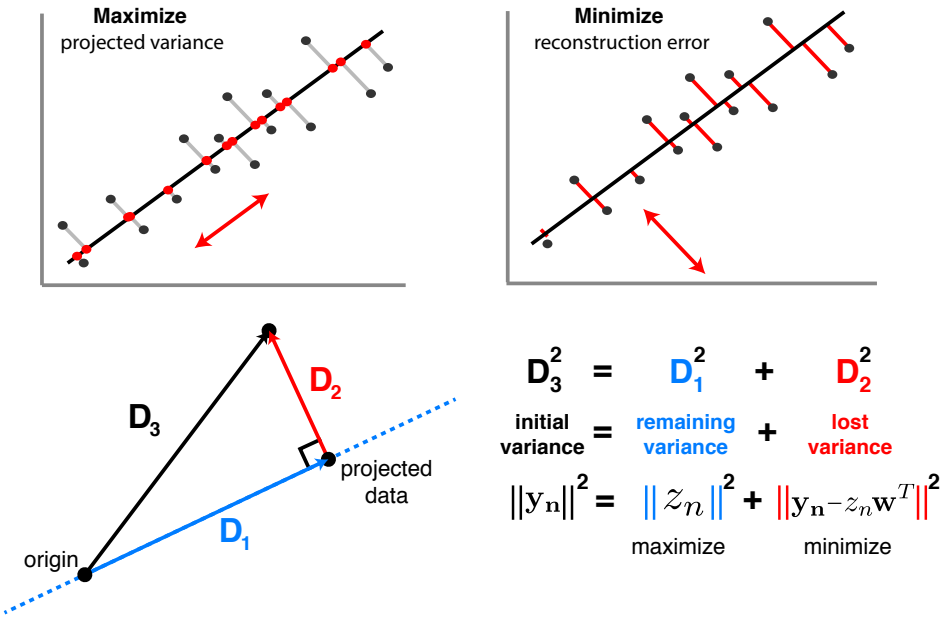


Figure 3.4: In the maximum variance formulation the aim is to maximise the variance of the projected data (blue line), whereas in the minimum error formulation the aim is to minimise the residual variance (red line). Given a fixed total variance (black line), both strategies are equivalent

The main strength of PCA relies on its simplicity and closed form solution. Additionally, the linear mapping has the advantage of yielding interpretable feature weights, so that inspection of $\mathbf{w}_k$ reveals which features are jointly affected by the k-th principal component.
However, PCA suffers from serious drawbacks when applying it to real datasets [171]. First, biological measurements are inherently noisy, and there is no explicit account of noise in PCA. In practice, high variance components are often asociated with signal whereas low-variance components are assumed to be noise, but an ideal model should explicitly disentangle the uncoordinated variability that is attributed to noise from the coordinated variability that is characterised as signal. Second, in its original formulation, no missing data is allowed [126]. Third, it does not offer a principled way of modelling prior information about the data.

3.1.9 Probabilistic Principal Component Analysis and Factor Analysis

A probabilistic version of PCA was initially proposed in [303]. It can be formulated by converting some (or all) fixed parameters into random variables and adding an explicit noise term to

Equation (3.9):

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T + \boldsymbol{\epsilon} \tag{3.12}$$

where the weights $\mathbf{W}$ are assumed to be non-probabilistic parameters, but the noise $\boldsymbol{\epsilon}$ and the latent variables $\mathbf{Z}$ (the principal components) are assumed to follow an isotropic normal distribution:

$$p(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right)$$
$$p(\epsilon) = \mathcal{N}\left(\epsilon \,|\, 0, \tau^{-1}\right)$$

where τ is the precision (inverse of the variance).

All together, this leads to a Gaussian likelihood:

$$p(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \tau) = \prod_{n=1}^{N}\prod_{d=1}^{D} \mathcal{N}\left(y_{n,d} \,|\, \mathbf{z}_{n,:}\mathbf{w}_{,:k}, \tau^{-1}\mathbf{I}\right) \tag{3.13}$$

The corresponding graphical model is:

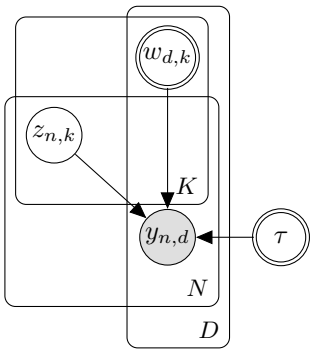


Figure 3.5: Graphical model for probabilistic PCA. The latent variables are modelled as random variables, whereas the weights and the noise are modelled as deterministic parameters.

Importantly, the choice of the distribution for ϵ implies that the noise of each feature is independent but restricted to have the same precision τ. In practice this is a limiting assumption, as different features are expected to show different degrees of noise, albeit this constraint can be relaxed and forms the basis of Factor Analysis [260, 31].

The inference procedures involves learning the parameters $\mathbf{W}$, and τ and a posterior probability distribution for $\mathbf{Z}$. As the model depends on latent variables, inference can be performed using the iterative Expectation-Maximisation (EM) algorithm [260, 31]. In the expectation step, the posterior distribution for $\mathbf{Z}$ is computed in closed form (due to conjugacy between the likelihood and the prior), given current estimates for the parameters $\mathbf{W}$, and τ. In the maximisation step, the parameters are calculated by maximising the expectation of the joint log likelihood under the posterior distribution of $\mathbf{Z}$ found in the E step [303].
Interestingly, the EM solution of probabilistic PCA lies in the same subspace as the traditional PCA solution [303], but the use of a probabilistic framework brings several benefits. First, model selection can be performed by comparing likelihoods across different settings of parameters. Second,

missing data can naturally be accounted for by ignoring the missing observations from the likelihood. Finally, the probabilistic formulation sets the core framework for a Bayesian treatment of PCA, enabling a broad range of principled extensions tailored different types of datasets.

3.1.10 Bayesian Principal Component Analysis and Bayesian Factor Analysis

The full Bayesian treatment of PCA requires the specification of prior probability distributions for all unobserved variables:

$$
\begin{aligned}
p(\mathbf{Z}) &= \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\
p(\mathbf{W}) &= \prod_{d=1}^{D} \prod_{k=1}^{K} \mathcal{N}\left(w_{dk} \,|\, 0, 1\right) \\
p(\epsilon) &= \mathcal{N}\left(\epsilon \,|\, 0, \tau^{-1}\right) \\
p(\tau) &= \mathcal{G}\left(\tau \,|\, a_0, b_0\right)
\end{aligned}
$$

A generalisation to Bayesian Factor Analysis follows by allowing a separate noise term per feature:

$$
\begin{aligned}
p(\boldsymbol{\epsilon}) &= \prod_{d=1}^{D} \mathcal{N}\left(\epsilon_d \,|\, 0, \tau_d^{-1}\right) \\
p(\boldsymbol{\tau}) &= \prod_{d=1}^{D} \mathcal{G}\left(\tau_d \,|\, a_0, b_0\right)
\end{aligned}
$$

where a_0 and b_0 are fixed hyperparameters. As in Equation (3.13), this results in a Normal likelihood:

$$
p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \boldsymbol{\tau}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(y_{nd} \,|\, \mathbf{w}_d^T \mathbf{z}_n, \tau_d\right)
$$

The corresponding graphical model is:

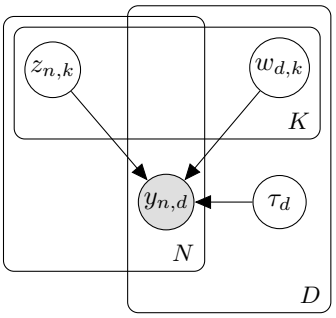


Figure 3.6: Graphical model for Bayesian Factor Analysis. All unobserved variables are modelled as random variables.

3.1.11 Hierarchical priors

A key advantage of the full Bayesian treatment is that it explicitly captures uncertainity on the estimation of all unobserved variables, as opposed to the probabilistic PCA model [30, 29]. Yet, more importantly, the use of (hierarchical) prior distributions allow different modelling assumptions to be encoded, providing a flexible and principled approach to extend PCA to a myriad of modelling scenarios, including multi-view generalisations [147, 318, 149, 45, 142, 340].

Automatic relevance determination

As an example, a major challenge in PCA is how to determine the dimensionality of the latent space (i.e. the number of principal components). As we will show, the use of hierarchical prior distributions allows the model to introduce sparsity assumptions on the weights in such a way that the model automatically learns the number of factors.
In the context of Factor Analysis, one of the first sparsity priors to be proposed was the Automatic Relevance determination (ARD) prior [217, 190, 30, 29].

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k} \,|\, 0, \frac{1}{\alpha_k}\mathbf{I}_D\right) \qquad p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k \,|\, a_0^{\alpha}, b_0^{\alpha}\right)$$

The aim of this prior is two-fold. First, the zero-mean normal distribution specifies that, *a priori*, no information is available and all features are *inactive*. When exposed to some data, the posterior distribution for $\mathbf{W}$ will be estimated by weighting the contribution from the likelihood, potentially allowing features to escape from the zero-centered prior (Figure 3.7).
Second, performing inference on the variable $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_k\}$ enables the model to discard inactive factors. To understand this, let us assume that only $K = 5$ true factors exist, but the model is initialised with $K = 20$ factors. In such case, inactive factors can be prunned out by driving the corresponding α_k to infinity. In turn, this causes the posterior $p(\mathbf{w}_{:,k}|\mathbf{Y})$ to be sharply peaked at zero, resulting in the inactivation of all its weights.

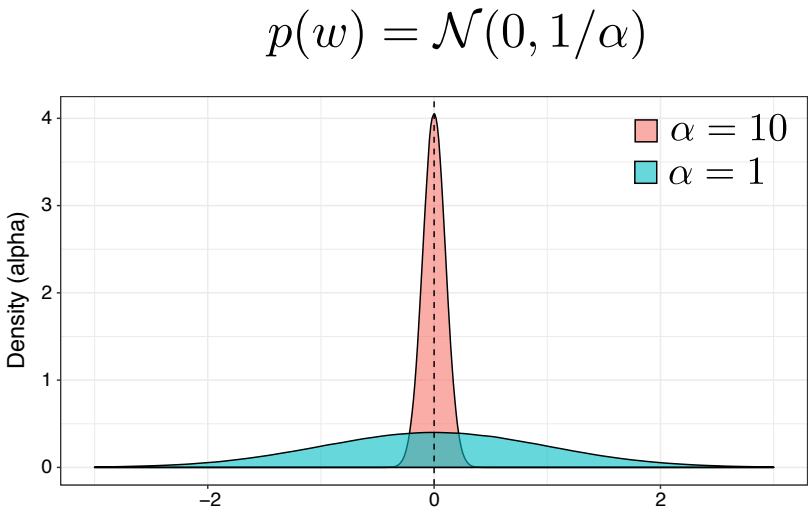


Figure 3.7: Visualisation of the sparsity-inducing Automatic Relevance Determination prior

Spike-and-slab prior

Sparse extensions of the Bayesian factor analysis model have been proposed as a regularisation mechanism but also to model inherent assumptions regarding the sparse nature of biological data [285, 96].
The variability observed in biological data is driven both by technical factors and biological factors. Technical factors (i.e. batch effects) tend to be relatively strong and alter the expression of a large proportion of genes, whereas the biological factors are potentially weak effects driven by changes in small gene regulatory networks [96]. Hence, a practical factor analysis model should be able to learn factors with different degrees of sparsity.
The ARD prior proposed in Section 3.1.11 allows entire factors to be dropped out from the model, but it provides a weak degree of regularisation when it comes to inactivating individual weights within the active factors.

A sparse generalisation of the Factor Analysis model proposed above can be achieved by combining the ARD prior with a spike-and-slab prior [205, 304]. For every weight $w_{d,k}$:

$$p(w_{d,k} \mid \alpha_k, \theta_k) = (1 - \theta_k)\mathbb{1}_0(w_{d,k}) + \theta_k \mathcal{N}\left(w_{d,k} \mid 0, \alpha_k^{-1}\right) \tag{3.14}$$

$$p(\theta_k) = \text{Beta}\left(\theta_k \mid a_0^\theta, b_0^\theta\right) \tag{3.15}$$

$$p(\alpha_k) = \mathcal{G}\left(\alpha_k \mid a_0^\alpha, b_0^\alpha\right) \tag{3.16}$$

The corresponding graphical model is:

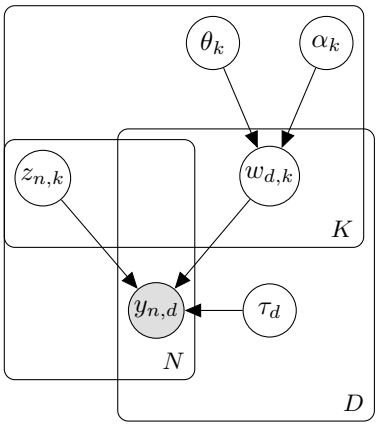


Figure 3.8: Graphical model for Bayesian sparse Factor Analysis. A double sparsity-inducing prior is used on the weights: an ARD prior to prune inactive factors and a spike-and-slab prior to inactive individual features within the active factors.

The spike-and-slab prior is effectively a mixture model where features are sampled from a zero-inflated Gaussian distribution, where $\theta_k \in (0, 1)$ dictates the level of sparsity per factor (i.e. how many active features). A value of θ_k close to 0 implies that most of the weights of factor k are shrunk to 0 (i.e. a sparse factor), whereas a value of θ_k close to 1 implies that most of the weights are non-zero (i.e. dense factors). By learning θ_k from the data, the model naturally accounts for combinations of sparse and dense factors.

3.1.12 Multi-view factor analysis models

Probabilistic PCA and Factor Analysis perform dimensionality reduction from a single input matrix. In some occasions data is collected from multiple data sources that exibit heterogeneous statistical properties, resulting in a structured data set where features are naturally partitioned into views [332, 170, 335]. A clear biological example is multi-omics data, where, for the same set of samples, multiple molecular layers are profiled. Each of the data modalities can be analysed separately using conventional (single-view) methods, but in the ideal strategy a single model should be used to leverage information across all molecular layers using a flexible and principled approach. This is referred to as the multi-view learning problem [332, 170].

A tempting approach to circumvent the multi-view learning problem is to simply concatenate all datasets before applying conventional (single-view) latent variable models [254]. However, this is prone to fail for several reasons. First, heterogeneous data modalities cannot always be modelled using the same likelihood function. For example, continuous measurements are often modelled using a normal distribution, but binary and count-based traits are not appropriately modelled by this distribution [234]. Second, even if all views are modelled with the same likelihood, differences in the scale and the magnitude of the variance can lead to some views being overrepresented in the latent space. Finally, in a multi-view data set we expect multiple sources of variation, some driven by a single view, whereas others could capture shared variability across multiple views. In other words, from a structured input space, one can also expect a structured latent representation. Not taking this behaviour into account can lead to challenges in the interpretability of the latent space.

A comprehensive review of multi-view machine learning methods can be found in [332] and a more genomics-oriented perspective in [254]. For the purpose of this thesis, I will describe only the use of latent variable models for multi-view data integration.

3.1.13 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between two datasets [121, 108].
Given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_2}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ with maximal cross-correlation. For the first pair of canonical variables, the optimisation problem is:

$$(\hat{\mathbf{u}_1}, \hat{\mathbf{v}_1}) = \arg\max_{\mathbf{u}_1, \mathbf{v}_1} corr(\mathbf{u}_1^T \mathbf{Y}_1, \mathbf{v}_1^T \mathbf{Y}_2)$$

As in conventional PCA, the linear components are constraint to be orthogonal. Hence, the first pair of canonical variables $\mathbf{u}_1$ and $\mathbf{v}_1$ contain the linear combination of variables that have maximal correlation. Subsequently, the second pair of canonical variables $\mathbf{u}_2$ and $\mathbf{v}_2$ is found from the residuals of the first canonical variables.

Given the similarity with PCA, both methods share statistical properties, including the linear mapping between the low-dimensional space and the high-dimensional space, and the closed-form solution using singular value decomposition [121, 108].
Because of its simplicity and efficient computation, CCA has widespread use as a dimensionality reduction technique [108]. Yet, as expected, CCA suffers from the same pitfalls as PCA: difficulties in selecting the number of components, lack of sparsity in the solutions and absence of probabilistic formulation. In addition, CCA have been shown to overfit for datasets where $D >> N$ [195, 103]. Hence, probabilistic versions with sparsity assumptions that reduce overfitting and improve interpretability followed.

Probabilistic Canonical Correlation Analysis

Following the derivation of probabilistic PCA [303], a similar effort enabled a probabilistic formulation of CCA as a generative model [20].
In this model, the two matrix of observations $\mathbf{Y}^1$ and $\mathbf{Y}^2$ are decomposed in terms of two weight matrices $\mathbf{W}^1$ and $\mathbf{W}^2$ but a joint latent matrix $\mathbf{Z}$:

$$\mathbf{Y}^1 = \mathbf{W}^1 \mathbf{Z} + \epsilon^1$$
$$\mathbf{Y}^2 = \mathbf{W}^2 \mathbf{Z} + \epsilon^2$$

With the following prior probability distributions:

$$
\begin{aligned}
p(z_{nk}) &= \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\
p(\epsilon^1) &= \mathcal{N}\left(\epsilon^1 \,|\, 0, \tau_1^{-1}\right) \\
p(\epsilon^2) &= \mathcal{N}\left(\epsilon^2 \,|\, 0, \tau_2^{-1}\right)
\end{aligned}
$$

As in [303], the weights and the variance of the noise are assumed to be non-probabilistic parameters, whereas the factors are probabilistic unobserved variables. This yields the following likelihood functions:

$$
\begin{aligned}
p(\mathbf{Y}^1|\mathbf{W}^1, \mathbf{Z}, \tau_1) &= \prod_{n=1}^{N} \prod_{d=1}^{D_1} \mathcal{N}\left(y_{n,d}^1 \,|\, (\mathbf{w}_{:,k}^1)^T \mathbf{z}_n, \tau_1^{-1}\right) \\
p(\mathbf{Y}^2|\mathbf{W}^2, \mathbf{Z}, \tau_2) &= \prod_{n=1}^{N} \prod_{d=1}^{D_2} \mathcal{N}\left(y_{n,d}^2 \,|\, (\mathbf{w}_{:,k}^2)^T \mathbf{z}_n, \tau_2^{-1}\right)
\end{aligned}
\tag{3.17}
$$

The corresponding graphical model is:

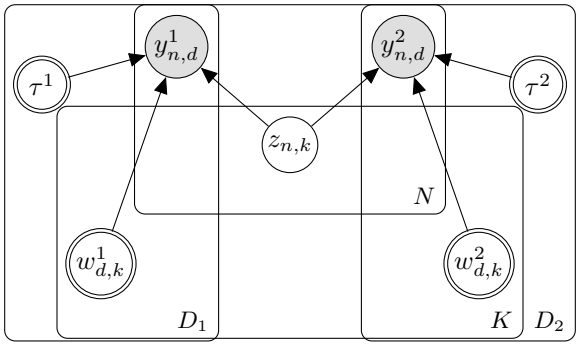


Figure 3.9: Graphical model for probabilistic Canonical Correlation Analysis

Notice that the observations for both datasets are generated from the same set of latent variables $\mathbf{Z}$. This ensures that the model is focused on capturing the variation associated with cross-correlated groups of features.

Analogously to probabilistic PCA, the expected value of the posterior distribution $p(\mathbf{Z}|\mathbf{Y}^1, \mathbf{Y}^2)$ span the same subspace as standard CCA [20]. Nonetheless, one of the many advantage of a probabilistic formulation is that it enables a broad range of principled extensions into larger graphical models.

Bayesian Canonical Correlation Analysis

A fully Bayesian treatment of CCA followed based on exactly the same principle presented in Section 3.1.10 by introducing prior distributions to all unobserved variables [320, 148]:

$$
\begin{aligned}
p(\mathbf{Z}) &= \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(z_{nk}\,|\,0,1\right) \\
p(\epsilon^1) &= \mathcal{N}\left(\epsilon^1\,|\,\sigma_1^2\right) \\
p(\epsilon^2) &= \mathcal{N}\left(\epsilon^2\,|\,\sigma_2^2\right) \\
p(\mathbf{W}^1|\boldsymbol{\alpha}) &= \prod_{k=1}^{K}\mathcal{N}\left(\mathbf{w}_{:,k}^1\,|\,0,\frac{1}{\alpha_k}\mathbf{I}_{D_1}\right) \\
p(\mathbf{W}^2|\boldsymbol{\alpha}) &= \prod_{k=1}^{K}\mathcal{N}\left(\mathbf{w}_{:,k}^2\,|\,0,\frac{1}{\alpha_k}\mathbf{I}_{D_2}\right) \\
p(\boldsymbol{\alpha}) &= \prod_{k=1}^{K}\mathcal{G}\left(\alpha_k\,|\,a_0^\alpha,b_0^\alpha\right)
\end{aligned}
$$

Resulting in the same likelihood model as in Equation (3.17). Yet, notice that an ARD is introduced per factor, allowing an automatic inference of the dimensionality in the latent subspace. Also, there is some flexibility in the definition of noise. Whereas an independent noise term can be defined per view, one can also model correlated noise by introducing a multivariate Gaussian distribution with full-rank covariance [320, 148].

The corresponding graphical model is:

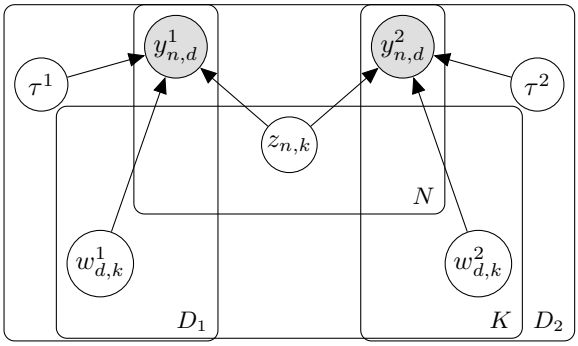


Figure 3.10: Graphical model for Bayesian Canonical Correlation Analysis

As expected, the sparsity priors yield a more sparse solution than traditional CCA, which is more appropriate for biological data analysis. However, this solution is still limited to $M = 2$ views, which leads us to the next model extension.

3.1.14 Group Factor Analysis

Group Factor Analysis (GFA) is the natural generalisation of Bayesian Canonical Correlation Analysis to an arbitrary number of views. The original idea was originally presented in [318] and a series of generalisations followed, tailored with specific assumptions for different applications [149, 169, 45, 142, 340, 250]. In this section we will outline the core principle of GFA.

Given a data set of M views $\mathbf{Y}_1, \cdots, \mathbf{Y}_M$, the task of GFA is to find K factors that capture the variability *within* as well as the variability *between* views. In other words, we want to capture factors that not only explain variance that is shared across all views but we also want to capture factors that explain variance within a single view or between different subsets of views.
The starting point is to generalise the Bayesian CCA model (Section 3.1.13) to M views:

$$\begin{aligned}\mathbf{Y}^1 &= \mathbf{W}^2\mathbf{Z} + \epsilon^1 \\ \mathbf{Y}^2 &= \mathbf{W}^2\mathbf{Z} + \epsilon^2 \\ &\cdots \\ \mathbf{Y}^M &= \mathbf{W}^M\mathbf{Z} + \epsilon^M\end{aligned}$$

Notice that there is a common factor space for all views, but there is a view-specific weight matrix. The key to disentangle the activity of each factor in each view lies on the sparsity structure imposed in the weights. Intuitively, if a factor k is not driving any variation in a specific view m we want all the individual weights to be pushed to zero. As shown before, this behaviour can be achieved using Automatic Relevance Determination (ARD) priors. However, if we were to use the same approach as in Bayesian CCA, where the ARD prior for factor k is shared across all views, then factors would be restricted to have the same activity across all views.
In GFA this is generalised as follows:

$$p(\mathbf{W}) = \prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{N}\left(\mathbf{w}_{:,k}^m \,|\, 0, \frac{1}{\alpha_k^m}\right) \tag{3.18}$$

$$p(\boldsymbol{\alpha}) = \prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{3.19}$$

This is effectively setting an ARD prior per factor k and view m. The matrix $\boldsymbol{\alpha} \in \mathbb{R}^{M\times K}$ defines four types of factors: (1) Inactive factors that do not explain variance in any view, which corresponds to all values $\boldsymbol{\alpha}_k$ being large. (2) Fully shared factors that explain variance across all views, which corresponds to all values $\boldsymbol{\alpha}_k$ being small. (3) Unique factors that explain variance in a single view, which corresponds to all values $\boldsymbol{\alpha}_k$ being large, except for one entry. (4) Partially shared factors that explain variance in a subsets views, which corresponds to a mixture of small and large values for $\boldsymbol{\alpha}_k$.

The corresponding graphical model is:

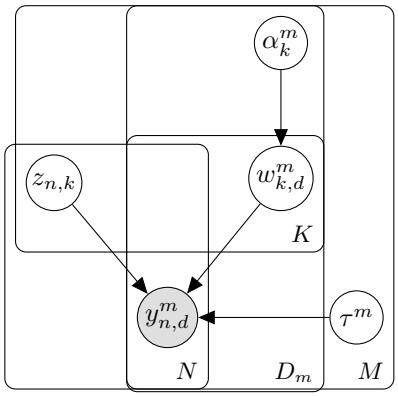


Figure 3.11: Graphical model for Bayesian Group Factor Analysis

Finally, notice that if $M = 1$ the model reduces to Bayesian PCA (Section 3.1.10), but when $M = 2$ the model does *not* reduce to Bayesian CCA because in the GFA setting factors are also allowed to capture both inter-specific variability (i.e. across views) and intra-specific variability (within a view). In Bayesian CCA, the views share a common ARD prior per factor to enforce the factors to explain variation in both views, at the expense of ignoring sources of variability that are specific to a single view.

3.2 MOFA Model description

3.2.1 Mathematical formulation

Multi-Omics Factor Analysis (MOFA) is a multi-view generalisation of conventional Factor Analysis to an arbitrary number of M data modalities (or views). It is inspired from the Group Factor Analysis framework discussed in Section 3.1.14.

The input data consists of M views $\mathbf{Y}^m \in \mathbb{R}^{N \times D_m}$ with non-overlapping features. Views, or data modalities, often represent different assays, but there is flexibility in their definition. Formally, the input data is factorised as:

$$\mathbf{Y}^m = \mathbf{Z}(\mathbf{W}^m)^T + \boldsymbol{\epsilon}^m \tag{3.20}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times K}$ is a matrix that contains the factor values and $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ are a set of M matrices (one per view) that contain the feature weights. Finally, $\boldsymbol{\epsilon}^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic:

$$p(\epsilon_d^m) = \mathcal{N}\left(\epsilon_d^m \,|\, 0, (\tau_d^m)^{-1}\right) \tag{3.21}$$

where τ corresponds to the precision (inverse of the variance). Altogether, this results in the following likelihood:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \mathbf{T}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{n=1}^{N} \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_n^T \mathbf{w}_d^m, (\tau_d^m)^{-1}\right) \tag{3.22}$$

Non-Gaussian noise models can also be defined (see Section 3.2.6), but unless otherwise stated, I will always assume Gaussian residuals.

Prior distributions for the factors

For the factors, we can define an isotropic Gaussian prior, as commonly done in most factor analysis models:

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{3.23}$$

This effectively assumes (1) a continuous latent space and (2) independence between samples and factors.

Prior distributions for the weights

The key determinant to ensure that the model is interpretable lies on the regularization structure imposed on the weights. Here we encode two levels of sparsity on their prior distributions, (1) a

view- and factor-wise ARD prior [190] and (2) a feature-wise spike-slab prior [205]:

$$p(w_{dk}^m) = (1 - \theta_k^m)\mathbb{1}_0(w_{dk}^m) + \theta_k^m \mathcal{N}\left(w_{dk}^m \,|\, 0, 1/\alpha_k^m\right)$$

The aim of the ARD prior is to disentangle the activity of factors to the different views, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not explain any variation in view m. The spike-and-slab prior encourages zero values within active factors at the level of individual features.

However, the standard formulation of the spike-and-slab prior contains a Dirac delta function, which is incompatible with the variational inference scheme. To solve this we adopt a re-parametrization of the weights w as a product of a Gaussian random variable $\hat{w}$ and a Bernoulli random variable s, [304] resulting in the following prior distribution:

$$p(\hat{w}_{dk}^m, s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, \frac{1}{\alpha_k^m}\right) \mathrm{Ber}(s_{dk}^m \,|\, \theta_k^m) \tag{3.24}$$

In this formulation α_k^m controls the activity of factor k in view m and θ_k^m controls the corresponding fraction of non-zero weights (i.e. the sparsity levels).

Finally, we define conjugate priors for θ and α:

$$p(\theta_k^m) = \mathrm{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{3.25}$$

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{3.26}$$

with hyper-parameters $a_0^\theta, b_0^\theta = 1$ and $a_0^\alpha, b_0^\alpha = 1e^{-5}$ to get uninformative priors. Posterior values of θ_k^m close to 0 implies that most of the weights of factor k in view m are shrunk to 0 (sparse factor). In contrast, a value of θ_k^m close to 1 implies that most of the weights are non-zero (non-sparse factor). A small value of α_k^m implies that factor k is active in view m. In contrast, a large value of α_k^m implies that factor k is inactive in view m.

All together, the joint probability density function of the model is given by

$$
\begin{aligned}
p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = & \prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{d,k}^m | \theta_k^m) \\
& \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right).
\end{aligned}
\tag{3.27}
$$

and the corresponding graphical model is shown below:

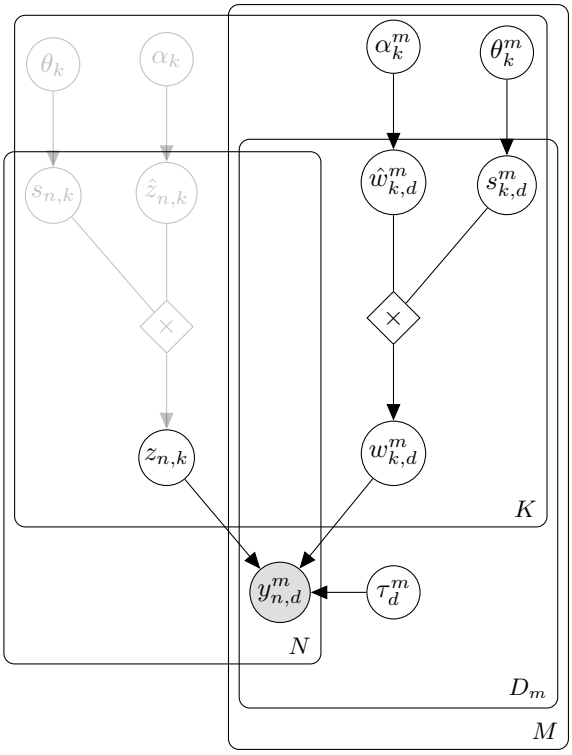


Figure 3.12: Graphical model for MOFA. Grey circles represent the observed variables whereas white circles represent hidden variables that are infered by the model. Each plate repsents a dimension of the model: M for the number of views, N for the number of samples, K for the number of factors and D_m for the number of features in the m-th view. The use of transparency in the top left nodes is intentional and becomes clear in Chapter 5.

This completes the definition of the MOFA model.

Inference

To make the model scalable to large datasets we adopt a Variational inference framework with a structured mean field approximation. A detailed overview is given in Section 3.1.4, and details on the variational updates for the MOFA model are given in Appendix A. To enable efficient inference for non-Gaussian likelihoods we employ local bounds [127, 271]. This is described in detail in Section 3.2.6.

Missing values

The probabilistic formulation naturally accounts for incomplete data matrices, as missing observations do not intervene in the likelihood. In practice, we implement this using memory-efficient binary masks $\mathcal{O}^m \in \mathbb{R}^{N \times D_m}$ for each view m, such that $\mathcal{O}_{n,d} = 1$ when feature d is observed for sample n, 0 otherwise.

3.2.2 Downstream analysis

Once trained, the MOFA model can be queried for a set of downstream analysis (Figure 3.13):

- **Variance decomposition**: calculate the variance explained (R^2) by each factor in each view. This is the first and arguably the most important plot to be inspected once the model is trained, as it summarises the variation (i.e. the signal) in a complex multi-view data set using a simple heatmap. With a quick visual inspection, this plot can be used to determine which factors are shared between multiple data modalities and which ones are exclusive to a single data modality.
- **Visualisation of the samples on the latent space**: the samples can be visualised in the latent space using beeswarm plots for individual factors or scatterplots for combinations of factors.
- **Inspection of weights**: the feature weights can be interpreted as an importance score for each feature on each factor. Inspecting the top weights for a given factor can reveal the molecular signatures that underlie each factor.
- **Association analysis between factors and external covariates**: multi-omic datasets typically consist of a large set of molecular readouts that are used for model training, and a small set of additionals covariates or response variables such as clinical outcome measurements. The external covariates are not used for model training but they can be linked to the factors *a posteriori* using a simple association analysis.
- **Imputation**: the latent factors capture a condensed low-dimensional representation of the data that can be used to generate (denoised) reconstructions of the input data. This can be valuable for the inspection of very sparse datasets.

- **Feature set enrichment analysis**: when a factor is difficult to characterise based only on the inspection of the top weights, one can compute a statistical test for enrichment of biological pathways using predefined gene-set annotations.

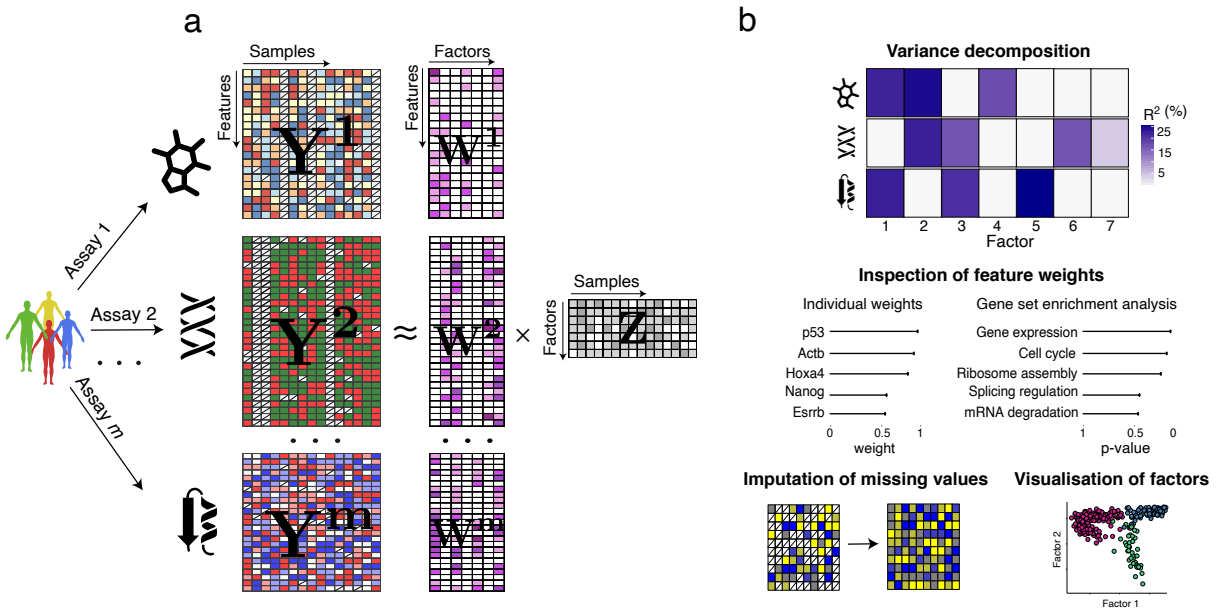


Figure 3.13: MOFA overview.
(a) The MOFA model takes as input M data matrices $(\mathbf{Y}^1, \cdots, \mathbf{Y}^M)$, each one representing a separate view or data modality. Samples must be co-occurrent but the features are not necessarily related across data modalities. MOFA performs a multi-view matrix decomposition that results in a matrix of factors ($\mathbf{Z}$) and M matrices of feature weights, one for each data modality $(\mathbf{W}^1, \cdots, \mathbf{W}^M)$.
(b) The trained MOFA model can be queried for different downstream analyses.

Interpretation of the factors

The interpretation of the factor values is intuitively similar to that of principal components in PCA. Each factor sorts cells along a one-dimensional axis with a mean of zero. Samples with different signs indicate opposite *effects* along this source of variation, with higher absolute value indicating a stronger effect.

For example, if the k-th factor captures the variability associated with commitment to cell type X, we could expect cells that belong to cell type X to be at one end of the factor (irrespective of the sign, only the relative positioning being of importance). In contrast, cells that do not belong to cell type X are expected to be at the other end of the factor.

Interpretation of the weights

The weights provide a score for each feature on each factor. Features with no association with the factor are expected to have values close to zero (as specified by the prior distributions). In contrast, features with strong association with the factor are expected to have large absolute values. The

sign of the weight indicates the direction of the effect such that a positive weight indicates that the feature is positively associated with the factor values.

Following the example above, genes that are upregulated cell type X are expected to have large positive weights, whereas genes that are downregulated in cell type X (or, equivalently, upregulated in the other cell types) are expected to have large negative weights. Genes that do not change in expression between the cell types are expected to have a value of zero.

Variance decomposition

The first step in the downstream analysis is to calculate the percentage of variance explained for each factor k in each view m ($R^2_{m,k}$), which can be visualised using a heatmap (see Figure 3.13b). This is done by adapting the coefficient of determination statistic that is traditionally used for linear regression analysis:

$$R^2_{m,k} = 100 \frac{(\sum_{n=1}^{N} \sum_{d=1}^{D_m} y^m_{n,d} - z_{nk} w^m_{dk})^2}{(\sum_{n=1}^{N} \sum_{d=1}^{D_m} y^m_{n,d})^2}$$

3.2.3 Model selection and consistency across random initilizations

The optimisation problem in MOFA is not convex and the resulting posterior distributions depend on the initialisation of the model. Thus, when doing random initialisation of the parameters and/or expectations it becomes mandatory to perform model selection and assess the consistency of the factors across different trials. The strategy we adopted in this work is to train several MOFA models under different parameter initialisations, where the expectation of each node is randomly sampled from its underlying distribution. After fitting, we select the model with the highest ELBO for downstream analysis. In addition, we evaluate the robustness of the factors by plotting the Pearson correlations between factors across all trials (Figure 3.14).

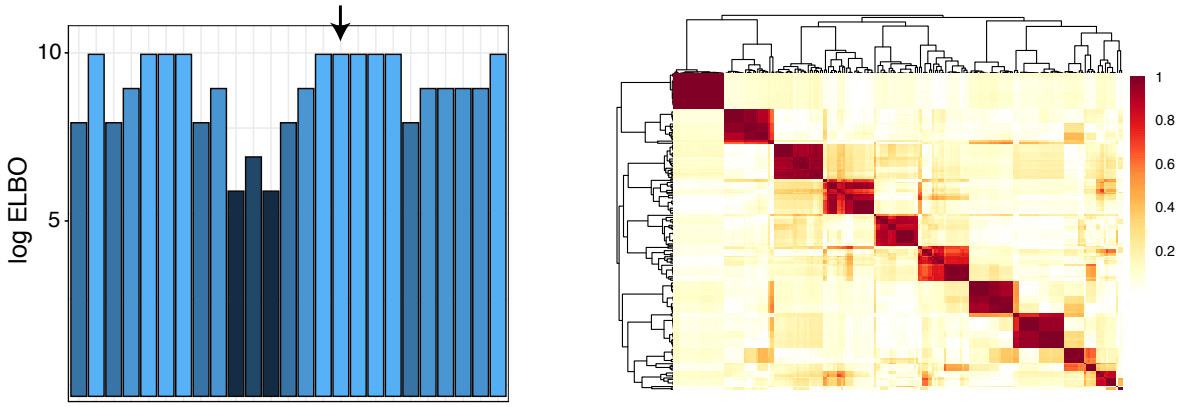


Figure 3.14: Model selection and robustness analysis in MOFA.
The left plot the log ELBO (y-axis) for 25 model instances (x-axis). The arrow indicates the model with the highest ELBO that would be selected for downstream analysis. The right plot displays the absolute value of the Pearson correlation coefficient between pairwise combinations of all factors across the 25 model instances. A block-diagonal matrix indicates that factors are robustly estimated regardless of the initialisation.

3.2.4 Learning the number of factors

As described in Section 3.1.11, the use of an ARD prior allows factors to be actively pruned by the model if their variance explained is negligible. In the implementation we control the pruning of factors by a hyperparameter that defines a threshold on the minimum fraction of variance explained by a factor (across all views). Additionally, because of the non-convexity of the optimisation problem, different model instances can potentially yield solutions with different number of active factors. Thus, the optimal number of factors can be selected by the model selection strategy outlined in Section 3.2.3.

3.2.5 Monitoring convergence

An attractive property of Variational inference is that the objective function (the ELBO) increases monotonically at every iteration. This provides a simple way of monitoring convergence:

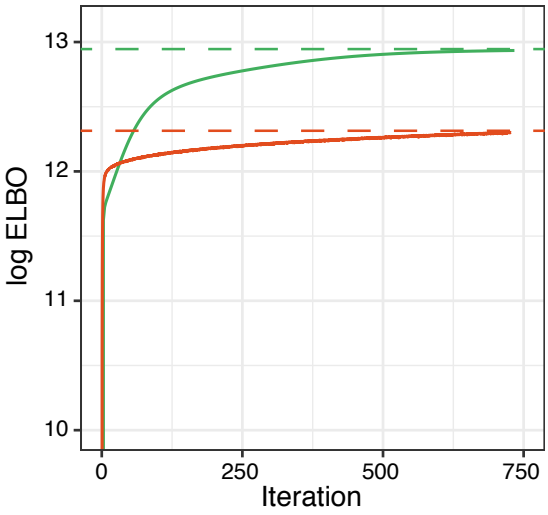


Figure 3.15: Training curve for two different instances of MOFA with random initialisations. The y-axis displays the log of the ELBO, with higher values indicating a better fit. The x-axis displays the iteration number. The horizontal dash lines mark the value of the ELBO upon convergence.

Training is stopped when the change in the lower bound becomes smaller than a predefined threshold.

3.2.6 Modelling and inference with non-Gaussian data

Gaussian likelihoods are sufficient to model the residuals of most types of continuous data. Thus, when possible, we advise the user to apply data transformations (i.e. *log* transformation for example) and use the Gaussian likelihood model. However, there are cases where Gaussian likelihoods are not appropriate, even after data transformations, namely binary and (low) count data. However, non-Gaussian likelihoods are problematic because they are not conjugated with prior distributions, and this prevents the use of the efficient variational inference scheme for Gaussian likelihoods (see Appendix A).

To implement efficient variational inference in conjunction with a non-Gaussian likelihood we adapt prior work from [271], where the full derivation can be found. The idea is to approximate non-Gaussian observations by a normally-distributed *pseudo-data* that is constructed using second-order Taylor expansions. This defines a lower bound that can be improved by adjusting parameters at each iteration. Denoting the parameters in the MOFA model as $\mathbf{X} = (\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\theta})$, recall that the variational framework approximates the posterior $p(\mathbf{X}|\mathbf{Y})$ with a distribution $q(\mathbf{X})$, which is indirectly optimised by optimising a lower bound of the log model evidence. The resulting optimization problem can be re-written as

$$\min_{q(\mathbf{X})}(-\mathcal{L}(\mathbf{X})) = \min_{q(\mathbf{X})} \mathbb{E}_q\big[-\log p(\mathbf{Y}|\mathbf{X})\big] + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})].$$

Let's now assume a general likelihood function $p(\mathbf{Y}|\mathbf{X}) = f(\mathbf{Y}|\mathbf{C})$ with $\mathbf{C} = \mathbf{Z}\mathbf{W}^T$ that we can write as

$$-\log p(\mathbf{Y}|\mathbf{X}) = \sum_{n=1}^{N}\sum_{d=1}^{D} f_{nd}(c_{nd})$$

with $f_{nd}(c_{nd}) = -\log p(y_{nd}|c_{nd})$. For simplicity in the derivation I will assume a single view and thus drop the index m for clarity.

Extending [271] to our heteroscedastic noise model, we require $f_{nd}(c_{nd})$ to be twice differentiable and upper bounded by a constant κ_d. I do not prove this here, but this property holds true in many important models as for example the Bernoulli and Poisson likelihoods, as demonstrated in [271]. Under this assumption a lower bound on the log likelihood can be defined using Taylor expansion:

$$f_{nd}(c_{nd}) \leq \frac{\kappa_d}{2}(c_{nd} - \zeta_{nd})^2 + f'(\zeta_{nd})(c_{nd} - \zeta_{nd}) + f_{nd}(\zeta_{nd}) := q_{nd}(c_{nd}, \zeta_{nd}),$$

where ζ_{nd} are additional variational parameters that determine the location of the Taylor expansion and have to be optimised to make the lower bound as tight as possible. Plugging the bounds into the optimization problem above, we obtain:

$$\min_{q(\mathbf{X}),\boldsymbol{\zeta}} \quad \sum_{d=1}^{D}\sum_{n=1}^{N} \mathbb{E}_q[q_{nd}(c_{nd}, \zeta_{nd})] + \mathrm{KL}[q(\mathbf{X})||p(\mathbf{X})]$$

The new objective function has two class of parameters to optimise: the parameters associated with the local bounds $\boldsymbol{\zeta}$, and the parameters associated with the model ($\mathbf{X}$). The algorithm proposed in [271] alternates between updates of $\boldsymbol{\zeta}$ and q($\boldsymbol{\Theta}$). The update for $\boldsymbol{\zeta}$ is given by

$$\zeta \leftarrow \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}]^T$$

where the expectations are taken with respect to the corresponding q distributions.
On the other hand, the updates for $q(\mathbf{X})$ are identical to the standard variational Bayesian updates with Gaussian likelihoods, but with the observed data $\mathbf{Y}$ replaced by the *pseudo-data* $\hat{\mathbf{Y}}$ and where the precisions τ_{nd} (which were treated as random variables) are replaced by the constant terms κ_d introduced above. This might seem a minor change, but it is very important. In the Gaussian case the model infers a variance parameter for each feature which means that MOFA explicitly models heteroscedastic noise, but when using non-Gaussian likelihoods this is no longer possible.

Finally, the general formula pseudodata is given by (derived in [271]):

$$\hat{y}_{nd} = \zeta_{nd} - f'(\zeta_{nd})/\kappa_d$$

where different log likelihood functions $f(\cdot)$ yield different κ_d values.

In MOFA we implemented a Bernoulli likelihood model for binary data and a Poisson likelihood model for (low) count data.

Bernoulli likelihood for binary data

When the observations are binary, $y \in \{0, 1\}$, they can be modelled using a Bernoulli likelihood:

$$p(y|c) = \frac{e^{yc}}{1 + e^c}$$

The second derivative of the log likelihood is bounded by:

$$f''(c) = \sigma(c)\sigma(-c) \leq 1/4 := \kappa$$

where σ is the sigmoid function $f(c) = 1/(1 + e^{-c})$.

The pseudodata updates are given by

$$\hat{y}_{nd} = \zeta_{nd} - 4 * (\sigma(\zeta_{nd}) - y_{nd})$$

Poisson likelihood for count data

When observations are natural numbers, such as count data $y \in \mathbb{N} = \{0, 1, \cdots\}$, they can be modelled using a Poisson likelihood:

$$p(y|c) = \lambda(c)^y e^{-\lambda(c)}$$

where $\lambda(c) > 0$ is a convex rate function. As done in [271], here we adopt the rate function $\lambda(c) = \log(1 + e^c)$, which yields the following upper bound of the second derivative of the log-likelihood:

$$f''_{nd}(c_{nd}) \leq \kappa_d = 1/4 + 0.17 * \max(\mathbf{y}_{:,d}).$$

The pseudodata updates are given by

$$\hat{y}_{nd} = \zeta_{nd} - \frac{\mathrm{S}(\zeta_{nd})(1 - y_{nd}/\lambda(\zeta_{nd}))}{\kappa_d}.$$

3.2.7 Theoretical comparison with published methods

A variety of latent variable models exist with the aim of perfoming multi-view data integration, most of them inspired by the Group Factor Analysis formulation. A summary is provided in the table below. MOFA is the only method that scales to large datasets (employs Variational Bayes inference instead of MCMC-based approaches), has a combination of ARD and spike-slab regularisation on the weights, and is also capable of handling non-gaussian modalities and missing values.

Publication	**Inference**	**View-wise sparsity**	**Feature-wise sparsity**	**Missing values**	**Likelihood**	**Noise model**
Shen2009	EM, grid search	L_1- penalties	L_1-penalty	No	Gaussian	Hetero-scedastic
Mo2013	EM, grid search	L_1- penalties	L_1-penalty	No	Gaussian, Poisson, Bernoulli	Hetero-scedastic
Virtanen2012	VB	ARD	None	No	Gaussian	Homo-scedastic
Klami2014	VB	ARD	None	No	Gaussian	Homo-scedastic
Bunte2016	Gibbs	ARD	Spike-Slab	No	Gaussian	Homo-scedastic
Hore2016	VB	None	Spike-Slab	Yes	Gaussian	Hetero-scedastic
Remes2016	VB	ARD	None	No	Gaussian	Homo-scedastic
Zhao2015	Gibbs	ARD	Three-parameter beta prior	No	Gaussian	Hetero-scedastic
Leppaaho2017	Gibbs	ARD	Spike-Slab	Yes	Gaussian	Homo-scedastic
MOFA	VB	ARD	Spike-Slab	Yes	Gaussian, Poisson, Bernoulli	Hetero-scedastic

Table 3.1: Overview of latent variable methods for multi-view data integration. Abbreviations used: VB (variational Bayes inference), Gibbs (Gibbs sampling based inference), ARD(Automatic Relevance Determination)

3.3 Model validation with simulated data

We used simulated data from the generative model to systematically test the technical capabilities of MOFA.

3.3.1 Recovery of simulated factors

First, we tested the ability of MOFA to recover simulated factors under varying number of views, features, factors and with different amounts of missing values.
For every simulation scenario we initialised a model with a high number of factors ($K = 100$), and inactive factors were automatically dropped during model training by the ARD prior. In addition, to test the robustness under different random initialisations, 10 model instances were trained for every simulation scenario.
We observe that in most settings the model accurately recovers the correct number of factors (Figure 3.16). Exceptions occur when the dimensionality of the latent space is too large (more than 50 factors) or when an excessive amount of missing values (more than 80%) is present in the data.

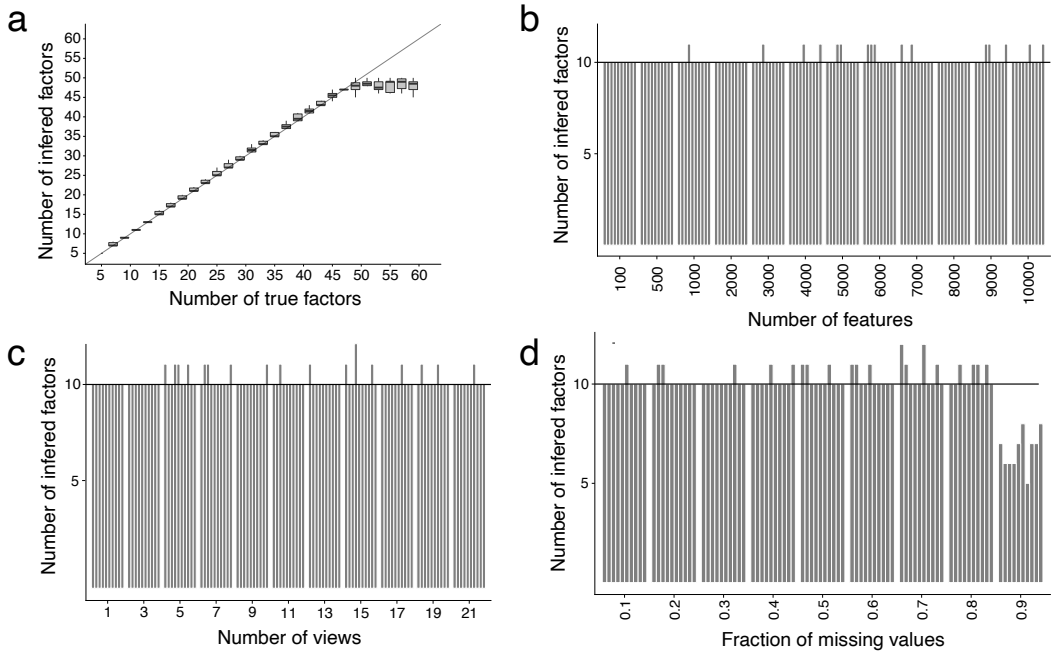


Figure 3.16: Assessing the ability to recover simulated factors.
In all plots the y-axis displays the number of infered factors. (a) x-axis displays the number of true factors, and boxplots summarise the distribution of inferred factors across 10 model instances. For (b-d) the true number of factors was set to $K = 10$ and each bar corresponds to a different model instance. (b) x-axis displays the number of features, (c) x-axis displays the number of views, (d) x-axis displays fraction of missing values.

View-wise sparsity on the weights

One of the essential features of MOFA is the use of an ARD prior aimed at disentangling the activity of factors across views (see Section 3.1.11 and Section 3.2).
We simulated data from the generative model such that the factors were set to be active or inactive in specific views by sampling α_k^m from a discrete distribution with values $\{1, 1e3\}$. We compared the performance with a popular integrative clustering method (iCluster) that is also formulated as a latent variable model [206]. In iCluster each factor shares the same sparsity constraint across all views, and hence the model is less accurate at detecting factors that show differential activity across different views:

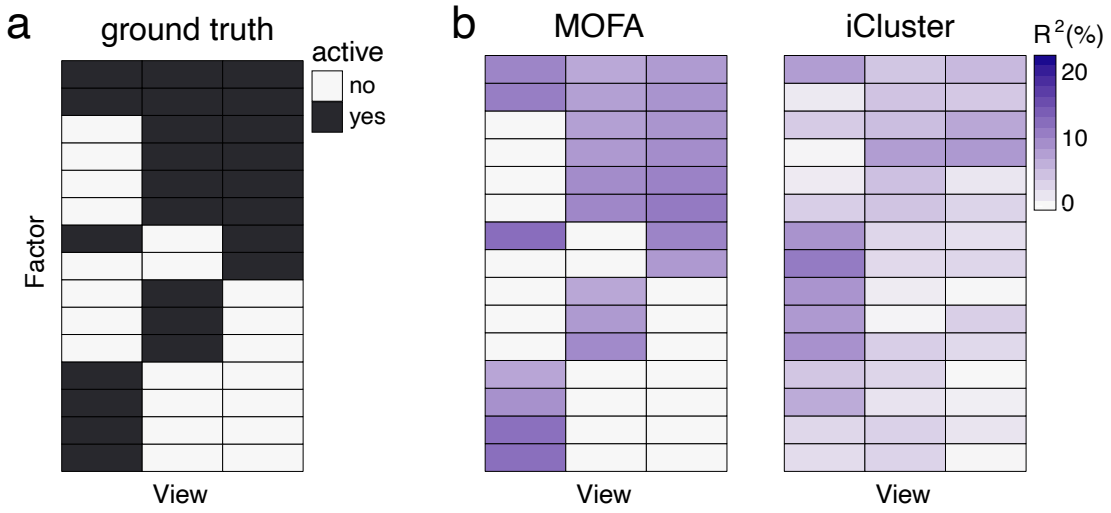


Figure 3.17: Evaluating the ability to recover differential factor activity across views. (a) The true activity pattern, with factors sampled to display differential activity across views. (b) Percentage of variance explained for each factor in each view, for MOFA and iCluster [206].

Feature-wise sparsity on the weights

In MOFA we implemented a spike-and-slab prior prior to enforce feature-wise sparsity on the weights with the aim of delivering a more interpretable solution (see Section 3.2.1).
To assess the effect of the spike-and-slab prior we trained a group of models with and without the spike-and-slab prior. Importantly, both models contain the ARD prior, which should provide some degree of regularisation. To compare both options to a non-sparse method, we also fit a Principal Component Analysis on the concatenated data set. As expected, we observe that the spike-and-slab prior induces more zero-inflated weights, although the ARD prior provided a moderate degree of regularisation. The PCA solution was notably more dense than both Bayesian models (Figure 3.18).

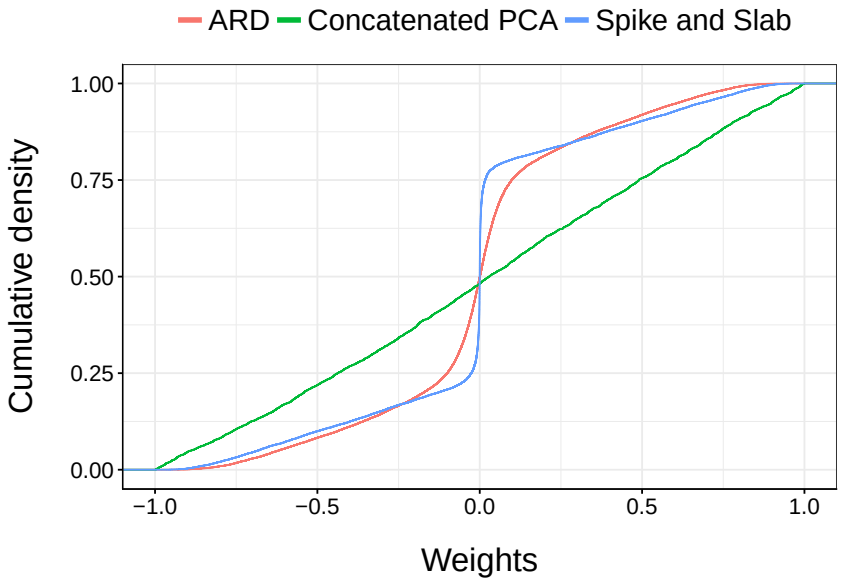


Figure 3.18: Assessing the sparsity priors on the weights.
The plot shows the empirical cumulative density function of the weights for an arbitrary factor in a single view. The weights were simulated with a sparsity level of $\theta_k^m = 0.5$ (50% of active features.)

3.3.2 Non-Gaussian likelihoods

A key improvement of MOFA with respect to previous methods is the use of non-Gaussian likelihoods to integrate data modalities with different types of readouts. In particular, as described in Section 3.2.6, we implemented a Bernoulli likelihood to model binary data and a Poisson likelihood to model count data.

To validate both likelihood models, we simulated binary and count data using the generative model and we fit two sets of models for each data type: a group of models with a Gaussian likelihood and a group of models with a Bernoulli or Poisson likelihood, respectively.

Reassuringly, we observe that although the Gaussian likelihood is also able to recover the true number of factors, the models with the non-Gaussian likelihoods result in a better fit to the data:

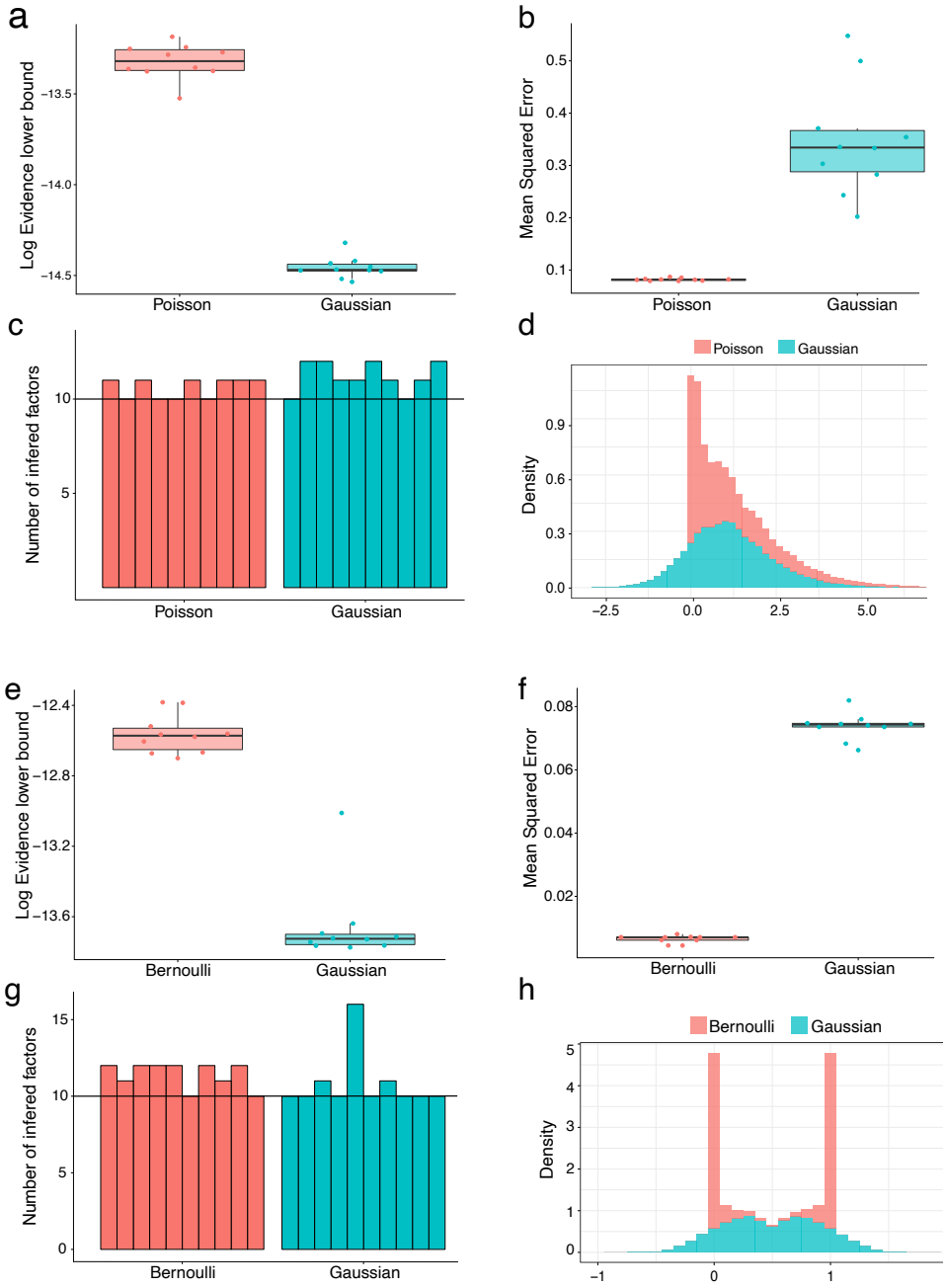


Figure 3.19: Validation of the non-Gaussian likelihood models using simulated data. (a-d) Comparison of Poisson and Gaussian likelihood models applied to count data. (e-h) Comparison of Bernoulli and Gaussian likelihood models applied to binary data. (a,e) The y-axis displays the ELBO for each model instance (x-axis). (b,f) The y-axis displays the mean reconstruction error for each model instance (x-axis). (c,g) The y-axis displays the number of estimated factrors for each model instance (x-axis). The horizontal dashed line marks the true number of factors $K = 10$. (d,h) Distribution of reconstructed data. Plotted are the expected values of the inferred posterior distributions, not samples from the corresponding posteriors. This is why reconstructed measurements are continuous and not discrete.

3.3.3 Scalability

Finally, we evaluated the scalability of the model when varying each of its dimensions independently, and we compared the speed with an implementation of GFA that uses Gibbs Sampling [169] and the popular Cluster+[206], which adopts a maximum-likelihood approach with grid search to optimise the hyperparameters. Overall, we observe that MOFA scales linear with respect to all dimensions and is significantly faster than any of the three evaluated techniques (Figure 3.20).

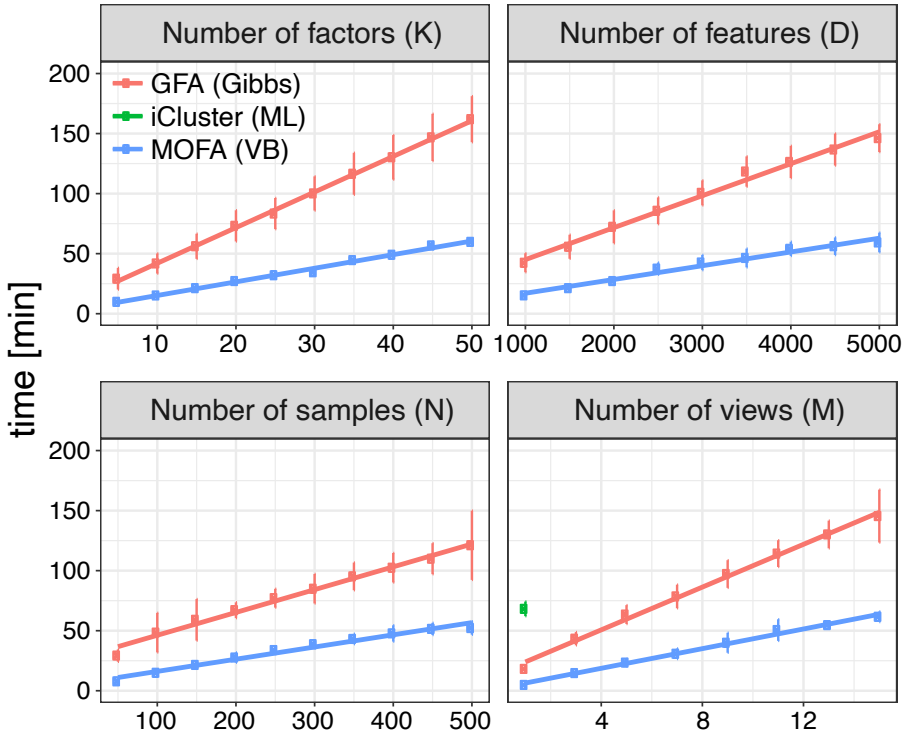


Figure 3.20: Evaluation of scalability in MOFA.
Shown is the time required for convergence (y-axis, in minutes). The x-axis displays the value of the dimension that was tested, either number of factors (K), number of features (D), number of samples (N) and number of views (M). Baseline parameters were $M = 3, K = 10, D = 1000, N = 100$. Each line represents a different model, GFA (red), MOFA (blue) and iCluster (green). Default convergence criteria where used for all methods. Each dot displays the average time across 10 trials with error bars denoting the standard deviation. iCluster is only shown for one value as all other settings required more than 200min for convergence.

As a real application showcase, the training on the CLL cohort that is described below (Figure 3.21) required 25 minutes using MOFA, 34 hours with GFA and 5-6 days with iCluster.

3.4 Application to a cohort of Chronic Lymphocytic Leukaemia patients

Personalised medicine is an attractive field for the use of multi-omics, as dissecting heterogeneity across patients is a major challenge in complex diseases, and requires data integration from multiple biological layers [56, 63, 7].

To demonstrate the potential of MOFA, we applied it to a publicly available study of 200 patient samples of Chronic Lymphocytic Leukaemia (CLL) profiled for somatic mutations, RNA expression, DNA methylation and *ex vivo* drug responses [75], all of them at the bulk level. We selected this data set for three main reasons: (1) The complex missing data structure, with nearly 40% samples having incomplete assays (Figure 3.21). As described in Section 3.2.1, the inference framework implemented in MOFA should cope with large amounts of missing values, including missing assays. (2) After data processing, three assays had continuous observations whereas for the somatic mutations the observations were binary. As described in Section 3.2.6, MOFA can combine different likelihood models. (3) The existence of clinical covariates provide an excellent test to evaluate whether the MOFA factors can capture the molecular variation that underlies clinically-relevant phenotypes.

3.4.1 Data overview and processing

Data processing and normalisation is essential for the model to work and it requires a few considerations. First, in the case of count-based assays such as RNA-seq one needs to remove differences in library size between samples. If not done correctly, the signal in the data will be dominated by this (undesired) source of variation, and more subtle heterogeneity will be harder to identify. Similarly, batch effects and other undesired technical sources of variation should be regressed out *a priori*, although this was not the case for this particular data set. Second, feature selection must be performed by selecting highly variable features. A proper feature selection will increase the signal-to-noise-ratio, it will simplify model selection and it will speed up the training procedure. Finally, as discussed above, the total number of features can influence the contribution of a data modality to the latent space. To mitigate this problem it is recommended to keep the number of features per view within the same order of magnitude, when possible.

Here we proceed to briefly describe the different data modalities and outline the basic data processing steps that we performed before applying MOFA:

- **RNA expression** was profiled using bulk RNA-seq. Genes with low counts were filtered out and the data was subsequently normalized using DESeq2 [181]. Feature selection was performed by considering the top 5,000 most variable genes.
- **DNA methylation** was profiled using Illumina 450K arrays. We converted the beta-values to M-values, as it has better statistical properties when modelled with a Gaussian distribution [79]. Feature selection was performed by considering the top 1% most variable CpG sites.

- ***Ex vivo* Drug response** was screened using the ATP-based CellTiter-Glo assay. Briefly, the asay includes a panel of 62 drugs at 5 different concentrations each, for a total of 310 measurements. The readout is a number proportional to the fraction of viable cells in culture based on quantitation of the ATP present, which signals the presence of metabolically active cells.

- **Somatic mutations** were profiled using a combination of targeted and whole exome sequencing. Feature selection was performed by considering only mutations that were present in at least three samples, which resulted in a total of 69 mutations.

3.4.2 Model overview

In this data set, MOFA recovered $K = 10$ factors, each one explaining a minimum of 3% of variance in at least one assay. Interestingly, MOFA detected factors which are shared across several data modalities (Factors 1 and 2, sorted by variance explained). Some factors captured sources of covariation between two data modalities (Factor 3 and 5, active in the RNA expression and drug response). In addition, some factors captured variation that is unique to a single data modality (Factor 4, active in the RNA expression data).
All together, the 10 MOFA factors explained 41% of variance in the drug response data, 38% in the mRNA expression, 24% in the DNA methylation and 24% in somatic mutations.

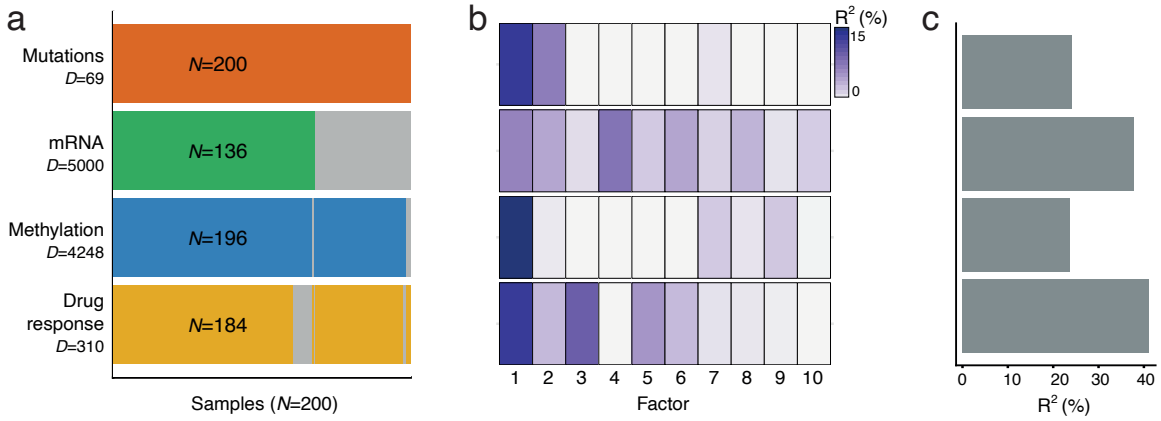


Figure 3.21: Application of MOFA to a study of chronic lymphocytic leukaemia. Model overview.
(a) Data overview. Assays are shown in different rows (D = number of features) and samples (N) in columns, with missing samples shown using grey bars. Notice that some samples are missing entire assays.
(b) Variance explained (%) by each Factor in each assay.
(c) Total variance explained (%) for each assay by all factors.

The first two factors are the most interesting from a molecular perspective, as they capture a phenotypic effect that is manifested across multiple molecular layers. To annotate Factors 1 and 2 we proceeded to visualise the feature weights, starting by the (binary) somatic mutation data, as it is the simplest data modality to interpret. Inspection of the top weights revealed that Factor 1 was associated with the mutation status of the immunoglobulin heavy-chain variable (IGHV) region,

while Factor 2 was aligned with trisomy of chromosome 12 (Figure 3.22).
Remarkably, in a completely unsupervised fashion, MOFA recovered the two most important clinical markers in CLL as the two major axes of molecular disease heterogeneity [86, 44, 66].

Next, we visualised the samples in the latent space spanned by Factors 1 and 2. A scatterplot based on these factors shows a clear separation of patients by their IGHV status on the first Factor and presence or absence of trisomy 12 on the second Factor (Figure 3.22). Interestingly, 24 patients lacked IGHV status measurements (grey crosses) due to quality control filtering in the DNA sequencing assay. Nonetheless, MOFA was able to pool information from the other molecular layers to map those samples to the latent space, and could be classified to the corresponding molecular subgroup.

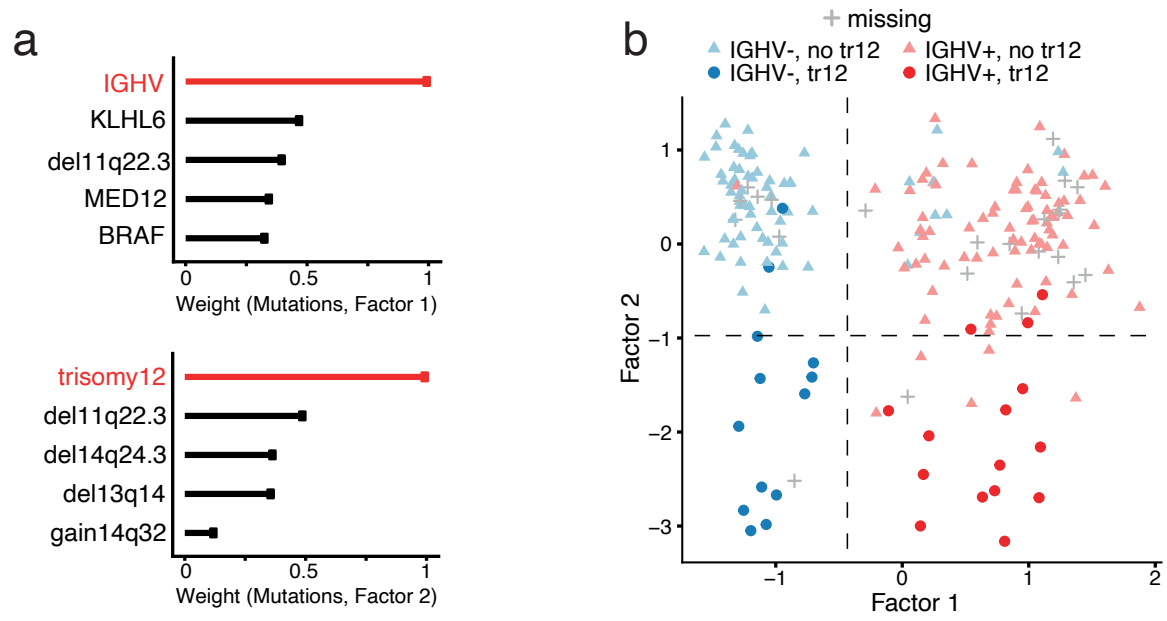


Figure 3.22: Visualisation of the genetic signature underlying Factor 1 and 2
(a) Weights of the top somatic mutations for Factors 1 and 2. (b) Scatterplots of Factors 1 and 2. Each dot corresponds to one sample and the colours denote the IGHV status of the tumours samples; symbol shape indicate chromosome 12 trisomy status.

IGHV status is currently the most important prognostic marker in CLL and has routinely been used to distinguish between two distinct subtypes of the disease[86]. Molecularly, it is a surrogate of the level of activation of the B-cell receptor, which is in turn related to the differentiation state of the tumoral cells. Multiple studies have associated mutated IGHV with a better response to chemotherapy, whereas unmutated IGHV patients have a worse prognosis [86, 44, 66].
In clinical practice, the IGHV status has been considered binary. Our results suggest that this is a fairly good approximation, but a more complex structure with at least three groups or a potential underlying continuum is supported (Figures 3.22 and 3.23), as also suggested in [240].

3.4.3 Molecular characterisation of Factor 1

An important step in the MOFA pipeline is the characterisation of the molecular signatures underlying each Factor. I will demonstrate this for Factor 1, although a similar strategy can be applied to Factor 2.

On the RNA expression, inspection of the top weights pinpoint genes that have been previously associated to IGHV status, some of which have been proposed as clinical markers[316, 209]. Heatmaps of the RNA expression levels for these genes reveals clear differences between samples when ordered according to the Factor 1 values.

On the drug response data the weights highlight kinase inhibitors targeting the B-cell receptor pathway. Splitting the patients into three groups based on k-means clustering shows clear separation in the drug response curves.

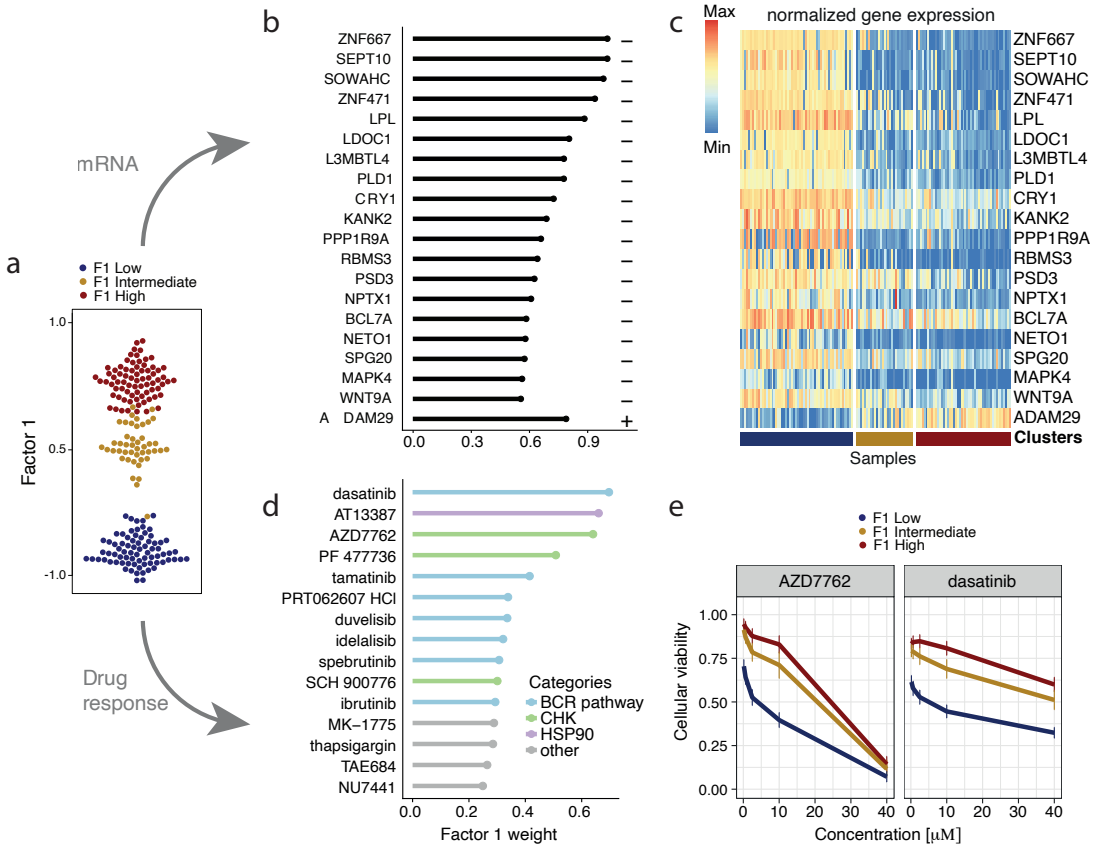


Figure 3.23: Characterization of MOFA Factor 1 as IGHV status.
(a) Beeswarm plot of Factor 1 values, where each dot corresponds to a patient sample. Colours denote three groups found by applying 3-means clustering on the Factor values.
(b) Genes with the largest weights (in absolute values) in the mRNA data. Plus or minus symbols on the right indicate the sign of the weight.
(c) Heatmap of gene expression values for the genes with the largest weights displayed in (b).
(d) Drugs with the largest weights (in absolute values) in the Drug response data, coloured by the drug's target category.
(e) Drug response curves for two of the drugs with top weights, stratified by the clusters displayed in (a).

3.4.4 Molecular characterisation of other factors

Despite their clinical importance, Factor 1 (IGHV status) and Factor 2 (chr12 trisomy) they explain less than 20% variability in each data modality, suggesting the existence of more subtle sources of variation. As an example, we will also characterise Factor 5, which explains 2% of the variance in the mRNA and 6% of variance in the drug response.
As mentioned in Section 3.2.2, instead of exploring the feature weights individually, factors can be annotated using gene set annotations. This procedure is particularly appealing for RNA expression data, as a rich amount of resources exist that have categorised genes into ontologies in terms of biological pathways, molecular function and cellular components [87, 17].

Briefly, the idea is to aggregate the weights using prior information to obtain a single statistic for each gene set, which can be tested against a competitive null hypothesis. Inspired from [94], in MOFA we implemented several scoring schemes and a variety of parametric and unparametric statistical tests. By default we use the weights as feature statistics and the average difference in the weight values as the feature set statistic. P-values are then obtained per feature set and factor via a simple t-test.

Gene Set Enrichment Analysis on the RNA weights using the Reactome annotations [87] reveals that Factor 2 is strongly enriched for oxidative stress and senescence pathways. Inspection of the top features highlights the importance of heat shock proteins (HSPs), a group of proteins that are essential for protein stability which are up-regulated upon stress conditions like high temperatures, pH shift or oxidative stress. Importantly, HSPs can be elevated in tumour cells and potentially contribute to prolonged tumour cell survival[72]. In agreement with the findings from the mRNA view, the drugs with largest weights on Factor 5 belong to clinical categories associated with stress response, such as target reactive oxygen species and DNA damage response (Figure 3.24)

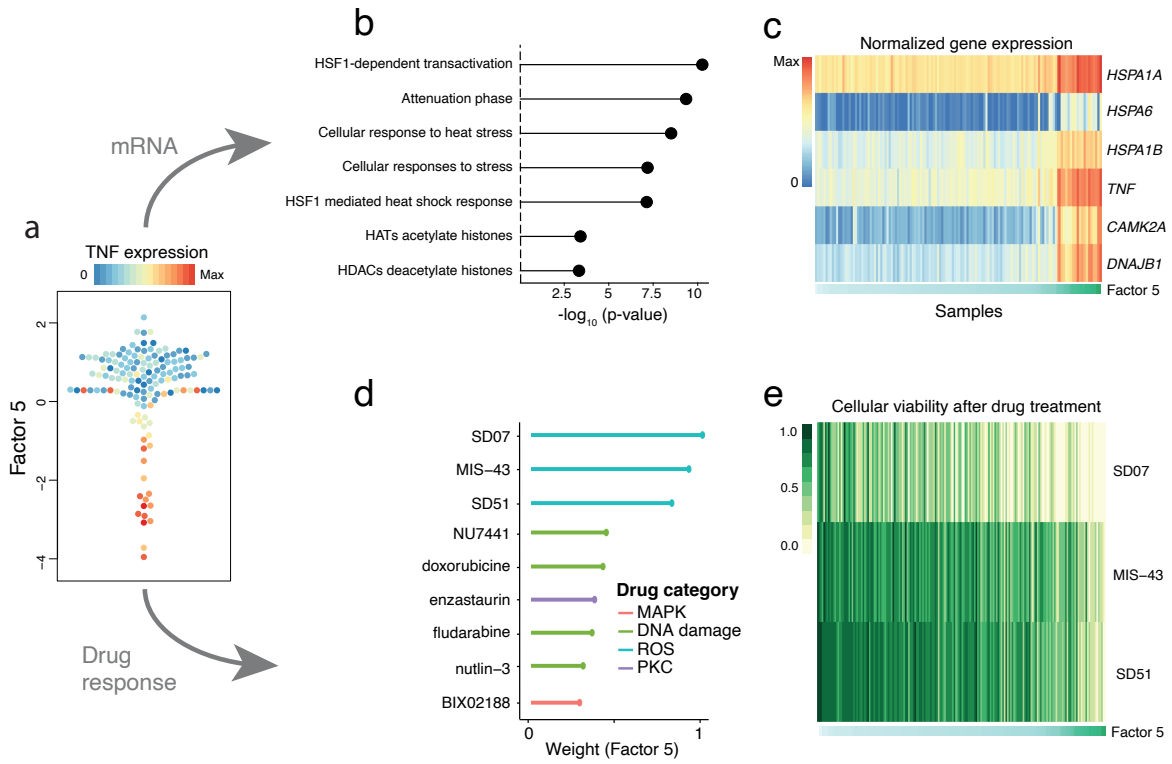


Figure 3.24: Characterization of Factor 5 in the CLL cohort as oxidative stress response.
(a) Beeswarm plot of Factor 5, where each dot corresponds to a patient sample. Colours represent the expression of TNF, an inflammatory stress marker that is present among the top mRNA weights.
(b) Gene set enrichment analysis results using Reactome pathways. Displayed are the top pathways with the strongest enrichment.
(c) Heatmap of mRNA expression values for representative genes among the top weights. Samples are ordered by their Factor 5 values.
(d) Weights for the top drugs, annotated by target category.
(e) Heatmap of drug response values for the top three drugs. Samples are ordered by their Factor 5 values, as in (c).

3.4.5 Prediction of clinical outcomes

We conjectured that the integration of multiple molecular layers could allow an improved prediction of the patients' clinical outcome. To evaluate the utility of the MOFA factors as predictors of clinical outcomes we fit Cox regression models [64] using the patients' time to next treatment (TTT) as a response variable. Two types of analysis were performed: a univariate analysis where each Factor was independently associated with TTT, and a multivariate analysis where the combination of all factors were used to predict TTT (Figure 3.25). In the univariate Cox models, we observe that Factor 1 (IGHV status), Factor 7 (associated with chemo-immunotherapy treatment prior to sample collection) and Factor 8 (enriched for Wnt signalling) were significant predictors of TTT. Accordingly, when splitting patients into binary groups based on the corresponding Factor values, we observe clear differences in the survival curves. In the multivariate Cox model, MOFA (Harrell's C-Index C=0.78) outperformed all other input settings, including PCA on single-omic data (C=0.68-0.72), individual genetic markers (C=0.66) as well PCA applied to the concatenated data matrix (C=0.74).

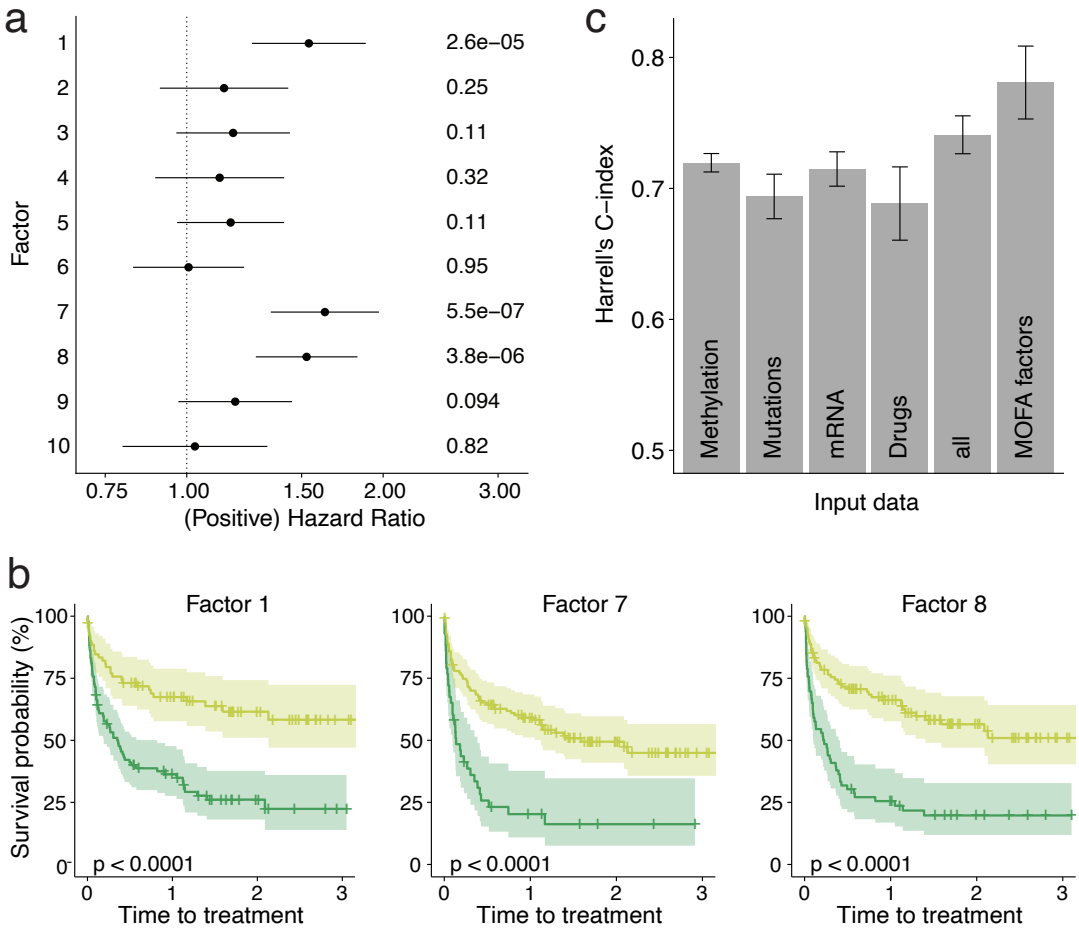


Figure 3.25: Association analysis between MOFA factors and clinical putcome.
(a) Association of MOFA factors to time to next treatment using a univariate Cox regression model Error bars denote 95% confidence intervals. Numbers on the right show p-values for each Factor.
(b) Kaplan-Meier plots for the three MOFA factors that show a significant association with time to next treatment.
(c) Prediction accuracy of time to treatment using multivariate Cox regression trained with the first 10 principal components applied to single data modalities, the full data set or the 10 MOFA factors. Shown are average values of Harrell's C-index from fivefold cross-validation. Error bars denote standard error of the mean.

3.4.6 Imputation of missing values

A promising application of MOFA is the imputation of missing values, including the potential to impute of entire assays.

The principle of imputation in MOFA follows the same logic as simulating from the generative model: if the factors and weights are known, the input data can be reconstructed by a simple matrix multiplication:

$$\hat{\mathbf{Y}} = \mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}]^T$$

where $\mathbb{E}[\mathbf{Z}]$ and $\mathbb{E}[\mathbf{W}]$ denote the expected values of the variational distributions for the factors and the weights, respectively. Notice that, when using the expectations of the posterior distributions, the noise ϵ (Equation (5.16)) has a mean of zero and does not contribute to the predictions.

The equation above results in point estimates, but it ignores the uncertainity on $\mathbf{Z}$ and $\mathbf{W}$. Instead of relying in point estimates, one could adopt a more Bayesian approach and calculate the posterior predictive distribution by propagating the uncertainity [97]. Nonetheless, due to the nature of the optimisation problem in variational inference, the variance of the posterior distributions can be underestimated (see Section 3.1.5). In addition, this would be substantially more complex to implement and would result in a significant increase in computational complexity, hence we did not implement this strategy.

To assess the imputation performance, we trained MOFA models using a data set of complete measurements (a total of N=121 samples) after masking parts of the drug response measurements. In a first experiment, we masked values at random, and in a second experiment we masked the entire drug response measurements. We compared the imputation accuracy of MOFA to some established imputation strategies, including imputation by feature-wise mean, SoftImpute [194], and a k-nearest neighbour method [311]. For both imputation tasks, MOFA consistently yielded more accurate predictions, albeit the differences are less pronounced in the imputation of full assays, a significantly more challenging task.

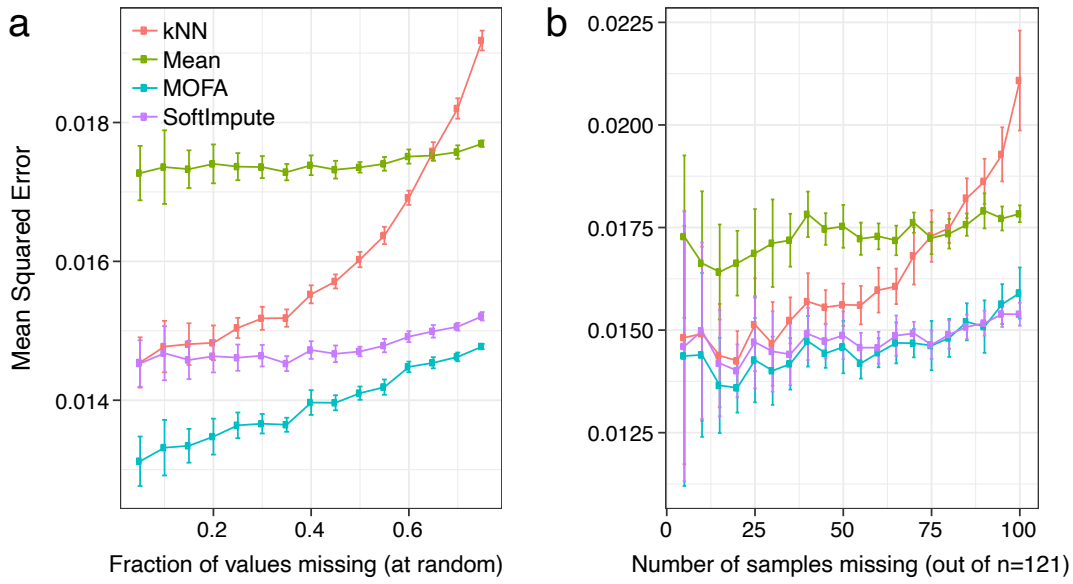


Figure 3.26: Evaluation of imputation performance in the drug response assay. The y-axis shows the mean-squared error (MSE) across 15 trials for increasing fractions of missing data (x-axis). Two experiments were considered: (a) values missing at random and (b) entire assays missing at random. Each point displays the mean across all trials and the error bars depict the corresponding standard deviations.

3.5 Application to single-cell multi-omics

The emergence of single-cell multi-modal techniques has created opportunities for the development of novel computational strategies [290, 61, 55].
To show case how MOFA can be used to integrate single-cell multi-omics data, we considered a simple data set that consists of 87 ESCs where RNA expression and DNA methylation were simultaneously measured using scM&T-seq[12]. Two populations of ESCs were profiled: the first one contains 16 cells grown in 2i media, which is known to induce a naive pluripotency state associated with genome-wide DNA hypomethylation [91]. The second population contains 71 cells grown in serum media, which contain stimuli that trigger a primed pluripotency state poised for differentiation [307].

3.5.1 Data processing

The RNA expression data was processed using *scran*[186] to obtain log normalised counts adjusted by library size. Feature selection was performed by selecting the top 5,000 most overdispersed genes[159]. A Gaussian likelihod was used for this data modality.
The DNA methylation data was processed as described in Chapter 2. Briefly, for each CpG site, we calculated a binary methylation rate from the ratio of methylated read counts to total read counts. Next, CpG sites were classified by overlapping with genomic contexts, namely promoters, CpG islands and enhancers (distal H3K27ac peaks). Finally, for each annotation we selected the top 5,000 most variable CpG sites with a minimum coverage of 10% across cells. Each of the resulting matrices was defined as a separate view for MOFA. A Bernoulli likelihod was used for this data modality.

3.5.2 Model overview

In this data set, MOFA inferred 3 factors with a minimum explained variance of 1% (Figure 3.27). Factor 1 captured the transition from naive to primed pluripotent states, which MOFA links to widespread coordinated changes between DNA methylation and RNA expression. Inspection of the gene weights for Factor 1 pinpoints important pluripotency markers including *Rex1/Zpf42* or *Essrb* [207]. As previously described both *in vitro* [12] and *in vivo* [19], the transition from naive to primed pluripotency state is concomitant with a genome-wide increase in DNA methylation levels. Factor 2 captured a second dimension of heterogeneity driven by the transition from a primed pluripotency state to a differentiated state, with RNA weights enriched with canonical differentiation markers including keratins and annexins [95].
Jointly, the combination of Factors 1 and 2 reconstruct the coordinated changes between the transcriptome and the epigenome along the differentiation trajectory from naive pluripotent cells to differentiated cells.

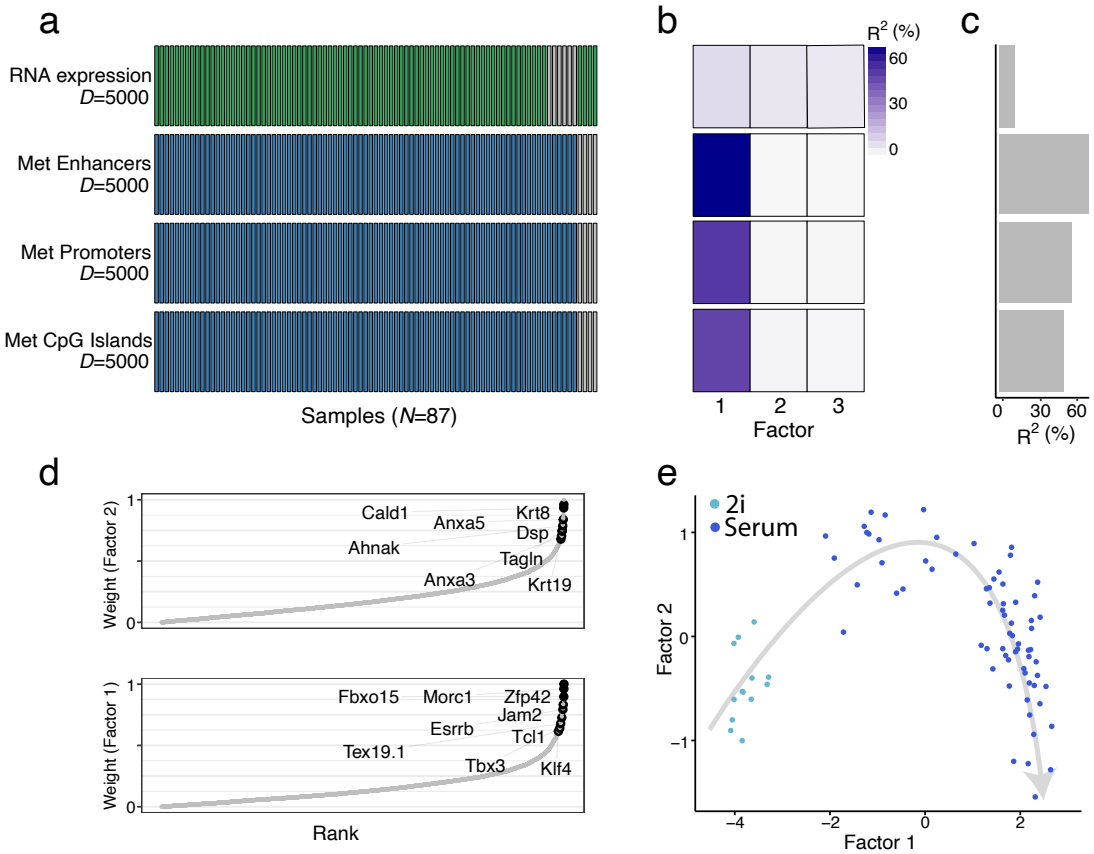


Figure 3.27: MOFA recovers a differentiation process from a single-cell multi-omics data set.
(a) Overview of the data modalities. Rows indicate number of features (D) and columns indicate number of samples (N). Grey bars denote missing samples.
(b) Fraction of variance explained per factor (column) and view (row).
(c) Cumulative fraction of variance explained per view (across all factors).
(d) mRNA weights of Factor 1 (bottom) and Factor 2 (top). The genes that are labelled are known markers of pluripotency (for Factor 1) or differentiation (for Factor 2).
(e) Scatter plot of Factor 1 (x-axis) against Factor 2 (y-axis). Cells are colored based on the culture condition. Grey arrow illustrates the differentiation trajectory from a naive pluripotency state to a differentiated state.

3.6 Limitations and open perspectives

MOFA solves important challenges for the integrative analysis of (single-cell) multi-omics datasets. Yet, the model is not free of limitations and there are open possibilities for future research:

- **Linearity**: this is an assumption that is critical for obtaining interpretable feature weights. Nonetheless, there is a trade-off between explanatory power and interpretability[158]. Non-linear approaches, including deep neural networks or variational autoencoders have shown promising results when it comes to dimensionality reduction [173, 77, 180], batch correction[180], denoising [85] or imputation [174]. Interestingly, very few multi-view factor analysis models exist that incorporate flexible non-linear assumptions, making it an interesting line of research to explore.

- **Scalability**: the size of biological datasets is rapidly increasing, particularly in the field of single cell sequencing [294, 49]. When comparing the inference framework to previous methods that make use of sampling-based MCMC approaches, the variatonal framework implemented in MOFA yields a vast improvement in scalability. Yet, in its vanilla form, variational inference also becomes prohibitively slow with very large datasets [117, 33, 118]. This has been recently addressed by a reformulation of the variational inference problem in terms of a gradient descent optimisation problem, which enables the full machinery of stochastic inference to be applied in the context of Bayesian inference. This line of research is followed in Chapter 5, with the development of a stochastic version of the variational inference algorithm.

- **Generalisations to multi-group structures**: the sparsity assumptions in MOFA are based on the principle that features are structured into non-overlapping views. As such, the activity of the latent factors is also expected to be structured, so that different factors explain variability in different subsets of views (Figure 3.13). Following the same logic, many studies contain structured samples, as either multiple experiments or conditions. A simple generalisation of MOFA would be to intuitively break the assumption of independent samples and introduce an additional prior that captures the group structure at the sample level. This line of research is followed in Chapter 5.

- **Bayesian treatment of predictions**: in the current implementation of MOFA, only the point estimates for the posterior distributions are used in the downstream analysis. While convienient for most operations, this ignores the uncertainity associated with the point estimates, which is a major strength of Bayesian modelling. Future extensions could attempt a more comprehensive Bayesian treatment that propagates uncertainity in the downstream analyses, mainly when it comes to making predictions and imputation [97].

- **Incorporation of prior information**: an unsupervised approach is appealing for discovering the principal axes of variation, but sometimes this can yield challenges in the interpretation of factors. Future extensions could exploit the rich information encoded in gene set ontologies, similar to the methodology proposed in [43].

- **View imbalance**: a property of MOFA is that the number of features can influence the contribution of a data modality to the latent space, such that bigger views tend to contribute more to the factors. This is because the objective function (the evidence lower bound, ELBO) does not weight the different data modalities according to their number of features. This is not necessarily a problem, as we have demonstrated in Section 3.4, where we extracted meaningful signal from small data modalities. In general, however, the signal that can be extracted from small data modalities will depend on the degree of structure within the dataset, the levels of noise and on how strong the feature imbalance is between data modalities. In practice we suggest users to try balance the number of features by subsetting highly variable features in the larger views. An alternative option would be to weight the contribution of each view on the ELBO, such that small views have a relatively higher contribution than large views. This is however a heuristic that does not arise from the Bayesian generative model and there would be no theoretical guarantees about its behaviour.

Chapter 4

Multi-omics profiling of mouse gastrulation at single-cell resolution

In this Chapter I will describe a study where we combined scNMT-seq (introduced in Chapter 2) and MOFA (introduced in Chapter 3) to explore the relationship between the transcriptome and the epigenome during mouse gastrulation. The work discussed results from a collaboration with the group of Wolf Reik (Babraham Institute, Cambridge, UK). It has been peer-reviewed and published in [14]. The experiments were carried out by Stephen Clark, Hisham Mohammed and Carine Stapel, with the help of Wendy Dean and Courtney Hanna for the collection of embryos. Tim Lohoff prepared the Embryoid Body *TET* TKO culture. Wei Xie and Yunlong Xiang shared the ChIP-seq data that was used to define germ layer-specific enhancers. Felix Krueger processed and managed sequencing data. Christel Krueger processed the ChIP-seq data. I performed the majority of the computational analysis, but with contributions from some authors. In particular, Stephen Clark calculated the transcription factor motif enrichment analysis, Carine Stapel explored the neuroectoderm and pluripotency signatures in ectoderm enhancers, and Ivan Imaz-Rosshandler performed the mapping to the gastrulation atlas. John C. Marioni and Wolf Reik supervised the project. The article was jointly written by Stephen Clark, Carine Stapel and me, with input from all authors.

4.1 Introduction

The human body is composed of a myriad of cell types with specialised structure, organisation and function; and yet, each cell in the body contains the same genetic information. The modulation of the genetic code by internal and external factors begin during embryonic development, giving rise to the formation of specialised molecular patterns that ultimately determines the complexity of adult organisms [133]. A key phase in mammalian embryonic development is gastrulation, when a single-layered blastula of pluripotent and relatively homogeneous cells is reorganised to form the three primordial germ layers: the ectoderm, mesoderm and endoderm [298, 277, 296].

The onset of gastrulation is determined by the formation of the primitive streak, which establishes the initial bilateral symmetry of the body. Involution of epiblast cells through the primitive streak gives rise to the mesoderm and endoderm, whereas epiblast cells establish the ectoderm [16, 296, 298, 297]. Although differences exist between species, the morphogenic process of gastrulation is evolutionary conserved throughout the animal kingdom [277]. In most cases, gastrulation is characterised by an epithelial to mesenchymal transition that brings mesodermal end endoderm progenitors beneath the future ectoderm. The epiblast cells that did not migrate through the

primitive streak differentiate towards ectoderm, which eventually gives rise to the nervous system (neural ectoderm) and epidermis (surface ectoderm). The embryonic endoderm gives rise to the interior linings of the digestive tract, the respiratory tract, the urinary bladder and part of the auditory system. The embryonic mesoderm gives rise to muscles, connective tissues, bone, cartilage, blood, kidneys, among others.

4.1.1 Transcriptomic studies

Significant research effort has been deployed to understand the molecular changes underlying gastrulation. Historically, microscopy was used to quantify gene expression at single cell resolution. However, constraints imposed by fluorophore emission spectra made this approach unsuitable for transcriptome-wide studies. Only after the breakthrough made by the introduction of single-cell sequencing technologies has it become possible to generate comprehensive molecular roadmaps of embryonic development [268, 232, 49, 249]. In a pioneer study, [232] generated the first high-resolution atlas of gastrulation and early somitogenesis by profiling the RNA expression of 116,312 cells from 411 whole mouse embryos collected between E6.5 and E8.5. This effort completed earlier attempts of reconstructing the transcriptomic landscape of post-implantation embryos [207, 270, 125, 326]. At the same time, another study employed a more scalable methodology to profile around 2 million cells from 61 embryos ranging from E9.5 and 13.5 days of gestation, spanning early organogenesis [49]. By constructing a densely sampled reference data set, both works have laid the ground for understanding transcriptomic variation during development.

4.1.2 Epigenetic studies

Transcriptomics is a central piece in the puzzle of understanding embryonic development, but still a single piece. The next step is to connect RNA expression to the accompanying epigenetic changes. In differentiated cell types, epigenetic marks confer stable characteristic patterns of cell type identity which have been extensively profiled using bulk sequencing approaches. Nevertheless, because of the low amounts of input material and the extensive cellular heterogeneity, the study of the epigenetic landscape during early development remains poorly understood [140].

Pre-implantation: establishment of the pluripotent state

The first efforts to interrogate the epigenetic dynamics of emrbyonic development using (bulk) next generation sequencing technologies have provided valuable insights for the pre-implantation stage. Multiple studies have described that, after fertilisation, there is a round of reprogramming that resets the epigenetic landscape to a ground state [276, 165]. DNA methylation is globally removed and the chromatin attains its highest levels of accessibility [329]. Consistently, Hi-C experiments have revealed a flexible chromatin landscape, with lack of topologically associating domains (TADs) or chromatin compartments [138, 80, 300], providing a plausible explanation for the remarkable plasticity of pluripotent ESCs.

In contrast to DNA methylation, the presence of post-translational modifications in histone marks are abundant at this stage, potentially providing the major mechanism of epigenetic regulation [107, 300]. Several histone modifications have been studied in ESCs, the most prominent being H3K27ac and H3K4me3, both (generally) activatory marks; and H3K27me3 and H3K9me3, both (generally) repressive marks [341]. Interestingly, many genes that are silenced in ESCs contain both activatory (H3K4me3) and repressive (H3K27me3) epigenetic marks. This distinctive signature of ESCs is thought to establish a bivalent or poised signature for a transcriptionally-ready state for genes that become expressed after gastrulation [27, 300].

Post-implantation: exit of pluripotency

In post-implantation development, cells exit pluripotency and undergo a set of critical cell fate decisions that will ultimately give rise to all somatic cell types. While multiple studies have profiled the epigenetic landscape in pre-implantation embryos, the epigenetic landscape of gastrulation and early mammalian organogenesis remains largely unexplored.

DNA methylation is one of the few epigenetic marks that has been profiled in a genome-wide manner in post-implantation embryos using both bulk and single-cell methodologies [19, 336, 70, 261]. All studies found that the hypomethylated state in the preimplantation blastocysts is followed by a *de novo* DNA methylation wave upon implantation that leads to a hypermethylation of most of the genome. The increase in DNA methylation is concomitant with the increased deposition of repressive histone marks, presumably with the aim of restricting the differentiation potential of early pluripotent cells [18]. The *de novo* methyltransferases (DNMT3A and DNMT3B) are the enzymes responsible for the insertion of DNA methylation marks. Both genes are highly expressed in early mouse embryos, and catalytically inactive mutants of both enzymes lacked *de novo* methylation activity [19, 221]. Interestingly, mouse ESCs remain viable despite complete loss of DNA methylation, but they are uncapable of undergoing cell fate commitment and escaping from the pluripotent state [313].

The interplay of histone marks during post-implantation development is complex. H3K4me3 is detected at transcription start sites after the zygotic genome activation, and remains remarkably stable even after differentiation [113]. H3K4me3 is thought to facilitate transcription by inducing a more efficient assembly of the transcriptional machinery [18, 317]. The other conventional activatory mark, H3K27ac, is deposted in different types of regulatory elements, including promoters and enhancers. It is significantly more dynamic than H3K4me3 in response to internal and external stimuli, and is hence a stronger candidate to regulate cell fate transitions [18, 241].
The inhibitory mark H3K27me3 shows a marked increase upon implantation, deposited by the Polycomb repressive complex 2 (PRC2) around multiple regulatory elements, including CpG-rich promoters of developmental genes. H3K27me3 is often present in transcriptionally inactive regions with low levels of DNA methylation, suggesting a potential antagonism between H3K27me and DNA methylation [38, 18]. Interestingly, inactivating PRC2 components in mouse embryos does not affect pre-implantation development, but the embryos become unviable after gastrulation [272].

This suggests that H3K27me3 has a critical role in regulating gene expression during cell fate commitment after germ layer specification.

Gastrulation: germ layer specification

The post-implantation blastocyst is a relatively homogeneous population of cells and can be characterised with some accuracy by bulk sequencing approaches. However, germ layer specification is uniquely heterogeneous and extremely challenging to study without single-cell technologies. Despite the technical difficulties, some studies attempted to manually dissect each germ layer, followed by bulk sequencing [337]. This revealed that the relatively homogeneous epigenetic landscape at the epiblast is succedeed by a more dynamic landscape where regulatory elements become activate in a lineage-specific manner [337, 166]. Consistent with a role of DNA methylation during gastrulation, perturbations that target the Ten-eleven translocation (TET) family of dioxygenases display developmental defects related to germ layer specification, ranging from impaired migration of primitive streak cells to failed maturation of the mesoderm layer [70].

The recent development of single-cell multi-modal technologies, where epigenomes can be unequivocally assigned to transcriptomes at single-cell resolution, unveils novel opportunities to study the cell fate commitment events during gastrulation [101, 321, 177].

4.2 Results

4.2.1 Data set overview

The aim of this project was to generate a multi-omics atlas of post-implantation mouse embryos at single-cell resolution. We applied scNMT-seq (described in Chapter 2) to jointly profile chromatin accessibility, DNA methylation and gene expression from 1,105 cells at four developmental stages (Embryonic Day (E) 4.5, E5.5, E6.5 and E7.5), spanning exit from pluripotency and germ layer commitment. Additionally, the transcriptomes of 1,419 additional cells from the relevant time points were also profiled:

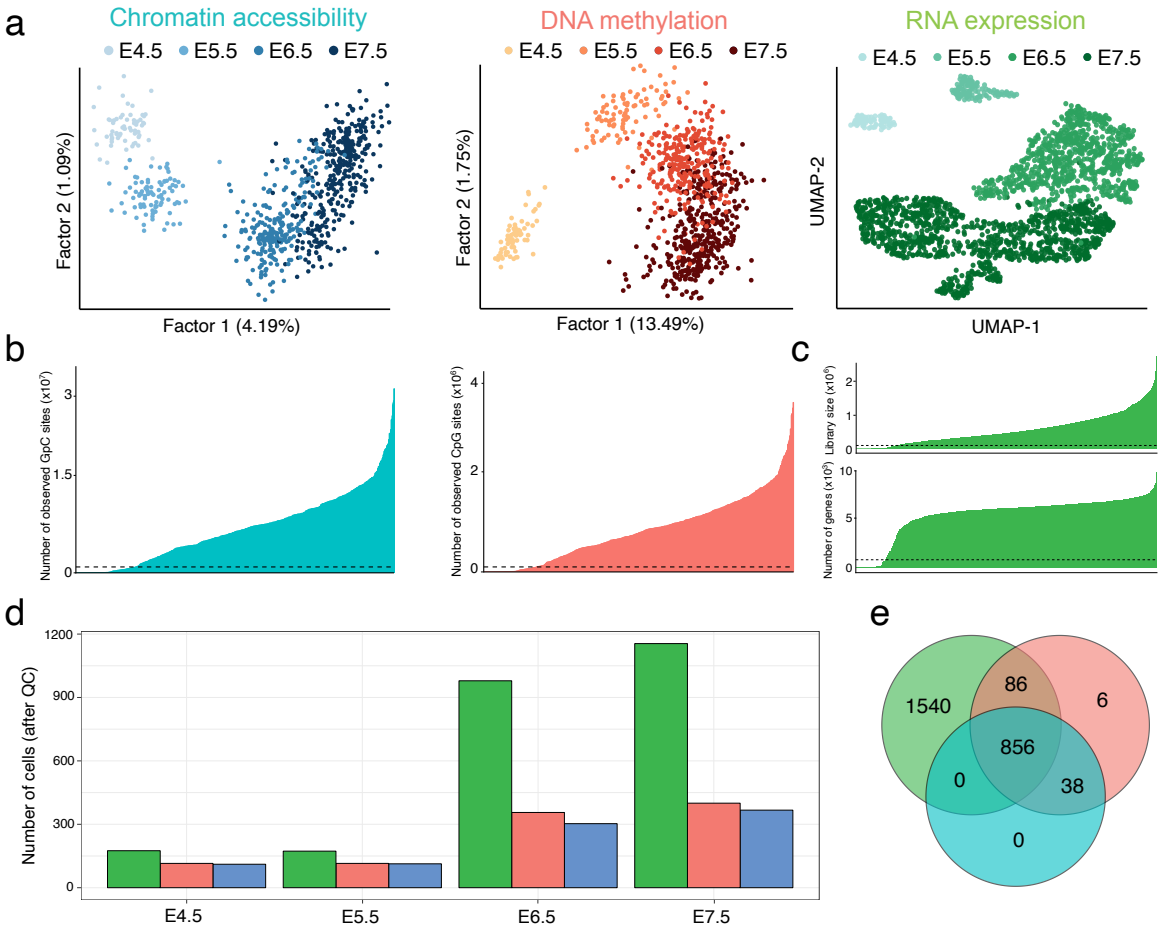


Figure 4.1: scNMT-seq gastrulation atlas. Data set overview.
(a) Dimensionality reduction for chromatin accessibility data (left, in blue), DNA methylation (middle, in red) and RNA expression (right, in green). For the gene expression data we used UMAP [199]. For chromatin accessibility and DNA methylation data we used Bayesian Factor Analysis [15].
(b) Number of observed cytosines in a GpC context (left, in blue) or (b) in a CpG context (right, in red). Each bar corresponds to a different cell. Cells are sorted by total number of GpC or CpG sites, respectively. Cells below the dashed line (50,000 CpG sites and 500,000 GpC sites, respectively) were removed on the basis of poor coverage.
(c) RNA library size (top) and number of expressed genes (bottom) per cell. Cells below the dashed line (10,000 reads and 500 expressed genes, respectively) were removed on the basis of poor coverage.
(d) Number of cells that pass quality control for each molecular layer, stratified by stage. Note that for 1,419 out of 2,524 cells only the RNA expression was sequenced.
(e) Venn Diagram displaying the number of cells that pass quality control for each data modality: RNA expression (green), DNA methylation (red), chromatin accessibility (blue).

4.2.2 Cell type assignment using the RNA expression data

To define cell type annotations we mapped the RNA expression profiles to the scRNA-seq gastrulation atlas [232] using a matching mutual nearest neighbours algorithm [105] (Figure 4.2). In short, the count matrices for both data sets were concatenated and normalised together. Then, Principal Component Analysis was applied to the joint normalised expression matrix. The resulting latent space was then used for the construction of a k-nearest neighbours graph. Finally, for each scNMT-

seq cell, we assigned a cell type using majority voting on the cell type distribution of the top 30 nearest neighbours in the atlas. We validated the cell type assignments by visualising the expression of known marker genes (Figure 4.3).

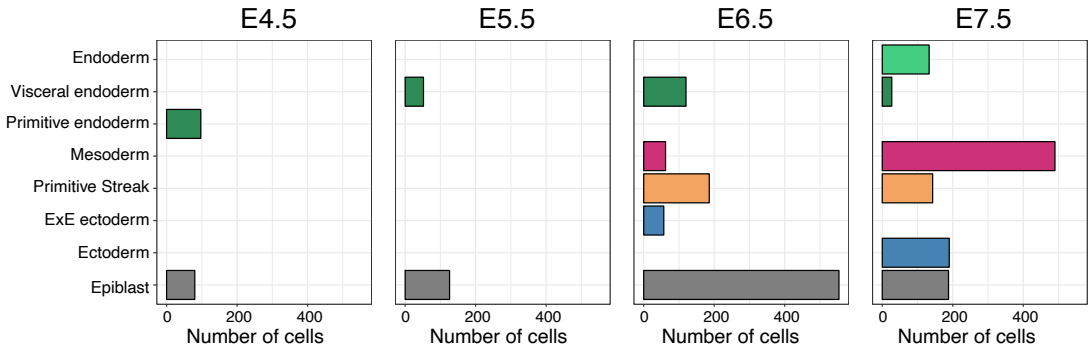


Figure 4.2: Cell type assignments using the RNA expression data. For each stage, the bar plots display the number of cells assigned to each lineage.

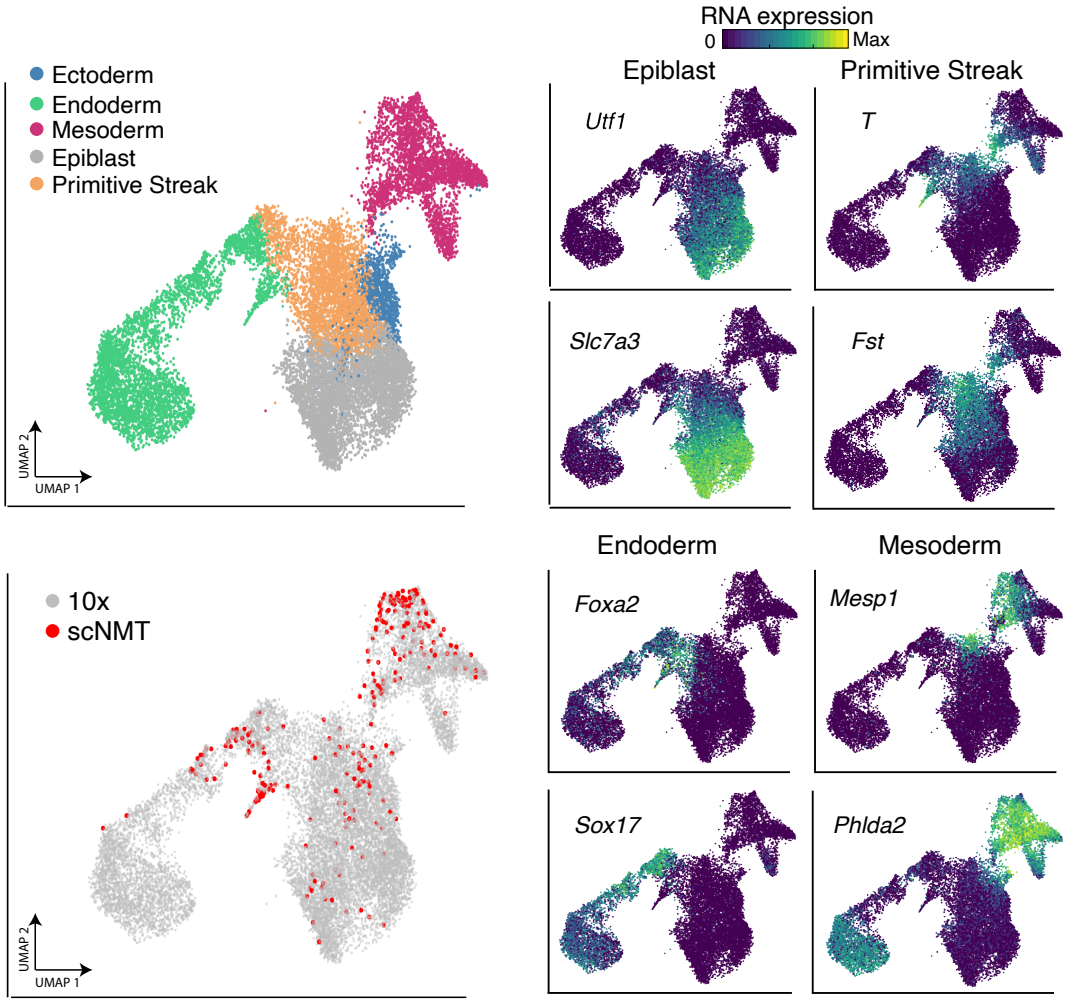


Figure 4.3: Validation of cell types by visualisation of marker genes.
UMAP projections of the atlas data set (stages E6.5 to E8.0). In the top left plot, cells are coloured by lineage assignment. In the bottom left plot, the cells coloured in red correspond to the nearest neighbours that were used to transfer labels to the scNMT-seq data set. The right plots display the RNA expression levels of marker genes for different cell types.

4.2.3 Validation of DNA methylation data and chromatin accessibility data

To validate the DNA methylation and chromatin accessibility data, we performed dimensionality reduction separately for both data modalities using two different settings: (1) with cells from all stages; and (2) separately at each stage. To handle the large amount of missing values that result from single-cell bisulfite data we adopted a Bayesian Factor Analysis model (i.e. MOFA with one view, as described in Chapter 3).

Reassuringly, we observe that for both modalities the model with all cells captures a developmental progression from E4.5 to E7.5 (Figure 4.1). When fitting a separate model for stages E4.5, E5.5 and E6.5, the largest source of variation (Factor 1) separates cells by embryonic versus extra-embryonic

origin, as expected (Figure 4.4). At E7.5 extra-embryonic cells were manually removed during the dissection and the first two latent factors discriminate the three germ layers (Figure 4.4).

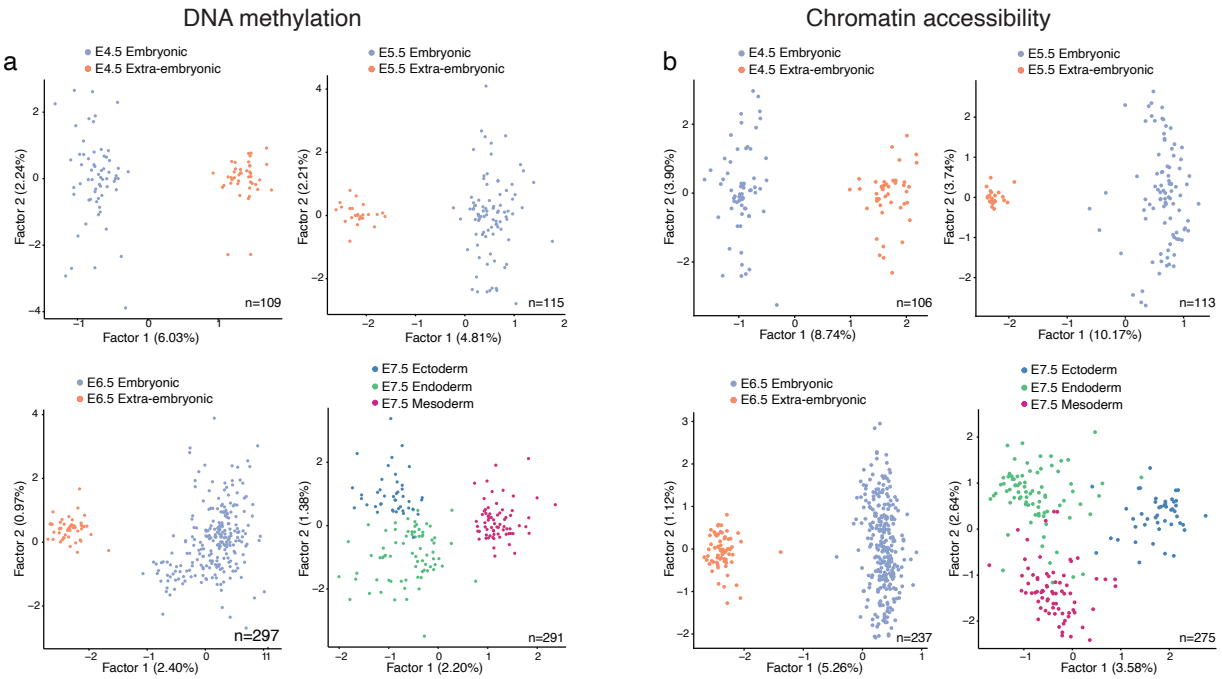


Figure 4.4: Dimensionality reduction of (a) DNA methylation and (b) chromatin accessibility data. Shown are scatter plots of the first two latent factors (sorted by variance explained) that result from applying Bayesian Factor Analysis. From E4.5 to E6.5 cells are coloured by embryonic or extra-embryonic origin. At E7.5, cells are coloured by their primary germ layer.

4.2.4 Exit from pluripotency is concomitant with the establishment of a repressive epigenetic landscape

First, we explored the dynamics of DNA methylation and chromatin accessibility associated with each stage transition. At the global level, CpG methylation levels increase from $\approx 25\%$ to $\approx 75\%$ in the embryonic tissue and $\approx 50\%$ in the extra-embryonic tissue. This is consistent with previous studies that described a *de novo* methylation wave from E4.5 to E5.5 that preferentially targets CpG-poor genomic loci [19, 336] (Figure 4.5 and Figure 4.6). In contrast to the sharp increase in DNA methylation, we observed a more gradual decline in global chromatin accessibility from $\approx 38\%$ at E4.5 to $\approx 29\%$ at E7.5, with no significant differences between embryonic and extra-embryonic tissues (t-test, Figure 4.6). Similar to the DNA methylation changes, CpG-rich regions remain more accessible than CpG-poor regions of the genome, as expected.

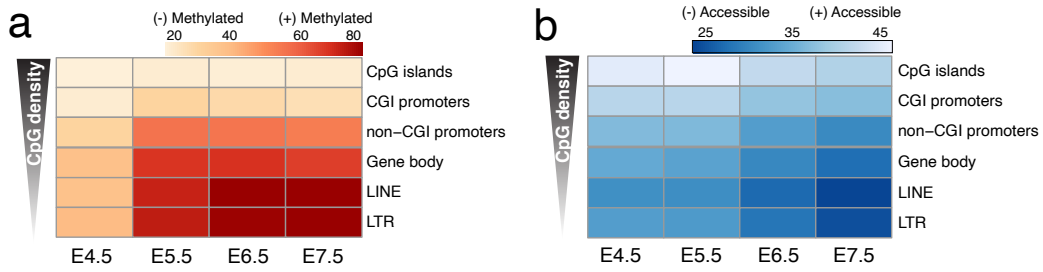


Figure 4.5: DNA methylation and chromatin accessibility levels per stage and genomic context.
Heatmaps display the mean levels across cells within a particular stage and across all loci within a particular genomic context.

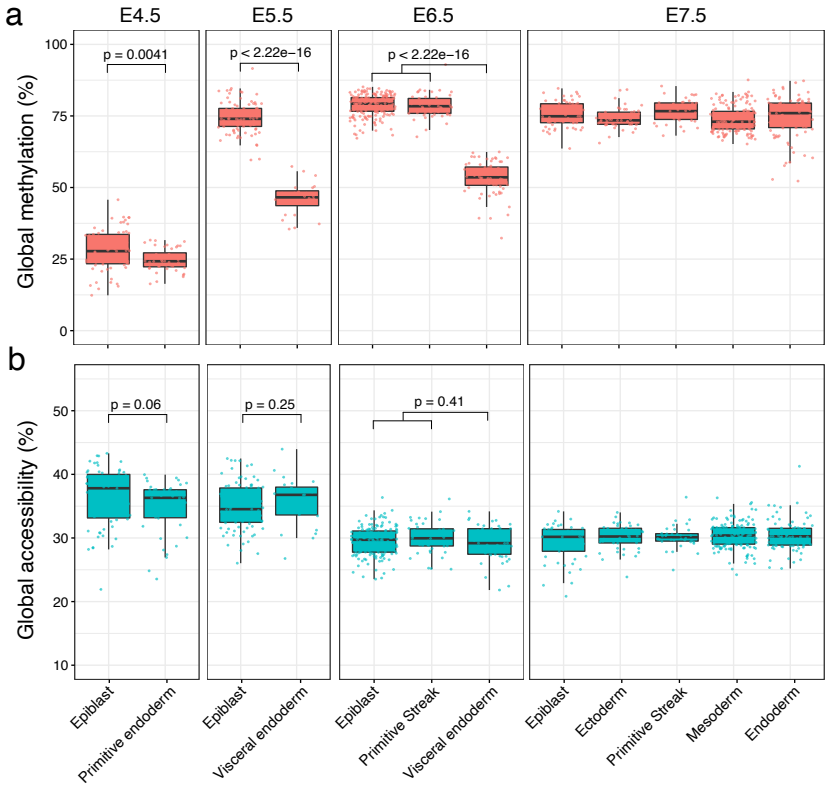


Figure 4.6: Global DNA methylation and chromatin accessibility levels per stage and lineage.
Box plots display the distribution of global (a) CpG methylation or (b) GpC accessibility per stage and lineage. Each dot represents a single cell.

Next we attempted to characterise the relationship between the transcriptome and the epigenome along differentiation. For simplicity we focused on gene promoters (defined as 2kb up and downstream from the transcription start site), as RNA expression and epigenetic readouts can be matched unambiguously. We calculated, for each gene, the correlation coefficient between RNA expression and the corresponding DNA methylation or chromatin accessibility levels. As a filtering criterion, we required a minimum number of 1 CpG (methylation) or 3 GpC (accessibility) measurements in at least 50 cells for each genomic feature. In addition, we restricted the analysis to the top 5,000 most variable genes, according to the rationale of independent filtering [35].

We identified 125 genes whose expression shows significant correlation with promoter DNA methylation and 52 that show a significant correlation with chromatin accessibility. Among the top hits that display significant associations for both comparisons we identified early pluripotency and germ cell markers, including *Dppa4*, *Dppa5a*, *Rex1*, *Tex19.1* and *Pou3f1* (Figure 4.7). Reassuringly, all of them have a negative association between RNA expression and DNA methylation and a positive association between RNA expression and chromatin accessibility. Inspection of the transcriptomic and epigenetic dynamics reveals that the repression of these early pluripotency markers are concomitant with the genome-wide trend of DNA methylation gain and chromatin closure.

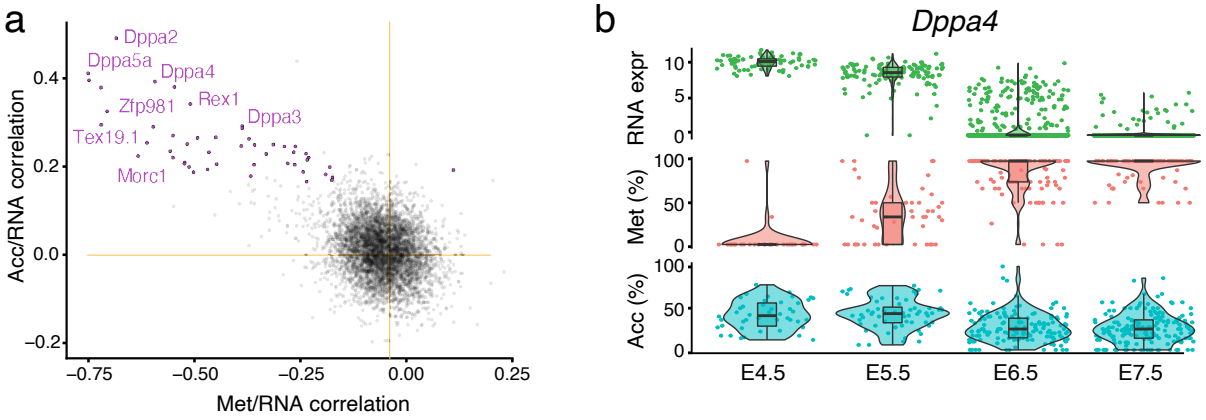


Figure 4.7: Genome-wide association analysis between RNA expression and the corresponding epigenetic status in gene promoters.
(a) Scatter plot of Pearson correlation coefficients between promoter DNA methylation versus RNA expression (x-axis); and promoter accessibility versus RNA expression (y-axis). Significant associations for both correlation modes (FDR<10%) are coloured in red. Examples of early pluripotency and germ cell markers among the significant hits are labeled in red.
(b) Illustrative example of epigenetic repression of the gene *Dppa4*. Box and violin plots display the distribution of RNA expression (log2 counts, green), DNA methylation (% levels, red) and chromatin accessibility (% levels, blue) per stage. Each dot corresponds to one cell.

4.2.5 MOFA reveals coordinated variability between the transcriptome and the epigenome during germ layer formation

In the previous section we demonstrated that exit from pluripotency is concomitant with the establishment of a repressive epigenetic landscape that is characterised by increasing levels of DNA methylation and decreasing levels of chromatin accessibility. Next, we sought to investigate the coordinated changes between RNA expression and epigenetic status that define germ layer commitment at the E7.5 stage. We performed an unsupervised integrative analysis using Multi-Omics Factor Anaysis (MOFA, introduced in Chapter 3). As a reminder for the reader, MOFA takes as input multiple data modalities and it exploits the covariation patterns between the features within and between modalities to learn a low-dimensional representation of the data in terms of a small number of latent factors (Figure 3.13). Each Factor captures a different source of cell-to-cell heterogeneity, and the corresponding weight vectors (one per data modality) provide a measure of feature importance. Importantly, MOFA relies on multi-modal measurements from the same cell to identify whether factors are unique to a single data modality or shared across multiple data

modalities, thereby providing a principled approach to reveal the extent of covariation between different data modalities.

Data preprocessing

As input to MOFA we used the RNA expression data quantified over genes and the DNA methylation and chromatin accessibility data quantified over putative regulatory elements. For this analysis, we selected distal H3K27ac sites (enhancers) and H3K4me3 (active transcription start sites). Both annotations were defined using an independently generated ChIP-seq data set, where each germ layer at E7.5 was manually dissected out prior to ChIP-seq [331]. An overview on the numbers and the overlap of the lineage-specific histone marks is given in the following figure:

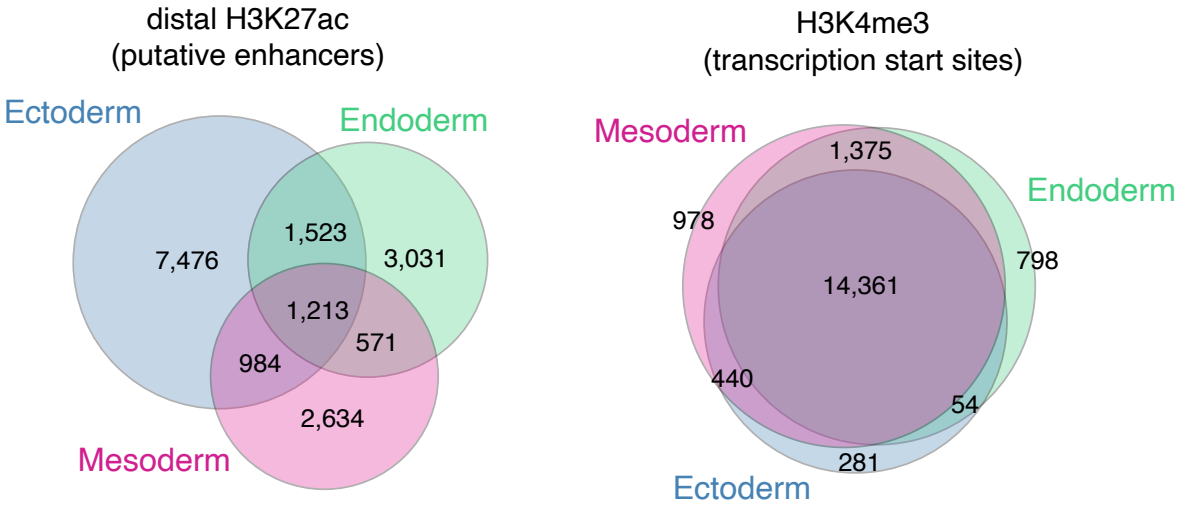


Figure 4.8: Venn diagrams showing overlap of peak calls for each lineage-specific histone mark, for distal H3K27ac (left) and all H3K4me3 (right). The figure shows that distal H3K27ac peaks (putative enhancer [65]) have moderate levels of overlap between the three germ layers. In contrast, H3K4me3 peaks (active transcription start sites [172]) are similar between the three germ layers.

Additionally, we quantified DNA methylation and chromatin accessibility in gene promoters, again defined as 2kb upstream and downstream of the transcription start sites.
To reduce computational complexity and to increase the signal-to-noise ratio we performed feature selection. First, we required for genomic features to have a minimum of 1 CpG (methylation) or 5 GpC (accessibility) observed in at least 25% of cells. Genes were required to be expressed in at least 25% of cells. Second, we subset the epigenetic modalities to the top 1,000 most variable features and the RNA expression to the top 2,500 most variable genes.

Summary of the MOFA output

MOFA identified 6 Factors capturing at least 1% of variance in the RNA expression data (Figure 4.9). The first two Factors (sorted by variance explained) captured the the emergence of the three germ layers, indicating that germ layer commitment is the largest source of variation across all molecular

layers at E7.5. Notably, for these two Factors, MOFA links the variation at the gene expression level to concerted DNA methylation and chromatin accessibility changes at lineage-specific enhancer marks. Surprisingly, these two Factors capture very small amounts of the variation in DNA methylation and chromatin accessibility at promoters. This suggests that epigenetic changes in promoters may not be linked to germ layer commitment, with distal regulatory elements (i.e. enhancers) playing a more prominent role. Yet, we cannot rule out important variation in other epigenetic layers such as histone marks or chromatin conformation.

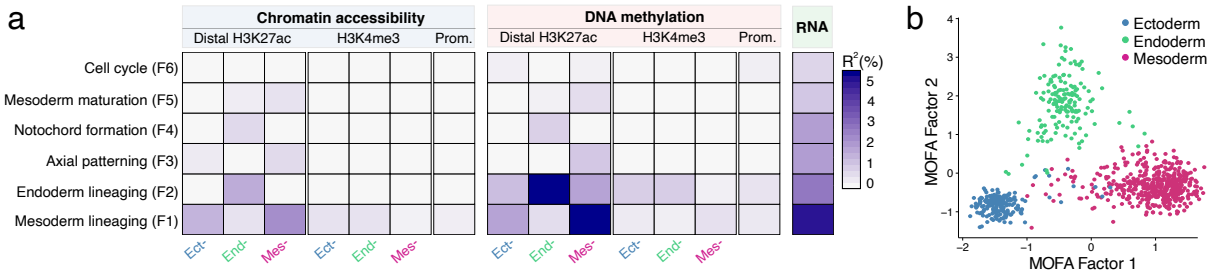


Figure 4.9: MOFA reveals coordinated epigenetic and transcriptomic variation at enhancer elements associated with germ layer commitment.
(a) Percentage of variance explained by each MOFA factor (rows) across data modalities (columns). Considered data modalities were RNA expression (green); DNA methylation (red) and chromatin accessibility (blue) quantified on promoters, lineage-specific H3K4me3-marked sites and distal H3K27ac-marked sites (putative enhancers). Factors are sorted by their total variance explained across all data modalities.
(b) Scatter plot of Factors 1 and 2. Cells are coloured according to their lineage assignment.

The four remaining factors correspond to mostly transcriptional signatures related to anterior-posterior axial patterning (Factor 3), lineaging events such as notochord formation (Factor 4) and mesoderm patterning (Factor 5); and cell cycle (Factor 6). Their characterisation is shown in Appendix B.

4.2.6 Differential DNA methylation and chromatin accessibility analysis

The MOFA analysis in the previous section reveals interesting genome-wide trends. We next attempted to pinpoint individual genomic elements that could be representative of the global patterns. This could be done by inspecting the feature weights in the MOFA model, but given that we can accurately classify cells into the three (discrete) germ layers, here we decided to adopt a more intuitive supervised approach. For each genomic element (with sufficient coverage), we calculated differential DNA methylation and chromatin accessibility between each germ layer versus the other two using a Fisher exact test for binomial proportions (Figures 4.10 and 4.11).

In general we observe that, consistent with the MOFA results, only enhancers display substantial amounts of epigenetic variation between the germ layers (Figure 4.10). As expected, endoderm enhancers seem to be more associated with endoderm commitment (more open and unmethylated in the endoderm cells) whereas mesoderm enhancers are more associated with mesoderm commitment

(again, more open and unmethylated in the mesoderm cells). Notably, for both endoderm and mesoderm commitment events, the effect sizes associated with regions that display differential demethylation and chromatin accessibility are moderate (less than ≈30% change in levels, Figure 4.11) but coordinated across multiple enhancers (between ≈10% and ≈25% of the distal H3K27ac peaks, Figure 4.10).

Intriguingly, ectoderm enhancers display less associations than their meso- and endoderm counterparts, even for ectoderm commitment. This indicates a potential asymmetric contribution of epigenetic modifications to germ layer commitment, a hypothesis which will be further explored below.

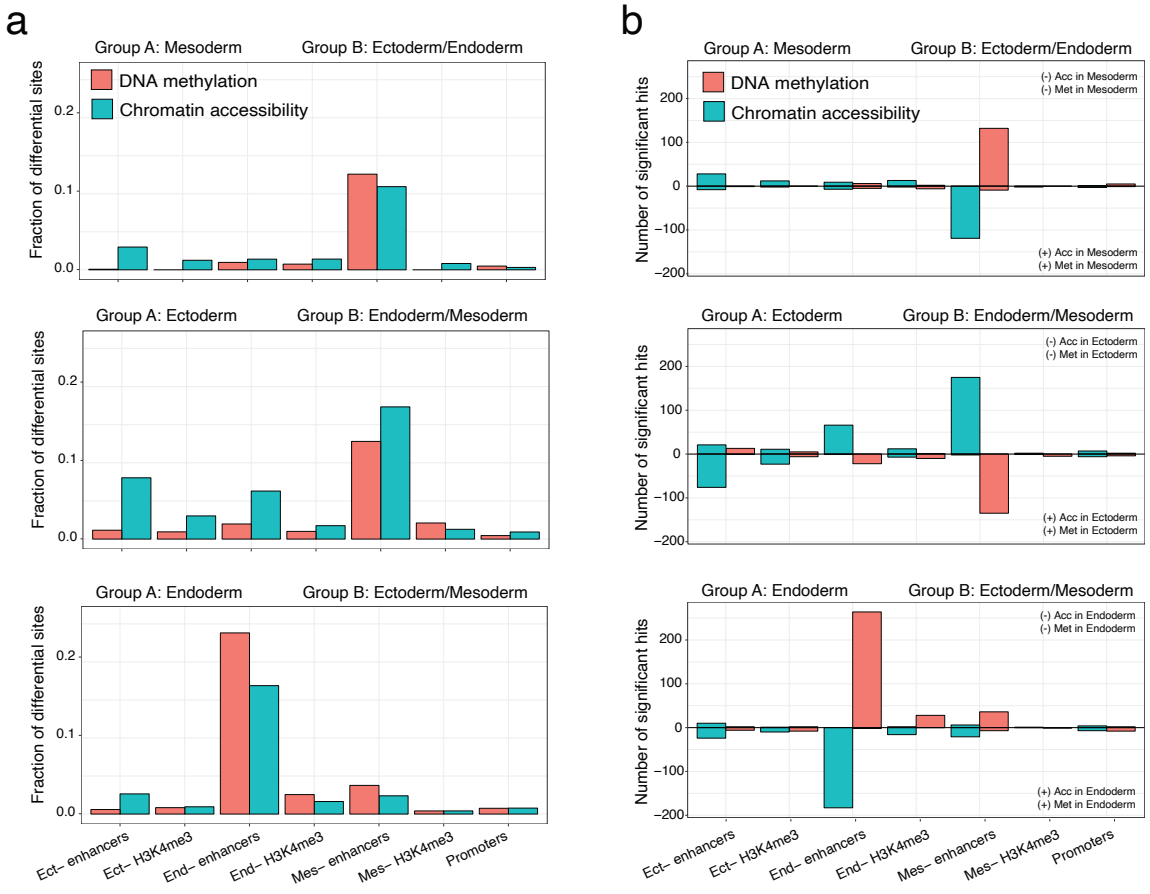


Figure 4.10: Differential DNA methylation and chromatin accessibility analysis between germ layers at E7.5
Bar plots display (a) the fraction and (b) the total number of differentially methylated (red) or accessible (blue) loci (FDR<10%, Fisher exact test for binomial proportions, y-axis) per genomic context (x-axis). Each panel corresponds to the comparison of cells from one germ layer (group A) against cells comprising the other two germ layers (Group B). For (b), positive values indicate increase in DNA methylation or chromatin accessibility in group A, whereas negative values indicate decrease in DNA methylation or chromatin accessibility.

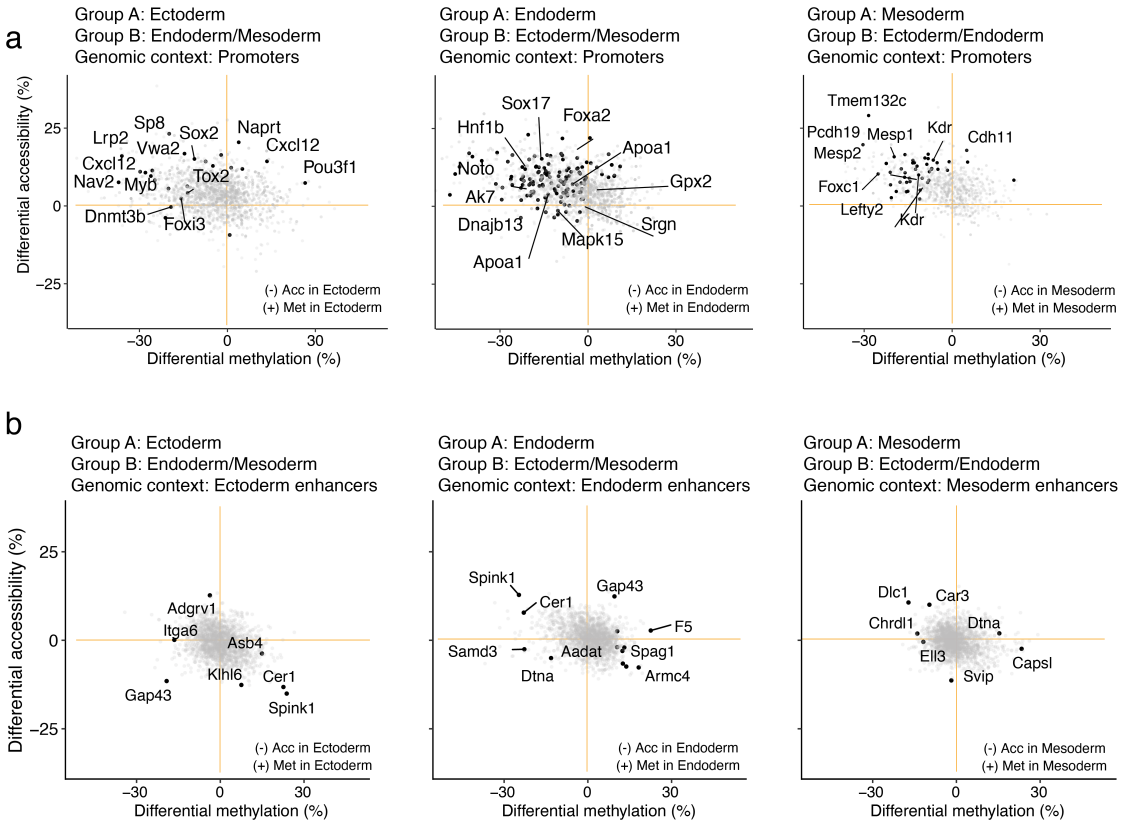


Figure 4.11: Differential DNA methylation and chromatin accessibility between germ layers at E7.5
Scatter plots display differential DNA methylation (%, x-axis) and chromatin accessibility (%, y-axis) at (a) lineage-defining enhancers and (b) promoters. Comparisons are ectoderm versus non-ectoderm cells (left), endoderm versus non-endoderm cells (middle) and mesoderm versus non-mesoderm cells (right). Black dots depict gene-enhancer or gene-promoter pairs with significant changes in RNA expression and DNA methylation or chromatin accessibility (FDR<10%). Genes were linked to enhancers by overlapping genomic coordinates with a maximum distance of 50kb.

Characterisation of individual enhancers

The results above suggest that the establishment of lineage-specific epigenetic profiles results from the coordinated action of multiple elements located all across the genome, and hence the identification of individual putative regulatory elements is not trivial and probably requires a much larger data set than the one we profiled. Nevertheless, when linking enhancers to genes by a maximum genomic distance of 25kb we identified some interesting gene-enhancer associations linked to key germ layer markers including *Snai1* and *Mesp2* for mesoderm, *Bmp2* and *Hnf1b* for endoderm, *Bcl11a* and *Sp8* for ectoderm (Figure 4.11).

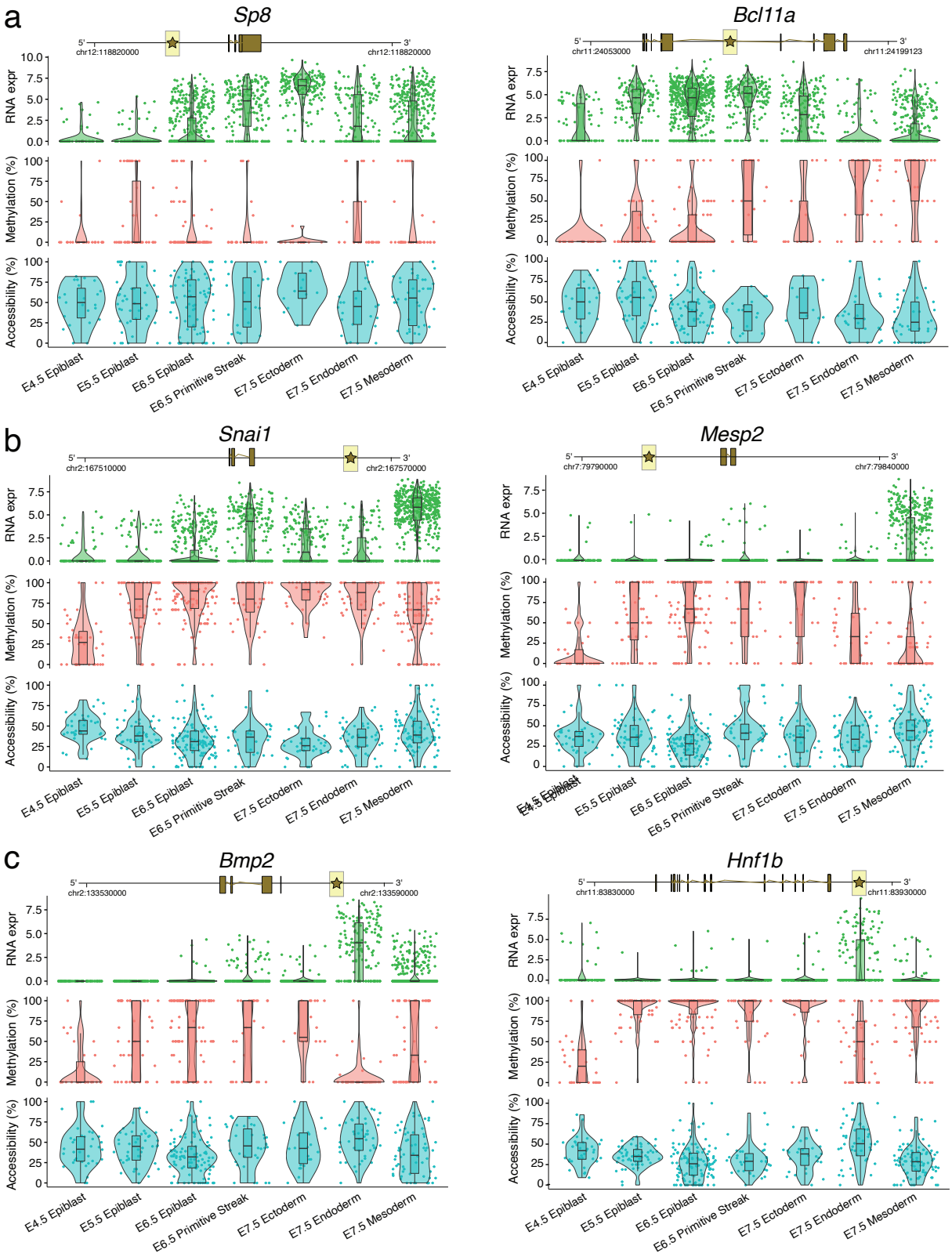


Figure 4.12: Illustrative examples of putative epigenetic regulation in enhancer elements during germ layer commitment.
Box and violin plots show the distribution of RNA expression (log normalised counts, green), DNA methylation (%, red) and chromatin accessibility (%, blue) levels per stage and lineage. Each dot corresponds to a cell. The enhancer region that is used to quantify DNA methylation and chromatin accessibility levels is represented with a star and highlighted in yellow in the genomic track above

4.2.7 Transcription factor motif enrichment analysis

To identify transcription factors (TFs) that could drive the epigenetic variation in lineage-defining enhancers during germ layer commitment, we integrated the chromatin accessibility and RNA information as follows. For every TF with an associated motif in the Jaspar core 95 vertebrates data base we extracted its position-specific weight matrix and we tested for enrichment in differentially accessible distal H3K27ac sites using a background of all distal H3K27ac sites. To assess statistical significance we used a Fisher exact test, as implemented in the *meme suite* (v4.10.1) [21]. This information was then integrated with differential RNA expression between germ layers for the same TFs, quantified using the gene-wise negative binomial generalised linear model with quasi-likelihood implemented in edgeR [196]. Reassuringly, this analysis revealed that lineage-defining enhancers are enriched for key developmental TFs, including POU3F1, SOX2, SP8 for ectoderm; SOX17, HNF1B, FOXA2 for endoderm; and GATA4, HAND1, TWIST1, for mesoderm (Figure 4.13).
Although this analysis serves as a good quality control for our results, it is important to keep in mind that using sequence information is only a proxy for true TF binding, and some essential TFs to not target specific motifs, including EOMES or BRACHYURY [305].

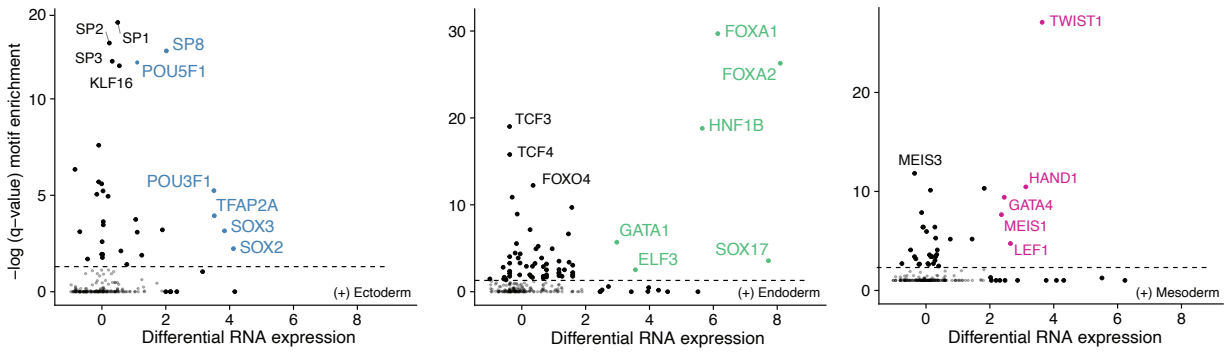


Figure 4.13: Transcription Factor motif enrichment analysis at lineage-defining distal H3K27ac sites. Shown is motif enrichment (-log10 q-value, y-axis) plotted against differential RNA expression (log fold change, x-axis) of the corresponding TF. The analysis is performed separately for each set of lineage-defining enhancers: ectoderm (left), endoderm (middle) and mesoderm (right). TFs with significant motif enrichment (FDR<1%) and differential RNA expression (FDR<1% and log-fold change higher than 2) are coloured and labelled.

4.2.8 Time resolution of the enhancer epigenome

In the previous section we have shown that distal regions marked with H3K27ac (i.e. putative enhancers) are the elements that drive or respond to germ layer specification at E7.5.
Next, we sought to explore how these epigenetic patterns are established. We visualised DNA methylation and chromatin accessibility levels at lineage-defining enhancers from E4.5 to E7.5 (Figure 4.14). Importantly, to interpret the visualisation, DNA methylation and chromatin accessibility values should be compared to the genome-wide background levels that are displayed as dashed lines.

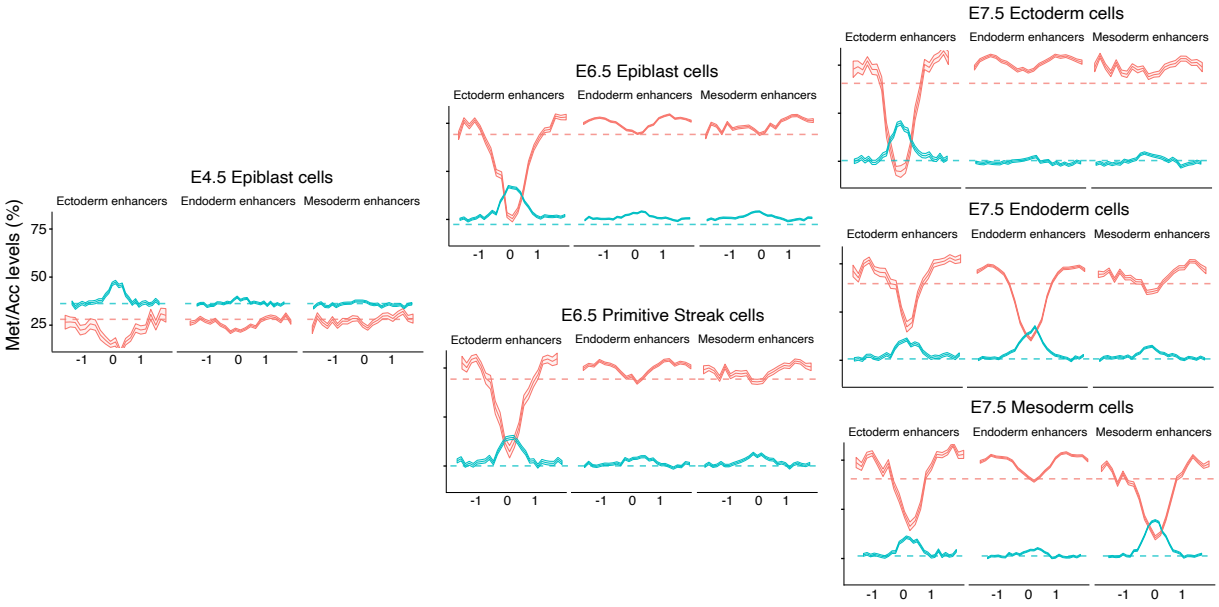


Figure 4.14: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers visualised at pseudobulk resolution.
DNA methylation (red) and chromatin accessibility (blue) levels at lineage-defining enhancers quantified over different lineages across development. Shown are running averages in consecutive 50bp windows around the center of the ChIP-seq peaks (1kb upstream and downstream). Solid lines display the mean across cells and shading displays the corresponding standard deviation. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue).

The DNA methylation and chromatin accessibility dynamics can also be visualised at the single-cell level (Figure 4.15).

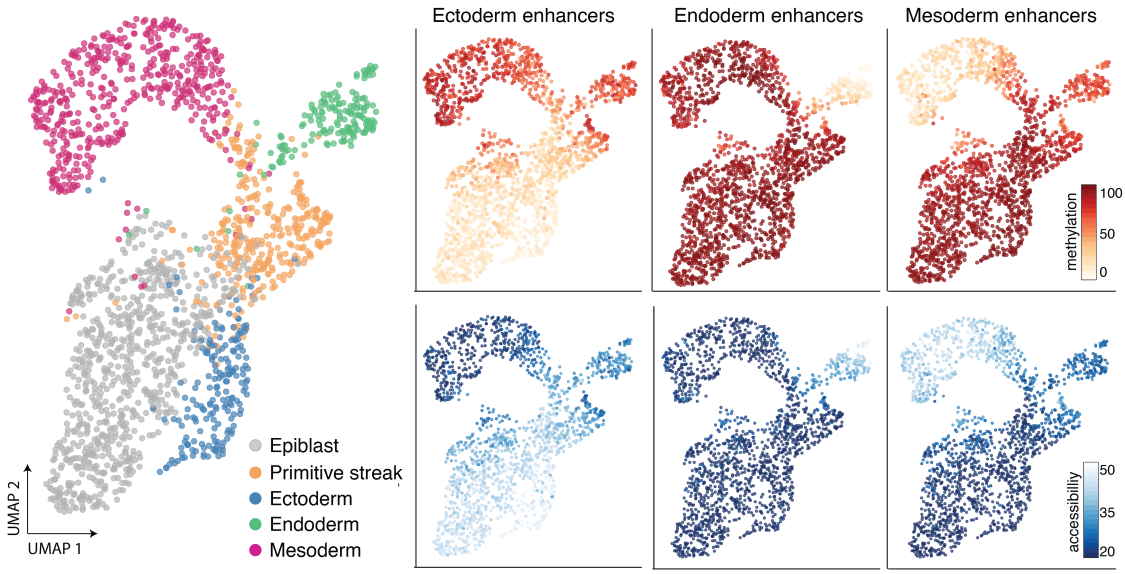


Figure 4.15: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers, visualised at single-cell resolution.
UMAP projection based on the MOFA factors inferred using all cells. In the left plot the cells are coloured according to their lineage assignment. In the right plots cells are coloured by average DNA methylation (top) or chromatin accessibility (bottom) at lineage-defining enhancers. For cells with only RNA expression data, the MOFA factors were used to impute the DNA methylation and chromatin accessibility values.

For clarity, the epigenetic dynamics for mesoderm and endoderm enhancers will be described first, followed by the ectoderm enhancers.

Mesoderm and endoderm enhancers undergo concerted demethylation and chromatin opening upon lineage specification

From E4.5 to E6.5, mesoderm and endoderm enhancers closely follow the genome-wide trend and undergo a dramatic increase in DNA methylation from an average of ≈25% to ≈80%. Consistently, the chromatin accessibility decreases from ≈35% to ≈ 25% (Figure 4.14 and Figure 4.15).
Upon germ layer specification at E7.5, mesoderm and endoderm enhancers undergo concerted demethylation from ≈ 80% to ≈50% in a lineage-specific manner (i.e. mesoderm enhancers demethylate in mesoderm cells, whereas endoderm enhancers demethylate in endoderm cells). Consistently, chromatin accessibility sharply increases from ≈ 25% to ≈45% upon lineage specification.

Ectoderm enhancers are primed in the early epiblast

In striking contrast to the mesoderm and endoderm enhancers, the ectoderm enhancers are open and demethylated as early as the E4.5 epiblast. Interestingly, the ectoderm cells share the same epigenetic profile (in enhancer elements) as the epiblast, characterised by demethylated and open ectoderm enhancers; and methylated and closed mesoderm and endoderm enhancers (Figure 4.14

and Figure 4.15). Upon commitment to mesoderm and endoderm, ectoderm enhancers become partially repressed.

Two hypotheses could explain this observation. The first hypothesis is that ectoderm enhancers are a mixture of pluripotency and proper ectoderm signatures, and hence the pluripotency signatures are driving the demethylation and chromatin opening in early stage, whereas the proper ectoderm signatures are driving the demethylation and chromatin opening upon commitment to ectoderm. The second hypothesis is that the ectoderm fate is epigenetically primed in the early epiblast (i.e. ectoderm is the default lineage), and hence the ectoderm enhancers remain demethylated and open all along from the epiblast to the ectoderm.

To investigate this, the first step is to disentangle the pluripotency and ectoderm signatures that may be confounded within the ectoderm enhancers. We selected the set of E7.5 ectoderm enhancers (n=2,039) and, at each element, we quantified the H3K27ac levels in ESCs and E10.5 midbrain, a tissue largely derived from the (neuro-)ectoderm layer. Both annotations were derived from the ENCODE project [333]. Remarkably, we observe that the E7.5 ectoderm enhancers consist of an almost exclusive mixture of pluripotent and neuroectoderm signatures, as indicated by the negative correlation between H3K27ac levels in ESCs versus E10.5 midbrain (Figure 4.16). This result supports the first hypothesis, but does not rule out the second hypothesis.

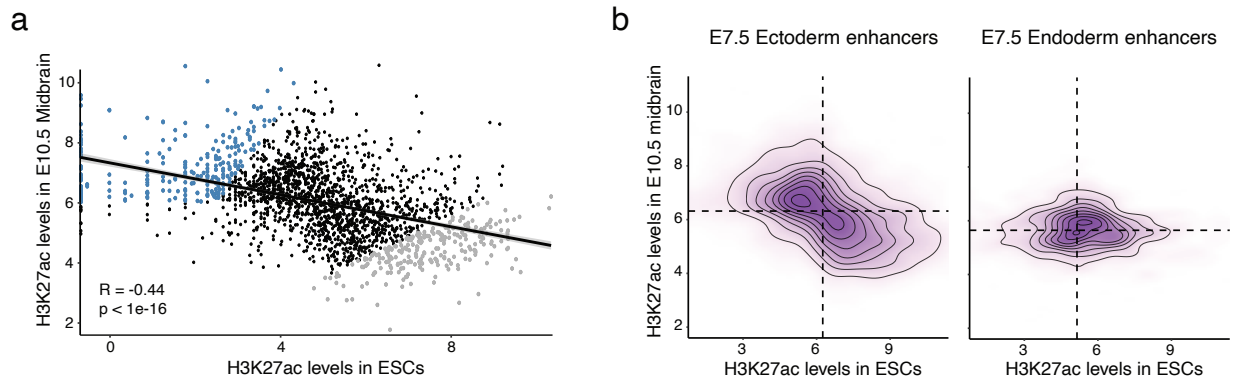


Figure 4.16: E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures.
(a) Scatter plot of ectoderm enhancers' H3K27ac levels quantified in ESCs (pluripotency enhancers, x-axis) and E10.5 midbrain (neuroectoderm enhancers, y-axis). Each dot corresponds to an ectoderm enhancer (Figure 4.8). Highlighted are the top 250 ectoderm enhancers that show the strongest differential H3K27ac levels between E10.5 midbrain and ESCs (blue for neuroectoderm enhancers and grey for pluripotency enhancers).
(b) Density plots of H3K27ac levels quantified in ESCs (x-axis) versus E10.5 midbrain (y-axis), for ectoderm enhancers (left) and endoderm enhancers (right). Endoderm enhancers were included as a control to show that the negative association is exclusive to ectoderm enhancers.

Next, among the E7.5 ectoderm enhancers we defined a set of 250 neuroectoderm enhancers (high H3K27ac levels in E10.5 midbrain) and a separate set of 250 pluripotency enhancers (high H3K27ac levels in ESCs) (blue and grey dots in Figure 4.16). Additionally, we also considered endoderm enhancers as a negative control.

For each class of enhancers, we quantified and visualised the DNA methylation and chromatin accessibility dynamics along the epiblast-ectoderm trajectory Figure 4.17). We plotted absolute levels in (a) and normalised levels to the genome-wide background in (b). We remind the reader that to interpret the plot below, it is critical to compare the absolute levels to the genome-wide background levels.

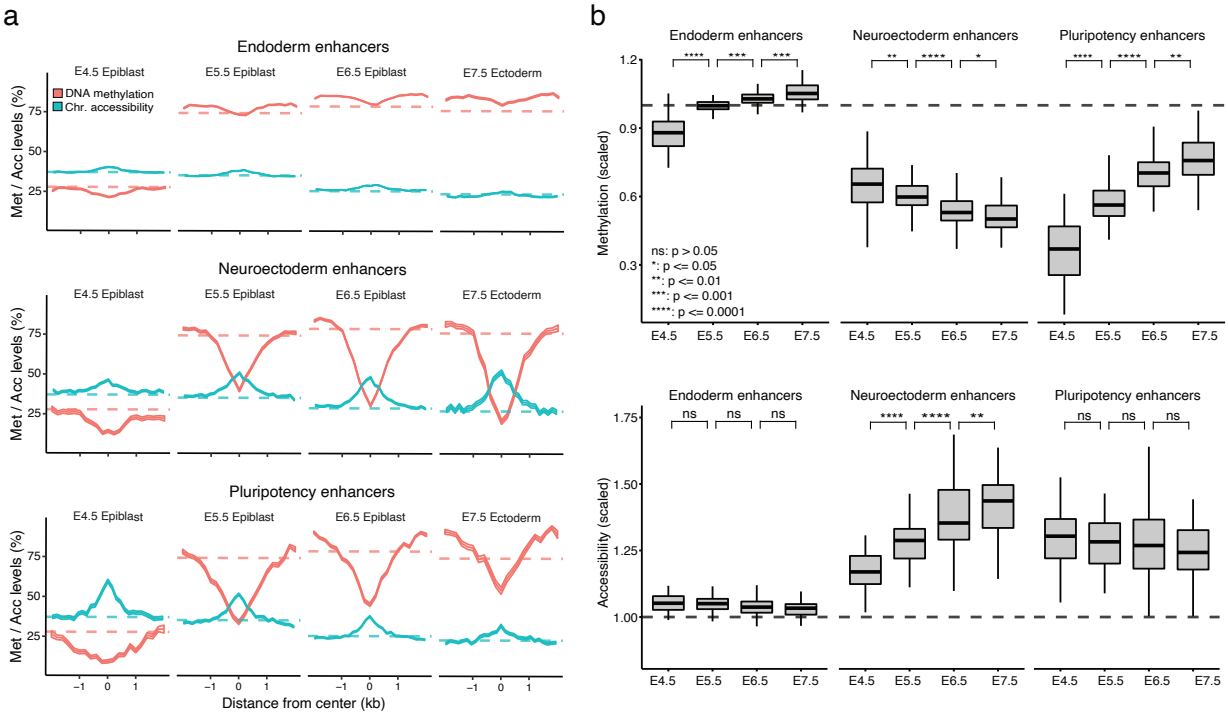


Figure 4.17: Pluripotency and neurectoderm enhancers display different DNA methylation and chromatin accessibility dynamics.
(a) Profiles of DNA methylation (red) and chromatin accessibility (blue) quantified along the epiblast-ectoderm trajectory. Each panel corresponds to a different genomic context. Profiles are quantified using running averages of 50-bp windows around the centre of the ChIP-seq peak for a total of 2 kb upstream and downstream. Solid lines display the mean across cells and shading displays the corresponding standard deviation. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue).
(b) Box plots of DNA methylation (top) and chromatin accessibility (bottom) levels quantified along the epiblast-ectoderm trajectory). Levels are scaled to the genome-wide background for each stage.

The three types of enhancers display very different epigenetic dynamics:

- Endoderm enhancers simply follow the genome-wide repressive dynamics, driven by a global increase in DNA methylation and a decrease in chromatin accessibility. Consistently, the relative levels for both measurements are close to ≈ 1.
- Pluripotency enhancers display an increase in DNA methylation from $\approx 15\%$ at E4.5 to $\approx 60\%$ at E7.5 and a decrease in chromatin accessibility from $\approx 50\%$ at E4.5 to $\approx 35\%$ at E7.5. This is similar to our previous result on the promoters dynamics of pluripotency genes (Figure 4.7). The relative levels show a steady decrease of DNA methylation and a moderate decrease in chromatin accessibility, consistent again with the global repressive dynamics.

- Neuroectoderm enhancers remain at ≈ 40% DNA methylation and ≈ 40% chromatin accessibility from E5.5 to E7.5. This is significantly higher methylation levels and lower chromatin accessibility levels than the genome-wide background. In addition, when looking at the relative values, neuroectoderm enhancers undergo steady decrease in DNA methylation and an increase in chromatin accessibility.

To our surprise, the results indicate that both hypotheses are correct. Ectoderm enhancers at E7.5 contain a mixture of pluripotency and neuroectoderm signatures, but both signatures display different epigenetic dynamics. Whereas pluripotency enhancers become repressed alongside the global repressive dynamics, neuroectoderm enhancers display a signature of active chromatin in the early epiblast.
We conclude that the epigenetic profile of neuroectoderm fate is primed as early as in the E4.5 epiblast. This finding supports the existence of a *default* pathway in the Waddington landscape of development, with the ectoderm being the default germ layer in the embryo. As we will discuss below, this model provides a potential explanation for the phenomenon of default differentiation of neuroectodermal tissue from ESCs *in vitro* [213, 114].

The following figure summarises our model for the epigenetic dynamics of germ layer commitment:

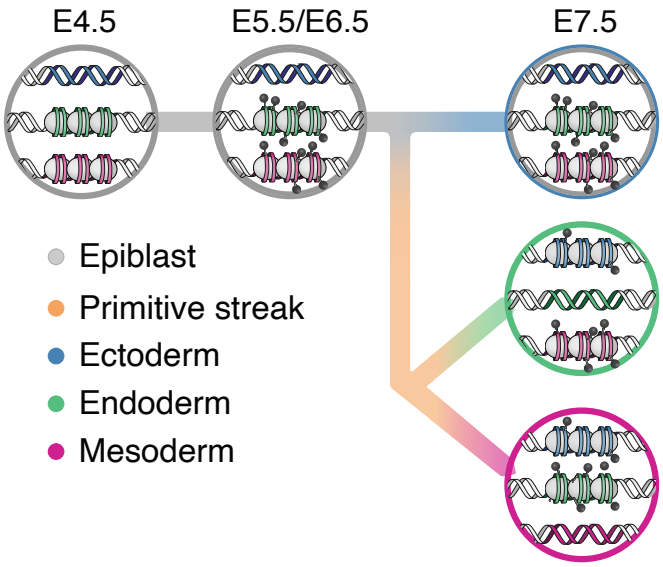


Figure 4.18: Schematic illustration of the hierarchical model for the epigenetic dynamics of germ layer commitment. Illustration designed by Veronique Juvin from SciArtWork.

Silencing of ectoderm enhancers precedes mesoderm and endoderm commitment

At E6.5, TGF-β and Wnt signalling in the posterior side of the embryo promote exit from pluripotency and induce the formation of the primitive streak, which is characterised by the expression of T-box factors such as *Eomes* and *Brachyury* [306]. This transient programme, also called the mesendoderm state, eventually gives rise to the embryonic endoderm and mesoderm lineages.

The triple-omics nature of scNMT-seq measurements prompted us to explore whether differences exist in the timing of onset of molecular events at the mesendoderm state. In particular, we explored whether the lineage-specific epigenetic profiles are remodelled prior or after the transcriptomic programme is activated.
Following recent successes in reconstructing trajectories from scRNA-seq data, we used the RNA expression profiles to order cells by their developmental state to generate two trajectories, corresponding to mesoderm and endoderm commitment (Figure 4.19a). Reassuringly, both pseudotime trajectories captured the transitiom from epiblast to either mesoderm or endoderm fates, with the primitive streak as a transient state. Subsequently, we plotted, for each cell, the average DNA methylation and chromatin accessibility for each class of lineage-defining enhancers (Figure 4.19b). We find that, as cells begin to display a primitive streak phenotype, ectoderm-defining enhancers progressively decrease in accessibility and gain methylation, a process that continues as cells differentiate into the mesoderm and endoderm. In contrast, mesoderm and endoderm-defining enhancers simultaneously become hypomethylated and accessible only after commitment to these cell fates. In both cases, changes in DNA methylation and chromatin accessibility co-occur, suggesting a tight regulation of the two epigenetic layers.

In conclusion, we observe a sequential process where the inactivation of ectoderm enhancers precedes the activation of the mesendoderm enhancers. Interestingly, this resembles reprogramming of induced pluripotent stem cells, where the differentiated programme is repressed prior to the activation of the pluripotency programme [224].

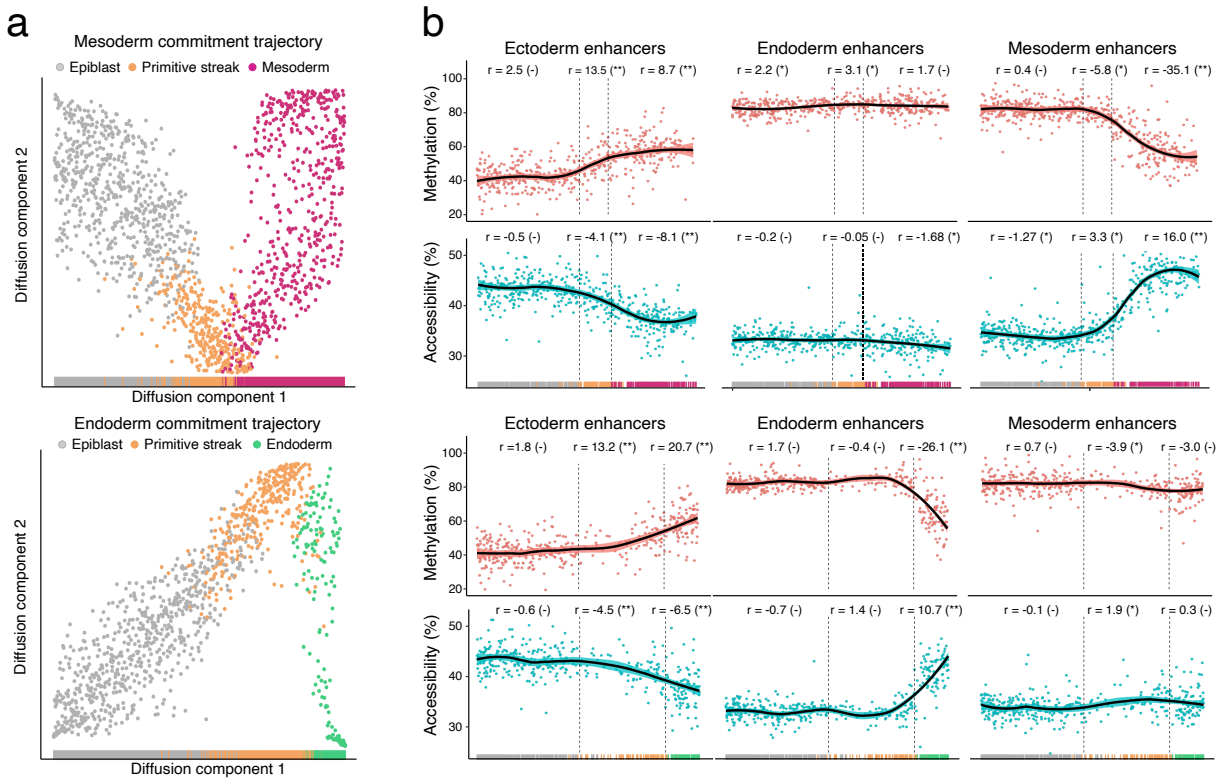


Figure 4.19: Silencing of ectoderm enhancers precedes activation of mesoderm and endoderm enhancers.
(a) Reconstructed mesoderm (top) and endoderm (bottom) commitment trajectories using a diffusion pseudotime method applied to the RNA expression data. Shown are scatter plots of the first two diffusion components, with cells coloured according to their lineage assignment. (b) DNA methylation (red) and chromatin accessibility (blue) dynamics of lineage-defining enhancers along the mesoderm (top) and endoderm (bottom) trajectories. Each dot denotes a single cell and black curves represent non-parametric loess regression fit. In addition, for each setting we fit a piece-wise linear regression model (vertical dashed lines indicate the thresholds used). For each model fit, the slope (r) and its significance level is displayed in the top (- for non-significant, $*$ for $0.01 < p < 0.1$ and $**$ for $p < 0.01$).

4.2.9 TET enzymes are required for efficient demethylation of lineage-defining enhancers in embryoid bodies

For a long time it was thought that DNA methylation was an irreversible epigenetic event, until a family of enzymes called ten eleven translocation proteins (TET)s were shown to erase DNA methylation marks via a succession of oxidative events [247]. This discovery fundamentally changed our understanding of DNA methylation, suggesting that it is not as static as previously assumed. In the context of development, TET enzymes have been implicated in enhancer demethylation, and loss-of-function experiments both *in vitro* and *in vivo* suggest that TET enzymes are vital for gastrulation [70, 265, 247, 170].

In our study, to test whether TET enzymes drive the lineage-specific demethylation events, we used an *in vitro* system where embryoid bodies were differentiated in serum conditions using both wild type (WT) mouse ESCs and cells that were deficient for all three TET enzymes (*TET TKO*). The

embryoid bodies were dissociated and subjected to scNMT-seq at days 2, 4-5, and 6-7 following the onset of differentiation.

Cell type assignment using the RNA expression

As in Figure 4.2, cell types were assigned by mapping the RNA expression profiles to the *in vivo* gastrulation atlas using a mutual nearest neighbours matching algorithm [105].
Notably, the WT cells from the EB differentiation protocol recapitulate the *in vivo* dynamics with remarkably accuracy (Figure 4.20). At day 2, most cells are in the pluripotent epiblast stage, which roughly corresponds to embryonic stages E4.5 to E5.5. At days 4-5, EBs begin the formation of primitive streak cells, as in embryonic stages E6.5 to E7.0. At days 6-7 of differentiation the primitive streak cells eventually commit to mesoderm (mostly) or endoderm fate, as in embryonic stages E7.0 to E8.0. In addition, at days 6-7 we observe the emergence of mature mesoderm structures including hematopoietic cell types.

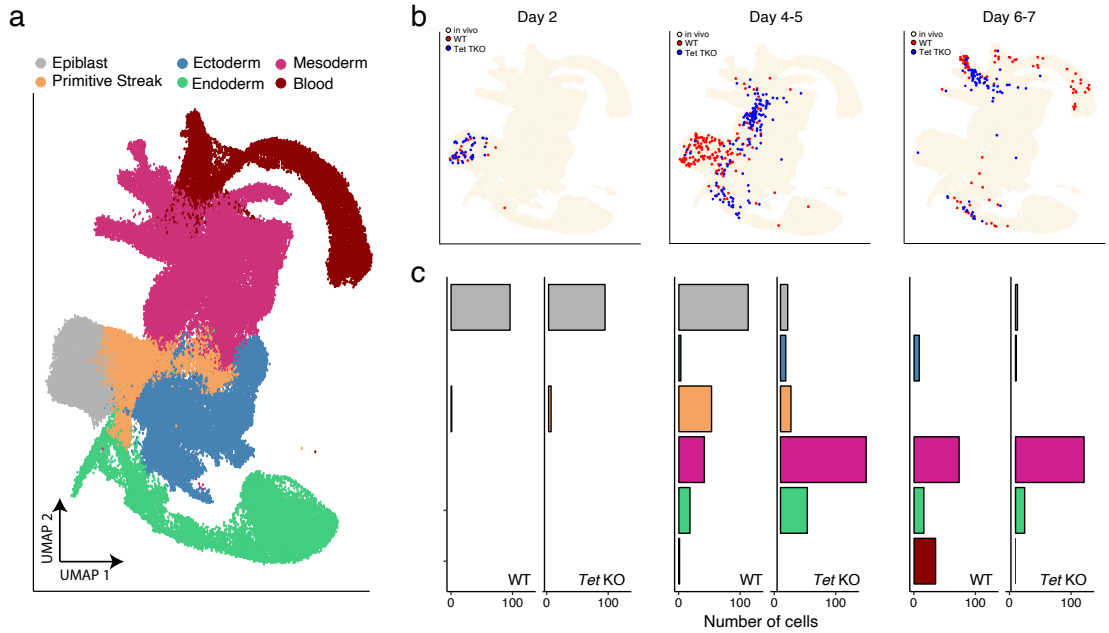


Figure 4.20: Cell type assignment for the Embryoid Body differentiation experiment.
(a) UMAP projection of the 10x atlas data set (stages E6.5 to E8.5, no extra-embryonic cells), where cells are coloured by lineage assignment.
(b) Same UMAP projection as in (a), but in this case, for each day of EB differentiation, cells are coloured by the the nearest neighbours that were used to assign cell type labels to the query cells. Cells from a WT genotype are shown in red and cells from a *TET* TKO genotype are shown in blue.
(c) Bar plots display the cell type numbers for each day of EB differentiation, grouped by WT or *TET* TKO genotype.

To validate the mapping results, we inspected the expression of marker genes for the different lineages. In general, we observe good consistency between cell type assignments and the corresponding expression profiles:

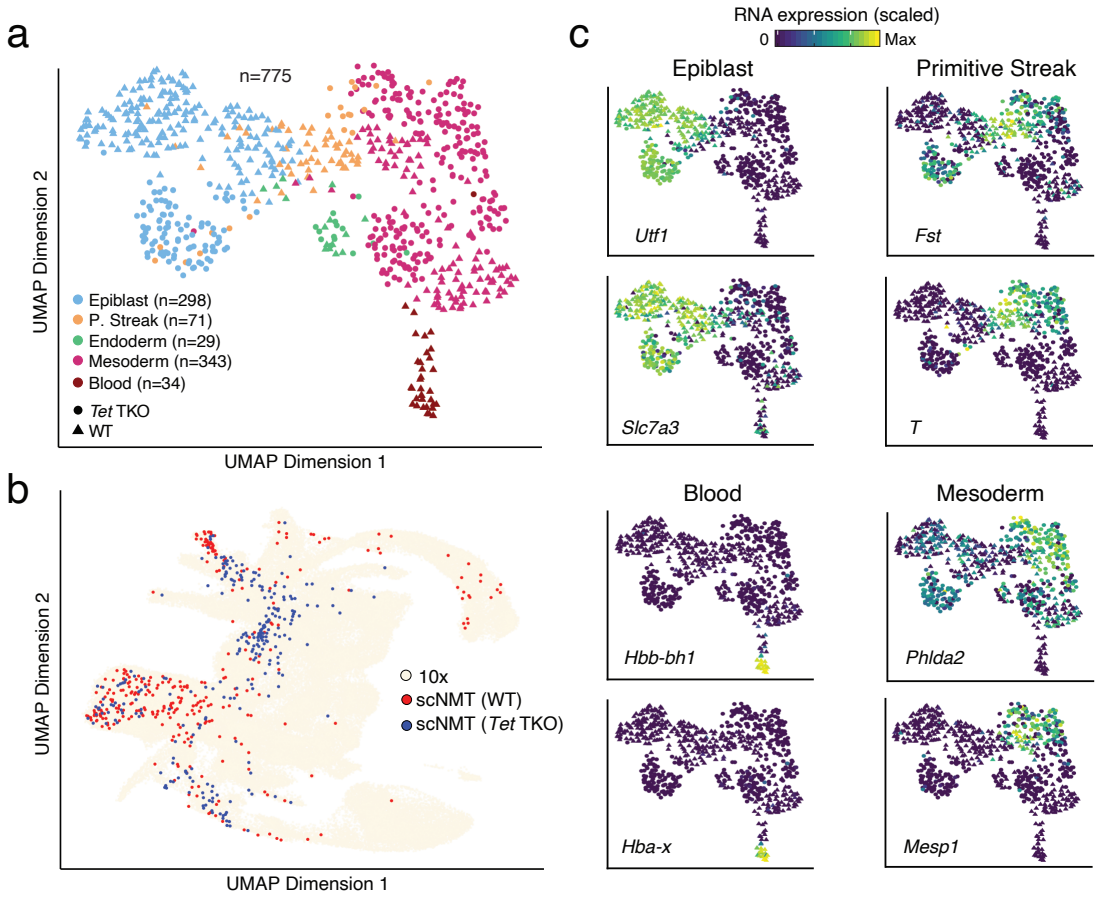


Figure 4.21: Embryoid bodies recapitulate the transcriptional heterogeneity of the mouse embryo.
(a) UMAP projection for the embryoid body dataset, where cells are coloured by lineage assignment and shaped by genotype (WT or *TET* TKO).
(b) UMAP projection of the atlas data set (stages E6.5 to E8.5, no extra-embryonic cells). Cells coloured correspond to the nearest neighbours that were used to assign cell type labels to the EB dataset, red for WT and blue for *TET* TKO.
(c) UMAP projection of embryoid body cells, as in (a), coloured by the relative RNA expression of marker genes.

Validation of epigenetic measurements

After validating the reproducibility of the EB system to capture the transcriptomics of post-implantation and early gastrulation, we proceed to validate the epigenetic measurements. At the global level, DNA methylation increases in WT cells from ≈55% at day 2 to ≈75% at day 7, whereas chromatin accessibility decreases from ≈20% at day 2 to ≈16% at day 7 (Figure 4.22).

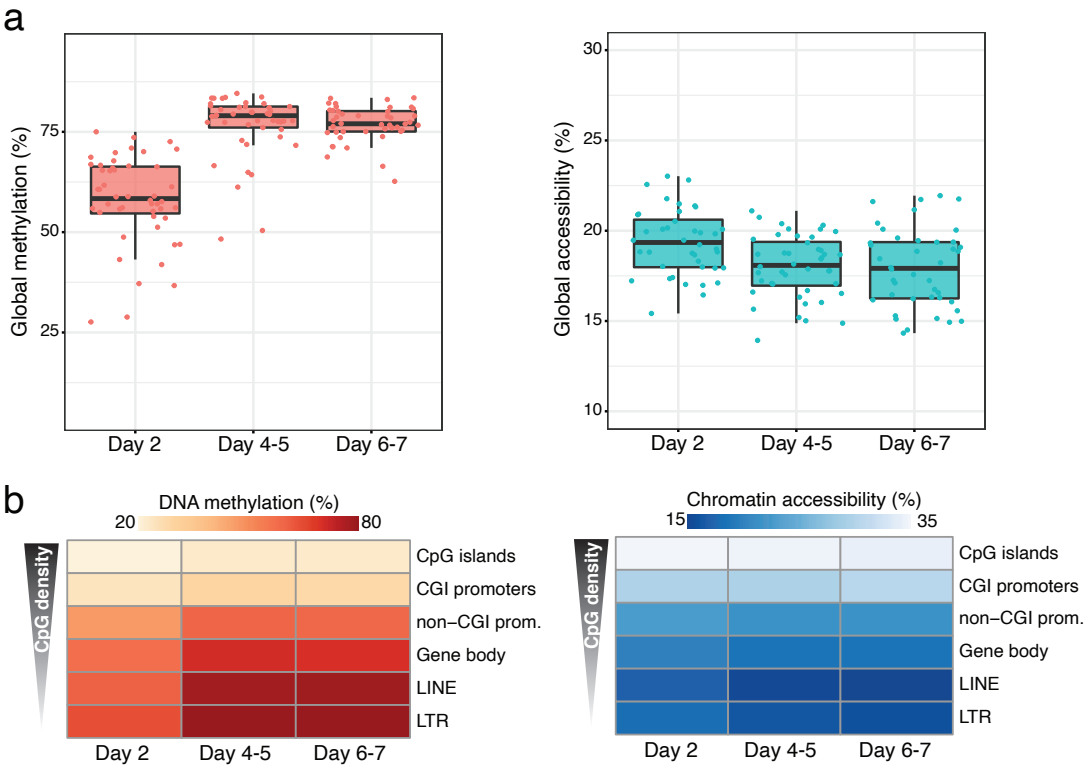


Figure 4.22: Global DNA methylation and chromatin accessibility levels during embryoid body differentiation (in WT cells).
(a) Box plots showing the distribution of genome-wide CpG methylation (left) or GpC accessibility levels (right) per stage and lineage. Each dot represents a single cell.
(b) Heatmap of DNA methylation (left) or chromatin accessibility (right) levels per stage and genomic context.

Critically, ectoderm-defining enhancers are protected from the global repressive dynamics in the epiblast-like cells. Upon mesoderm commitment, mesoderm-defining enhancers demethylate from ≈85% to ≈70% and increase in accessibility from ≈19% to ≈30% (Figure 4.23).

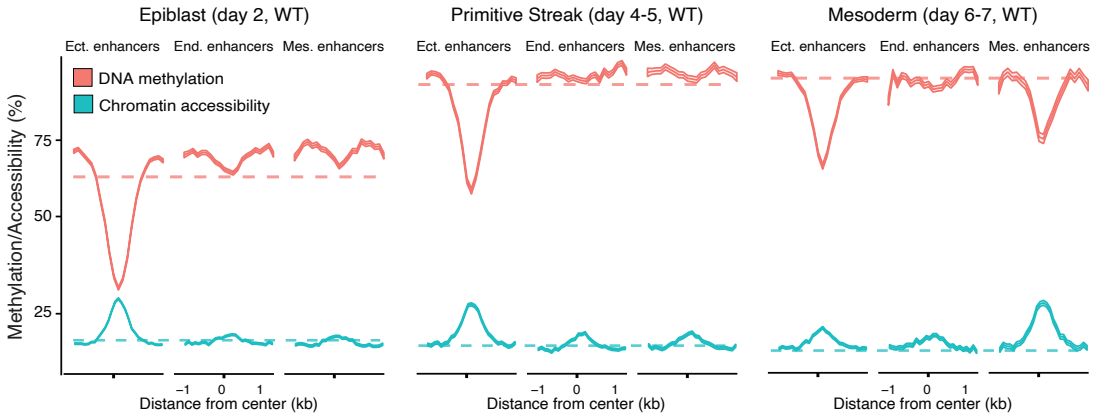


Figure 4.23: Profiles of DNA methylation (red) and chromatin accessibility (blue) at lineage-defining enhancers quantified along EB differentiation using WT cells. Shown are running averages in consecutive 50bp windows around the centre of the ChIP-seq peaks (1kb upstream and downstream). Solid lines display the mean across cells and shading displays the corresponding standard deviation. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue).

In conclusion, although the absolute numbers differ with the *in vivo* data, the relative changes in DNA methylation and chromatin accessibility in WT EBs substantially mirror the *in vivo* results.

Characterisation of the *TET* TKO phenotype

Having validated the EB system from a transcriptomic and epigenetic perspective, we proceed to compare the WT and the *TET* TKO cells. At the epigenetic level, *TET* TKO epiblast-like cells (day 2) display higher levels of DNA methylation in ectoderm enhancers, but no differences in mesoderm or endoderm enhancers (Figure 4.24). No significant differences are observed between WT and *TET* TKO for chromatin accessibility. Interestingly, the *TET* TKO cells also display an increased proportion of cells undergoing mesendoderm transition (days 4-5, 95% versus 51% in the WT). This is suggestive of an early induction of gastrulation.

After the mesendoderm transition (days 4-5), mesoderm-committed *TET* TKO cells (days 6-7) failed to properly demethylate mesoderm-specific enhancers (Figure 4.24). This indicates that (1) enhancer demethylation is not required for early mesoderm commitment, and (2) demethylation of lineage-defining enhancers results from an active process that is at least partially driven by TET proteins.

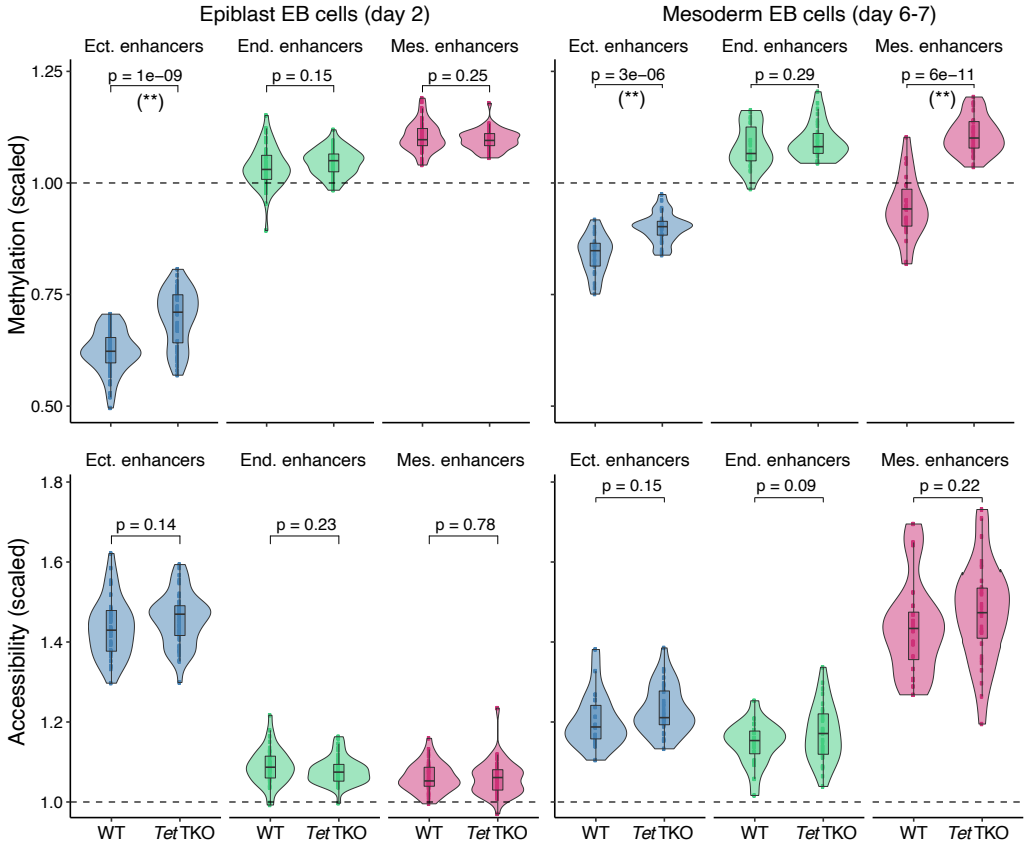


Figure 4.24: Distribution of DNA methylation (top) or chromatin accessibility values for lineage-defining enhancers in the epiblast-like cells at day 2 and the mesoderm-like cells at days 6-7 of EB differentiation.
The y-axis shows the DNA methylation or chromatin accessibility levels (%) scaled to the genome-wide levels. P-values resulting from comparisons of group means (t-test) are displayed above each pair of box plots. Asterisks denote significant differences at a significance threshold of 1% FDR.

Finally, at days 6-7 we observe a systematic loss of hematopoietic cell types in the *TET* TKO (Figure 4.20). This suggests that TET-mediated demethylation events, although not crucial for early mesendoderm commitment, seem to be important for subsequent cell fate decisions. Notably, our observations are concordant with findings from previous studies *in vivo* [70], which demonstrated that *TET* TKO embryos are able to initiate gastrulation, but by E8.5 they display defective mesoderm migration with no recognisable mature mesoderm structures.

4.3 Conclusions, limitations and future perspectives

In this work we have employed scNMT-seq to generate a multi-omics atlas of mouse gastrulation at single-cell resolution. We find that the initial exit from pluripotency coincides with the establishment of a repressive epigenetic landscape, characterised by increasing levels of DNA methylation and decreasing levels of chromatin accessibility. This gradual lock-down of the genome is followed by the emergence of distal regulatory elements that become unmethylated and accessible upon germ layer commitment. Most notably, when tracing back the epigenetic dynamics for the lineage-defining

enhancers to the early epiblast stage, we observe that post-implantation cells display epigenetic priming for an ectoderm fate. This finding supports the existence of a default path in the Waddington landscape of development, with the ectoderm being the default germ layer in the embryo. In contrast, commitment to endoderm and mesoderm fates occurs by an active diversion from the default path likely driven by signalling cues in the primitive streak transient state.

Experimental evidence exist to support this hypothesis. Several groups have shown that, in the absence of external stimuli, ESCs differentiate to neurons [213, 114], a phenomenon that still remains largely unexplained. We believe that the epigenetic priming of neuroectoderm enhancers that we identified in this study could provide the molecular logic for a hierarchical emergence of the primary germ layers.

Our study is not free of limitations that we hope to address in the future:

- Scalability: in its current form, scNMT-seq is a laborious and expensive protocol, unsuitable for the profiling of large numbers of cells. In this study, we had to rely on pseudobulk approaches to obtain sufficient statistical power for some of our results. Also, it is likely that we have been underpowered to detect subtle yet important epigenetic variation. As discussed in Chapter 2, some optimisations can be implemented to make the assay more high-throughput, with the eventual goal of applying it to study organogenesis.

- Coverage: single-cell bisulfite sequencing technologies yield very sparse measurements, particularly for small regulatory elements. Hence, it is very likely that we have missed important regulatory elements in our analysis.

- Further experimental support for the default pathway: the default pathway hypothesis is appealing and supported by independent experiments. Nonetheless, further investigation is required to understand how it works. How are ectoderm enhancers epigenetically primed (i.e. what protects them from DNA methylation in the pluripotent stages)? How could we perturb the default pathway? Is there a way to artifically methylate ectoderm enhancers by precise genome targeting?

- Further experimental validation for the role of *TET* TKO in lineage commitment: our experiments using EBs have yielded promising insights, but as a next step we should verify whether this can be reproduced in an *in vivo* setting. However, obtaining knock out mice is challenging and time-consuming, and more importantly, the phenotypic effects of the mutation can be masked by gross developmental defects or embryo lethality. For this reason, we are going to explore the usage of chimeric embryos by injecting fluorescence-labelled ESCs cells with a *TET* TKO background into wild-type blastocysts. If the injection is successful, the adult will contain a mixture of WT and *TET* TKO cells that can be separated by FACS and studied independently [232]. A major benefit of this experimental system is that any function impaired in the *TET* TKO cells should be compensated by the WT cells and, in contrast a full knock out, the embryo can develop (almost) normally.

Chapter 5

MOFA+: a statistical framework for the integration of large-scale structured datasets

In Chapter 3 we developed Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of multi-omics data. MOFA addresses key challenges in data integration, including overfitting, noise reduction, handling of missing values and improved interpretation of the model output. However, when applied to increasingly-large single-cell genomics datasets, the variational inference scheme offers limited scalability. In addition, the increased experimental throughput has facilitated the simultaneous study of multiple conditions within the same experimental design [251]. However, MOFA makes strong independence assumptions about the dependencies across samples and it provides no principled strategy to model datasets where the samples are structured into multiple groups (i.e. batches, donors or even independent studies). In this Chapter we improve the model formulation with the aim of performing integrative analysis of large-scale datasets where the features are structured into multiple data modalities (views) and the samples (or cells) are structured into different groups.

The work discussed in this Chapter has been peer-reviewed and published in [13]. The project was conceived by Damien Arnol and me. The mathematical derivations and the implementation of the stochastic variational inference scheme were done together by Damien Arnol, Yonatan Deloro and me. I implemented the downstream analysis package, but with significant contributions from Danila Bredikhin. I generated most figures and I wrote the manuscript with feedback from all authors. John C. Marioni and Oliver Stegle supervised the project.

5.1 Theoretical foundations

5.1.1 Exponential family distributions

Exponential family distributions are a parametric class of probability distributions that have characteristic mathematical properties which make them amenable for probabilistic modelling.
The majority of probability distributions that are commonly used in statistics belong to the exponential family, including the normal or Gaussian, Gamma, Poisson, Bernoulli, Exponential, etc.
Formally, exponential family distributions can be represented in the following form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\{\eta(\boldsymbol{\theta})T(\mathbf{x}) - A(\boldsymbol{\theta})\} \tag{5.1}$$

where $\mathbf{x}$ is a multivariate random variable and $\boldsymbol{\theta}$ are the distribution's parameters. Each term has a common notation: T($\mathbf{x}$): sufficient statistics; $\eta(\boldsymbol{\theta})$: natural parameters; $h(\mathbf{x})$: base measure; $A(\eta)$: the log-partition function (or the normaliser).

The exponential family form for the probability distributions that are frequently used in this thesis are shown below:

Univariate normal distribution:

$$\begin{aligned}
&\eta(\mu, \sigma) = [\frac{\mu}{\sigma^2}; -\frac{1}{2\sigma^2}] \\
&h(x) = \frac{1}{\sqrt{2\pi}} \\
&T(x) = [x; x^2] \\
&A(\mu, \sigma) = \frac{\mu^2}{2\sigma^2} + \log \|\sigma\|
\end{aligned}$$

Multivariate normal distribution:

$$\begin{aligned}
&\eta(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = [\Sigma^{-1}\mu; -0.5\Sigma^{-1}] \\
&T(x) = [x; xx^T] \\
&h(x) = (2\pi)^{-\frac{k}{2}} \\
&A(\theta) = -0.25\eta_1^T\eta_2 - 1\eta_1 - 0.5\log(\| - 2\eta_2\|)
\end{aligned}$$

Gamma distribution:

$$\begin{aligned}
&\eta = [\alpha - 1; -\beta] \\
&T(x) = [\log x; x] \\
&h(x) = 1 \\
&A(\theta) = \log(\Gamma(\eta_1 + 1)) - (\eta_1 + 1)\log(-\eta_2)
\end{aligned}$$

Beta distribution:

$$\begin{aligned}
&\eta = [\alpha; \beta] \\
&T(x) = [\log x; \log(1 - x)] \\
&h(x) = \frac{1}{x(1 - x)} \\
&A(\theta) = \log(\Gamma(\eta_1)) + \log(\Gamma(\eta_2)) - \log(\Gamma(\eta_1 + \eta_2))
\end{aligned}$$

In the context of Bayesian inference, the main property that make exponential family distributions indispensable is that they have conjugate priors. That is, the combination of likelihood and prior distributions ensure a closed-form posterior distribution which is of the same form as the prior. As we have discussed in Chapter 3, this property is essential for enabling efficient statistical inference,

otherwise posterior distributions must be computed using expensive and approximate numerical methods.

5.1.2 Gradient ascent

Gradient ascent is a first-order optimization algorithm for finding the maximum of a function [31, 214]. Formally, for a differentiable function $F(x)$, the iterative scheme of gradient ascent is:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \nabla F(\mathbf{x}^{(t)}) \tag{5.2}$$

In short, the algorithm works by taking steps proportional to the gradient ∇F evaluated at each iteration t. Importantly, the step size $\rho^{(t)}$ is typically adjusted at each iteration t such that it satisfies the Robbins-Monro conditions: $\sum_t \rho^{(t)} = \infty$ and $\sum_t (\rho^{(t)})^2 < \infty$. F is guaranteed to converge to the *global* maximum if the objective function is convex [256]. If F is not convex, the algorithm is sensitive to the initialisation $\mathbf{x}^{t=0}$ and can converge to local optima.

Stochastic gradient ascent

Gradient ascent becomes prohibitively slow with large datasets, mainly because of the computational cost involved in the iterative calculation of gradients [280].
A simple strategy to speed up gradient ascent is to replace ∇F by an estimate $\hat{\nabla} F$ using a random subset of the data (minibatch). The iterative scheme is then defined in the same way as in standard gradient ascent:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \hat{\nabla} F(\mathbf{x}^{(t)}) \tag{5.3}$$

Natural gradient ascent

Gradient ascent becomes problematic when applied to probabilistic models. To give the intuition, consider a probabilistic model with a hidden variable x and corresponding parameter θ, with a general objective function $\mathcal{L}(\theta)$. From the definition of a derivative:

$$\nabla \mathcal{L}(\theta) = \lim_{||h|| \to 0} \frac{\mathcal{L}(\theta + h) - \mathcal{L}(\theta)}{||h||}$$

where h represents an infinitesimally small positive step in the space of θ.
To find the direction of steepest ascent, one would need to search over all possible directions d in an infinitely small distance h, and select the $\hat{d}$ that gives the largest gradient:

$$\nabla \mathcal{L}(\theta) = \lim_{h \to 0} \frac{1}{h} \arg\max_{d \, s.t. \|d\| = h} \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

Importantly, this operation requires a distance metric to quantify what a *small* distance h means. In standard gradient ascent, this is measured using the Euclidean norm, and the direction of steepest ascent is hence dependent on the Euclidean geometry of the θ space. This problematic when doing

probabilistic modelling because it does not consider the uncertainty that underlies probability distributions. When θ is a random variable with an underlying probability distribution, a small step from $\theta^{(t)}$ to $\theta^{(t+1)}$ does not guarantee an equivalently small change from $\mathcal{L}(\theta^{(t)})$ to $\mathcal{L}(\theta^{(t+1)})$. To illustrate this, consider the following example of four random variables:

$$\begin{aligned} \mathcal{N}(\psi_1 \,|\, 0, 5) \qquad & \mathcal{N}(\psi_3 \,|\, 0, 1) \\ \mathcal{N}(\psi_2 \,|\, 10, 5) \qquad & \mathcal{N}(\psi_4 \,|\, 10, 1) \end{aligned} \tag{5.4}$$

Using the Euclidean metric, the distance between ψ_1 and ψ_2 is the same as the distance between ψ_3 and ψ_4. However, the distance in distribution space (measured for example by the KL divergence) is much larger between ψ_3 and ψ_4 than between ψ_1 and ψ_2 (Figure 5.1).

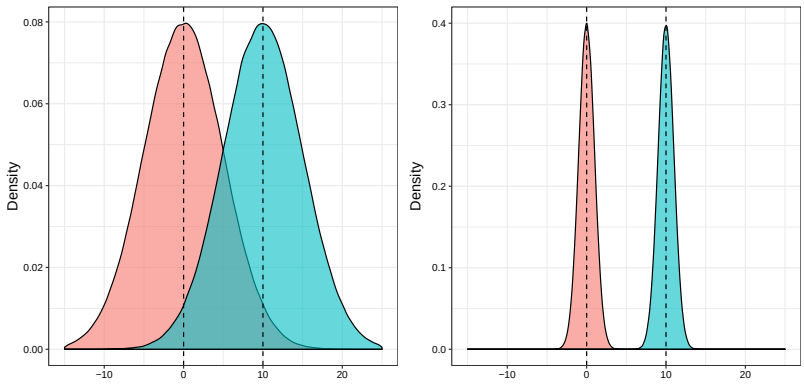


Figure 5.1: Illustration of the problem of using the Euclidean norm as a distance measure between parameters of probability distributions.
In both plots, the red and blue distributions are separated by the same Euclidean distance of 100. Yet, the distance in probability space between the two distributions is higher in the right.

This basic simulation suggests that replacing the Euclidean distance by the KL divergence as a distance metric may be more appropriate in the context of probabilistic modelling:

$$\nabla_{KL}\mathcal{L}(\theta) = \lim_{h \to 0} \frac{1}{h} \underset{d\, s.t. KL[p_\theta || p_{\theta+d}]=h}{\arg\max} \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

The direction of steepest ascent measured by the KL divergence is called the natural gradient [8, 192]. To find the optimal $\hat{d}_{KL}$, one needs to solve the following optimisation problem:

$$\underset{d}{\arg\min}\, \mathcal{L}(\theta + d) \qquad \text{subject to} \quad KL[p_\theta || p_{\theta+d}] < c$$

where c is an arbitrary constant. Previous works have shown that this can be solved by introducing Lagrange multipliers and Taylor expansions [8, 155]. The solution corresponds to the standard (Euclidean) gradient pre-multiplied by the inverse of the Fisher Information Matrix of $q(x|\theta)$:

$$\hat{d}_{KL} \propto \mathbf{F}^{-1}(\theta) \nabla_\theta \mathcal{L}(\theta) \tag{5.5}$$

where $\mathbf{F}(\theta)$ is defined as

$$\mathbf{F}(\theta) = \mathbb{E}_{q(x|\theta)}[(\nabla_\theta \log q(x|\theta))(\nabla_\theta \log q(x|\theta))^T]$$

In conclusion, while the standard gradient points to the direction of steepest ascent in Euclidean space, the natural gradient points to the direction of steepest ascent in a space where distances are defined by the KL divergence [155, 8, 119].

5.1.3 Stochastic variational inference

In this section I will demonstrate how to derive a stochastic variational inference algorithm for general Bayesian models. This work is inspired and adapted from [119]. A comprehensive mathematical derivation of the algorithm is not sought in this Chapter, instead I will describe a modified and simplified derivation that captures the gist of the original. For a complete mathematical derivation I refer the reader to [119].

This section builds upon three theoretical foundations that have been introduced before: Variational inference (Section 3.1.4), exponential family distributions (Section 5.1.1) and (natural) gradient ascent (Section 5.1.2).

Model definition

Consider a probabilistic model with a set of unobserved random variables, observations and (non-random) parameters. We begin by classifying the variables of the model into four different categories:

- observations ($\mathbf{Y}$): N different vectors $\mathbf{y}_n$, each one containing the observed variables for the n-th sample.
- local (hidden) variables ($\mathbf{Z}$): N different vectors $\mathbf{z}_n$, each one containing K hidden variables associated with the n-th sample.
- *global* (hidden) variables ($\boldsymbol{\beta}$): one vector that contains B hidden variables not indexed by n.
- parameters (non-random) for the *global* variables ($\boldsymbol{\alpha}_\beta$).
- parameters (non-random) for the *local* variables ($\boldsymbol{\alpha}_z$).

First, let us assume the following factorisation of the joint distribution:

$$p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}_\beta, \boldsymbol{\alpha}_z) = p(\mathbf{Z}|\boldsymbol{\alpha}_z)p(\boldsymbol{\beta}|\boldsymbol{\alpha}_\beta)\prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\beta}) \tag{5.6}$$

and the corresponding graphical model representation:

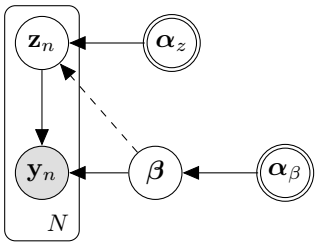


Figure 5.2: Graphical model for a general probabilistic model where unobserved variables are classified as *global* and *local*.
The dashed line indicates that the connection between *global* and *local* variables is optional, not used in the MOFA model.

Notice that the difference between *local* and *global* variables lies on the conditional dependency assumptions. The *local* variables for the n-th sample $\mathbf{z}_n$ are conditionally independent from any other observation $\mathbf{y}_j$ or *local* variable $\mathbf{z}_j$ (where $j \neq n$), given that the *global* variables $\boldsymbol{\beta}$ are observed:

$$p(\mathbf{y}_n, \mathbf{z}_n | \mathbf{y}_j, \mathbf{z}_{nj}, \boldsymbol{\beta}, \boldsymbol{\alpha}_{z_n}, \boldsymbol{\alpha}_{z_j}) = p(\mathbf{y}_n, \mathbf{z}_n | \boldsymbol{\beta}, \boldsymbol{\alpha}_{z_n})$$

To relate this formulation to the MOFA model, the *local* variables would contain the factors whereas the *global* variables would contain the feature weights.

For simplicity in the derivation, we will assume the existence of a single *global* variable β, a single parameter α_β for the *global* variables and a single parameter $\alpha_{z_{nk}}$ for each *local* variable.

The first assumption in the model is that the prior distributions of the *local* and *global* variables are members of the exponential family (see Equation (5.1))

$$\begin{aligned} p(\beta|\alpha_\beta) &= h(\beta) \exp\{\eta_g(\alpha_\beta) t(\beta) - a_g(\alpha_\beta)\} \\ p(z_{nk}|\alpha_z) &= h(z_{nk}) \exp\{\eta_l(\alpha_z) t(z_{nk}) - a_l(\alpha_z)\} \end{aligned} \tag{5.7}$$

The second assumption is that the complete conditionals of the unobserved variables are also members of the exponential family:

$$\begin{aligned} p(\beta|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}) &= h(\beta) \exp\{\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha})^T t(\beta) - a_g(\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}))\} \\ p(\mathbf{z}_n|\mathbf{y}_{nj}, \mathbf{z}_{nj}, \beta) &= h(\mathbf{z}_n) \exp\{\eta_l(\mathbf{y}_{nj}, \mathbf{z}_{nj}, \beta)^T t(\mathbf{z}_n) - a_l(\eta_l(\mathbf{y}_{nj}, \mathbf{z}_{nj}, \beta))\} \end{aligned} \tag{5.8}$$

Setting up the inference problem

First, we set up the variational distributions for both the *local* variables and the *global* variables. Here we are going to assume that all unobserved variables are independent (mean-field assumption)

$$q(\mathbf{z}, \beta) = q(\beta|\lambda) \prod_{n=1}^{N} \prod_{k=1}^{K} p(z_{nk}|\phi_{nk})$$

and belong to the same exponential family as the corresponding prior distribution:

$$q(\beta|\lambda) = h(\beta)\exp\{\eta_g(\lambda)t(\beta) - a_g(\lambda)\} \tag{5.9}$$

$$q(z_{nk}|\phi_{nk}) = h(z_{nk})\exp\{\eta_l(\phi_n)t(z_{nk}) - a_l(z_{nk})\} \tag{5.10}$$

where λ are the parameters governing the variational distribution for the *global* variables and ϕ_{nk} are the parameters governing the variational distribution for the k-th *local* variable and the n-th sample.

From the assumptions above, the ELBO (the objective function in variational inference, introduced in Chapter 3) factorises as:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q(\mathbf{Z},\beta)}[\log p(\mathbf{Y},\mathbf{Z},\beta)] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\beta)}[\log q(\beta)] \\ &= \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}_n,\beta)}[\log p(\mathbf{y}_n,\mathbf{z}_n,\beta)] - \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{q(z_{nk})}[\log q(z_{nk})] - \mathbb{E}_{q(\beta)}[\log q(\beta)]\end{aligned} \tag{5.11}$$

Notice that the objective decomposes into *global* terms (not involving N) and *local* terms (involving N). Importantly, the *local* terms can be approximated using estimates of the gradient by subsampling the data set. Assuming a mini-batch of size S:

$$\hat{\mathcal{L}} = \frac{N}{S}\sum_{n=1}^{S}\mathbb{E}_{q(\mathbf{z}_n,\beta)}[\log p(\mathbf{y}_n,\mathbf{z}_n,\beta)] - \frac{N}{S}\sum_{s=1}^{S}\sum_{k=1}^{K}\mathbb{E}_{q(z_{nk})}[\log q(z_{nk})] - \mathbb{E}_{q(\beta)}[\log q(\beta)]$$

If the samples are independent then the expectation of this noisy gradient is equal to the true gradient. This is the main principle of stochastic optimisation. The next step is to derive an iterative algorithm to find the values of the variational parameters that maximise the ELBO.

Calculating the gradient for the *global* parameters

To derive the updates for the *global* parameters we first write the ELBO in terms of λ:

$$\mathcal{L}(\lambda) = \mathbb{E}_{q(z,\beta)}[\log p(\beta|\mathbf{Y},\mathbf{Z})] - \mathbb{E}_{q(\beta)}[\log q(\beta)] + \text{const.}$$

where the constant term captures all quantities that do not depend on λ. Then, from the assumption that the complete conditionals and the variational distributions belong to the exponential family (Equations (5.8) to (5.9)):

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{E}_{q(z,\beta)}[\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})^T t(\beta)] - \mathbb{E}_{q(\beta)}[\lambda^T t(\beta) - a_g(\lambda)] + \text{const.} \\ &= \mathbb{E}_{q(z)}[\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})^T]\nabla a(\lambda) - \lambda^T\nabla a_g(\lambda) - a_g(\lambda) + \text{const.}\end{aligned}$$

where we have used the exponential family identity $\mathbb{E}_{q(\beta)}[t(\beta)] = \nabla a_g(\lambda)$.

Taking the gradient with respect to λ:

$$\nabla_\lambda\mathcal{L}(\lambda) = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_{q(z)}[\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})] - \lambda) \tag{5.12}$$

and setting it to zero leads to the solution:

$$\lambda = \mathbb{E}_{q(z)}[\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha})] \tag{5.13}$$

Calculating the gradient for the *local* parameters

Turning to the *local* parameters, as a function of ϕ_{nk} the ELBO becomes:

$$\mathcal{L}(\phi_{nk}) = \mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\log p(\mathbf{z}_{nj}|\mathbf{y}_n, \mathbf{z}_{nj}, \beta)] - \mathbb{E}_{q(z_{nk})}[\log q(z_{nk})] + \text{const.}$$

Again, from the assumption that the complete conditionals and the variational distributions belong to the exponential family (Equations (5.8) to (5.9)):

$$\begin{aligned}\mathcal{L}(\phi_{nk}) &= \mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)^T t(\mathbf{z}_{nj})] - \mathbb{E}_{q(z_{nk})}[\phi_{nk} t(z_{nk}) - a_l(\phi_{nk})] + \text{const.} \\ &= \mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)]^T \nabla a_l(\phi_{nk}) - \phi_{nk} \nabla a_l(\phi_{nk}) - a_l(\phi_{nk}) + \text{const.}\end{aligned}$$

Taking the gradient with respect to ϕ_{nk}:

$$\nabla_\phi \mathcal{L}(\phi_{nk}) = \nabla^2_\phi a_l(\phi_{nk})(\mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)] - \phi_{nk}) \tag{5.14}$$

and setting it to zero leads to the following solution:

$$\phi_{nk} = \mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)] \tag{5.15}$$

Coordinate ascent variational inference algorithm

Now that we have the gradients for both the *local* and the *global* parameters, we can define a gradient ascent algorithm to optimise the model:

Algorithm 1 Coordinate ascent variational inference algorithm

1: Initialise the *global* parameters $\boldsymbol{\lambda}^{(t=0)}$
2: **repeat**
3: **for** each *local* variational parameter ϕ_{nk} **do**
4: $\phi_{nk}^{(t+1)} \leftarrow \mathbb{E}_{q(\boldsymbol{\beta}, \mathbf{z}_{nj})^t}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \boldsymbol{\beta})]$
5: **end for**
6: **for** each *global* variational parameter λ **do**
7: $\lambda^{(t+1)} = \mathbb{E}_{q(z)^t}[\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha})]$
8: **end for**
9: **until** Convergence

However, as discussed in Section 5.1.2, the use of Euclidean-based gradients ignores important information about the geometry of the distribution and is thus not optimal for the optimisation of probabilistic models. Next, we will derive a similar coordinate ascent algorithm but using instead the natural gradient.

Deriving the natural gradients for the *global* variational parameters

From Equation (5.12), the gradient of the ELBO with respect to the *global* parameters λ is:

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_{q(z)}[\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha})] - \lambda)$$

Premultiplying by $\mathbf{F}(\beta)^{-1} = \nabla^2_\lambda a_g(\lambda)$ gives the natural gradient for the *global* parameters:

$$\hat{\nabla}_\lambda \mathcal{L}(\lambda) = \mathbb{E}_{q(z)}[\eta_g(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha})] - \lambda$$

Deriving the natural gradients for the *local* variational parameters

From Equation (5.14), the gradient of the ELBO with respect to the *local* parameters $\boldsymbol{\phi}$ is:

$$\nabla_\phi \mathcal{L}(\phi_{nk}) = \nabla^2_\phi a_l(\phi_{nk})(\mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)] - \phi_{nk})$$

Premultiplying by $\mathbf{F}(z_{nk})^{-1} = \nabla^2_\phi a_l(\phi_{nk})$ gives the natural gradient for the *global* parameters:

$$\hat{\nabla}_\phi \mathcal{L}(\phi_{nk}) = \mathbb{E}_{q(\beta, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \beta)] - \phi_{nk}$$

Remarkably, the natural gradient for both the *local* and *global* variational parameters is simply the standard gradient subtracting the current value of the parameters. Thus, the Fisher Information matrix does *not* need to be explicitly computed at each iteration, which leads to a considerable simplification of the problem.

Stochastic variational inference algorithm using natural gradients

After replacing the Euclidean gradient with the natural gradients, the model can be trained using the following stochastic algorithm based on gradient descent (Algorithm 2). Notice that the stochastic variational inference algorithm introduces additional hyperparameters:

- **Batch size**: controls the number of samples that are used to compute the gradients at each iteration. A trade off exists where large batch sizes lead to a more expensive computation of the gradient but yield a less noisy estimate.
- **Learning rate**: The learning rate $p(t)$ controls the step size in the direction of the natural gradient, with high learning rates leading to higher steps. In the natural gradient setting, the learning rate also controls how much memory from previous iterations is translated to the current updates. The particular case of a constant learning rate of 1 yields no memory from previous iterations (thus simplifies to standard gradient ascent). To ensure proper convergence, the learning rate has to be decayed during training. Several strategies exist [246], here we used the simple function $\rho(t) = \frac{\rho_0}{(1+\kappa t)^{3/4}}$, which introduces two extra hyperparameters: (1) The forgetting rate κ, which controls the decay of the learning rate, and ρ_0 which determines the initial learning rate.

Algorithm 2 Stochastic variational inference algorithm using natural gradients

Initialise the *global* parameters $\boldsymbol{\lambda}^{(t=0)}$.
Initialise step size $\rho^{(t=0)}$
repeat
    sample $\mathcal{B}$ a mini-batch of samples of size S
    for each *local* variational parameter ϕ_{nk} such that n is in batch $\mathcal{B}$ **do**

$$\phi_{nk}^{(t+1)} = \mathbb{E}_{q^{(t)}(\boldsymbol{\beta}, \mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n, \mathbf{z}_{nj}, \boldsymbol{\beta})]$$

    end for
    for each *global* variational parameter λ **do**

$$\begin{aligned}\lambda^{(t+1)} &= (1-\rho^{(t)})\lambda^{(t)} + \rho^{(t)}\hat{\nabla}_\lambda \mathcal{L}^S(\lambda) \\ &= (1-\rho^{(t)})\lambda^{(t)} + \rho^{(t)}\mathbb{E}_{q^{(t+1)}(z)}\left[\frac{N}{S}\eta_g(\mathbf{Y}_{[n\in\mathcal{B}],:}, \mathbf{Z}_{[n\in\mathcal{B}],:}, \boldsymbol{\alpha})\right]\end{aligned}$$

        where $[n \in \mathcal{B}]$ denotes the subset of indices corresponding to the samples in $\mathcal{B}$
    end for
until Convergence

5.2 Model description

In MOFA+ we introduce two key novelties, both in the model aspect and in the inference scheme. In the model side we introduce a principled approach for modelling multi-omic data set where the samples are structured into non-overlapping groups, where groups typically correspond to batches, donors or experimental conditions. In the inference side we implement a stochastic inference algorithm to improve scalability and enable inference with large single-cell datasets.

Formally, we generalise the model to a disjoint set of M input views (i.e. groups of features) and G input groups (i.e. groups of samples). The data is factorised according to the following model:

$$\mathbf{Y}_g^m = \mathbf{Z}_g(\mathbf{W}^m)^T + \boldsymbol{\epsilon}_g^m \tag{5.16}$$

where $\mathbf{Z}_g \in \mathbb{R}^{N_g \times K}$ are a set of G matrices that contain the factor values for the g-th group and $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ are a set of M matrices that define the feature weights for the m-th view. $\boldsymbol{\epsilon}_g^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise for each feature in each group. Notice that if $G = 1$ then the model simplifies to the MOFA framework presented in Chapter 3.

It is important to get the intuition for the multi-group formulation right. In the factor analysis setting, the aim is not to capture differential changes in *mean* levels between the groups but rather to exploit the covariation patterns of the features to identify which sources of variability (i.e. latent Factors) are consistently found across multiple groups and which ones are exclusively found within a single group. This is symmetric to the interpretation of the multi-view framework in MOFA v1: the absolute levels of the features are not compared across views, only the covariation patterns are of interest. To achieve this, the features are centered per view and also per group before fitting the model. Figure 5.3 summarises the MOFA+ pipeline.

As in MOFA v1, the linearity assumptions leads to an interpretable latent space that be visualised and employed for a range of downstream analyses, including clustering, inference of non-linear differentiation trajectories, denoising and feature selection, among others. The most important extension is the generalisation of the variance decomposition analysis, where a value of variance explained per view and group is obtained for every factor. For example, imagine that Factor 1 in Figure 5.3b corresponds to cell cycle variation, the variance decomposition analysis indicates that cell cycle is a driver of cell-to-cell heterogeneity largely in views 2 and 3, but with only minor influence in view 1. Also, this effect is manifested in groups 1 and 2, but not in group 3. This simple visualisation provides a very intuitive approach to understand variability in complex experimental designs where observations are structured into multiple views and multiple groups of cells.

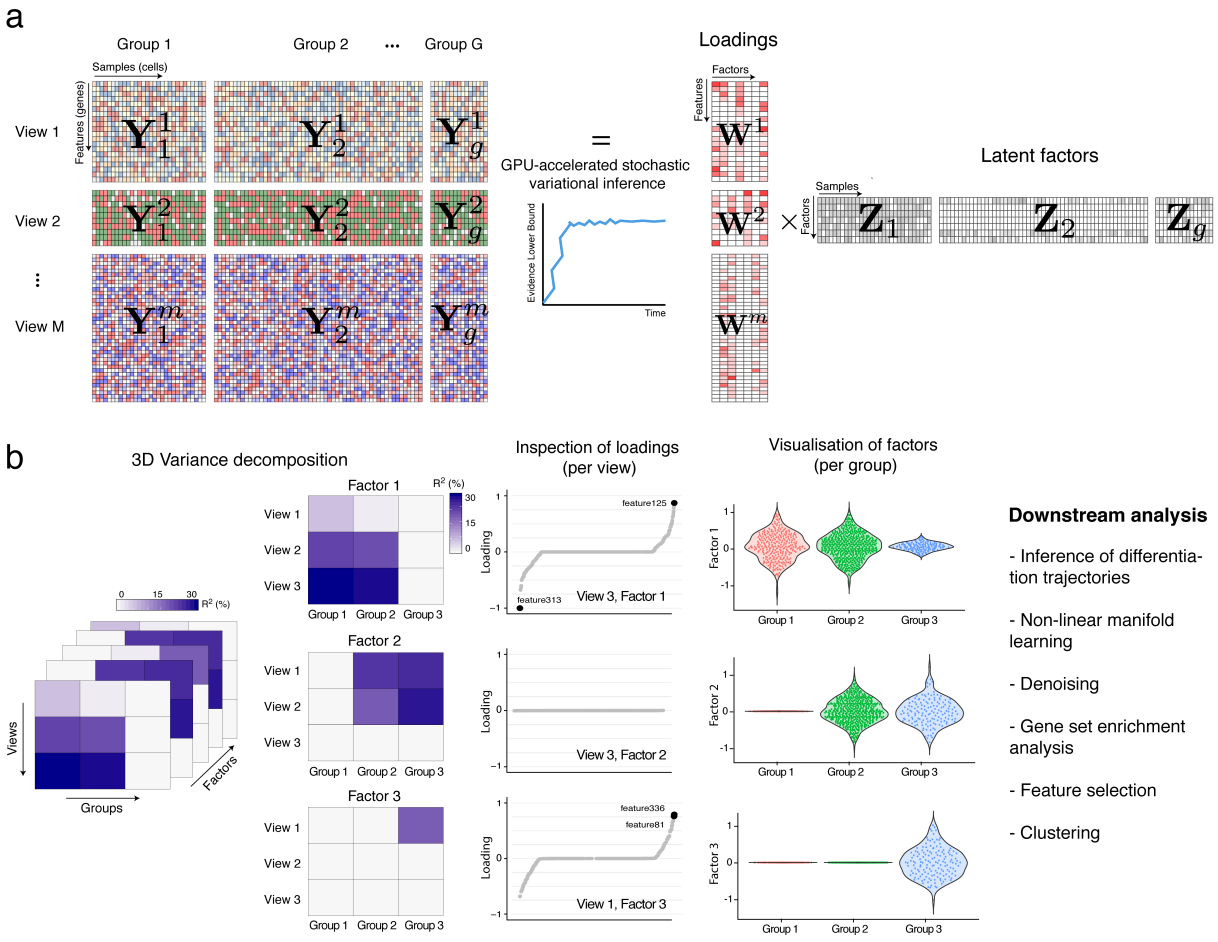


Figure 5.3: Multi-Omics Factor Analysis v2 (MOFA+) provides an unsupervised framework for the integration of multi-group and multi-view (single-cell) data.
(a) Model overview: the input data consists of multiple datasets structured into M views and G groups. Views consist of non-overlapping sets of features that often represent different assays. Analogously, groups consist of non-overlapping sets of samples that often represent different conditions or experiments. Missing values are allowed in the input data. MOFA+ exploits the covariation between the features to learn a low-dimensional representation of the data ($\mathbf{Z}$) defined by K latent factors that capture the global sources of variability. The weights ($\mathbf{W}$) provide a measure of feature importance. Model inference is performed using (GPU-accelerated) stochastic variational inference. (b) The trained MOFA+ model can be queried for a range of downstream analyses: variance decomposition, inspection of feature weights, visualisation of factors and other applications such as clustering, inference of non-linear differentiation trajectories, denoising and feature selection.

5.2.1 Model priors and likelihood

Prior on the weights

This remains the same as in MOFA v1. We adopt a two-level sparsity prior with an Automatic Relevance Determination per factor and view, and a (reparametrised [304]) feature-wise spike-and-slab prior:

$$p(\hat{w}_{dk}^m, s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{dk}^m \,|\, \theta_k^m) \tag{5.17}$$

with the corresponding conjugate priors for θ and α:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{5.18}$$

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{5.19}$$

As discussed in Chapter 3, the aim of the ARD prior is to encourage sparse associations between factors and views, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not explain any variation in view m. The aim of the spike-and-slab prior is to push individual weights to zero to yield a more interpretable solution.

Prior on the factors

In MOFA v1 we adopted an isotropic Gaussian prior which assumes an unstructured latent space *a priori*:

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{5.20}$$

This is the assumption that we want to break. Following the same logic as for the weights, the integration of multiple groups of samples requires a flexible prior distribution that defines the existence of non-overlapping groups, such that the model encourages sparse linkages between factors and groups. To formalise this intuition we simply need to extrapolate the sparsity prior from the weights to the factors:

$$p(\hat{z}_{nk}^g, s_{nk}^g) = \mathcal{N}(\hat{z}_{nk}^g \,|\, 0, 1/\alpha_k^g)\,\text{Ber}(s_{nk}^g \,|\, \theta_k^g) \tag{5.21}$$

$$p(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, a_0^\theta, b_0^\theta\right) \tag{5.22}$$

$$p(\alpha_k^g) = \mathcal{G}\left(\alpha_k^g \,|\, a_0^\alpha, b_0^\alpha\right), \tag{5.23}$$

where g is the index of the sample groups.

Prior on the noise

The variable $\boldsymbol{\epsilon}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic. In MOFA v2 we generalise the noise to have an estimate per feature and per group. This is important to capture the case where some features may be highly variable in one group but not variable in other groups.

$$p(\boldsymbol{\epsilon}_g^m) = \mathcal{N}\left(\boldsymbol{\epsilon}_g^m \,|\, 0, (\boldsymbol{\tau}_g^m)^{-1}\mathbf{I}\right) \tag{5.24}$$

$$p(\tau_g^m) = \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_g^m \,|\, a_0^\tau, b_0^\tau\right) \tag{5.25}$$

In addition, as in MOFA v1, non-Gaussian noise models can also be defined, but unless otherwise stated, we will always assume Gaussian residuals.

Graphical model

In summary, the updated model formulation introduces symmetric two-level sparsity priors in both the weights and the factors. The corresponding graphical model is shown below:

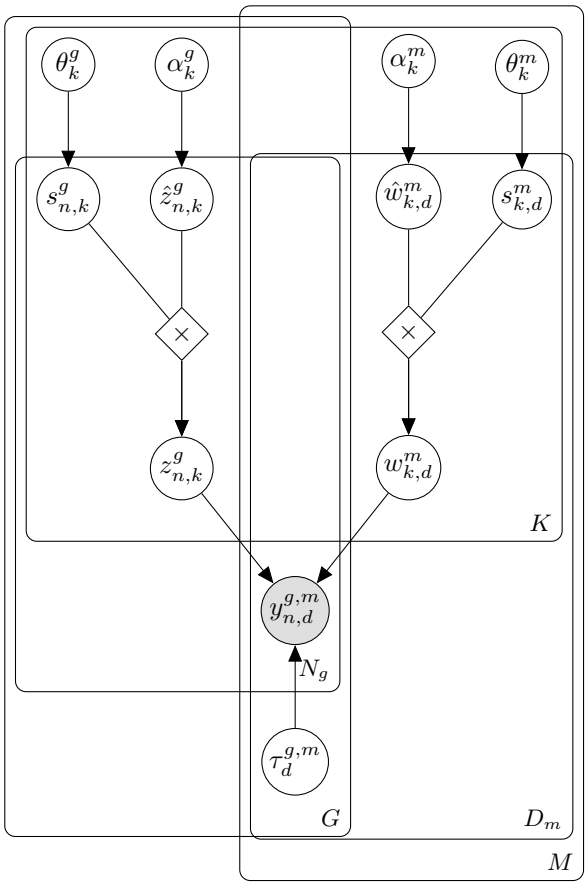


Figure 5.4: Graphical model for MOFA+.
The white circles represent hidden variables that are inferred by the model, whereas the grey circles represent the observed variables. There are a total of five plates, each one representing a dimension of the model: M for the number of views, G for the number of groups, K for the number of factors, D_m for the number of features in the m-th view and N_g for the number of samples in the g-th group.

Guidelines on the definition of views and groups

- **Views**: views typically correspond to different assays, but there is flexibility in their definition and the user can explore different definitions of views. For example, one could divide the RNA expression data into three views corresponding to mRNA, rRNA and miRNA. Similarly, one can quantify DNA methylation and chromatin accessibility data over different genomic context (enhancers, promoters, etc.).

- **Groups**: groups are generally motivated by the experimental design, but the user can also explore data-driven formulations. There is no *right* or *wrong* definition of groups, depending on the hypothesis that is sought to explore some definitions will be more useful than others.

Model selection

As discussed in Section 3.2.3, the inference procedure depends on the parameter initialisation. When using random initialisation, the Factors can vary between different model instances and a model selection step is advised. I realised that this was not a user-friendly solution and it requires a lot of computational resources when applying the model to large datasets. To simplify model training in MOFA+ we initialise the Factors using the principal components from the concatenated data set. In practice, we observe faster convergence times and better ELBO estimates when initialising with the PCA solution (Figure 5.5).

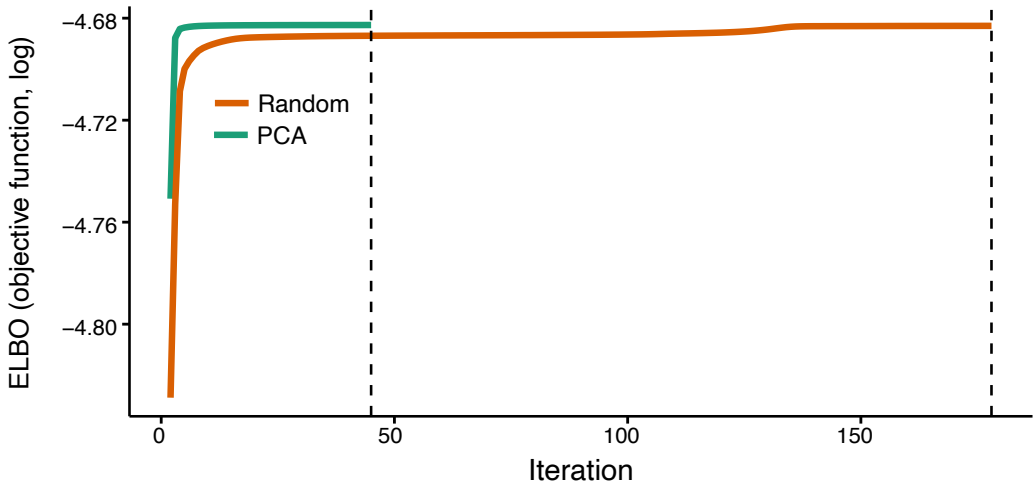


Figure 5.5: Comparison of PCA and Random initialisation in MOFA.
Data was simulated from the generative model with the following dimensions: $M = 2$ modalities, $G = 2$ groups, $D = 1000$ features, $N = 1000$ samples and $K = 10$ factors. The dashed lines mark the iteration at which the model converged.

5.2.2 A note on the implementation

The core of MOFA+ is implemented in Python, and the downstream analysis and visualisations are implemented in R. GPU acceleration is implemented using CuPy [222], an open-source matrix library accelerated with NVIDIA CUDA. To facilitate adoption of the method, we deployed MOFA+ as open-source software[1] with multiple tutorials and a web-based analysis workbench[2].

[1] https://github.com/bioFAM/MOFA2
[2] http://www.ebi.ac.uk/shiny/mofa/

5.3 Model validation

We validated the new features of MOFA+ using simulated data drawn from its generative model.

5.3.1 Stochastic variational inference

We simulated data with varying sample sizes, with the other dimensions fixed to $M = 3$ views, $G = 3$ groups, $D = 1000$ features (per view), and $K = 25$ factors.

We trained a set of models with (deterministic) variational inference (VI) and a set of models with stochastic variational inference (SVI). Overall, we observe that SVI yields Evidence Lower Bounds that are comparable to those obtained from VI across a range of batch sizes, learning rates and forgetting rates (Figure 5.6). In terms of speed, GPU-accelerated SVI inference is up to $\approx$ 20x faster than VI, with speed differences becoming more pronounced with increasing number of cells (Figure 5.7). For completeness, we also compared the convergence time estimates for SVI when using CPU versus GPU. We observe that for large sample sizes there is a speed improvement even when using CPUs, although these advantages become more prominent when using GPUs.

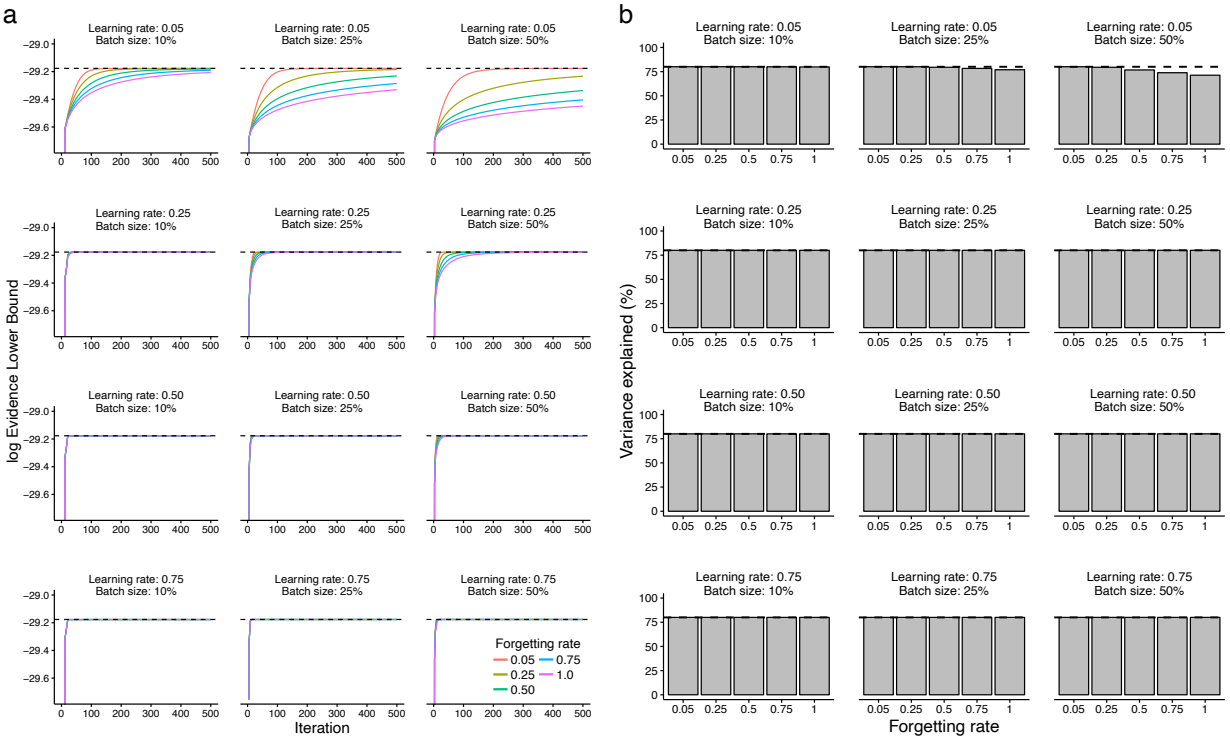


Figure 5.6: Validation of stochastic variational inference using simulated data.
(a) Line plots display the iteration number of the inference (x-axis) and the Evidence Lower Bound on the y-axis. Panels correspond to different values of batch sizes (10%, 25%, 50% of the data) and initial learning rates (0.05, 0.25, 0.5, 0.75). Colours correspond to different forgetting rates (0.05, 0.25, 0.5, 0.75, 1.0). The dashed horizontal line indicates the ELBO achieved using VI.
(b) Bar plots display the forgetting rate (x-axis) and the total variance explained (%) in the y-axis. Panels correspond to different values of batch sizes (10%, 25%, 50% of the data) and initial learning rates (0.05, 0.25, 0.5, 0.75). The dashed line indicates the variance explained achieved using standard VI.

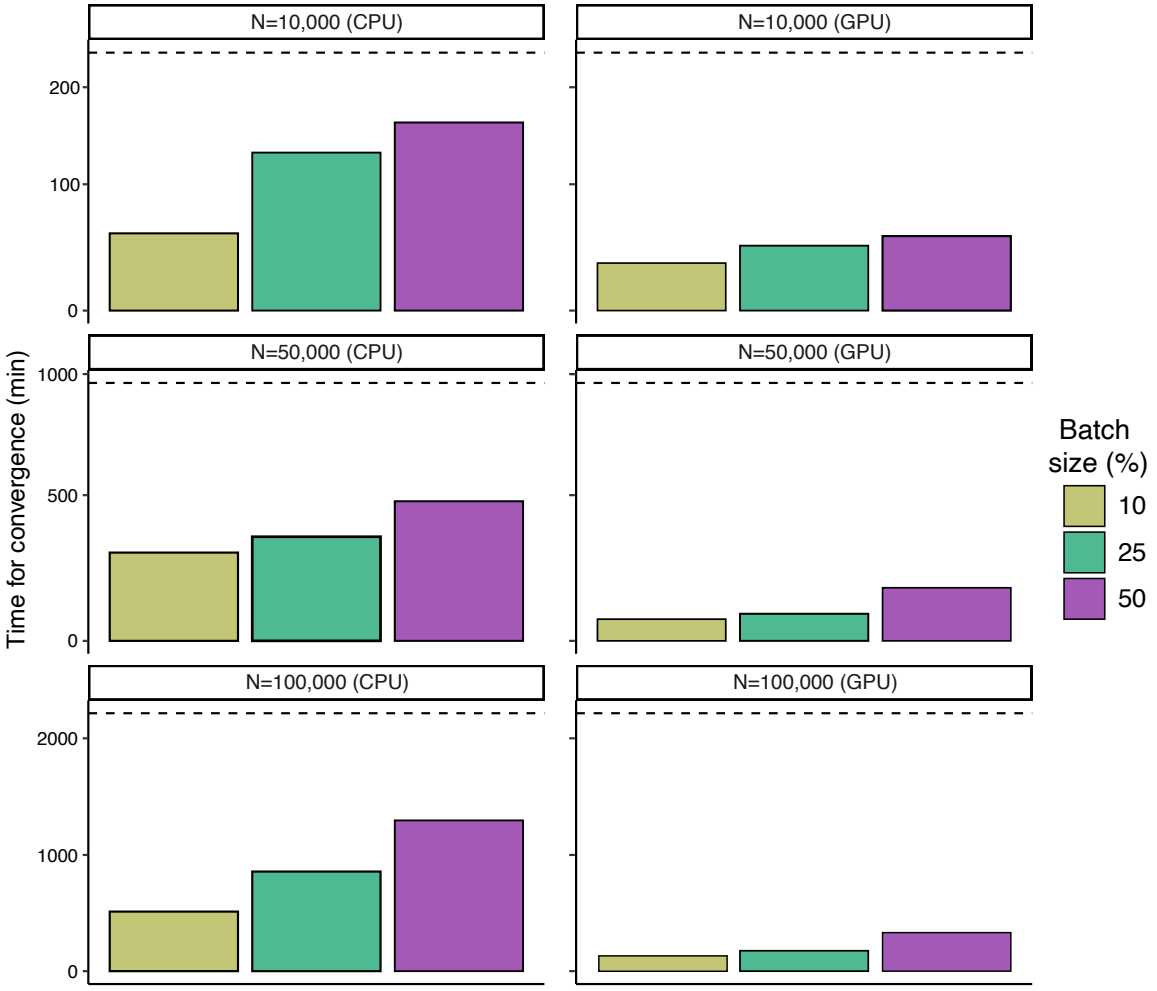


Figure 5.7: Evaluation of convergence speed for stochastic variational inference using simulated data.
Bar plots show the time elapsed for training MOFA+ models usign SVI. Colours represent different batch sizes (10%, 25% or 50%). The dashed line indicates the training time for standard VI. CPU models were trained using a single E5-2680v3. GPU models were trained an Nvidia GTX 1080Ti (second column).

5.3.2 Multi-group inference

We evaluated whether the new model formulation improves the detection of factors with differential activity across groups and views. We simulated data with the following parameters: $M = 2$ views, $G = 2$ groups, $D = 1000$ features, $N = 1000$ samples and $K = 10$ factors. Differential factor activities are incorporated in the simulation process by turning some factors off in random sets of views and groups (Figure 5.8, see ground truth). The task is to recover the true factor activity structure given a random initialisation. We compared three models: Bayesian Factor Analysis (no sparsity priors), MOFA v1 (only view-wise sparsity prior) and MOFA+ (view-wise and group-wise sparsity prior). Indeed, we observe that when having factors that explain varying amounts of variance across groups and across views, MOFA+ was able to more accurately reconstruct the true factor activity patterns (Figure 5.8).

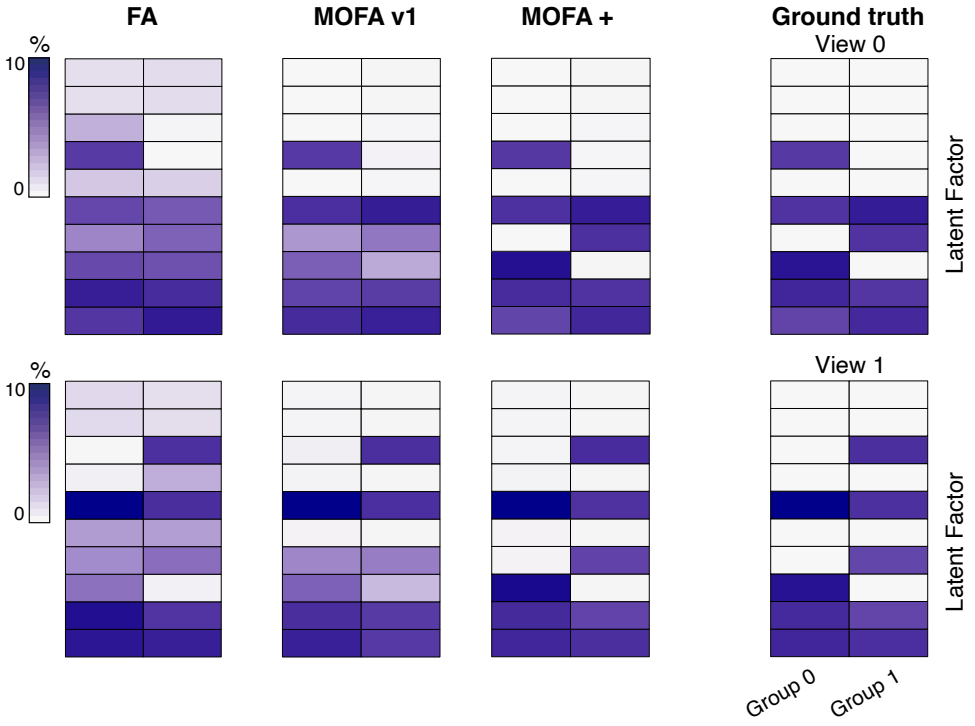


Figure 5.8: Recovering complex factor activity patterns using simulated data. Representative example of the resulting variance explained patterns. The first row of heatmaps correspond to view 0 and the second row to view 1. In each heatmap, the first column corresponds to group 0 and the second column to group 1. Rows correspond to the inferred factors. The colour scale displays the percentage of variance explained by a given factor in a given view and group. The heatmaps displayed in columns one to three show the solutions yielded by different models (Bayesian Factor Analysis; MOFA; MOFA+). The ground truth is shown in the right panel.

5.4 Applications

5.4.1 Integration of a heterogeneous time-course single-cell RNA-seq dataset

To demonstrate the novel multi-group integration framework, we considered a time course scRNA-seq dataset comprising 16,152 cells that were isolated from a total of 8 mouse embryos from developmental stages E6.5, E7.0 and E7.25 (two biological replicates per stage), encompassing post-implantation and early gastrulation [232]. This data set, which has been introduced in Chapter 4, consists of a single view but with a clear group structure where cells belongs to different biological replicates at different developmental time points. Different embryos are expected to contain similar subpopulations of cells but also some differences due to developmental progression. As a proof of principle, we used MOFA+ to disentangle stage-specific transcriptional signatures from signatures that are shared across all stages. Although in principle one could employ the MOFA+ factors for clustering and cell type annotation, here we adopted the cell type definitions described in [232].

MOFA+ identified 7 Factors that explained at least 1% of variance in a group and all together captured between 35% and 55% of the total transcriptional heterogeneity per embryo (Figure 5.9).

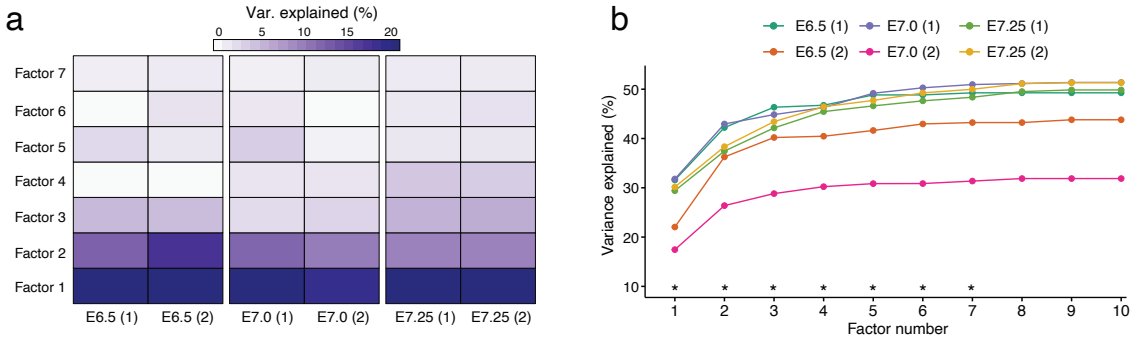


Figure 5.9: MOFA+ applied to gastrulation scRNA-seq atlas: variance decomposition analysis
(a) Heatmap displays the variance explained (%) for each factor (rows) in each group (mouse embryos at a specific developmental stage, columns).
(b) Cumulative variance explained (per group, y-axis) versus factor number (x-axis). Asterisks indicate the factors that are selected for downstream analysis (minimum of 1% variance explained in at least one group).

Characterisation of individual factors

Some factors recover the existence of post-implantation developmental cell types, including extra-embryonic (ExE) tissue (Factor 1 and Factor 2), and the emergence of mesoderm cells from the primitive streak (Factor 4). Consistently, the top weights for these factors are enriched for lineage-specific gene expression markers, including *Ttr* and *Apoa1* for ExE endoderm (Figure 5.10); *Rhox5* and *Bex3* for ExE ectoderm (Figure 5.11); *Mesp1* and *Phlda2* for nascent mesoderm (Figure 5.13). Other factors captured technical variation due to metabolic stress that affects all batches in a similar fashion (Factor 3, Figure 5.12).

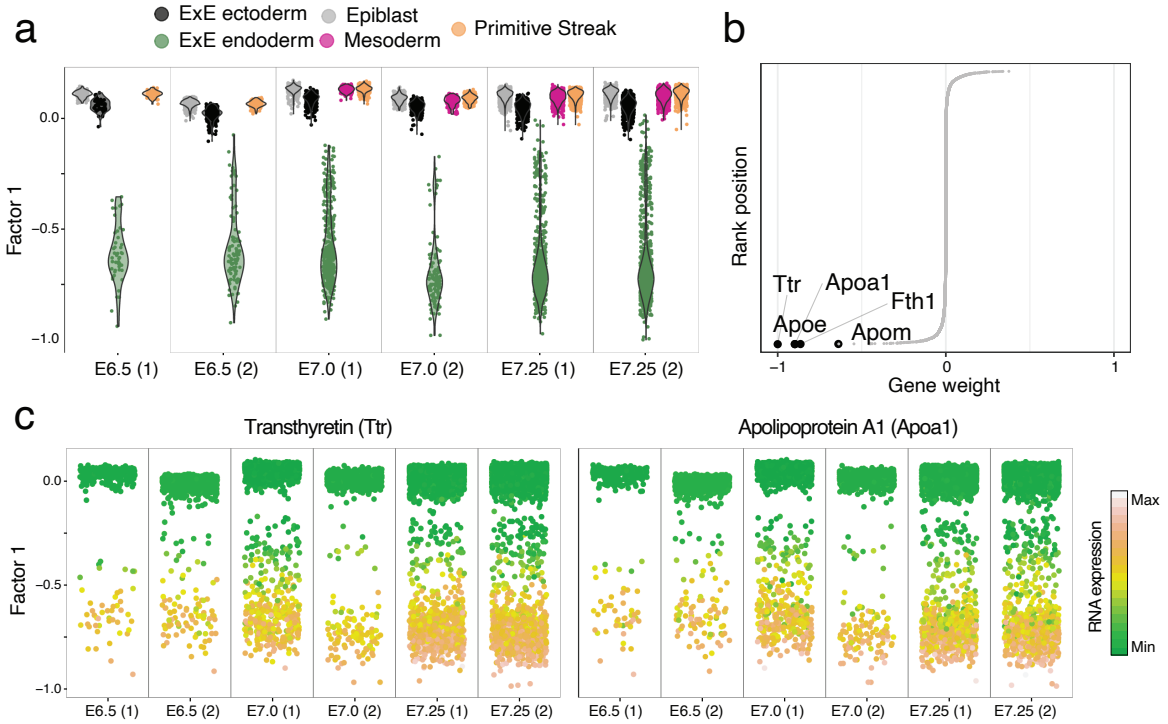


Figure 5.10: Characterisation of Factor 1 as extra-embryonic (ExE) endoderm formation.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type. (b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values) (c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

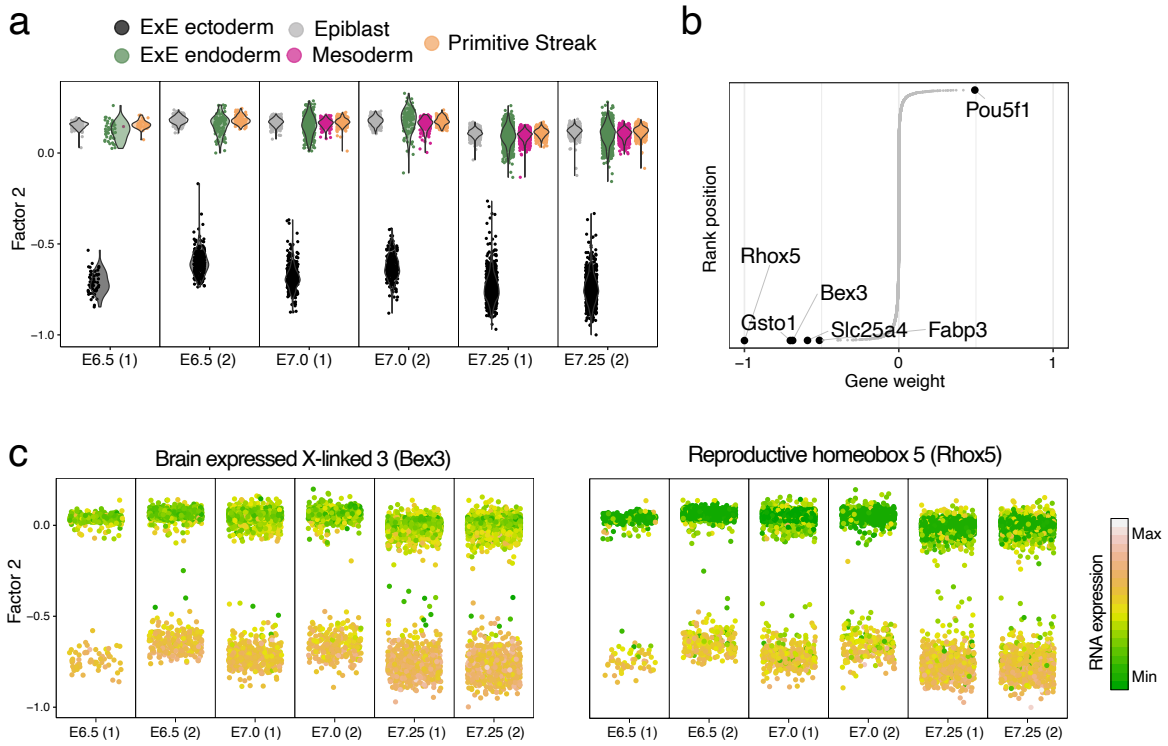


Figure 5.11: Characterisation of Factor 2 as extra-embryonic (ExE) ectoderm formation.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

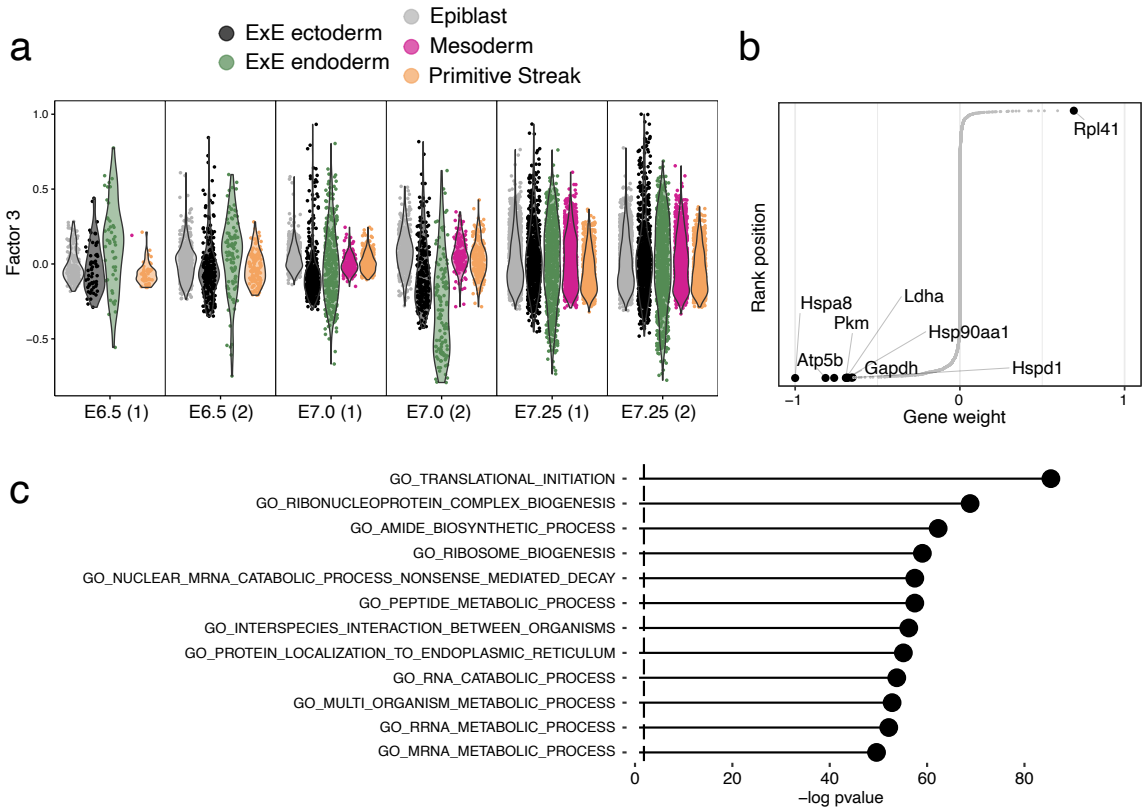


Figure 5.12: Characterisation of Factor 3 as cell-to-cell differences in metabolic activity.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top seven genes with largest weight (in absolute values)
(c) Gene set enrichment analysis applied to the gene weights using the Reactome gene sets [87]. Significance is assessed via a parametric. Resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

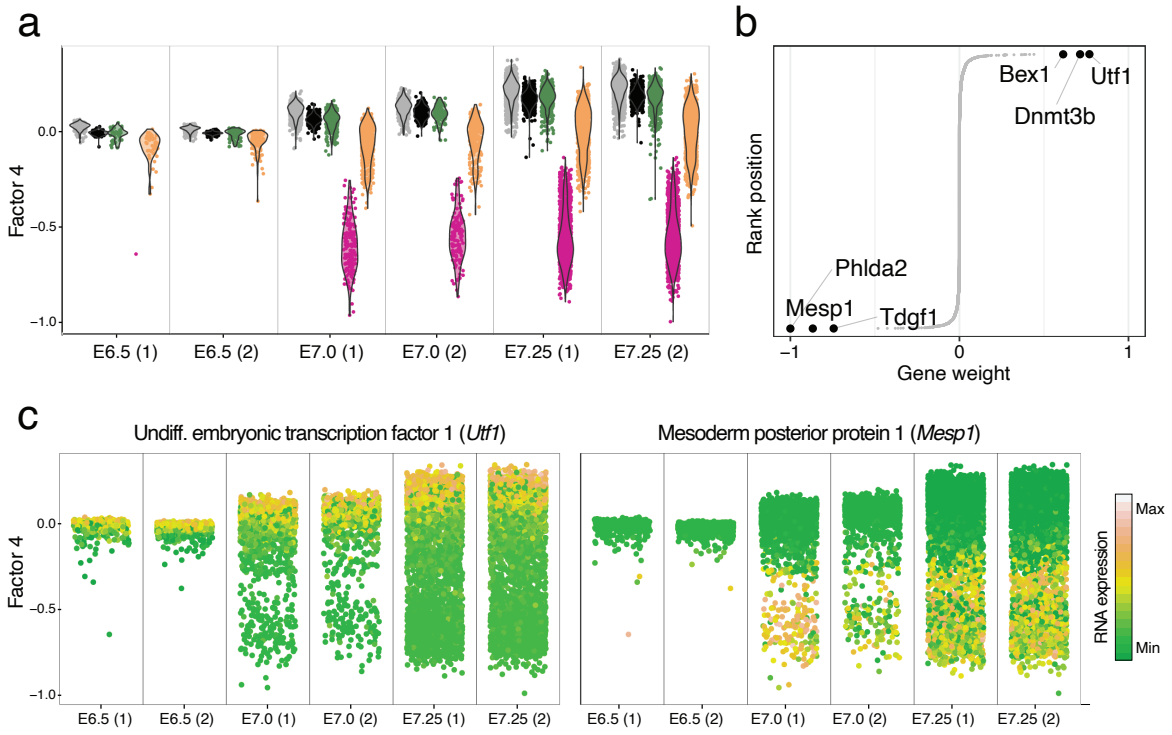


Figure 5.13: Characterisation of Factor 4 as mesoderm commitment.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

Interestingly, Factors displayed different signatures of activity (variance explained values) across developmental stages. For example, Factors 1 and 2 remain constant across the developmental progression (Figure 5.9), indicating that commitment to ExE ectoderm and ExE endoderm fate occurs early in the embryo and the proportion of this cell type remains relatively constant. In contrast, the activity of Factor 4 increases with developmental progression, consistent with a higher proportion of cells committing to mesoderm after ingression through the primitive streak.

In conclusion, this application shows how MOFA+ can identify biologically relevant structure in multi-group scRNA-seq datasets.

5.4.2 Identification of molecular signatures of lineage commitment during mammalian embryogenesis

As a second application, I considered the multi-omic atlas of mouse gastrulation introduced in Chapter 4, where scNMT-seq was used to profile RNA expression, DNA methylation and chromatin accessibility in 1,828 cells at multiple stages of development [14]. The main difference with respect to the MOFA analysis presented in Chapter 4 (Section 4.2.5) is that MOFA+ can employ the multi-group functionality to perform a simultaneous analysis across multiple stages, instead of focusing only on stage E7.5.

Data processing

As input to the model we quantified DNA methylation and chromatin accessibility values over two sets of regulatory elements: gene promoters and enhancer elements (distal H3K27ac sites). RNA expression was quantified over protein-coding genes. More details on the feature quantification and data processing steps are described in Chapter 4. We defined separate views for the RNA expression and for each combination of genomic context and epigenetic readout. Cells were grouped according to their developmental stage (E5.5, E6.5 and E7.5), reflecting the underlying experimental design [14] (Figure 5.14). Note that the CpG methylation (endogenous DNA methylation) and GpC methylation (proxy for chromatin accessibility) readouts result in very sparse matrices that are challenging to analyse with standard statistical methods.

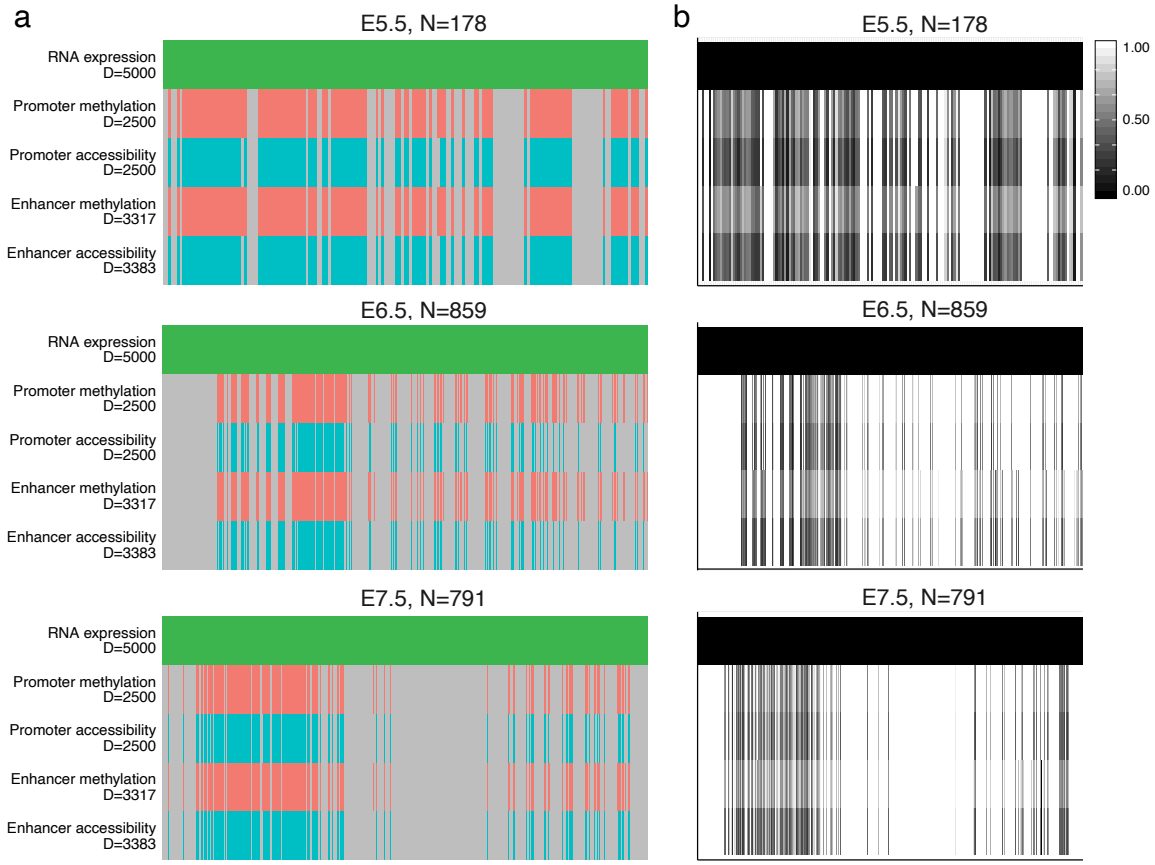


Figure 5.14: Overview of the scNMT-seq mouse gastrulation data set used as input for MOFA+.
(a) Structure of the input data in terms of views (y-axis) versus samples (x-axis). Each panel corresponds to a different group (embryonic stage). Grey bars represent missing views.
(b) Structure of the missing values in the data. For each cell and view, the colour displays the fraction of missing values.

Model overview

In this data set MOFA+ identified 8 largely orthogonal factors with a minimum variance explained of 1% in the RNA expression (in at least one group, Figure 5.15). The model explains little amounts

of variance in chromatin accessibility, both for promoters (≈15%) and enhancers (≈18%), mostly driven by Factors 1 and 2. In contrast, the model explains larger amounts of variation in DNA methylation (≈23% for promoters and ≈59% for enhancers). However, as in chromatin accessibility, this variation is mostly driven by the first two Factors. Finally, for RNA expression there is a steady increase in the variance explained per Factor, suggesting that the (small) sources of variation captured beyond Factor 2 are largely driven by RNA expression alone.

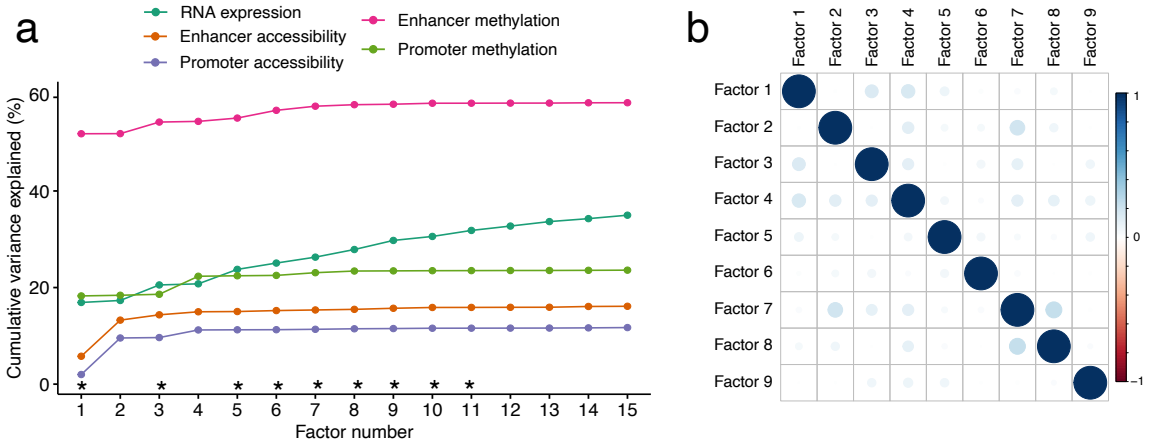


Figure 5.15: MOFA+ application to the scNMT-seq gastrulation data set: model overview.
(a) Cumulative variance explained (per view, y-axis) versus factor number (x-axis). Asterisks indicate the factors that are selected for downstream analysis (minimum of 1% variance explained in the RNA expression). Note that the variance estimates shown here are the sum across all groups. (b) Pearson correlation coefficients between selected factors. In MOFA+ there are no orthogonality constraints, but the factors are expected to be largely uncorrelated.

Characterisation of the MOFA+ Factors

Factor 1 captured the formation of ExE endoderm, a cell type that is present across all stages (Figure 5.16a), in agreement with our previous results using the independently generated scRNA-seq atlas of mouse gastrulation (Figure 5.10). This Factor is associated to widespread changes across al molecular layers, most notably DNA methylation (up to 15% variance explained). For both promoters and enhances, the distribution of weights for DNA methylation are skewed towards negative values. This suggests that most features are uniformly affected by this Factor, such that lower methylation levels are observed in ExE endoderm cells. This is consistent with previous studies that have shown that ExE endoderm cells are characterised by a state of global demethylation[336, 14]. The weights for chromatin accessibility are not skewed towards one direction, indicating that accessibility changes are not uniform and the state of global demethylation in ExE endoderm cells is not necessarily associated with a (globally) more open chromatin state.

The next two factors, Factor 2 and Factor 3, captured the molecular variation associated with the formation of the primary germ layers at E7.5: mesoderm (Factor 2, Figure 5.16b), and embryonic endoderm (Factor 3, not shown). As with Factor 1, MOFA+ connects transcriptome variation to changes in DNA methylation and chromatin accessibility, but only at stage E7.5, when the germ layers are known to emerge. Nevertheless, there is a striking difference between Factor 1 and Factor

2. The variance decomposition analysis and the distribution of weights indicate that the epigenetic dynamics are mostly driven by enhancer elements. This is consistent with our results from Chapter 4 where we showed that little coordinated variation is observed in promoters, even for genes that show strong differential expression between germ layers.

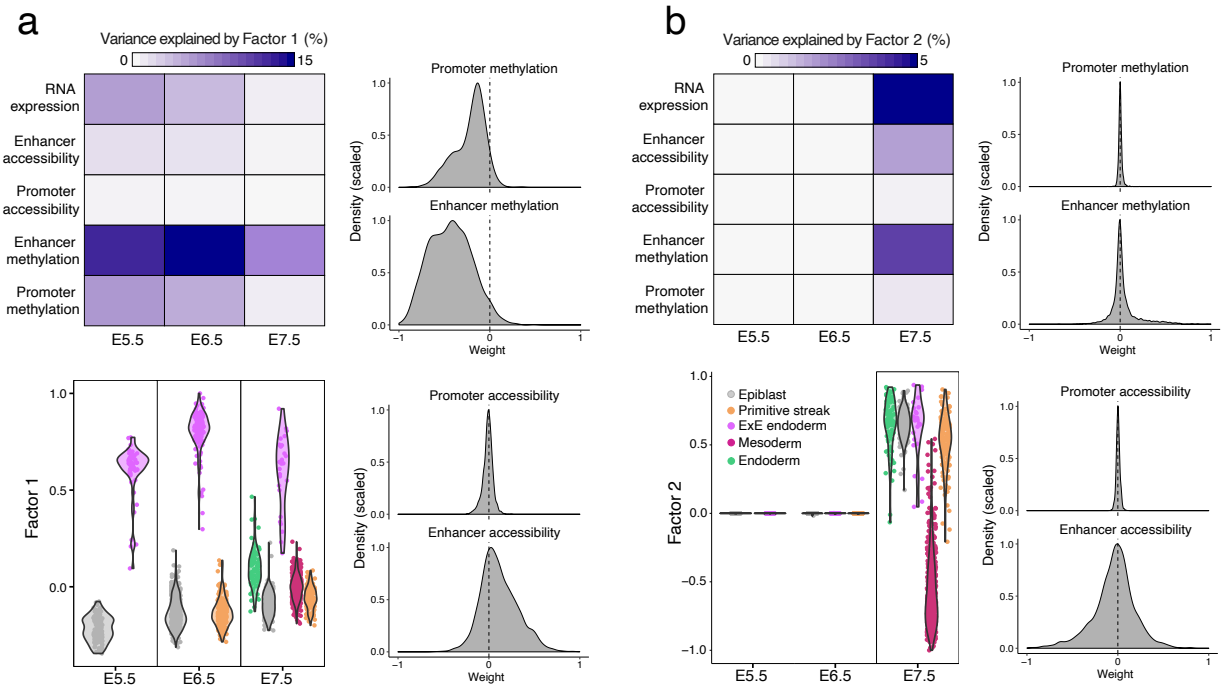


Figure 5.16: MOFA+ integrates multi-modal scNMT-seq experiments to reveal epigenetic signatures associated with lineage commitment during mammalian embryogenesis.
Characterisation of (a) Factor 1 as ExE endoderm formation and (b) Factor 2 as Mesoderm commitment. Top left plot shows the percentage of variance explained by the factor across the different views (rows) and groups (embryonic stages, as columns). Bottom left plot shows the distribution of factor values for each stage, coloured by cell type assignment. Histograms display the distribution of DNA methylation and chromatin accessibility weights for promoters and enhancer elements.

As suggested by the variance decomposition analysis (Figure 5.15), the remaining MOFA+ Factors explain significantly less variance than Factors 1 and 2, and they are mostly driven by RNA expression alone (Figure 5.15). Their aetiology can be identified by the inspection of gene weights and by gene set enrichment analysis. For simplicity, I will only display the characterisation of Factor 6, which captures cell-cycle variation that is consistently found across all three embryonic stages (Figure 5.17).

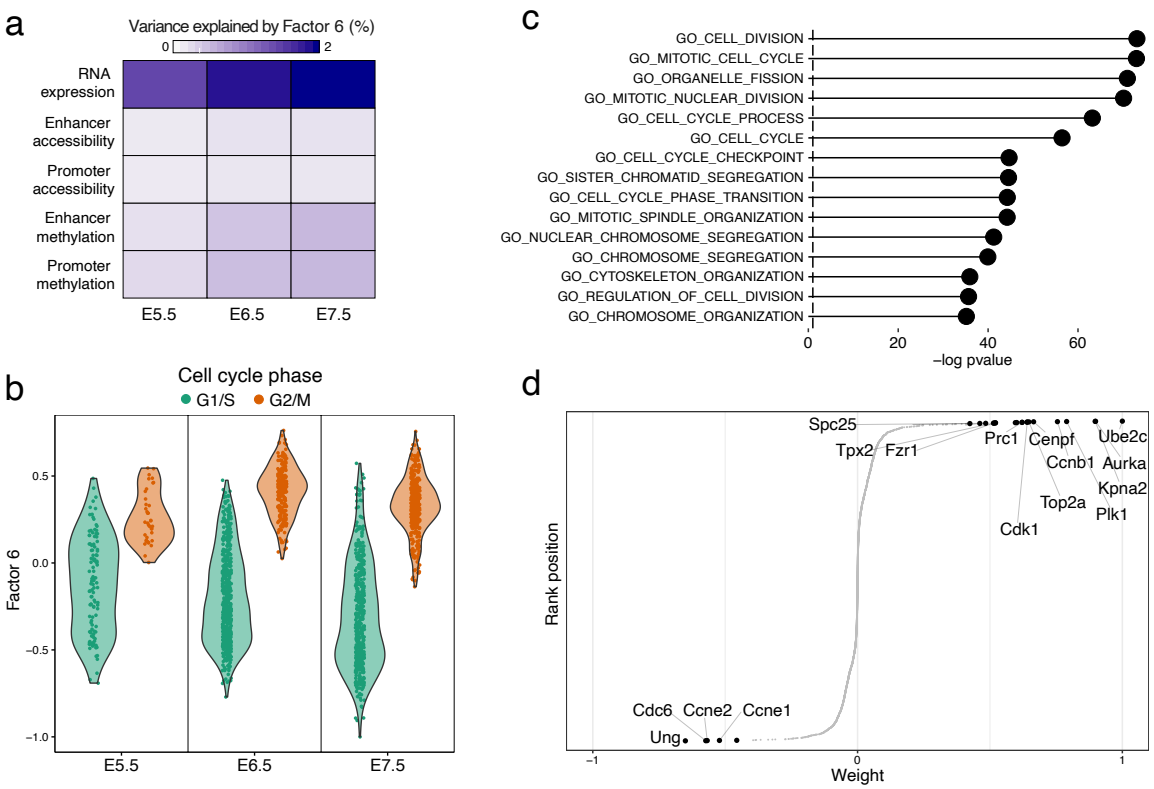


Figure 5.17: Characterisation of Factor 6 as cell cycle variation.
(a) Variance explained by Factor 6 in each group (embryonic stage, columns) and view (rows).
(b) Distribution of Factor 6 values per group (embryonic stage, x-axis), with cells coloured by the inferred cell cycle state using *cyclone*.
(c) Gene set enrichment analysis applied to the Factor 6 weights.
(d) Cumulative distribution of RNA weights for Factor 6. The top genes with the highest (absolute) weight are labelled.

5.5 Conclusions, limitations and open perspectives

In this Chapter I introduced a generalisation of the MOFA model for the principled analysis of large-scale structured datasets. Although we have emphasised single-cell applications, the model remains applicable to bulk datasets. MOFA+ solves some of the limitations of the MOFA model presented in Chapter 3, but a significant number of challenges remain unsolved and could be addressed in future research:

- **Linearity**: this is arguably the major limitation of MOFA. Although it is critical for obtaining interpretable feature weights, this results in a significant loss of explanatory power. Deep generative models have proven successful in modelling complex observations. Their principle is the use of non-linear maps via neural networks to encode the parameters of probability distributions. Among this class of methods, variational autoencoders provide a rigorous and scalable non-linear generalisation of factor models [3].
- **Improving the stochastic inference scheme**: a common extension of stochastic gradient descent is the addition of a *momentum* term, which has been widely adopted in the training of artificial neural networks [334, 239]. The idea is to take account of past updates when calculating the present step, using for example a moving average calculation. This has been shown to improve the stability of gradients vectors, thus leading to a faster convergence.
- **Modelling dependencies between groups**: often groups are not independent and have some type of structure among themselves. A clear example are time course experiments. Explicit modelling of these dependencies, when known, could help on model inference and interpretation.
- **Modelling continuous dependencies between samples and/or features**: in the MOFA framework the views and the groups correspond to discrete and non-overlapping sets. An interesting improvement would be to model continuous dependencies using Gaussian Process priors [53]. A clear application for this is spatial transcriptomics, where one could build a covariance matrix using spatial distances which can then be imposed in the prior distribution of the latent factors (recall that in MOFA and MOFA+ the prior distribution for the factors assumes independence between samples). This would improve the detection of sources of variation with a spatial component.

Chapter 6

Concluding remarks

The last few years have seen an explosion of single-cell sequencing technologies, which have provided new directions and opportunities for the study of biological complexity. The ultimate goal of single-cell sequencing is to move from descriptive snapshots to comprehensive multi-modal roadmaps of biological processes mapped across time and space. Unifying molecular variation across these two will be at the forefront of scientific research.

The first stones on this path have been laid. Experimental designs that include single-cell genomics technologies have now become ubiquitious, and the computational pipelines are gradually becoming standarised. In addition, multi-modal measurements have been successfully collected in single cells, although this remains in pilot stages and at the time of this writing very few commercial platforms are available, thus limiting its widespread use by the community.

In this Chapter we discuss current and future perspectives of experimental and computational methods that will lay the foundations for an unprecedented and exciting era for studying biological complexity using single-cell multi-omics.

6.1 Experimental perspectives

6.1.1 Recording space

The dissociation and pooling of cells from their native location, with the consequent loss of information of their spatial coordinates, is one of the biggest limitations of current single-cell technologies. The positioning and the interaction of cells in their native tissue is essential to understand biological function, as complex tissues arise following organized events in space and time [193]. Pioneering work in single-molecule fluorescence *in situ* hybridization demonstrated that mRNA molecules can be measured, but quantifying multiple genes at the same time has been a major challenge, mostly hindered by physical limitations on the optical resolution and the high density of transcripts within each cell [83]. Remarkbly, recent technological breakthroughts have scaled these assays to transcriptome-wide measurements while maintaining high accuracy. One of these methods, seqFISH [182, 83], employs a multiplexed strategy to overcome the diffraction limit where multiple rounds of sequential probe hybridization and imaging are applied. Although most of the spatially-resolved methods have been focused on transcriptomics, epigenomic measurements have also been successfully recorded by adapting the ATAC-seq protocol [301], thus paving the way for future groundbreaking multi-modal spatially-resolved assays.

6.1.2 Recording time

Recording the past

In the timespan of a few days, a mammalian embryo expands from a handful of cells to millions of them, experiencing in the process a set of cell fate commitments that will eventually generate a myriad of specialised cell types. Besides knowing the spatial location of cells, recording the timing and the hierarchy of the events at high-resolution is an essential step to understand complex biological dynamics. Static single-cell experiments can provide useful snapshots that can be used to reconstruct dynamic differentiation processes [323], but the past story of the cells remains elusive.

A fundamental biological principle is that each cell originates from another existing cell. Thus, each adult cell has an associated cell lineage tree that unambiguously defines its past history and can potentially be recorded by tracking the progeny of single cells. Historically, imaging-based techniques have been used to perform lineage tracing in a low-throughput manner by employing fluorescent protein markers. However, classical fluorescence-based experiments are limited when it comes to both temporal and molecular resolution. The current generation of lineage tracing techniques introduce inheritable genetic marks that can be read in terminal cells through next-generation sequencing [24, 141, 200]. For example, one can induce double-stranded breaks using CRISPR-Cas9 at target genomic sites, which after repair results in random genomic insertions and deletions (indels) that become inherited in different hierarchical combinations by the progeny cells [24, 141, 200]. Notably, this strategy can be combined with single-cell sequencing for the simultaneous quantification of clonal history (applying phylogenetic methods on the indel profiles) and cell type identity (applying clustering methods on the gene expression profiles). This strategy was employed to map cell fates in several model organisms, including adult zebrafish [5] and mouse embryos [54].

One of the most exciting opportunities that will soon become a reality is combining lineage tracing technologies with single-cell multi-modal readouts. An ideal system would measure multiple biological layers *in situ* and would recover each cell's past history at the same time [200]. In the context of this thesis, the genomic modality of scNMT-seq could be employed to track not only CRISPR-Cas9-induced indels but also naturally occurring mitochondrial mutations. The latter in particular would be critical to identify clones in complex human tissues, where genetic manipulation is not an option [183].

In conclusion, multi-modal lineage tracing would be a major step towards the ultimate goal of constructing dynamic models of cell fate commitment and to understand how they become dysregulated in disease. Notably, each of these tools is already available, but combining them in coassays will be one of the big challenges ahead.

Recording the future

scRNA-seq offers transcriptome-wide snapshots of the present status of each cell. However, simultaneously measuring some property of its future state would provide extremely valuable information to understand cell fate decisions.

A milestone of scRNA-seq analysis was achieved with the inference of RNA velocity in single cells [160]. This method leverages the small quantity of intronic reads that are present in the sequencing library to calculate a relative ratio of unspliced (intronic) and spliced (exonic) mRNAs to infer gene expression kinetics. In turn, by pooling information across multiple genes and under some assumptions one can estimate the nascent transcriptional state for each cell. Interestingly, similar information can be obtained from spatial transcriptomics data. However, instead of identifying intronic reads, which are lowly abundant due to polyA selection, one can exploit the subcellular resolution of the molecules to distinguish between nuclear (unspliced) and cytoplasmic (spliced) reads [330].

RNA velocity is appealing because it can be applied to most conventional scRNA-seq protocols. However, intronic coverage is sparse, particularly for droplet-based assays, thus limiting the reliability of the method in some datasets [278]. Moreover, the inference of gene expression kinetics is restricted to genes that undergo splicing events. Alternative strategies to record the future state of individual cells are aimed at directly monitoring newly synthesised RNA. For example, NASC-seq [115] relies on chemical modifications of the nascent RNA that can be read out by sequencing at much greater sensitivity and better temporal resolution than RNA velocity estimates.

6.2 Computational perspectives

None of the biological insights offered by multi-modal assays would be possible without concomitant development of computational methods. Each new data modality presents distinct challenges, from low level processing, quality control and normalisation through to downstream challenges such as quantifying sources of biological variability and using these to generate testable biological hypotheses. Additionally, in the context of single-cell multi-omics datasets there exist specific challenges that need to be overcome, as discussed in Section 1.3.2. Here I outline some challenges that need to be addressed in order to optimally leverage the power of single-cell multi-omics.

6.2.1 Mechanistic insights

One of the most promising aspects of multi-modal sequencing is the opportunity to move from descriptive snapshot to a more mechanistic understanding of gene regulation. By incorporating prior knowledge about the hierarchical relationship between molecular layers, we envision that multi-modal assays will play an important role in identifying causal chains of events in gene regulatory networks. However, to construct such mechanistic models it will be essential to combine multi-modal readouts with perturbations assays.

Single-cell perturbation studies using the CRISPR technology is one of most exciting experiments that single-cell genomics has brougth us [78, 71, 128, 4]. In the first step, cells are infected with a pool of lentiviral constructs that contain guide RNAs that target (typically by inactivation) specific genes. Notably, increasing the multiplicity of infection can be used to target multiple genes at once and thus study epistatic effects. After stimulation or differentiation, cells are sequenced

using single-cell methodologies. The unique combination of barcodes within each cell enables the computational identification of the gene(s) targeted. Remarkably, this strategy of pooling guides and cells together and later deconvoluting them enables this protocol to be done in a massively parallel fashion.

While most of the perturbation studies have been focused on RNA expression, the profiling of epigenomic layers is receiving increased interest [259]. Yet, to my knowledge, no multi-modal single-cell CRISPR screening has performed to date. This is a matter of time, as all the ingredients are already available, and I envision that this will become the state-of-the-art for the characterisation of gene regulatory networks.

6.2.2 Benchmarking of methods

Benchmarking of methods has been extensively performed on *horizontal* data integration strategies, particularly in the context of batch correction [185, 309]. However, benchmarking *vertical* and *diagonal* integration strategies is notorously difficult, as the ground truth is rarely known. In the context of MOFA, for example, it is difficult to assess the quality of the output. There are useful quality control metrics, such as the number of factors or the total variance explained, but it is not clear how to assess whether the latent Factors are a reliable representation of biological variation or whether they are just arbitrary (but useful) mathematical representations.

In this context, having gold-standard truth datasets will prove essential to benchmark integrative methods. Here, we discuss two existing biological systems that are well suited to benchmark data integration tasks. The first one is Peripheral Blood Mononuclear Cells (PBMCs), which is the *de facto* dataset to validate single-cell technologies developed by 10x Genomics, owing to its simplicity and well-characterised subpopulations of cells. Multiple assays have been profiled on PBMCs across multiple human donors and different species, including scRNA-seq, scATAC-seq, CITE-seq and [T/B]-cell receptor sequencing. Moreover, some horizontal and vertical integration strategies have already been successfully applied [291].

The second biological system is mammalian embryonic development, a significantly more complex system with branching differentiation trajectories and where the regulation between molecular layers is less well understood when compared to somatic cell types. In addition, unlike PBMCs, the solid tissue enables the integration of molecular readouts with spatial context information. A large variety of single-cell technologies have been applied to mouse embryonic development, including scRNA-seq [232, 207, 270, 100, 220], scATAC-seq [233], scNMT-seq [14] and even spatial measurements [229, 37].

As the Human Cell Atlas project [249] matures, we expect many more biological systems to be suitable datasets for benchmarking data integration strategies, not only across data modalities but also across individuals and even across different species.

6.2.3 Mosaic integration

A major challenge in the next few years will be to integrate independent experiments with the goal of building self-consistent multi-modal datasets of biological processes. However, given how difficult it is to simultaneously capture multiple molecular layers in an efficient and scalable manner, this task will require computational integration of independent uni-modal and multi-modal experiments from the same biological system. There is an urgent need to develop a unifying integrative strategy that selectively exploits cells and features as common coordinate frameworks to perform transfer learning of molecular signatures across experiments (Figure 6.1). I coin the term *mosaic* integration for this combination of *vertical*, *horizontal* and *diagonal* integration tasks. Undeniably, this task will not be solved by linear matrix factorisation frameworks such as MOFA, and it will require non-linear strategies probably in the form of multi-view variational autoencoders [180].

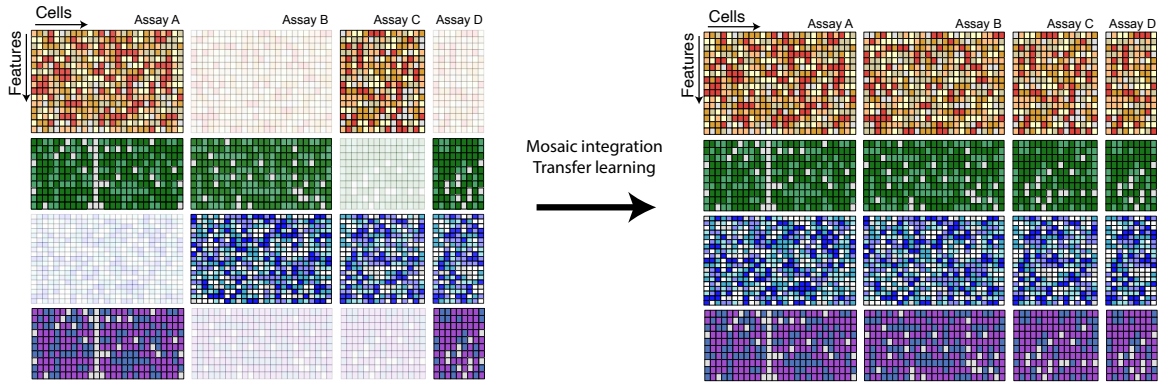


Figure 6.1: Mosaic integration The aim is to integrate a complex experimental design that consists of uni-modal and multi-modal datasets by selectively exploiting cells and features as common coordinate frameworks. The output would be a self-consistent data set where all missing data modalities have been imputed across all experimental conditions.

6.2.4 Software infrastructure

Open-source software, data sharing platforms and reproducible analysis pipelines are essential elements in computational biology. Generic frameworks that can contain increasingly large complex experimental designs from single-cell genomics are urgently needed. Popular tools for scRNA-seq analysis such as *SingleCellExperiment*, *Seurat* and *Scanpy* are continuously being extended to handle novel multi-modal assays. There are however important challenges that need to be overcome.

The first one is data standarisation and interoperabiliy between platforms. Conversion between R-based objects is relatively easy, but connecting Python and R is still a significant challenge that involves tedious configurations [6].

Second, the large scale of single-cell genomics requires optimisation of memory usage. Just to give a sense of how important this is, one of datasets released by 10x Genomics contains transcriptome-wide measurements for 1.3 million cells from the mouse brain [1]. Storing this data set in an ordinary

[1] https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons

integer matrix requires more than 100 GB of memory. Clearly, the conventional approach of loading the entire data set to memory on a conventional laptop becomes prohibitive. One of the most efficient approaches to handle vast amounts of data is to use on-disk operations, where common array operations are performed using a block processing mechanism, thus preventing the entire object from being loaded into memory at once. This is a strategy that we implemented for the downstream analysis in MOFA+ by adapting the *DelayedArray* framework [116], albeit the matrix operations during model training require loading the entire data set in memory.

Finally, we need a centralised data sharing platform where curated landmark datasets are made available with common data structures alongside reproducible analysis vignettes. This would be of immense help to both beginners and method developers alike to cope with the vast amounts of invaluable data that single-cell genomics promises to generate in the future.

6.3 Thesis summmary

In this Thesis I have described the work I performed throughout my PhD where I sought to develop and apply computational strategies for data integration in the context of single-cell multi-omics.

First, I worked together with experimental collaborators to devise scNMT-seq, an experimental protocol for the genome-wide profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. I developed the entire computational analysis pipeline and after validating the quality of the readouts I demonstrated how scNMT-seq can be used to study coordinated epigenetic and transcriptomic heterogeneity along a simple differentiation process.

Second, motivated by the need to discover biologically meaningful insights from such complex data I developed Multi-Omics Factor Analysis (MOFA), a statistical framework for the integration of multi-omics datasets. Briefly, MOFA is a statistically rigorous generalisation of Principal Component Analysis for multi-view data. It provides a systematic approach to explore, in an unsupervised manner, the underlying sources of sample heterogeneity in a multi-omics data set. Owing to its linear formulation, interpretability is an essential property of this model that permits a series of useful downstream analyses. Before applying MOFA to single-cell datasets, we benchmarked it using a large multi-omics cohort of chronic lymphocytic leukaemia patients and demonstrated how MOFA can be used to capture multiple dimensions of disease heterogeneity, enhance data interpretation and build predictive models for clinical outcomes.

Third, I aimed to leverage the experimental and computational frameworks described above to study embryonic development and specifically germ layer commitment. Together with experimental collaborators we employed scNMT-seq to simultaneously profile mRNA expression, DNA methylation and chromatin accessibility for more than 1000 cells, spanning four time points between the exit from pluripotency and primary germ layer specification. This data set represents the first multi-omics roadmap of mouse gastrulation at single-cell resolution, which enabled us to perform an integrative study that revealed novel insights into the dynamics of the epigenome during gastrulation. Notably, we show that cells committed to mesoderm and endoderm undergo widespread epigenetic rearrangements, driven by demethylation in enhancer marks and by concerted changes in chromatin accessibility. In contrast, the epigenetic landscape of ectodermal cells remains in a *default* state, resembling earlier stage epiblast cells. This work provides a comprehensive insight into the molecular logic for a hierarchical emergence of the primary germ layers, revealing underlying molecular constituents of Waddington's landscape.

Finally, after having benchmarked MOFA using high-quality bulk multi-omics and relatively small single-cell genomics dat asets, I developed a second version of the software aimed at the scalable analysis of datasets with thousands of samples and more complex experimental designs. Key methodological improvements included a fast stochastic variational inference framework and a flexible structure of the prior distributions that enable integration of multiple groups of samples.

Appendix A

Mathematical derivations of MOFA+

A.1 Deriving the variational inference algorithm

The theoretical foundations for the variational inference scheme are described in Section 3.1.4. Just to brief, we need to define a variational distribution of a factorised form and subsequently look for the member of this family that most closely resembles the true posterior using the KL divergence as a *distance* metric. Following the mean-field principle, in MOFA+ we factorised the variational distribution as follows:

$$
\begin{aligned}
q(\mathbf{X}) &= q\left(\{\widehat{\mathbf{Z}^{\mathbf{g}}}, \mathbf{S}^{\mathbf{g}}, \alpha^{\mathbf{g}}, \theta^{\mathbf{g}}\}, \{\widehat{\mathbf{W}^{\mathbf{m}}}, \mathbf{S}^{\mathbf{m}}, \alpha^{\mathbf{m}}, \theta^{\mathbf{m}}\}, \{\tau^{\mathbf{gm}}\}\right) \\
&= \prod_{g=1}^{G}\prod_{n=1}^{N_g}\prod_{k=1}^{K} q(\hat{z}_{nk}^{g}, s_{nk}^{g}) \prod_{g=1}^{G}\prod_{k=1}^{K} q(\alpha_k^g) \prod_{g=1}^{G}\prod_{k=1}^{K} q(\theta_k^g) \\
&\times \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{kd}^{m}, s_{kd}^{m}) \prod_{m=1}^{M}\prod_{k=1}^{K} q(\alpha_k^m) \prod_{m=1}^{M}\prod_{k=1}^{K} q(\theta_k^m) \\
&\times \prod_{g=1}^{G}\prod_{m=1}^{M}\prod_{d=1}^{D_m} q(\tau_d^{gm})
\end{aligned}
\tag{A.1}
$$

However, inspired by [304], we did not adopt a fully factorised distribution as $\hat{w}_k^m$ and s_k^m can hardly be assumed to be independent.

To derive the variational updates we can proceed in two ways, as described in Section 3.1.4. One option is to use exploit the mean-field assumption and use calculus of variations to find the optimal distribution $q(\mathbf{X})$ that maximises the lower bound $\mathcal{L}(\mathbf{X})$[31, 214]. The alternative and possibly easier approach is to define a parametric form for the distribution $q(\mathbf{X})$ with some parameters $\mathbf{\Theta}$ to be of the same form as the corresponding prior distribution $p(\mathbf{X})$. Then, one can find the gradients with respect to the parameters to obtain the coordinate ascent optimisation scheme. In our derivations we followed the first approach, but because we used conjugate priors the second one should converge to the same result.
Below we give the explicit update equations for every hidden variable in the MOFA+ model which are applied at each iteration of the variational inference algorithm.

A.2 Variational update equations

Factors For every group g, sample n and factor k:

Prior distribution $p(\hat{z}^g_{nk}, s^g_{nk})$:

$$p(\hat{z}^g_{nk}, s^g_{nk}) = \mathcal{N}(\hat{z}^g_{nk} \,|\, 0,\ 1/\alpha^g_k)\,\mathrm{Ber}(s^g_{nk} \,|\, \theta^g_k) \tag{A.2}$$

Variational distribution $q(\hat{z}^g_{nk}, s^g_{nk})$:

Update for $q(s^g_{nk})$:

$$q(s^g_{nk}) = \mathrm{Ber}(s^g_{nk}|\gamma^g_{nk}) \tag{A.3}$$

with

$$\begin{aligned}
\gamma^g_{nk} &= \frac{1}{1+\exp(-\lambda^g_{nk})} \\
\lambda^g_{nk} &= \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle} - 0.5 \ln \left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle} \right) \\
&+ \frac{\langle \tau^{gm}_d \rangle}{2} \frac{\left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} y^{gm}_{nd} \langle w^m_{kd} \rangle - \sum_{j \neq k} \langle s^g_{nj} \hat{z}^g_{nj} \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w^m_{kd} \rangle \langle w^m_{jd} \rangle \right)^2}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}}
\end{aligned} \tag{A.4}$$

Update for $q(\hat{z}^g_{nk})$:

$$\begin{aligned}
q(\hat{z}^g_{nk}|s^g_{nk} = 0) &= \mathcal{N}\left(\hat{z}^g_{nk} \middle| 0, 1/\alpha^g_k\right) \\
q(\hat{z}^g_{nk}|s^g_{nk} = 1) &= \mathcal{N}\left(\hat{z}^g_{nk} \middle| \mu_{z^g_{nk}}, \sigma^2_{z^g_{nk}}\right)
\end{aligned} \tag{A.5}$$

with

$$\begin{aligned}
\mu_{z^g_{nk}} &= \frac{\sum_{m=1}^{M} \sum_{d=1}^{D_m} y^{m,g}_{nd} \langle w^m_{kd} \rangle - \sum_{j \neq k} \langle s^g_{nj} \hat{z}^g_{nj} \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w^m_{kd} \rangle \langle w^m_{jd} \rangle}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}} \\
\sigma^2_{z^g_{nk}} &= \frac{\langle \tau^{gm}_d \rangle^{-1}}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}}
\end{aligned} \tag{A.6}$$

ARD prior on the factors For every group g and factor k:

Prior distribution:

$$p(\alpha^g_k) = \mathcal{G}\left(\alpha^g_k \,|\, a^\alpha_0, b^\alpha_0\right) \tag{A.7}$$

Variational distribution $q(\alpha^g_k)$:

$$q(\alpha^g_k) = \mathcal{G}\left(\alpha^g_k \,|\, \hat{a}^\alpha_{gk}, \hat{b}^\alpha_{gk}\right) \tag{A.8}$$

where:

$$\begin{aligned} \hat{a}_{gk}^{\alpha} &= a_0^{\alpha} + \frac{N_g}{2} \\ \hat{b}_{gk}^{\alpha} &= b_0^{\alpha} + \frac{\sum_{n=1}^{N_g} \langle (\hat{z}_{nk}^{g})^2 \rangle}{2} \end{aligned} \tag{A.9}$$

Sparsity parameter of the Factors For every group g and factor k:

Prior distribution:

$$p(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, a_0^{\theta}, b_0^{\theta}\right) \tag{A.10}$$

Variational distribution:

$$q(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, \hat{a}_{gk}^{\theta}, \hat{b}_{gk}^{\theta}\right) \tag{A.11}$$

where

$$\begin{aligned} \hat{a}_{gk}^{\theta} &= \sum_{n=1}^{N_g} \langle s_{nk}^{g} \rangle + a_0^{\theta} \\ \hat{b}_{gk}^{\theta} &= b_0^{\theta} - \sum_{n=1}^{N_g} \langle s_{nk}^{g} \rangle + N_g \end{aligned} \tag{A.12}$$

Feature weights For every view m, feature d and factor k:

Prior distribution $p(\hat{w}_{kd}^m, s_{kd}^m)$:

$$p(\hat{w}_{kd}^m, s_{kd}^m) = \mathcal{N}(\hat{w}_{kd}^m \,|\, 0,\ 1/\alpha_k^m)\, \text{Ber}(s_{kd}^m \,|\, \theta_k^m) \tag{A.13}$$

Variational distribution $q(\hat{w}_{kd}^m, s_{kd}^m)$:

Update for $q(s_{kd}^m)$:

$$q(s_{kd}^m) = \text{Ber}(s_{kd}^m | \gamma_{kd}^m) \tag{A.14}$$

with

$$\begin{aligned} \gamma_{kd}^m &= \frac{1}{1+\exp(-\lambda_{kd}^m)} \\ \lambda_{kd}^m &= \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} - 0.5 \ln \left(\sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} \right) \\ &+ \frac{\langle \tau_d^{gm} \rangle}{2} \frac{\left(\sum_{g=1}^{G} \sum_{n=1}^{N_g} y_{nd}^{gm} \langle z_{nk}^g \rangle - \sum_{j \neq k} \langle s_{jd}^m \hat{w}_{jd}^m \rangle \sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle z_{nk}^g \rangle \langle z_{nj}^g \rangle \right)^2}{\sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle}} \end{aligned} \tag{A.15}$$

Update for $q(\hat{w}_{kd}^m)$:

$$\begin{aligned} q(\hat{w}_{kd}^m|s_{kd}^m=0) &= \mathcal{N}\left(\hat{w}_{kd}^m|0,1/\alpha_k^m\right) \\ q(\hat{w}_{kd}^m|s_{kd}^m=1) &= \mathcal{N}\left(\hat{w}_{kd}^m\middle|\mu_{w_{kd}^m},\sigma^2_{w_{kd}^m}\right) \end{aligned} \tag{A.16}$$

with

$$\begin{aligned} \mu_{w_{kd}^m} &= \frac{\sum_{g=1}^G\sum_{n=1}^{N_g} y_{nd}^{gm}\langle z_{nk}^g\rangle - \sum_{j\neq k}\langle s_{jd}^m\hat{w}_{jd}^m\rangle\sum_{g=1}^G\sum_{n=1}^{N_g}\langle z_{nk}^g\rangle\langle z_{nj}^g\rangle}{\sum_{g=1}^G\sum_{n=1}^{N_g}\langle (z_{nk}^g)^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^{gm}\rangle}} \\ \sigma^2_{w_{kd}^m} &= \frac{\langle\tau_d^{gm}\rangle^{-1}}{\sum_{g=1}^G\sum_{n=1}^{N_g}\langle (z_{nk}^g)^2\rangle + \frac{\langle\alpha_k^m\rangle}{\langle\tau_d^{gm}\rangle}} \end{aligned} \tag{A.17}$$

ARD prior on the weights For every view m and factor k:

Prior distribution $p(\alpha_k^m)$:

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right)$$

Variational distribution $q(\alpha_k^m)$:

$$q(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, \hat{a}_{mk}^\alpha, \hat{b}_{mk}^\alpha\right) \tag{A.18}$$

where:

$$\begin{aligned} \hat{a}_{mk}^\alpha &= a_0^\alpha + \frac{D_m}{2} \\ \hat{b}_{mk}^\alpha &= b_0^\alpha + \frac{\sum_{d=1}^{D_m}\langle(\hat{w}_{kd}^m)^2\rangle}{2} \end{aligned} \tag{A.19}$$

Sparsity parameter of the weights For every view m and factor k:

Prior distribution:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right)$$

Variational distribution:

$$q(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, \hat{a}_{mk}^\theta, \hat{b}_{mk}^\theta\right) \tag{A.20}$$

where

$$\begin{aligned} \hat{a}_{mk}^\theta &= \sum_{d=1}^{D_m}\langle s_{kd}^m\rangle + a_0^\theta \\ \hat{b}_{mk}^\theta &= b_0^\theta - \sum_{d=1}^{D_m}\langle s_{kd}^m\rangle + D_m \end{aligned} \tag{A.21}$$

Noise (Gaussian) For every view m, group g and feature d:

Prior distribution $p(\tau_d^{gm})$:

$$p(\tau_d^{gm}) = \mathcal{G}\left(\tau_{dg}^m \,|\, a_0^\tau, b_0^\tau\right),$$

Variational distribution $q(\tau_d^{gm})$:

$$q(\tau_d^{gm}) = \mathcal{G}\left(\tau_d^{gm} \mid \hat{a}_d^{gm}, \hat{b}_d^{gm}\right) \tag{A.22}$$

where:

$$\begin{aligned} \hat{a}_d^{gm} &= a_0^{\tau} + \frac{N_g}{2} \\ \hat{b}_d^{gm} &= b_0^{\tau} + \frac{1}{2}\sum_{n=1}^{N_g}\langle\left(y_{nd}^{gm} - \sum_{k}^{K} w_{kd}^{m} z_{nk}^{g}\right)^2\rangle \end{aligned} \tag{A.23}$$

A.3 Evidence Lower Bound

Although computing the ELBO is not necessary in order to estimate the posterior distribution of the parameters, it is used to monitor the convergence of the algorithm. As shown in Equation (3.2), the ELBO can be decomposed into a sum of two terms: (1) the expected log likelihood under the current estimate of the posterior distribution of the parameters and (2) the KL divergence between the prior and the variational distributions of the parameters:

$$\mathcal{L} = \mathbb{E}_{q(X)} \ln p(Y|X) - \mathrm{KL}\left(q(X)||p(X)\right) \tag{A.24}$$

Log likelihood term Assuming a Gaussian likelihood:

$$\begin{aligned} \mathbb{E}_{q(X)} \ln p(Y|X) = &- \sum_{m=1}^{M} \frac{N D_m}{2} \ln(2\pi) + \sum_{g=1}^{G} \frac{N_g}{2} \sum_{m=1}^{M}\sum_{d=1}^{D_m} \langle \ln(\tau_d^{gm})\rangle \\ &- \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{D_m} \frac{\langle \tau_d^{gm}\rangle}{2} \sum_{n=1}^{N_g} \left(y_{nd}^{m,g} - \sum_{k=1}^{K} \langle s_{kd}^{m} \hat{w}_{kd}^{m}\rangle \langle z_{nk}^{g}\rangle\right)^2 \end{aligned} \tag{A.25}$$

KL divergence terms Note that $\mathrm{KL}\left(q(X)||p(X)\right) = \mathbb{E}_q(q(X)) - \mathbb{E}_q(p(X))$. Below, we will write the analytical form for these two expectations.

Weights

$$\begin{aligned} \mathbb{E}_q[\ln p(\hat{W}, S)] = &- \sum_{m=1}^{M} \frac{K D_m}{2} \ln(2\pi) + \sum_{m=1}^{M} \frac{D_m}{2} \sum_{k=1}^{K} \ln(\alpha_k^m) - \sum_{m=1}^{M} \frac{\alpha_k^m}{2} \sum_{d=1}^{D_m}\sum_{k=1}^{K} \langle (\hat{w}_{kd}^m)^2\rangle \\ &+ \langle \ln(\theta)\rangle \sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K} \langle s_{kd}^m\rangle + \langle \ln(1-\theta)\rangle \sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K} (1 - \langle s_{kd}^m\rangle) \end{aligned} \tag{A.26}$$

$$\begin{aligned}\mathbb{E}_q[\ln q(\hat{W},S)] = &-\sum_{m=1}^{M}\frac{KD_m}{2}\ln(2\pi)+\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\ln(\langle s_{kd}^m\rangle\sigma_{w_{kd}^m}^2+(1-\langle s_{kd}^m\rangle)/\alpha_k^m)\\ &+\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}(1-\langle s_{kd}^m\rangle)\ln(1-\langle s_{kd}^m\rangle)-\langle s_{kd}^m\rangle\ln\langle s_{kd}^m\rangle\end{aligned} \tag{A.27}$$

Factors

$$\begin{aligned}\mathbb{E}_q[\ln p(\hat{Z},S)] = &-\sum_{g=1}^{G}\frac{N_gK}{2}\ln(2\pi)+\sum_{g=1}^{G}\frac{N_g}{2}\sum_{k=1}^{K}\ln(\alpha_k^g)-\sum_{g=1}^{G}\frac{\alpha_k^g}{2}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\langle(\hat{z}_{nk}^g)^2\rangle\\ &+\langle\ln(\theta)\rangle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\langle s_{nk}^g\rangle+\langle\ln(1-\theta)\rangle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}(1-\langle s_{nk}^g\rangle)\end{aligned} \tag{A.28}$$

$$\begin{aligned}\mathbb{E}_q[\ln q(\hat{Z},S)] = &-\sum_{g=1}^{G}\frac{N_gK}{2}\ln(2\pi)+\frac{1}{2}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\ln(\langle s_{nk}^g\rangle\sigma_{z_{nk}^g}^2+(1-\langle s_{nk}^g\rangle)/\alpha_k^g)\\ &+\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}(1-\langle s_{nk}^g\rangle)\ln(1-\langle s_{nk}^g\rangle)-\langle s_{nk}^g\rangle\ln\langle s_{nk}^g\rangle\end{aligned} \tag{A.29}$$

ARD prior on the weights

$$\begin{aligned}\mathbb{E}_q[\ln p(\boldsymbol{\alpha})] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(a_0^\alpha\ln b_0^\alpha+(a_0^\alpha-1)\langle\ln\alpha_k\rangle-b_0^\alpha\langle\alpha_k\rangle-\ln\Gamma(a_0^\alpha)\Big)\\ \mathbb{E}_q[\ln q(\boldsymbol{\alpha})] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(\hat{a}_k^\alpha\ln\hat{b}_k^\alpha+(\hat{a}_k^\alpha-1)\langle\ln\alpha_k\rangle-\hat{b}_k^\alpha\langle\alpha_k\rangle-\ln\Gamma(\hat{a}_k^\alpha)\Big)\end{aligned} \tag{A.30}$$

Sparsity parameter of the weights

$$\begin{aligned}\mathbb{E}_q\left[\ln p(\boldsymbol{\theta})\right] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\big((a_0-1)\times\langle\ln(\pi_{d,k}^m)\rangle+(b_0-1)\langle\ln(1-\pi_{d,k}^m)\rangle-\ln(\mathrm{B}(a_0,b_0))\big)\\ \mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\big((a_{k,d}^m-1)\times\langle\ln(\pi_{d,k}^m)\rangle+(b_{k,d}^m-1)\langle\ln(1-\pi_{d,k}^m)\rangle-\ln(\mathrm{B}(a_{k,d}^m,b_{k,d}^m))\big)\end{aligned} \tag{A.31}$$

ARD prior on the Factors

$$\begin{aligned}\mathbb{E}_q[\ln p(\boldsymbol{\alpha})] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\Big(a_0^\alpha\ln b_0^\alpha+(a_0^\alpha-1)\langle\ln\alpha_k\rangle-b_0^\alpha\langle\alpha_k\rangle-\ln\Gamma(a_0^\alpha)\Big)\\ \mathbb{E}_q[\ln q(\boldsymbol{\alpha})] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\Big(\hat{a}_k^\alpha\ln\hat{b}_k^\alpha+(\hat{a}_k^\alpha-1)\langle\ln\alpha_k\rangle-\hat{b}_k^\alpha\langle\alpha_k\rangle-\ln\Gamma(\hat{a}_k^\alpha)\Big)\end{aligned} \tag{A.32}$$

Sparsity parameter of the Factors

$$\mathbb{E}_q\left[\ln p(\boldsymbol{\theta})\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\Big((a_0-1)\times\langle\ln(\pi_{n,k}^g)\rangle + (b_0-1)\langle\ln(1-\pi_{n,k}^g)\rangle - \ln(\mathrm{B}(a_0,b_0))\Big)$$
$$\mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\Big((a_{k,n}^g-1)\times\langle\ln(\pi_{n,k}^g)\rangle + (b_{k,n}^g-1)\langle\ln(1-\pi_{n,k}^g)\rangle - \ln(\mathrm{B}(a_{k,n}^g,b_{k,n}^g))\Big) \tag{A.33}$$

Noise

$$\mathbb{E}_q[\ln p(\boldsymbol{\tau})] = \sum_{m=1}^{M} D_m a_0^\tau \ln b_0^\tau + \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm}(a_0^\tau-1)\langle\ln\tau_d^{gm}\rangle - \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm} b_0^\tau\langle\tau_d^{gm}\rangle - \sum_{m=1}^{M} D_m\ln\Gamma(a_0^\tau)$$
$$\mathbb{E}_q[\ln q(\boldsymbol{\tau})] = \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\Big(\hat{a}_{dgm}^\tau\ln\hat{b}_{dgm}^\tau + (\hat{a}_{dgm}^\tau-1)\langle\ln\tau_d^{gm}\rangle - \hat{b}_{dgm}^\tau\langle\tau_d^{gm}\rangle - \ln\Gamma(\hat{a}_{dgm}^\tau)\Big) \tag{A.34}$$

Appendix B

Characterisation of MOFA factors in the scNMT-seq gastrulation data set

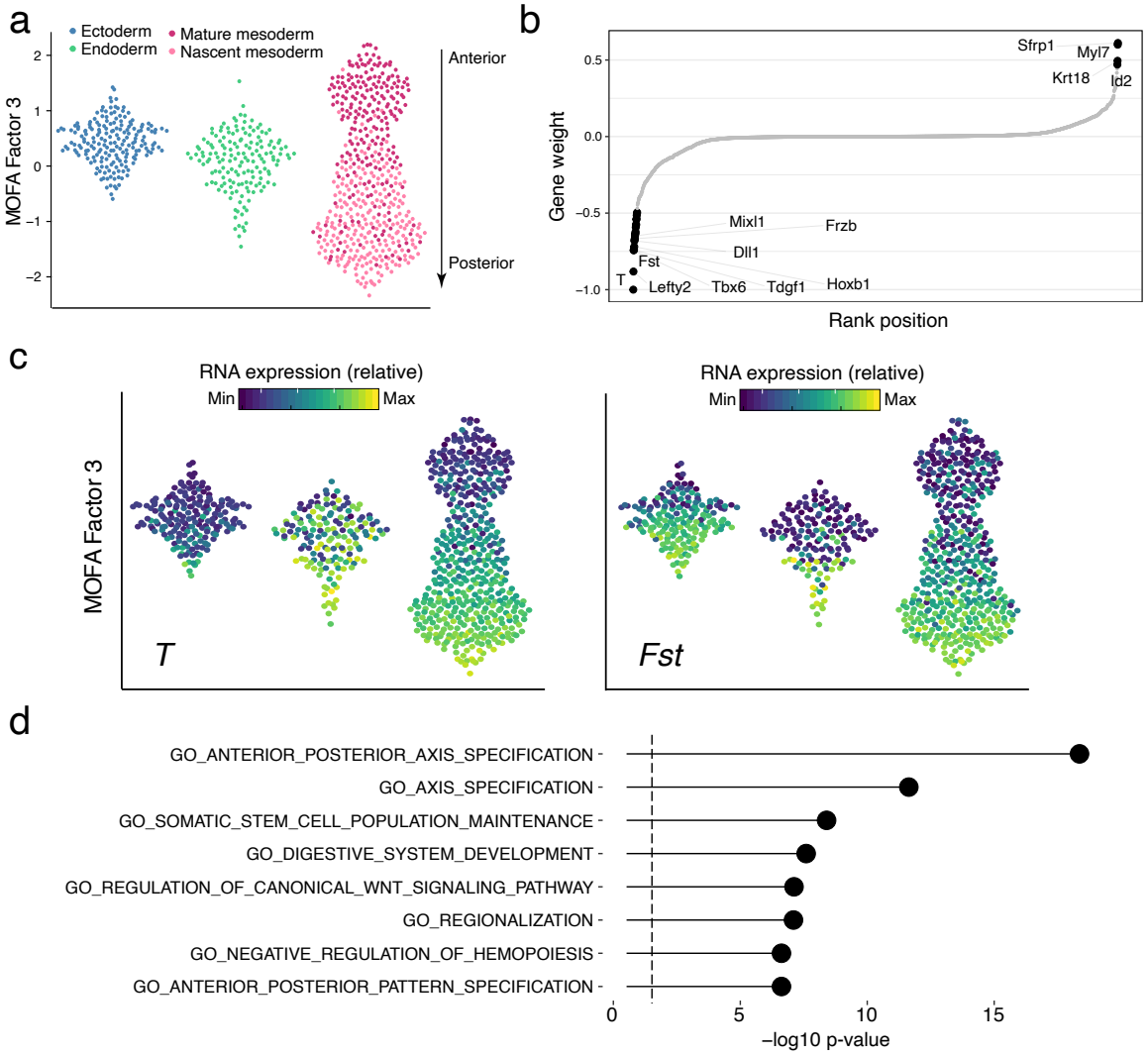


Figure B.1: Characterisation of MOFA Factor 3 as antero-posterior axial patterning and mesoderm maturation.
(a) Beeswarm plot of Factor 3 values, grouped and coloured by cell type. The mesoderm cells are subclassified into nascent and mature mesoderm.
(b) RNA expression weights for Factor 3. Genes with large positive weights increase expression in the positive factor values (more anterior), whereas genes with negative weights increase expression in the negative factor values (more posterior).
(c) Same beeswarm plots as in (a), coloured by the relative RNA expression of genes with the highest positive (top) or negative (bottom) weight.
(d) Gene set enrichment analysis of the gene weights of Factor 3. Shown are the top most significant pathways from MSigDB C2 [292, 17].

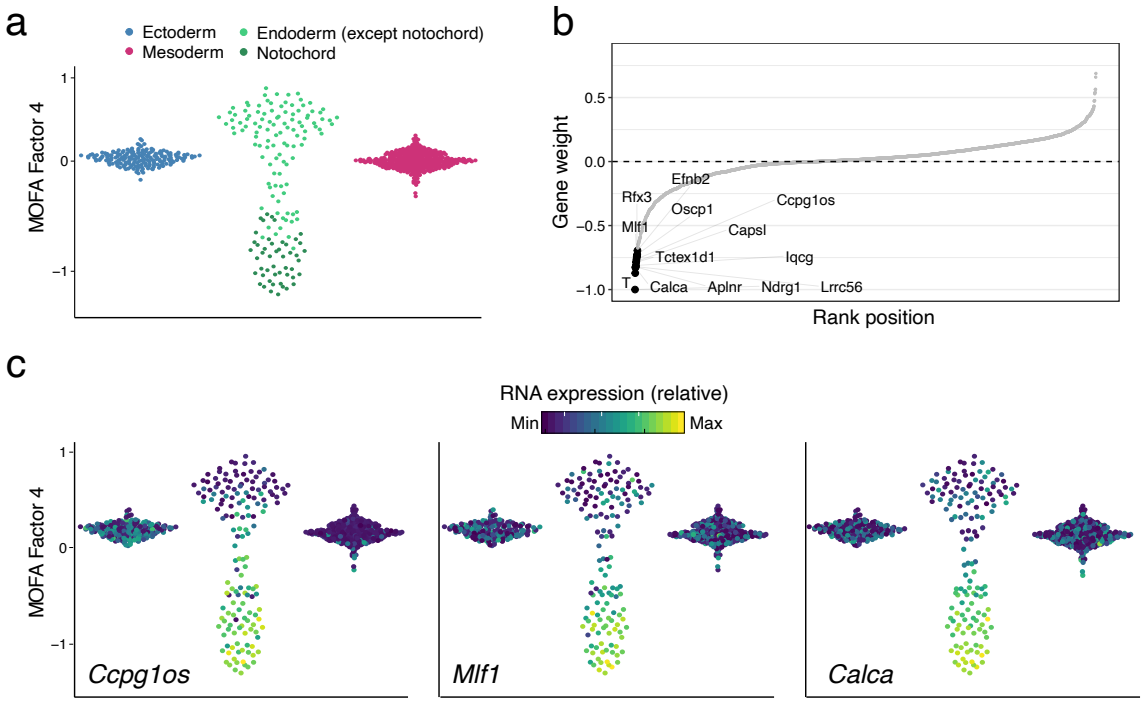


Figure B.2: Characterisation of MOFA Factor 4 as notochord formation.
(a) Beeswarm plot of Factor 4 values, grouped and coloured by cell type. The endoderm cells are subclassified into notochord (dark green) and not notochord (green) (see Figure S2).
(b) RNA expression weights for Factor 4. Genes with large positive weights increase expression in the positive factor values (endoderm cells), whereas genes with negative weights increase expression in the negative factor values (notochord cells).
(c) Same beeswarm plots as in (a), coloured by the relative RNA expression of genes with the highest negative weight (notochord markers).

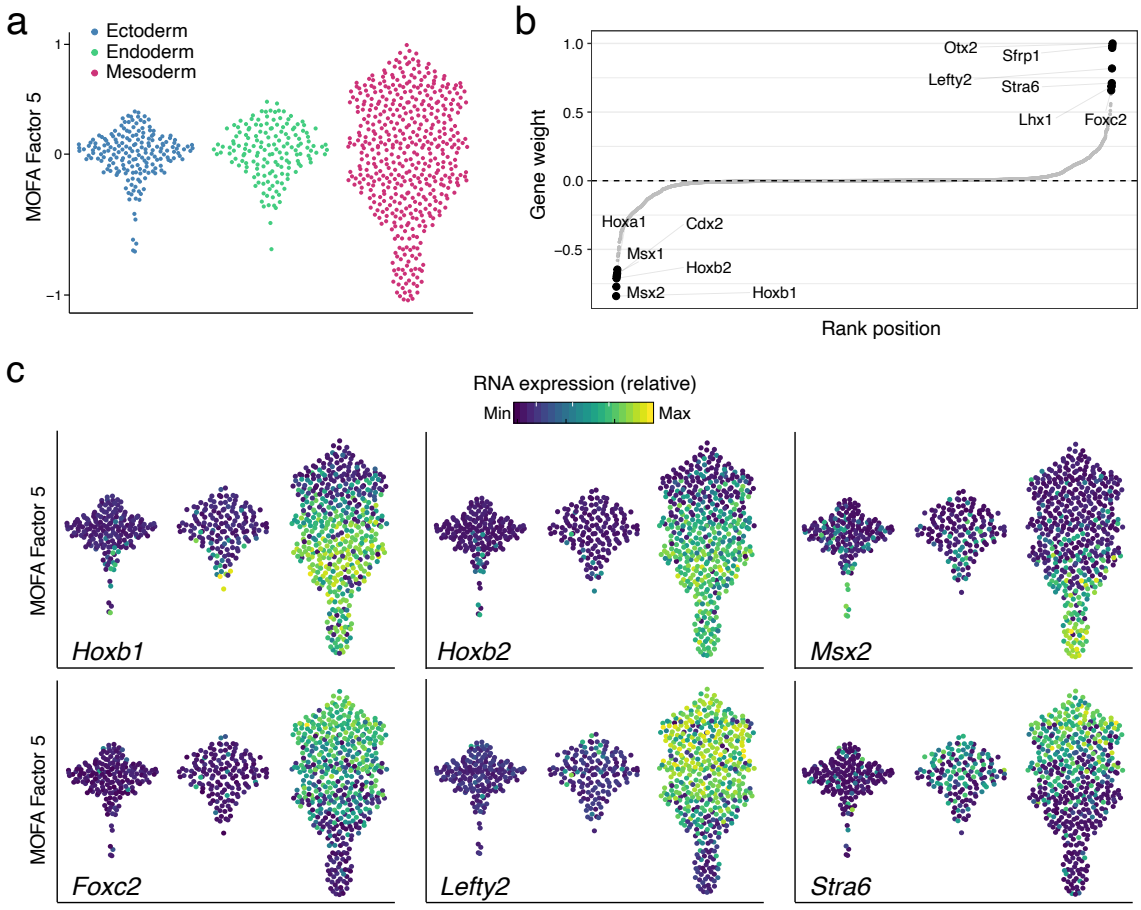


Figure B.3: Characterisation of MOFA Factor 5 as mesoderm patterning.
(a) Beeswarm plot of Factor 5 values, grouped and coloured by cell type.
(b) RNA expression weights for Factor 5. A higher absolute value indicates higher feature importance.
(c) Same beeswarm plots as in (a), coloured by the relative RNA expression of genes with the highest weight on this factor.

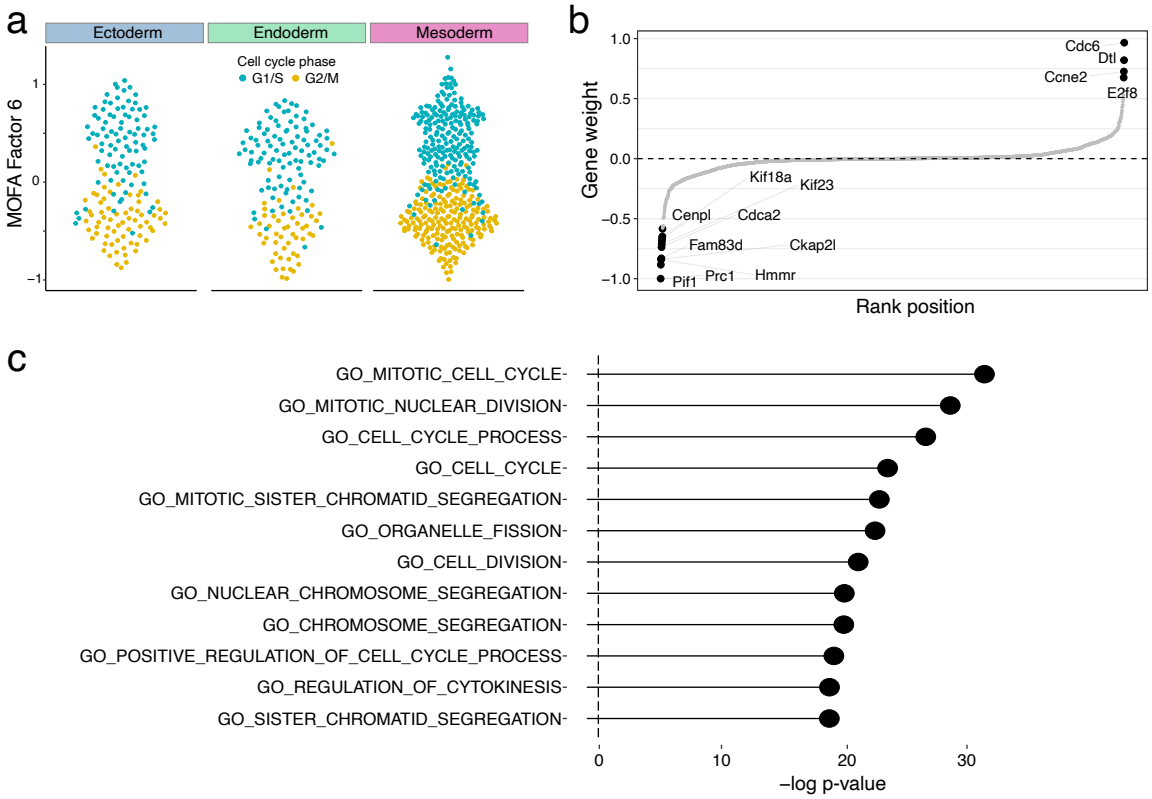


Figure B.4: Characterisation of MOFA Factor 6 as cell cycle.
(a) Beeswarm plot of Factor 6 values, grouped by cell type and coloured by infered cell cycle state using *cyclone*[269].
(b) RNA expression weights for Factor 6. Genes with large positive weights increase expression in the positive factor values (G1/S phase), whereas genes with negative weights increase expression in the negative factor values (G2/M phase).

Bibliography

[1] T. Abdelaal et al. "A comparison of automatic cell identification methods for single-cell RNA sequencing data". In: *Genome Biology* 20.1 (2019), p. 194.

[2] S. Aibar et al. "SCENIC: single-cell regulatory network inference and clustering". In: *Nature Methods* 14.11 (2017).

[3] S. Ainsworth et al. *Interpretable VAEs for nonlinear group factor analysis*. 2018. arXiv: 1802.06765 [cs.LG].

[4] C. Alda-Catalinas et al. "A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program". In: *Cell Systems* 11.1 (2020), 25–41.e9.

[5] A. Alemany et al. "Whole-organism clone tracing using single-cell sequencing". In: *Nature* 556.7699 (2018), pp. 108–112.

[6] J. Allaire et al. *reticulate: R Interface to Python*. 2017.

[7] A. Alyass, M. Turcotte, and D. Meyre. "From big data analysis to personalized medicine for all: challenges and opportunities". In: *BMC Medical Genomics* 8.1 (2015), p. 33. ISSN: 1755-8794. DOI: 10.1186/s12920-015-0108-y.

[8] S.-I. Amari. "Natural Gradient Works Efficiently in Learning". In: *Neural Comput.* 10.2 (1998), pp. 251–276.

[9] R. A. Amezquita et al. "Orchestrating single-cell analysis with Bioconductor". In: *Nature Methods* 17.2 (2020).

[10] R. E. Amir et al. "Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2". In: *Nature Genetics* 23 (1999).

[11] C. Angermueller et al. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning". In: *Genome Biology* 18.1 (2017).

[12] C. Angermueller et al. "Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity". In: *Nature Methods* 13 (2016).

[13] R. Argelaguet et al. "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1 (2020), p. 111.

[14] R. Argelaguet et al. "Multi-omics profiling of mouse gastrulation at single-cell resolution". In: *Nature* 576.7787 (2019).

[15] R. Argelaguet et al. "Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets". In: *Mol Syst Biol* 14.6 (2018), e8124. ISSN: 1744-4292 (Electronic) 1744-4292 (Linking). DOI: 10.15252/msb.20178124.

[16] S. J. Arnold and E. J. Robertson. "Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo". en. In: *Nat. Rev. Mol. Cell Biol.* 10.2 (2009), pp. 91–103.

[17] M. Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nat Genet* 25.1 (2000), pp. 25–9. DOI: 10.1038/75556.

[18] Y. Atlasi and H. G. Stunnenberg. "The interplay of epigenetic marks during stem cell differentiation and development". In: *Nature Reviews Genetics* 18 (2017).

[19] G. Auclair et al. "Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse". In: *Genome biology* 15.12 (2014), p. 545. ISSN: 1474-760X.

[20] F. R. Bach and M. I. Jordan. *A probabilistic interpretation of canonical correlation analysis*. Tech. rep. 2005.

[21] T. L. Bailey et al. "MEME Suite: tools for motif discovery and searching". In: *Nucleic Acids Research* 37 (2009). ISSN: 0305-1048. DOI: 10.1093/nar/gkp335.

[22] M. N. Bainbridge et al. "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach". In: *BMC Genomics* 7.1 (2006), p. 246. ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-246.

[23] N. Barkas et al. "Joint analysis of heterogeneous single-cell RNA-seq dataset collections". In: *Nature Methods* 16.8 (2019), pp. 695–698.

[24] C. Baron and A. van Oudenaarden. "Unravelling cellular relationships during development and regeneration using genetic lineage tracing". In: *Nature Reviews Molecular Cell Biology* 20.12 (2019), pp. 753–765.

[25] T. Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S". In: *Philosophical Transactions of the Royal Society of London* 53 (1763), pp. 370–418. DOI: 10.1098/rstl.1763.0053.

[26] S. B. Baylin and P. A. Jones. "A decade of exploring the cancer epigenome —biological and translational implications". In: *Nature Reviews Cancer* 11 (2011).

[27] B. E. Bernstein et al. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells". In: *Cell* 125.2 (2006), pp. 315–326.

[28] P. Bheda and R. Schneider. "Epigenetics reloaded: the single-cell revolution". In: *Trends in Cell Biology* 24.11 (2014), pp. 712–723. DOI: https://doi.org/10.1016/j.tcb.2014.08.010.

[29] C. Bishop. "Variational Principal Components". In: *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*. Vol. 1. 1999, pp. 509–514.

[30] C. M. Bishop. "Bayesian PCA". In: *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 382–388. ISBN: 0-262-11245-0.

[31] C. M. Bishop. "Pattern recognition". In: *Machine Learning* 128 (2006), pp. 1–58.

[32] D. M. Blei and M. I. Jordan. "Variational inference for Dirichlet process mixtures". In: *Bayesian Anal.* 1.1 (2006), pp. 121–143. DOI: 10.1214/06-BA104.

[33] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *arXiv e-prints*, arXiv:1601.00670 (2016), arXiv:1601.00670. arXiv: 1601.00670 [stat.CO].

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. ISSN: 1532-4435.

[35] R. Bourgon, R. Gentleman, and W. Huber. "Independent filtering increases detection power for high-throughput experiments". In: *Proceedings of the National Academy of Sciences* 107.21 (2010).

[36] M. Braun and J. McAuliffe. "Variational inference for large-scale models of discrete choice". In: *arXiv e-prints*, arXiv:0712.2526 (2007), arXiv:0712.2526. arXiv: 0712.2526 [stat.ME].

[37] S. C. van den Brink et al. "Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids". In: *Nature* 582.7812 (2020), pp. 405–409.

[38] A. B. Brinkman et al. "Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk". In: *Genome Research* 22.6 (2012), pp. 1128–1138. DOI: 10.1101/gr.133728.111.

[39] J. D. Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current Protocols in Molecular Biology* 109.1 (2015).

[40] J. D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523 (2015).

[41] J. D. Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10 (2013).

[42] F. Buettner et al. "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". In: *Nature Biotechnology* 33 (Jan. 2015).

[43] F. Buettner et al. "f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq". In: *Genome Biol.* 18.1 (2017), p. 212.

[44] P. Bulian et al. "Mutational status of IGHV is the most reliable prognostic marker in trisomy 12 chronic lymphocytic leukemia". In: *Haematologica* 102.11 (2017), e443–e446. ISSN: 0390-6078. DOI: 10.3324/haematol.2017.170340. eprint: http://www.haematologica.org/content/102/11/e443.full.pdf.

[45] K. Bunte et al. "Sparse group factor analysis for biclustering of multiple data sources". In: *Bioinformatics* 32.16 (2016), pp. 2457–2463.

[46] A. Butler et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature Biotechnology* 36.5 (2018), pp. 411–420.

[47] J. Cao et al. "Comprehensive single-cell transcriptional profiling of a multicellular organism". In: *Science* 357.6352 (2017), pp. 661–667. ISSN: 0036-8075. DOI: 10.1126/science.aam8940.

[48] J. Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409 (2018), p. 1380.

[49] J. Cao et al. "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745 (2019), pp. 496–502. DOI: 10.1038/s41586-019-0969-x.

[50] J. Cao et al. "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745 (2019), pp. 496–502.

[51] K. Cao et al. "Unsupervised topological alignment for single-cell multi-omics integration". In: *Bioinformatics* 36 (July 2020), pp. i48–i56. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa443.

[52] P. Carbonetto and M. Stephens. "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies". In: *Bayesian Anal.* 7.1 (2012), pp. 73–108. DOI: 10.1214/12-BA703.

[53] F. P. Casale et al. *Gaussian Process Prior Variational Autoencoders*. 2018. arXiv: 1810.11738.

[54] M. M. Chan et al. "Molecular recording of mammalian embryogenesis". In: *Nature* 570.7759 (2019).

[55] L. Chappell, A. J. Russell, and T. Voet. "Single-Cell (Multi)omics Technologies". In: *Annual Review of Genomics and Human Genetics* 19.1 (2018), pp. 15–41. DOI: 10.1146/annurev-genom-091416-035324.

[56] R. Chen and M. Snyder. "Promise of personalized omics to precision medicine". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 5.1 (2013), pp. 73–82. DOI: 10.1002/wsbm.1198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.1198.

[57] X. Chen et al. "A rapid and robust method for single cell chromatin accessibility profiling". In: *Nature Communications* 9.1 (2018), p. 5345.

[58] S. J. Clark et al. "Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq)". In: *Nature Protocols* 12 (2017).

[59] S. J. Clark et al. "scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells". In: *Nature Communications* 9.1 (2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-03149-4.

[60] S. J. Clark et al. "Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity". In: *Genome Biology* 17.1 (2016), p. 72.

[61] M. Colome-Tatche and F. Theis. "Statistical single cell multi-omics integration". In: *Current Opinion in Systems Biology* 7 (2018). Future of systems biology Genomics and epigenomics, pp. 54–59. DOI: https://doi.org/10.1016/j.coisb.2018.01.003.

[62] T. E. P. Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489 (2012).

[63] J. C. Costello et al. "A community effort to assess and improve drug sensitivity prediction algorithms". In: *Nature Biotechnology* 32 (June 2014).

[64] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246.

[65] M. P. Creyghton et al. "Histone H3K27ac separates active from poised enhancers and predicts developmental state". In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21931–21936. DOI: 10.1073/pnas.1016071107.

[66] J. Crombie and M. S. Davids. "IGHV mutational status testing in chronic lymphocytic leukemia". In: *American Journal of Hematology* 92.12 (2017), pp. 1393–1397. DOI: 10.1002/ajh.24808. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.24808.

[67] A. S. E. Cuomo et al. "Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression". In: *Nature Communications* 11.1 (2020), p. 810.

[68] D. A. Cusanovich et al. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility". In: *Cell* 174.5 (2018), 1309–1324.e18. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2018.06.052.

[69] D. A. Cusanovich et al. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237 (2015), pp. 910–914. ISSN: 0036-8075. DOI: 10.1126/science.aab1601.

[70] H.-Q. Dai et al. "TET-mediated DNA demethylation controls gastrulation by regulating Lefty-Nodal signalling". en. In: *Nature* 538.7626 (2016), pp. 528–532.

[71] P. Datlinger et al. "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature Methods* 14.3 (2017), pp. 297–301.

[72] N. C. Dempsey et al. "Differential heat shock protein localization in chronic lymphocytic leukemia". In: *Journal of Leukocyte Biology* 87.3 (2010), pp. 467–476.

[73] Q. Deng et al. "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells". In: *Science* 343.6167 (2014), pp. 193–196. ISSN: 0036-8075. DOI: 10.1126/science.1245316.

[74] S. S. Dey et al. "Integrated genome and transcriptome sequencing of the same cell". In: *Nature Biotechnology* 33 (2015).

[75] S. Dietrich et al. "Drug-perturbation-based stratification of blood cancer". In: *J. Clin. Invest.* 128.1 (2018), pp. 427–445.

[76] L. Dietz. *Directed Factor Graph Notation for Generative Models*. 2010.

[77] J. Ding, A. Condon, and S. P. Shah. "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models". In: *Nature Communications* 9.1 (2018), p. 2002.

[78] A. Dixit et al. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". In: *Cell* 167.7 (2016), 1853–1866.e17.

[79] P. Du et al. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". en. In: *BMC Bioinformatics* 11 (2010), p. 587.

[80] Z. Du et al. "Allelic reprogramming of 3D chromatin architecture during early mammalian development". In: *Nature* 547 (2017).

[81] M. Eisenstein. "The secret life of cells". In: *Nature Methods* 17.1 (2020), pp. 7–10.

[82] M. Emtiyaz Khan and W. Lin. "Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models". In: *arXiv e-prints*, arXiv:1703.04265 (2017), arXiv:1703.04265. arXiv: 1703.04265.

[83] C.-H. L. Eng et al. "Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+". In: *Nature* 568.7751 (2019), pp. 235–239.

[84] M. Enge et al. "Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns". In: *Cell* 171.2 (2017), 321–330.e14.

[85] G. Eraslan et al. "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature Communications* 10.1 (2019), p. 390.

[86] G. Fabbri and R. Dalla-Favera. "The molecular pathogenesis of chronic lymphocytic leukaemia". In: *Nat. Rev. Cancer* 16.3 (2016), pp. 145–162.

[87] A. Fabregat et al. "The reactome pathway knowledgebase". In: *Nucleic acids research* 44.D1 (2015), pp. D481–D487.

[88] C. Faes, J. T. Ormerod, and M. P. Wand. "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 959–971. ISSN: 01621459.

[89] J. Fan et al. "Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data". In: *Genome Research* 28.8 (2018), pp. 1217–1227.

[90] M. Farlik et al. "Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics". In: *Cell Reports* 10.8 (2015), pp. 1386–1397. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2015.02.001.

[91] G. Ficz et al. "FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency". In: *Cell Stem Cell* 13.3 (2013), pp. 351–359.

[92] R. Fleischmann et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223 (1995), pp. 496–512. ISSN: 0036-8075. DOI: 10.1126/science.7542800.

[93] M. Frommer et al. "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." In: *Proceedings of the National Academy of Sciences* 89.5 (1992), pp. 1827–1831. ISSN: 0027-8424. DOI: 10.1073/pnas.89.5.1827.

[94] H. R. Frost, Z. Li, and J. H. Moore. "Principal component gene set enrichment (PCGSE)". In: *BioData mining* 8.1 (2015).

[95] E. Fuchs. "Keratins as biochemical markers of epithelial differentiation". In: *Trends in Genetics* 4.10 (1988), pp. 277–281. ISSN: 0168-9525. DOI: https://doi.org/10.1016/0168-9525(88)90169-2.

[96] C. Gao, C. D. Brown, and B. E. Engelhardt. "A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects". In: *arXiv e-prints*, arXiv:1310.4792 (2013), arXiv:1310.4792. arXiv: 1310.4792 [stat.AP].

[97] A. Gelman et al. *Bayesian Data Analysis, Third Edition*. Hardcover. 2013.

[98] S. Gravina et al. "Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome". In: *Genome Biology* 17.1 (2016), p. 150.

[99] J. A. Griffiths, A. Scialdone, and J. C. Marioni. "Using single-cell genomics to understand developmental processes and cell fate decisions". In: *Molecular Systems Biology* 14.4 (2018). DOI: 10.15252/msb.20178046.

[100] S. Grosswendt et al. "Epigenetic regulator function through mouse gastrulation". In: *Nature* 584.7819 (2020), pp. 102–108.

[101] F. Guo et al. "Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells". In: *Cell Research* 27 (2017).

[102] H. Guo et al. "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing". In: *Genome Research* 23.12 (2013), pp. 2126–2135.

[103] Y. Guo et al. "Sufficient Canonical Correlation Analysis". In: *Trans. Img. Proc.* 25.6 (2016), pp. 2610–2619. ISSN: 1057-7149. DOI: 10.1109/TIP.2016.2551374.

[104] C. Hafemeister and R. Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1 (2019), p. 296.

[105] L. Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature Biotechnology* 36 (2018).

[106] L. Haghverdi et al. "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nature Methods* 13 (2016).

[107] C. W. Hanna, H. Demond, and G. Kelsey. "Epigenetic regulation in development: is the mouse a good model for the human?" In: *Human Reproduction Update* 24.5 (2018), pp. 556–576. ISSN: 1355-4786. DOI: 10.1093/humupd/dmy021.

[108] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2007, pp. 321–30. ISBN: 978-3-540-72243-4.

[109] T. Hashimshony et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (2012), pp. 666–673.

[110] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83. DOI: 10.1186/s13059-017-1215-1.

[111] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman, 2015. ISBN: 9781498712163.

[112] H. H. He et al. "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification". In: *Nature Methods* 11 (2013).

[113] N. D. Heintzman et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression". In: *Nature* 459.7243 (2009).

[114] A. Hemmati-Brivanlou and D. Melton. "Vertebrate Embryonic Cells Will Become Nerve Cells Unless Told Otherwise". In: *Cell* 88.1 (1997).

[115] G.-J. Hendriks et al. "NASC-seq monitors RNA synthesis in single cells". In: *Nature Communications* 10.1 (2019), p. 3138.

[116] w. c. f. P. H. Herve Pages and A. Lun. *DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets*. 2020.

[117] M. D. Hoffman et al. "Stochastic variational inference". In: *The Journal of Machine* (2013).

[118] M. D. Hoffman and D. M. Blei. "Structured Stochastic Variational Inference". In: *arXiv e-prints*, arXiv:1404.4114 (2014), arXiv:1404.4114. arXiv: 1404.4114.

[119] M. Hoffman et al. "Stochastic Variational Inference". In: *arXiv e-prints*, arXiv:1206.7051 (2012), arXiv:1206.7051. eprint: 1206.7051.

[120] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.

[121] H. Hotelling. "Relations between two sets of variates". In: *Biometrika* 28.3-4 (1936), pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: http://oup.prod.sis.lan/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf.

[122] Y. Hou et al. "Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas". In: *Cell Research* 26 (2016).

[123] Y. Hu et al. "Simultaneous profiling of transcriptome and DNA methylome from a single cell". In: *Genome Biology* 17.1 (2016), p. 88.

[124] S. Huang, K. Chaudhary, and L. X. Garmire. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods". In: *Frontiers in Genetics* 8 (2017), p. 84. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00084.

[125] X. Ibarra-Soria et al. "Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation". en. In: *Nat. Cell Biol.* 20.2 (2018), pp. 127–134.

[126] A. Ilin and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values". In: *J. Mach. Learn. Res.* 11 (2010), pp. 1957–2000. ISSN: 1532-4435.

[127] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1 (2000), pp. 25–37.

[128] D. A. Jaitin et al. "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq". In: *Cell* 167.7 (2016), 1883–1896.e15.

[129] E. Jaynes. "Prior Probabilities". In: *IEEE Transactions on Systems Science and Cybernetics* 4.3 (1968), pp. 227–241.

[130] C. Jiang and B. F. Pugh. "Nucleosome positioning and gene regulation: advances through genomics". In: *Nature Reviews Genetics* 10 (2009).

[131] W. Jin et al. "Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples". In: *Nature* 528 (2015).

[132] Z. Jin and Y. Liu. "DNA methylation in human diseases". In: *Genes & Diseases* 5.1 (2018), pp. 1–8. DOI: https://doi.org/10.1016/j.gendis.2018.01.002.

[133] R. M. John and C. Rougeulle. "Developmental Epigenetics: Phenotype and the Flexible Epigenome". In: *Frontiers in Cell and Developmental Biology* 6 (2018), p. 130. ISSN: 2296-634X. DOI: 10.3389/fcell.2018.00130.

[134] W. E. Johnson, C. Li, and A. Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2006), pp. 118–127. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxj037. eprint: https://academic.oup.com/biostatistics/article-pdf/8/1/118/25435561/kxj037.pdf.

[135] P. A. Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13 (2012).

[136] C.-A. Kapourani and G. Sanguinetti. "Melissa: Bayesian clustering and imputation of single cell methylomes". In: *bioRxiv* (2018). DOI: 10.1101/312025.

[137] H. S. Kaya-Okur et al. "CUT-Tag for efficient epigenomic profiling of small samples and single cells". In: *bioRxiv* (2019), p. 568915.

[138] Y. Ke et al. "3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis". In: *Cell* 170.2 (2017).

[139] T. K. Kelly et al. "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules". In: *Genome Research* 22.12 (2012), pp. 2497–2506.

[140] G. Kelsey, O. Stegle, and W. Reik. "Single-cell epigenomics: Recording the past and predicting the future". In: *Science* 358.6359 (2017), pp. 69–75. DOI: 10.1126/science.aan6826.

[141] L. Kester and A. van Oudenaarden. "Single-Cell Transcriptomics Meets Lineage Tracing". In: *Cell Stem Cell* 23.2 (2018), pp. 166–179.

[142] S. A. Khan et al. "Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis". In: *Bioinformatics* 30.17 (2014), pp. i497–i504.

[143] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. "Bayesian approach to single-cell differential expression analysis". In: *Nature Methods* 11.7 (2014), pp. 740–742.

[144] J. A. Kilgore et al. "Single-molecule and population probing of chromatin structure using DNA methyltransferases". In: *Methods* 41.3 (2007). Methods Related to the Structure and Function of Eukaryotic Chromatin, pp. 320–332. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2006.08.008.

[145] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20.5 (2019).

[146] V. Y. Kiselev et al. "SC3: consensus clustering of single-cell RNA-seq data". In: *Nature Methods* 14.5 (2017).

[147] A. Klami and S. Kaski. "Probabilistic approach to detecting dependencies between data sets". In: *Neurocomputing* 72.1 (2008), pp. 39–46.

[148] A. Klami, S. Virtanen, and S. Kaski. "Bayesian Canonical Correlation Analysis". In: *J. Mach. Learn. Res.* 14.1 (2013), pp. 965–1003. ISSN: 1532-4435.

[149] A. Klami et al. "Group factor analysis". In: *IEEE transactions on neural networks and learning systems* 26.9 (2015), pp. 2136–2147.

[150] A. M. Klein et al. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5 (2015), pp. 1187–1201.

[151] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* (2019).

[152] A. A. Kolodziejczyk et al. "The Technology and Biology of Single-Cell RNA Sequencing". In: *Molecular Cell* 58.4 (2015), pp. 610–620. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2015.04.005.

[153] S. Komili and P. A. Silver. "Coupling and coordination in gene expression processes: a systems biology view". In: *Nat. Rev. Genet.* 9 (2008), p. 38.

[154] I. Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature Methods* 16.12 (2019), pp. 1289–1296.

[155] A. Kristiadi. *Natural Gradient Descent*. https://wiseodd.github.io/techblog/2018/03/14/natural-gradient/. Blog. 2019.

[156] F. Krueger and S. R. Andrews. "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications". In: *Bioinformatics* 27.11 (2011), pp. 1571–1572. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr167.

[157] W. L. Ku et al. "Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification". In: *Nature Methods* 16.4 (2019), pp. 323–325.

[158] M. Kuhn and J. Kjell. "Applied Predictive Modeling". In: (2013).

[159] A. T. L. Lun, K. Bach, and J. C. Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome Biology* 17.1 (2016), p. 75.

[160] G. La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (2018), pp. 494–498.

[161] A. Lafzi et al. "Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies". In: *Nature Protocols* 13.12 (2018), pp. 2742–2757.

[162] D. Lahnemann et al. "Eleven grand challenges in single-cell data science". In: *Genome Biology* 21.1 (2020), p. 31.

[163] G. R. G. Lanckriet et al. "A statistical framework for genomic data fusion". In: *Bioinformatics* 20.16 (2004).

[164] N. D. Lawrence et al. "Efficient inference for sparse latent variable models of transcriptional regulation". In: *Bioinformatics* 33.23 (2017), pp. 3776–3783. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx508. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/33/23/3776/25168082/btx508.pdf.

[165] H. J. Lee, T. A. Hore, and W. Reik. "Reprogramming the Methylome: Erasing Memory and Creating Diversity". In: *Cell Stem Cell* 14.6 (2014).

[166] H. J. Lee et al. "Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos". In: *Nature Communications* 6.1 (2015).

[167] I. Lee et al. "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". In: *bioRxiv* (2018), p. 504993.

[168] J. T. Leek and J. D. Storey. "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis". In: *PLoS Genet.* 3.9 (2007), e161.

[169] E. Leppäaho and S. Kaski. "GFA: exploratory analysis of multiple data sources with group factor analysis". In: *Journal of Machine Learning Research* 18 (2017), pp. 1–5.

[170] X. Li et al. "Tet proteins influence the balance between neuroectodermal and mesodermal fate choice by inhibiting Wnt signaling". In: *Proceedings of the National Academy of Sciences* 113.51 (2016).

[171] Z. Li, S. E. Safo, and Q. Long. "Incorporating biological information in sparse principal component analysis with application to genomic data". In: *BMC Bioinformatics* 18.1 (2017), p. 332.

[172] G. Liang et al. "Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome". In: *Proc Natl Acad Sci U S A* 101.19 (2004), pp. 7357–62. DOI: 10.1073/pnas.0401866101.

[173] C. Lin et al. "Using neural networks for reducing the dimensions of single-cell RNA-Seq data". In: *Nucleic Acids Research* 45.17 (2017), e156–e156. ISSN: 0305-1048. DOI: 10.1093/nar/gkx681.

[174] D. Lin et al. "An integrative imputation method based on multi-omics datasets". In: *BMC Bioinformatics* 17.1 (2016), p. 247.

[175] R. Lister et al. "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". In: *Cell* 133.3 (2008), pp. 523–536. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2008.03.029.

[176] J. Liu et al. "Jointly embedding multiple single-cell omics measurements". In: *bioRxiv* (2019). DOI: 10.1101/644310.

[177] L. Liu et al. "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity". In: *Nature Communications* 10.1 (2019).

[178] Y. Liu et al. "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis". In: *Nature Biotechnology* 31 (2013).

[179] E. F. Lock et al. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types". In: *Ann. Appl. Stat.* 7.1 (Mar. 2013). DOI: 10.1214/12-AOAS597.

[180] R. Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (2018), pp. 1053–1058.

[181] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biol.* 15.12 (2014), p. 550.

[182] E. Lubeck et al. "Single-cell in situ RNA profiling by sequential hybridization". In: *Nature Methods* 11.4 (2014), pp. 360–361.

[183] L. S. Ludwig et al. "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics". In: *Cell* 176.6 (2019), 1325–1339.e22.

[184] M. D. Luecken and F. J. Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (2019), e8746.

[185] M. Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *bioRxiv* (2020). DOI: 10.1101/2020.05.22.111161. eprint: https://www.biorxiv.org/content/early/2020/05/23/2020.05.22.111161.full.pdf.

[186] A. Lun, D. McCarthy, and J. Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; referees: 3 approved, 2 approved with reservations]". In: *F1000Research* 5.2122 (2016). DOI: 10.12688/f1000research.9501.2.

[187] S. Ma et al. "Chromatin potential identified by shared single cell profiling of RNA and chromatin". In: *bioRxiv* (2020).

[188] L. van der Maaten and G. Hinton. "Visualizing Data using tSNE". In: *Journal of Machine Learning Research* 9 (2008).

[189] I. C. Macaulay et al. "G&T-seq: parallel sequencing of single-cell genomes and transcriptomes". In: *Nature Methods* 12 (2015).

[190] D. J. MacKay. "Bayesian methods for backpropagation networks". In: *Models of neural networks III*. Springer, 1996, pp. 211–254.

[191] E. Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5 (2015), pp. 1202–1214.

[192] J. Martens. "New insights and perspectives on the natural gradient method". In: *arXiv e-prints*, arXiv:1412.1193 (2014), arXiv:1412.1193. arXiv: 1412.1193 [cs.LG].

[193] U. Mayr, D. Serra, and P. Liberali. "Exploring single cells in space and time during tissue development, homeostasis and regeneration". In: *Development* 146.12 (2019), dev176727.

[194] R. Mazumder, T. Hastie, and R. Tibshirani. "Spectral Regularization Algorithms for Learning Large Incomplete Matrices". In: *J. Mach. Learn. Res.* 11.Aug (2010), pp. 2287–2322.

[195] S. D. McCabe, D.-Y. Lin, and M. I. Love. "MOVIE: Multi-Omics VIsualization of Estimated contributions". In: *bioRxiv* (2018). DOI: 10.1101/379115. eprint: https://www.biorxiv.org/content/early/2018/07/29/379115.full.pdf.

[196] D. J. McCarthy, Y. Chen, and G. K. Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". In: *Nucleic Acids Research* 40.10 (2012), pp. 4288–4297. ISSN: 0305-1048. DOI: 10.1093/nar/gks042. eprint: https://academic.oup.com/nar/article-pdf/40/10/4288/25335174/gks042.pdf.

[197] D. J. McCarthy et al. "Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants". In: *bioRxiv* (2018), p. 413047.

[198] C. S. McGinnis, L. M. Murrow, and Z. J. Gartner. "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors". In: *Cell Systems* 8.4 (2019).

[199] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018. arXiv: 1802.03426 [stat.ML].

[200] A. McKenna and J. A. Gagnon. "Recording development with single cell dynamic lineage tracing". In: *Development* 146.12 (2019), dev169730.

[201] C. Meng et al. “A multivariate approach to the integration of multi-omics datasets”. In: *BMC Bioinformatics* 15.1 (2014).

[202] C. Meng et al. “Dimension reduction techniques for the integrative analysis of multi-omics data”. In: *Brief. Bioinform.* 17.4 (2016), pp. 628–641.

[203] “Method of the Year 2019: Single-cell multimodal omics”. In: *Nature Methods* 17.1 (2020), pp. 1–1.

[204] T. P. Minka. “Expectation Propagation for approximate Bayesian inference”. In: *arXiv e-prints*, arXiv:1301.2294 (2013), arXiv:1301.2294. arXiv: 1301.2294.

[205] T. J. Mitchell and J. J. Beauchamp. “Bayesian variable selection in linear regression”. In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[206] Q. Mo et al. “Pattern discovery and cancer gene identification in integrated cancer genomic data”. In: *Proceedings of the National Academy of Sciences* 110.11 (2013), pp. 4245–4250. DOI: 10.1073/pnas.1208949110.

[207] H. Mohammed et al. “Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation”. In: *Cell Reports* 20.5 (), pp. 1215–1228.

[208] R. Moore et al. “A linear mixed-model approach to study multivariate gene–environment interactions”. In: *Nature Genetics* 51.1 (2019), pp. 180–186.

[209] F. Morabito et al. “Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment”. In: *Leuk. Res.* 39.8 (2015), pp. 840–845.

[210] A. Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature Methods* 5 (2008).

[211] A. Moudgil et al. “Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells”. In: *bioRxiv* (2019), p. 538553.

[212] R. M. Mulqueen et al. “Highly scalable generation of DNA methylation profiles in single cells”. In: *Nature Biotechnology* 36 (2018).

[213] I. Munoz-Sanjuan and A. H. Brivanlou. “Neural induction, the default model and embryonic stem cells”. In: *Nature Reviews Neuroscience* 3.4 (2002).

[214] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 9780262018029.

[215] U. Nagalakshmi et al. “The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing”. In: *Science* 320.5881 (2008), pp. 1344–1349. ISSN: 0036-8075. DOI: 10.1126/science.1158441.

[216] S. Nakajima and S. Watanabe. “Variational Bayes Solution of Linear Neural Networks and Its Generalization Performance”. In: *Neural Computation* 19.4 (2007), pp. 1112–1153.

[217] R. M. Neal. *Bayesian learning for neural networks*. 1995.

[218] A. C. Nica and E. T. Dermitzakis. “Expression quantitative trait loci: present and future”. In: *Philosophical Transactions of the Royal Society Biological Sciences* 368.1620 (2013).

[219] K. Nordstrom et al. "Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data". In: *bioRxiv* (2019). DOI: 10.1101/547596.

[220] S. Nowotschin et al. "The emergent landscape of the mouse gut endoderm at single-cell resolution". In: *Nature* 569.7756 (2019).

[221] M. Okano et al. "DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development". In: *Cell* 99.3 (1999), pp. 247–257. ISSN: 0092-8674. DOI: https://doi.org/10.1016/S0092-8674(00)81656-6.

[222] R. Okuta et al. "CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations". In: *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)*. 2017.

[223] E. Papalexi and R. Satija. "Single-cell RNA sequencing to explore immune cell heterogeneity". In: *Nature Reviews Immunology* 18 (2017).

[224] B. Papp and K. Plath. "Epigenetics of Reprogramming to Induced Pluripotency". In: *Cell* 152.6 (2013), pp. 1324–1343.

[225] B. Papp and K. Plath. "Pluripotency re-centered around Esrrb". In: *The EMBO Journal* 31.22 (2012), pp. 4255–4257. ISSN: 0261-4189. DOI: 10.1038/emboj.2012.285.

[226] A. Parle-Mcdermott and A. Harrison. "DNA Methylation: A Timeline of Methods and Applications". In: *Frontiers in Genetics* 2 (2011), p. 74. ISSN: 1664-8021. DOI: 10.3389/fgene.2011.00074.

[227] A. P. Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–1401. ISSN: 0036-8075. DOI: 10.1126/science.1254257.

[228] F. Paul et al. "Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors". In: *Cell* 163.7 (2015), pp. 1663–1677.

[229] G. Peng et al. "Molecular architecture of lineage allocation and tissue organization in early mouse embryo". In: *Nature* 572.7770 (2019), pp. 528–532.

[230] V. M. Peterson et al. "Multiplexed quantification of proteins and transcripts in single cells". In: *Nature Biotechnology* 35 (2017).

[231] S. Picelli et al. "Full-length RNA-seq from single cells using Smart-seq2". In: *Nature Protocols* 9 (2014).

[232] B. Pijuan-Sala et al. "A single-cell molecular map of mouse gastrulation and early organogenesis". In: *Nature* 566.7745 (2019), pp. 490–495.

[233] B. Pijuan-Sala et al. "Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis". In: *Nature Cell Biology* 22.4 (2020), pp. 487–497.

[234] M. Pilling. "Handbook of Applied Modelling: Non-Gaussian and Correlated Data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4 (2018), pp. 1264–1265. DOI: 10.1111/rssa.12402. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12402.

[235] O. Poirion et al. "Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage". In: *Nature Communications* 9.1 (2018), p. 4892.

[236] K. Polanski et al. "BBKNN: fast batch alignment of single cell transcriptomes". In: *Bioinformatics* 36.3 (Aug. 2019), pp. 964–965. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz625. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/3/964/32369783/btz625.pdf.

[237] S. Pott. "Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells". In: *eLife* 6 (2017). DOI: 10.7554/eLife.23203.

[238] I. Pournara and L. Wernisch. "Factor analysis for gene regulatory networks and transcription factor activity profiles". In: *BMC Bioinformatics* 8.1 (2007), p. 61.

[239] N. Qian. "On the momentum term in gradient descent learning algorithms". In: *Neural Networks* 12.1 (1999), pp. 145–151.

[240] A. C. Queirós et al. "A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact". en. In: *Leukemia* 29.3 (2015), pp. 598–605.

[241] A. Rada-Iglesias et al. "A unique chromatin signature uncovers early developmental enhancers in humans". In: *Nature* 470.7333 (2011).

[242] H. Raiffa and R. Schlaifer. *Applied statistical decision theory*. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961. ISBN: 9780875840178.

[243] A. Raj, M. Stephens, and J. K. Pritchard. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". In: *Genetics* 197.2 (2014), pp. 573–589. ISSN: 0016-6731. DOI: 10.1534/genetics.114.164350. eprint: http://www.genetics.org/content/197/2/573.full.pdf.

[244] D. Ramskold et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature Biotechnology* 30 (2012).

[245] R. Ranganath, S. Gerrish, and D. M. Blei. "Black Box Variational Inference". In: *arXiv e-prints*, arXiv:1401.0118 (2013), arXiv:1401.0118. arXiv: 1401.0118 [stat.ML].

[246] R. Ranganath et al. "An Adaptive Learning Rate for Stochastic Variational Inference". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. JMLR.org, 2013, pp. II-298–II-306.

[247] K. D. Rasmussen and K. Helin. "Role of TET enzymes in DNA methylation, development, and cancer". In: *Genes & Development* 30.7 (2016).

[248] M. Rattray et al. "Inference algorithms and learning theory for Bayesian sparse factor analysis". In: *Journal of Physics: Conference Series* 197 (2009), p. 012002. DOI: 10.1088/1742-6596/197/1/012002.

[249] A. Regev et al. "Science Forum: The Human Cell Atlas". In: *eLife* 6 (2017), e27041. DOI: 10.7554/eLife.27041.

[250] S. Remes, T. Mononen, and S. Kaski. "Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers". In: *arXiv preprint arXiv:1512.05610* (2015).

[251] J. M. Replogle et al. "Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing". In: *Nature Biotechnology* (2020).

[252] M. Ringner. "What is principal component analysis?" In: *Nat. Biotechnol.* 26 (2008), p. 303.

[253] D. Risso et al. "A general and flexible method for signal extraction from single-cell RNA-seq data". In: *Nature communications* 9.1 (2018), pp. 1–17.

[254] M. D. Ritchie et al. "Methods of integrating data to uncover genotype–phenotype interactions". In: *Nature Reviews Genetics* 16 (2015).

[255] M. E. Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015), e47–e47. ISSN: 0305-1048. DOI: 10.1093/nar/gkv007. eprint: https://academic.oup.com/nar/article-pdf/43/7/e47/7207289/gkv007.pdf.

[256] H. Robbins and S. Monro. "A stochastic approximation method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[257] M. D. Robinson and A. Oshlack. "A scaling normalization method for differential expression analysis of RNA-seq data". In: *Genome Biology* 11.3 (2010), R25.

[258] A. B. Rosenberg et al. "Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding". In: *Science* 360.6385 (Apr. 2018), p. 176.

[259] A. J. Rubin et al. "Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks". In: *Cell* 176.1 (2019), 361–376.e17.

[260] D. B. Rubin and D. T. Thayer. "EM algorithms for ML factor analysis". In: *Psychometrika* 47.1 (1982), pp. 69–76.

[261] S. Rulands et al. "Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency". In: *Cell Systems* 7.1 (2018).

[262] W. Saelens et al. "A comparison of single-cell trajectory inference methods". In: *Nature Biotechnology* 37.5 (2019).

[263] W. Saelens et al. "A comparison of single-cell trajectory inference methods: towards more accurate and robust tools". In: *bioRxiv* (2018), p. 276907.

[264] G. Sanguinetti, N. D. Lawrence, and M. Rattray. "Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities". In: *Bioinformatics* 22.22 (2006), pp. 2775–2781. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl473. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/22/22/2775/16851948/btl473.pdf.

[265] J. L. Sardina et al. "Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate". In: *Cell Stem Cell* 23.5 (2018).

[266] A. Sarkar and M. Stephens. "Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis". In: *bioRxiv* (Jan. 2020), p. 2020.04.07.030007.

[267] L. K. Saul, T. Jaakkola, and M. I. Jordan. "Mean Field Theory for Sigmoid Belief Networks". In: *arXiv e-prints*, cs/9603102 (1996), cs/9603102. arXiv: cs/9603102.

[268] N. Schaum et al. "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". In: *Nature* 562.7727 (2018).

[269] A. Scialdone et al. "Computational assignment of cell-cycle stage from single-cell transcriptome data". In: *Methods* 85 (2015).

[270] A. Scialdone et al. "Resolving early mesoderm diversification through single-cell expression profiling". en. In: *Nature* 535.7611 (2016), pp. 289–293.

[271] M. Seeger and G. Bouchard. "Fast variational Bayesian inference for non-conjugate matrix factorization models". In: *Artificial Intelligence and Statistics*. 2012, pp. 1012–1018.

[272] Y. Shan et al. "PRC2 specifies ectoderm lineages and maintains pluripotency in primed but not naive ESCs". In: *Nature Communications* 8.1 (2017).

[273] A. Singh et al. "DIABLO: from multi-omics assays to biomarker discovery, an integrative approach". In: *bioRxiv* (2018), p. 067611.

[274] S. A. Smallwood et al. "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity". In: *Nature Methods* 11 (2014).

[275] Z. D. Smith and A. Meissner. "DNA methylation: roles in mammalian development". In: *Nature Reviews Genetics* 14 (Feb. 2013).

[276] Z. D. Smith et al. "A unique regulatory phase of DNA methylation in the early mammalian embryo". In: *Nature* 484.7394 (2012), pp. 339–344.

[277] L. Solnica-Krezel and D. S. Sepich. "Gastrulation: making and shaping germ layers". en. In: *Annu. Rev. Cell Dev. Biol.* 28 (2012), pp. 687–717.

[278] C. Soneson et al. "Preprocessing choices affect RNA velocity results for droplet scRNA-seq data". In: *bioRxiv* (2020), p. 2020.03.13.990069.

[279] L. Song and G. E. Crawford. "DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells". In: *Cold Spring Harbor Protocols* 2010.2 (2010), pdb.prot5384.

[280] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. Hoboken, N.J: J. Wiley, 2003.

[281] R. Spektor et al. "methyl-ATAC-seq measures DNA methylation at accessible chromatin". In: *bioRxiv* (2018), p. 445486.

[282] S. G. Stark et al. "SCIM, Universal Single-Cell Matching with Unpaired Feature Sets". In: *bioRxiv* (2020). DOI: 10.1101/2020.06.11.146845.

[283] O. Stegle, S. Teichmann, and J. Marioni. "Computational and analytical challenges in single-cell transcriptomics". In: *Nat Rev Genet* 16 (3 2015), pp. 133–45.

[284] O. Stegle et al. "A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies". In: *PLOS Computational Biology* 6.5 (May 2010). DOI: 10.1371/journal.pcbi.1000770.

[285] O. Stegle et al. “Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses”. en. In: *Nat. Protoc.* 7.3 (2012), pp. 500–507.

[286] G. L. Stein-O’Brien et al. “Enter the Matrix: Factorization Uncovers Knowledge from Omics”. In: *Trends in Genetics* 34.10 (2018), pp. 790–805.

[287] S. M. Stigler. “The Epic Story of Maximum Likelihood”. In: *arXiv e-prints*, arXiv:0804.2996 (2008), arXiv:0804.2996. arXiv: 0804.2996 [stat.ME].

[288] M. Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. In: *Nature Methods* 14 (2017).

[289] K. Struhl and E. Segal. “Determinants of nucleosome positioning”. In: *Nature Structural Molecular Biology* 20 (2013).

[290] T. Stuart and R. Satija. “Integrative single-cell analysis”. In: *Nature Reviews Genetics* (2019).

[291] T. Stuart et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7 (2019), 1888–1902.e21. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.05.031.

[292] A. Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (2005).

[293] V. Svensson. “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38.2 (2020), pp. 147–150.

[294] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. “Exponential scaling of single-cell RNA-seq in the past decade”. In: *Nature Protocols* 13 (2018).

[295] V. Svensson et al. “Power analysis of single-cell RNA-sequencing experiments”. In: *Nature Methods* 14 (2017).

[296] P. P. L. Tam and D. A. F. Loebel. “Gene function in mouse embryogenesis: get set for gastrulation”. en. In: *Nat. Rev. Genet.* 8.5 (2007), pp. 368–381.

[297] P. P. L. Tam, E. A. Williams, and W. Y. Chan. “Gastrulation in the mouse embryo: Ultrastructural and molecular aspects of germ layer morphogenesis”. In: *Microscopy Research and Technique* 26.4 (1993), pp. 301–328. DOI: 10.1002/jemt.1070260405. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jemt.1070260405.

[298] P. P. Tam and R. R. Behringer. “Mouse gastrulation: the formation of a mammalian body plan”. In: *Mechanisms of Development* 68.1 (1997), pp. 3–25. ISSN: 0925-4773. DOI: https://doi.org/10.1016/S0925-4773(97)00123-8.

[299] A. Tambe and L. Pachter. “Barcode identification for single cell genomics”. In: *BMC Bioinformatics* 20.1 (2019), p. 32.

[300] W.-W. Tee and D. Reinberg. “Chromatin features and the epigenetic regulation of pluripotency states in ESCs”. In: *Development* 141.12 (2014), pp. 2376–2390. ISSN: 0950-1991. DOI: 10.1242/dev.096982. eprint: https://dev.biologists.org/content/141/12/2376.full.pdf.

[301] C. A. Thornton et al. “Spatially-mapped single-cell chromatin accessibility”. In: *bioRxiv* (2019), p. 815720.

[302] R. E. Thurman et al. “The accessible chromatin landscape of the human genome”. In: *Nature* 489 (2012).

[303] M. Tipping and C. Bishop. “Probabilistic Principal Component Analysis”. In: *Journal of the Royal Statistical Society* 61(3) (1999), pp. 611–22.

[304] M. K. Titsias and M. Lázaro-Gredilla. “Spike and slab variational inference for multi-task and multiple kernel learning”. In: *Advances in neural information processing systems*. 2011, pp. 2339–2347.

[305] J. Tosic et al. “Eomes and Brachyury control pluripotency exit and germ-layer segregation by changing the chromatin state”. In: *Nature Cell Biology* 21.12 (2019), pp. 1518–1531.

[306] J. Tosic et al. “Eomes and Brachyury control pluripotency exit and germ-layer segregation by changing the chromatin state”. In: *Nature Cell Biology* 21.12 (2019).

[307] M. Tosolini and A. Jouneau. “Acquiring Ground State Pluripotency: Switching Mouse Embryonic Stem Cells from Serum/LIF Medium to 2i/LIF Medium”. In: *Embryonic Stem Cell Protocols*. Springer New York, 2016, pp. 41–48. ISBN: 978-1-4939-2954-2. DOI: 10.1007/7651_2015_207.

[308] F. W. Townes et al. “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1 (2019), p. 295.

[309] H. T. N. Tran et al. “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”. In: *Genome Biology* 21.1 (2020), p. 12.

[310] C. Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nature Biotechnology* 32 (2014).

[311] O. Troyanskaya et al. “Missing value estimation methods for DNA microarrays”. In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[312] M. Tsompana and M. J. Buck. “Chromatin accessibility: a window into the genome”. In: *Epigenetics & Chromatin* 7.1 (2014), p. 33.

[313] A. Tsumura et al. “Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b”. In: *Genes to Cells* 11.7 (2006).

[314] C. A. Vallejos, J. C. Marioni, and S. Richardson. “BASiCS: Bayesian Analysis of Single-Cell Sequencing Data”. In: *PLOS Computational Biology* 11.6 (June 2015), e1004333–.

[315] K. Van den Berge et al. “Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications”. In: *Genome Biology* 19.1 (2018), p. 24.

[316] Y. Vasconcelos et al. “Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes”. en. In: *Leukemia* 19.11 (2005), pp. 2002–2005.

[317] N. L. Vastenhouw et al. “Chromatin signature of embryonic pluripotency is established during genome activation”. In: *Nature* 464.7290 (2010).

[318] S. Virtanen et al. “Bayesian group factor analysis”. In: *Artificial Intelligence and Statistics*. 2012, pp. 1269–1277.

[319] B. Wang et al. “Similarity network fusion for aggregating data types on a genomic scale”. In: *Nature Methods* 11 (Jan. 2014).

[320] C. Wang. “Variational Bayesian Approach to Canonical Correlation Analysis”. In: *IEEE Trans Neural Netw* 3.18 (2007).

[321] Y. Wang et al. “Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos”. In: *bioRxiv* (2019). DOI: 10.1101/803890. eprint: https://www.biorxiv.org/content/early/2019/10/14/803890.full.pdf.

[322] Y. J. Wang et al. “Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues”. In: *bioRxiv* (2019), p. 541433.

[323] C. Weinreb et al. “Fundamental limits on dynamic inference from single-cell snapshots”. In: *Proceedings of the National Academy of Sciences* 115.10 (2018), E2467.

[324] J. D. Welch, A. J. Hartemink, and J. F. Prins. “MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics”. In: *Genome Biology* 18.1 (2017), p. 138.

[325] J. D. Welch et al. “Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity”. In: *Cell* 177.7 (2019), 1873–1887.e17. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2019.05.006.

[326] J. Wen et al. “Single-cell analysis reveals lineage segregation in early post-implantation mouse embryos”. In: *J. Biol. Chem.* 292.23 (2017), pp. 9840–9854.

[327] M. G. P. van der Wijst et al. “Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs”. In: *Nature Genetics* 50.4 (2018), pp. 493–497.

[328] F. A. Wolf, P. Angerer, and F. J. Theis. “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19.1 (2018), p. 15.

[329] J. Wu et al. “The landscape of accessible chromatin in mammalian preimplantation embryos”. In: *Nature* 534 (2016).

[330] C. Xia et al. “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression”. In: *Proceedings of the National Academy of Sciences* 116.39 (2019), p. 19490.

[331] Y. Xiang et al. “Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency”. In: *Nature Genetics* 52.1 (2020).

[332] C. Xu, D. Tao, and C. Xu. “A Survey on Multi-view Learning”. In: *arXiv e-prints*, arXiv:1304.5634 (2013), arXiv:1304.5634. arXiv: 1304.5634 [cs.LG].

[333] F. Yue et al. “A comparative encyclopedia of DNA elements in the mouse genome”. In: *Nature* 515.7527 (2014), pp. 355–364.

[334] M. D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. 2012. arXiv: 1212.5701 [cs.LG].

[335] I. S. L. Zeng and T. Lumley. “Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science)”. In: *Bioinformatics and Biology Insights* 12 (2018).

[336] C. Zhang et al. “Advances in Variational Inference”. In: *arXiv e-prints*, arXiv:1711.05597 (2017), arXiv:1711.05597. arXiv: 1711.05597.

[337] C. Zhang et al. “A Study on Overfitting in Deep Reinforcement Learning”. In: *arXiv e-prints*, arXiv:1804.06893 (2018), arXiv:1804.06893. arXiv: 1804.06893.

[338] X. Zhang et al. “Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems”. In: *Molecular Cell* 73.1 (2019), 130–142.e5. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2018.10.020.

[339] Z. Zhang et al. “Opening the black box of neural networks: methods for interpreting neural network models in clinical applications.” In: *Annals of translational medicine* 6 11 (2018), p. 216.

[340] S. Zhao et al. “Bayesian group factor analysis with structured sparsity”. In: *Journal of Machine Learning Research* 17.196 (2016), pp. 1–47.

[341] Y. Zhao and B. A. Garcia. “Comprehensive Catalog of Currently Documented Histone Modifications”. In: *Cold Spring Harbor Perspectives in Biology* 7.9 (2015).

[342] G. X. Y. Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nature Communications* 8 (2017).

[343] C. Ziegenhain et al. “Comparative Analysis of Single-Cell RNA Sequencing Methods”. In: *Molecular Cell* 65.4 (2017), 631–643.e4.

[344] R. Zilionis et al. “Single-cell barcoding and sequencing using droplet microfluidics”. In: *Nature Protocols* 12 (2016).

[345] I. Zvetkova et al. “Global hypomethylation of the genome in XX embryonic stem cells”. In: *Nature Genetics* 37 (2005).