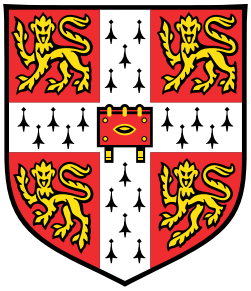
Statistical methods for the integrative analysis of single-cell multi-omics data



Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

April 2020

0.1 Introduction

0.1.1 Multi-omics at single cell resolution

Next-generation sequencing technologies have revolutionised the study of biological systems by enabling the genome-wide profiling of molecular layers in un unbiased manner, including the genome[72] the epigenome[73] and the transcriptome[149, 14, 177, 172], among others. However, bulk sequencing approaches rely on the pool of large number of cells to report an average molecular readout, and are hence limited for the study of complex biological processes where heterogeneity is expected at single cell resolution [80, 183, 187].
The progressive development of low-input sequencing techniques resulted in an explosion of single-cell sequencing technologies, mostly for the transcriptome. In contrast to bulk protocols, single-cell techniques provide an unprecedented opportunity to study the molecular variation associated with cellular heterogeneity, lineage diversification and cell fate commitment [128].

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, driven by incremental optimisations of reagent volumes and consumable costs, as well as profound changes in the nature of the technology [**Svensson2019**]. The earliest high-throughput 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 [39]. 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 [128, 80, 183, 187]. 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[208].

While the large majority of single-cell studies are focused on capturing RNA expression information, transcriptomic readouts provide a single dimension of cellular heterogeneity and hence contain limited information to characterise the molecular determinants of phenotypic variation [213]. 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 [80, 117, 19].

To get a better insight into the epigenetics of cell fate commitment, significant effort has been placed to obtain epigenetic measurements at single-cell resolution by adapting bulk methods to low-input material. This has been particularly successful for chromatin accessibility. Due to its cost-effective strategy, single-cell ATAC-seq (scATAC-seq) has become the most popular technique to map open chromatin, and is also available in an efficient commercial platform. [56, 40, 45].
Other molecular layers have also been queried through the lens of single cells, including DNA methylation[226], histone modifications [132], chromatin conformation [132], proteomics [231] and lipidomics [246].

Despite is profuse success in studying molecular variation, no single "-omics" technology can capture the intricacy of complex biological mechanisms. Nonetheless, the collective information has the potential to draw a more comprehensive picture of biological processes [90, 213]. In particular, multi-omics (or multi-modal) assays have the potential to go beyond snapshots and provide a more dynamic, perhaps even mechanistic, understanding of the connection between molecular layers. Motivated by this, multi-omic data sets are receiving increasing interest across a wide range of biological domains, including cancer biology [3, 78], regulatory genomics [43], microbiology [121] or host-pathogen interactions [228].

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 [213]. This strategy is also used with single-cell assays, but it has the important downside that the different molecular layers cannot be unambiguously linked, hence limiting the insights that can be inferred from the data. The ultimate goal in single-cell sequencing is to obtain different molecular readouts from the same cell.

Multi-modal measurements at single-cell resolution can be obtained using a variety of strategies, some of which will be discussed in this thesis. 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 a more comprehensive insights into the complex variation that underlie cellular populations [239, 49].

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*. However, their development is still in pilot stages and there is no commercial platform available, limiting its widespread use by the community. Furthermore, common challenges in (uni-modal) single-cell assays such as low coverage and high levels of technical noise become exacerbated when doing multi-modal profiling. Quoting a befitting sentence from 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.* Thus, one of the biggest challenge in integrative multi-modal analysis is to develop statistical frameworks that are capable of uncovering biological signal across multiple data modalities while overcoming all the technical biases and missing information that are inherent to single-cell experiments.

0.1.2 Integrative analysis of multi-modal assays

From the computational perspective, the rapid development of single-cell technologies is introducing unprecedented challenges for the statistical community, and novel computational methods need to be developed (or adapted) for interrogating the data generated [233].
The vast majority of methods for single-cell data analysis are focused on scRNA-seq. These include normalisation[**Lun2018**], feature selection[252], differential expression [119], clustering[122], cell type recognition [1], pseudotime inference [86], detection of gene regulatory networks and batch correction [85], among others. Analysis tools have been wrapped into popular platforms such as

Seurat [36], Scater [163] and Scanpy [265].
Despite the explosion of statistical methods for scRNA-seq data analysis, to date very few methods have been published with the aim to perform data integration of single-cell multi-modal assays. This is probably due to the lack of large-scale data sets to apply and benchmark methods. But also, given the high levels of missing information, the inherent amounts of technical noise and the potentially large number of cells, the integrative analysis of multi-modal measurements is arguably one of the most challenging problems in single-cell biology.

The first step when performing data integration is to consider (or not) a common coordinate framework to anchor the different data modalities. This defines three broad types of strategies for single-cell data integration:

- Samples are the common coordinate framework: when the different data modalities are derived from the same cell. We call this *matched* multi-omics and the main advantage is that here unambiguous assignment between the molecular profiles. Nevertheless, as mentioned above, such assays are difficult, expensive and currently less adopted by the community. In this case, the aim of the computational strategies is to distinguish the axis of cell-to-cell heterogeneity that are coordinated versus the variation that is uncoordinated across molecular layers.

- Genomic features are the common coordinate framework: when the different data modalities are not derived from the same cell. We call this *non-matched* multi-omics and the main advantage is that it is significantly easier and cheaper to obtain, and as a result most of the data sets to date belong to this category. In this case, the aim of computational strategies is generally to find a common manifold and identify cell anchors between the two modalities. An example of this type of data integration is when having scRNA-seq and scATAC-seq experiments from different sets of cells. Genes can be defined as the common coordinate framework by quantifying mRNA expression and summarising chromatin accessibility at the gene promoter or gene body level.

- No common coordinate framework (in the high-dimensional space): when the different data modalities are not anchored by any of the two axis in the high-dimensional space (i.e. cells or genes). In this scenario, methods could exploit the assumption of a common manifold in a potential *low-dimensional* space, for example when cells are sampled from the same differentiation trajectory. In the general and multivariate case, this is arguably the most complex data integration task and, to my current knowledge, no convincing and principled methods have been published to date.

Once the common coordinate framework is defined, the actual data integration strategy can be classified into two classes that can be categorised as *local* and *global*. This notation is inspired from integrative approaches that have been pursuit over the last years at the bulk level[213], but this classification remains applicable to the single-cell domain. In fact, as we shall demonstrate in this thesis, some methods designed for bulk data sets can be applicable to single-cell data with minor modifications.
Local analysis refers to the study of hierarchical associations between individual features from

different molecular layers. Prominent examples are genome-wide association studies (GWAS) in combination with expression quantitative trait loci (eQTL), methylation QTLs or protein QTLs [264, 43, 191, 26]. While eminently useful for characterising genetic variants, such association studies are inherently local and have a limited capacity to discover global maps of molecular heterogeneity that typically result from complex interactions between features. In addition, such approaches are challenging in the multi-omics setting due to the massive multiple testing problem [240].
Global analysis on the other hand try to extract patterns from the full data set. This can be done by direct concatenation of all data modalities followed by the use of traditional statistical methods. Additional 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 [137, 261]. Nonetheless, both of these approaches entail important setbacks that will be discussed and addressed in this thesis.

0.1.3 Thesis overview

In this PhD thesis I sought to develop computational strategies for data integration using single-cell multi-omics data. In particular my research focued on the *matched* case, when cells are the common coordinate framework.

In Chapter 1 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 readouts using a simple prototypic experiment, and we show how scNMT-seq can be used to study coordinated epigenetic and transcriptomic heterogeneity along a simple differentiation process.

In Chapter 2 I present Multi-Omics Factor Analysis (MOFA), a statistical framework for the integration of multi-omics data sets. 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 a multi-omics data set. Once the model is trained, the inferred low-dimensional space can be interpreted using a tool-kit of downstream analysis that include multiple visualisations, clustering, imputation or prediction of clinical outcomes. First, we validate the different model features using simulated data. Second, we apply MOFA to a multi-omics study of 200 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 3 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 ectoderm cells remains in a *default* state, resembling earlier stage epiblast cells is epigenetically established in the early epiblast. This work provides a comprehensive insight into the molecular logic for a hierarchical emergence of the primary germ layers, revealing underlying molecular constituents of the Waddington's landscape.

In Chapter 4 I propose an improved formulation of the MOFA framework presented in Chapter 2 with the aim of performing integrative analysis of large-scale (single-cell) data sets across multiple studies/conditions as well as data modalities. To tailor MOFA to the statistical challenges of single-cell data, 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 single-cell studies. First, we benchmark the new features of the model using simulated data. Next, we use a single-cell DNA methylation data set of neurons from mouse frontal cortex to demonstrate how from a seemingly unimodal data set, one can investigate hypothesis using a multi-group and multi-view setting. Finally, we apply MOFA to the scNMT-seq data set generated in Chapter 3, revealing underlying sources of molecular variation associated with early cell fate decisions.

Finally, Chapter 5 summarises this thesis and provides an outlook of future research.

Chapter 1

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

1.1 Introduction to single-cell (multi-) omics sequencing

Single-cell profiling techniques have provided an unprecented opportunity to study cellular heterogeneity at multiple molecular levels. The maturation of single-cell RNA-sequencing technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment [128, 80, 183, 187]. Yet, the accompanying epigenetic changes and the role of other molecular layers in driving cell fate decisions still remains poorly understood. Consequently, the profiling the epigenome at the single-cell level is receiving increasing attention, but without associated transcriptomic readouts, the conclusions that can be extracted from epigenetic measurements are limited [239, 117, 80].

1.1.1 Single-cell RNA sequencing

single-cell RNA sequencing (scRNA-seq) protocols differ extensively in terms of scalability, costs and sensitivity [241, 136]. Broadly speaking, they can be classified into plate-based and droplet-based methods. In plate-based methods such as CEL-seq [89] and Smart-seq[203, 190], 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 splice variants[103], allele-specific fractions[58] and RNA velocity information [135].

Droplet-based methods are based on the use of droplet microfluidics technology [274]. 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. These class of methods

include InDrop [126, 281], Drop-seq[158] and the comercial 10x Genomics Chromium [279]. As a trade-off, the increased high throughput of droplet-based approaches comes at the expense of reduced sensitivity[280, 263, 242].

More recently, a third type of scRNA-seq methodology emerged based on a combinatorial cellular indexing strategy [37, 215, 39], which has permitted the sequencing of more than a milion cells in a single experiment for a fraction of the cost of other methods, yet with much lower sensitivity.

1.1.2 Single-cell sequencing of the epigenome

While the large majority of single-cell studies are focused on capturing the mRNA expression, transcriptomic readouts provide a single dimension of cellular heterogeneity and hence contain limited information to characterise the molecular determinants of phenotypic variation [213]. 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 [80, 117, 19].

1.1.2.1 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 [109, 186, 139, 227]. Its classical roles include the silencing of repeating elements, inactivation of the X chromosome, gene imprinting, and repression of gene expression [111]. Consistently, the disruption of the DNA methylation machinery is associated with multiple dysfunctions, including cancer [16], autoimmune diseases [152] and neurological disorders [6].
In mammalian genomes, DNA methylation predominantly occurs at CpG dinucleotides (mCG). The presence of DNA methylation in non-CpG contexts (mCH) has been confirmed, albeit its functional role remains controversial [93, 202, 150].

Alongside developments in scRNA-seq technologies, protocols for the profiling of DNA methylation in single cells also emerged from its bulk counterparts (Figure 1.1), most notably bisulfite sequencing (BS-seq) [226, 83, 79, 70]. 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 eventually to thymine (T), after PCR amplification), leaving 5-methylcytosine residues intact. The resulting C→T transitions can then be detected by DNA sequencing [73, 48, 46]. 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, [226] adapted the post-bisulfite adaptor tagging (PBAT) protocol with multiple rounds of 3' random primer amplification (**??**). When the bisulfite treatment is performed before ligation of adaptors, rather than afterwards, loss of adapter-tagged molecules is minimised, unveiling the potential to use scBS-seq from low-input material. In a proof of concept study, [226] applied scBS-seq on ovulated metaphase II oocytes and mouse ESCs, reporting an average coverage of 3.7 million CpG dinucleotides (17.7%) per cell.

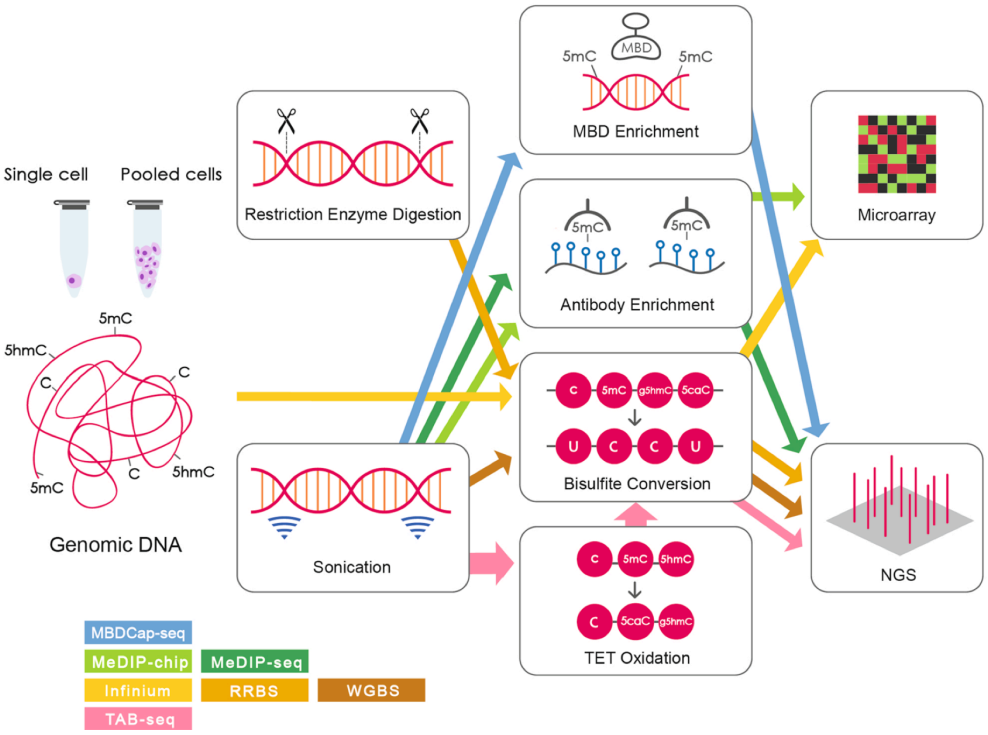


Figure 1.1: Workflow of DNA methylation profiling protocols. Reprinted from [269]

Alongside scBS, other bulk sequencing methods were also adapted to the single cell resolution, with different trade-offs between coverage and costs. For instance, [82] adapted the reduced-representation bisulfite sequencing (RRBS-seq) to low starting material by performing all experimental steps before PCR amplification into a single tube. The key principle behind RRBS-seq is to digest the DNA with a restriction endonuclease, followed by a size-selection strategy to enrich for CpG-dense areas [165]. This approach significantly reduces sequencing costs at the expense of low coverage in CpG-poor genomic areas, which include repetitive elements, gene bodies and enhancer elements.

1.1.2.2 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 unit being the nucleosome, which consists on a string of $\approx$ 150bp of DNA wrapped around histone proteins, with linker DNA of $\approx$ 80bp connecting them [127, 255]. The positioning of the nucleosomes in the nucleus provide an important layer of gene regulation, mostly by exposing or sheltering transcription factors binding sites [107]. In general, active regulatory regions tend to have low occupancy of nucleosomes, whereas inactive regions show a high density of nucleosomes [238]. 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 map chromatin accessibility in a genome-wide and high-throughput manner (Figure 1.2): DNase sequencing (DNase-seq) [229], transposase-accessible chromatin followed by sequencing (ATAC-seq) [32] and Nucleosome

Occupancy and Methylome-sequencing (NOMe-seq) [116]. A systematic comparison with a controlled experimental design can be found in [180].

- **DNase-seq**: the chromatin is incubated with DNAse I, an enzyme that in low concentrations cuts nucleosome-free regions. Hence accessible sites are released and sequenced [229]. Although this methodology became one of the gold standards to map chromatin accessibility by the ENCODE consortium [50, 247], it has now been reported that DNase I introduces significant cleavage biases, thus affecting its reliability to infer transcription factor footprints [92].

- **ATAC-seq**: the chromatin is incubated with 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 arguably displaced DNase-seq as the *de facto* method for profiling chromatin accessibility due to its fast and sensitive protocol [30, 255, 180].

- **NOMe-seq**: follows a very different strategy than the previous technologies. The idea is to incubate cells 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 [120]. Hence, after M.CviPI treatment followed by bisulfite sequencing, GpC methylation marks can be interpreted as direct read outs for chromatin accessibility. [116]. 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 ($\approx$ 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 [116, 195, 180]. Third, missing data 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%) and CGC positions (27%), as we will discussed later on.

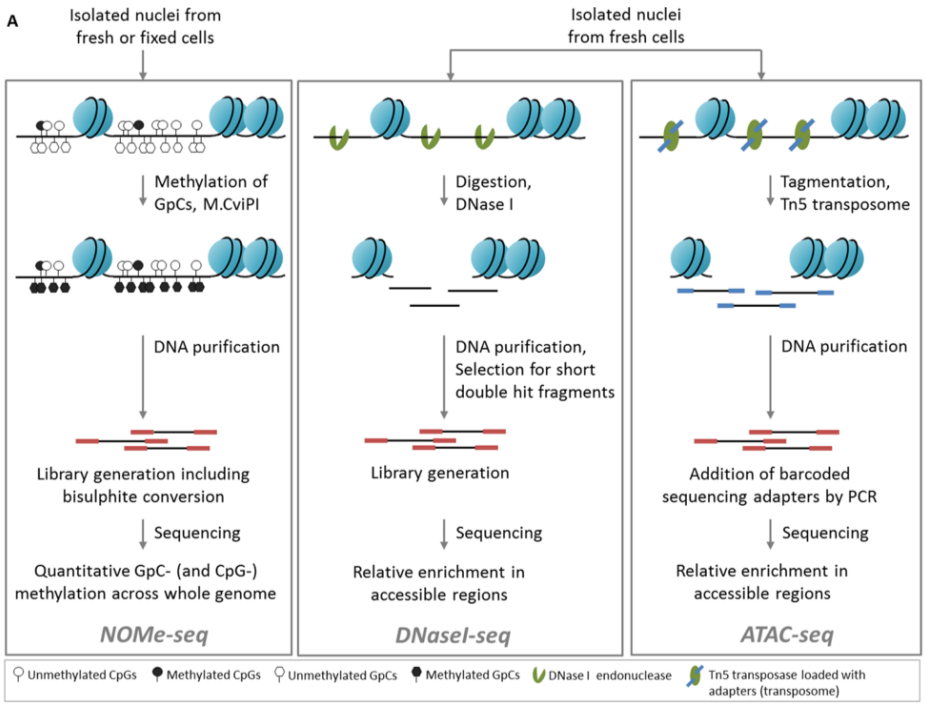


Figure 1.2: High-level overview of the workflows for the three main chromatin accessibility assays: NOMe-seq, DNase-seq and ATAC-seq. Reprinted from [180].

As with DNA methylation, single-cell profiling methods for chromatin accessibility also emerged from its bulk counterparts, including ATAC-seq[31], NOMe-seq [195] and DNase-seq [108]. Due to its cost-effective strategy, single-cell ATAC-seq (scATAC-seq) has become the most popular technique to map open chromatin [56, 40, 45]. Compared to bulk ATAC-seq, scATAC-seq libraries are notably sparse. In a saturated library, [56] reported a range of $\approx$ 500 to $\approx$ 70,000 mapped reads per cell, with a median of $\approx$ 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, [55] generated an atlas of chromatin accessibility for different mouse tissues, defining the first *in vivo* landscape of the regulatory genome single-cell resolution.

1.1.3 Multi-modal single-cell sequencing

Cellular phenotypes result from the combination of multiple sources of biological information. 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 [90, 213]. In addition, multi-omics assays have the potential to go beyond snapshots to provide a more dynamic, perhaps even mechanistic, understanding of the connection between molecular layers.

Interestingly, recent technological advances have enabled the profiling of multiple omics in the same single cell. As reviewed in [239, 42], multi-modal measurements can be obtained using four broad strategies:

- **Application of a non-destructive assay before a destructive assay**: a prominent example is the sorting of cells based on protein surface markers using (multiparameter) fluorescence-activated cell sorting (FACS) followed by high-throughput sequencing [188]. 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) [156]. 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 the simultaneous profiling of transcriptomic measurements with (epi)-genomic measurements, including DNA sequence, copy number variation, DNA methylation or chromatin accessibility [156, 101, 8, 102].

- **Conversion of different molecular layers to a common format that can be measured using the same readout**: prominent examples are the simultaneous measure of surface proteins and mRNA expression as in Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq[237]) and RNA expression and protein sequencing assay (REAP-seq[189]). 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.
  A second prominent example is NOMe-seq, described in Section 1.1.2.2. 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 have proven successful, they still face numerous difficulties, both from the experimental and the computational front, including limited scalability, low coverage and high levels of technical noise. These difficulties, also inherent to single-cell uni-modal techniques, generally get exacerbated when doing multi-modal profiling. Quoting Cole Trapnell: *When you do a multi-omic assay, you are combining all the bad things from multiple protocols.*

A clear example is sci-CAR [38], 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 approaches that reported, for the first time, the profiling of both modalities in thousands of cells. However, the chromatin accessibility modality yielded $\tilde{1}$0-fold less complexity than (already sparse) scATAC-seq experiments.

I envision that a significant effort will be placed in the next years 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 use of computational methods that integrate multi-modal measurements from different sets of cells will be a cornerstone of single-cell analysis.

1.2 scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation and RNA expression 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 [47].
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 thesis, except for the non-linear chromatin accessibility profiles, which was done by Andreas Kapourani. 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.

1.2.1 Description of the experimental protocol

scNMT-seq builds upon two previous multi-modal protocols: single-cell Methylation and Transcriptome sequencing (scM&T-seq) [8] and Nucleosome Occupancy and Methylation sequencing (NOMe-seq) [116, 195]. An overview of the protocol is shown in Figure 1.3.

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[120, 116]. 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[120, 116].

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[226], whereas the RNA fraction undergoes Smart-seq2 [190].

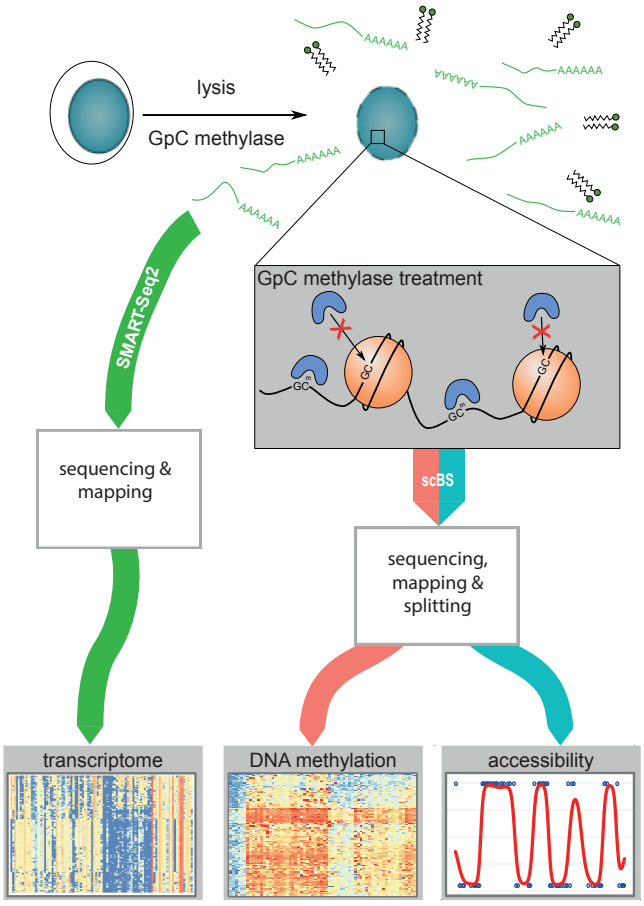


Figure 1.3: scNMT-seq protocol overview.
In the first step, cells are isolated and lysed. Second, cells are incubated with a GpC methyltransferase. Third, the RNA fraction is separated using oligo-dT probes and sequenced using Smart-seq2. The DNA fraction undergoes scBS-seq library preparation and sequencing. Finally, CpG Methylation and GpC chromatin accessibility data are separated computationally.

As discussed in Section 1.1.2.2, 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 [116, 195, 180]. 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 [180].

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%), 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 [116]. This filtering step reduces the number of genome-wide cytosines that can be assayed from 22 million to 11 million.

1.2.2 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 [131] 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. After mapping, a new round of quality control is performed per cell based on mapping efficiency, bisulfite conversion efficiency and library size.
Finally, methylation calls for each CpG and GpC site are calculated after removal of duplicate alignments. Following the approach of [226], individual CpG or GpC sites in each cell are modelled using a binomial model where the number of successes is the number of methylated reads and the number of trials is the total number of reads. A CpG methylation or GpC accessibility rate for each site and cell is calculated by maximum likelihood.

1.2.3 Validation

1.2.3.1 Coverage

We validated scNMT-seq in 70 EL16 mouse embryonic stem cells (ESCs), together with three 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 [226, 8, 71].

First, we compared the theoretical maximum coverage that could be achieved with the empirical coverage (Figure 1.4). 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.

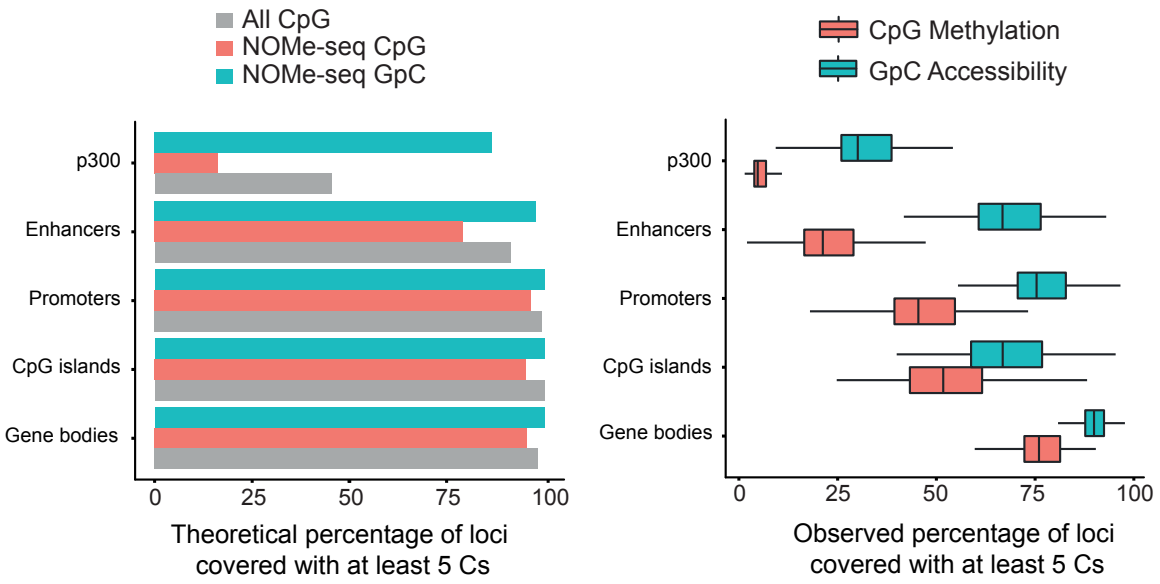


Figure 1.4: Coverage statistics for CpG DNA methylation and GpC chromatin accessibility.
(a) Fraction of loci with at least 5 CpG (red) or GpC (blue) dinucleotides (y-axis) per genomic context (x-axis), after exclusion of the conflictive trinucleotides. The grey bar shows the total number of CpGs without exclusion of trinucleotides. (b) Empirical coverage (y-axis) per genomic context (x-axis) 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 [8] (Figure 1.4), where the the conflictive 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 downsampling the number of reads.

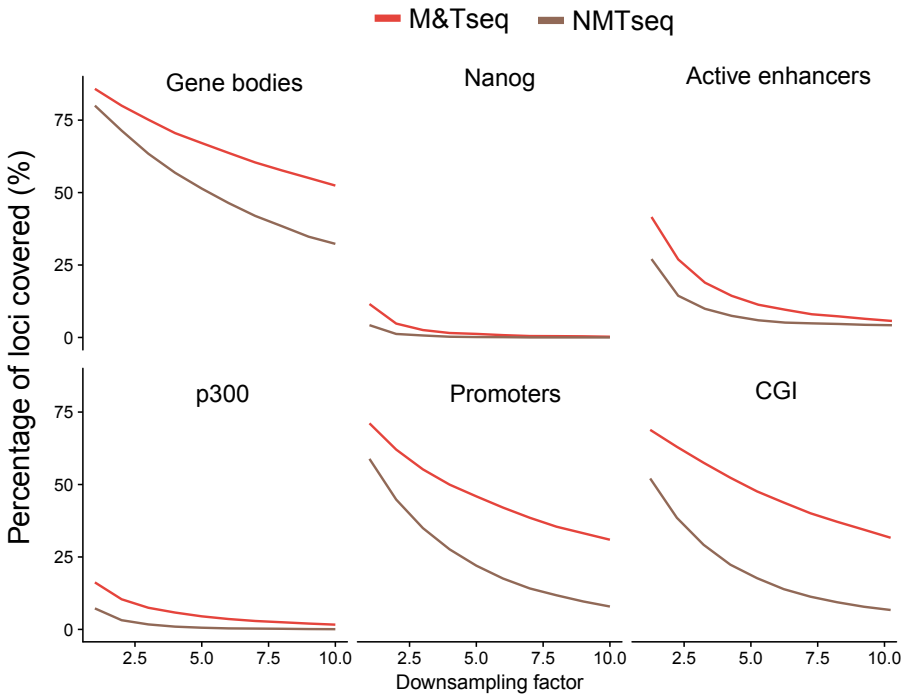


Figure 1.5: Comparison of the empirical coverage of DNA methylation with scM&T-seq [8].
The y-axis displays the fraction of loci covered with at least 5 CpG sites. The x-axis displays the downsampling factor. To facilitate the comparison, we selected two cells that were sequenced at equivalent depth.

1.2.3.2 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 data sets from the same cell lines profiled with similar technologies, including scM&T-seq[8], scBS-seq[226] and bulk BS-seq[71]. We find that most of the variation is not attributed to the technology but to differences in culture condition (2i vs serum, captured by PC1). This result is expected, as cells grown in 2i media remain in a native pluripotency state that is associated with genome-wide DNA hypomethylation [71]. Interestingly, the serum-cultured cells processed in this study overlapped with 2i-cultured cells from previous data sets. 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 [8, 226, 71]). Previous studies have sown that female ESCs tend to show lower levels of mean global methylation, which is consistent with a more pluripotent phenotype [282].

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 ($R = 0.74$).

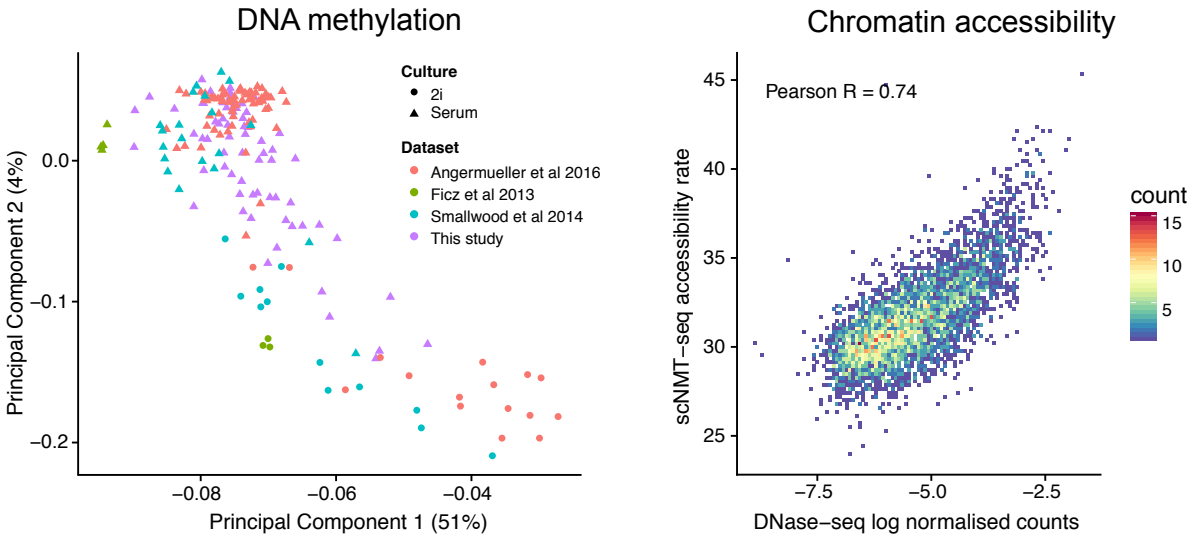


Figure 1.6: Comparison of unsupervised genome-wide quantifications to published data sets.
(a) Principal Component Analysis of 1kb running windows. Missing values were imputed using the average methylation rate per locus.
(b) Scatter plot of chromatin accessibility quantified over 10kb running windows 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.

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

We pseudobulked the data across all cells and we 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 [195]. As a control, we observe that cells which did not receive M.CviPI treatment showed globally low GpC methylation levels ($\approx 2\%$, **??**).

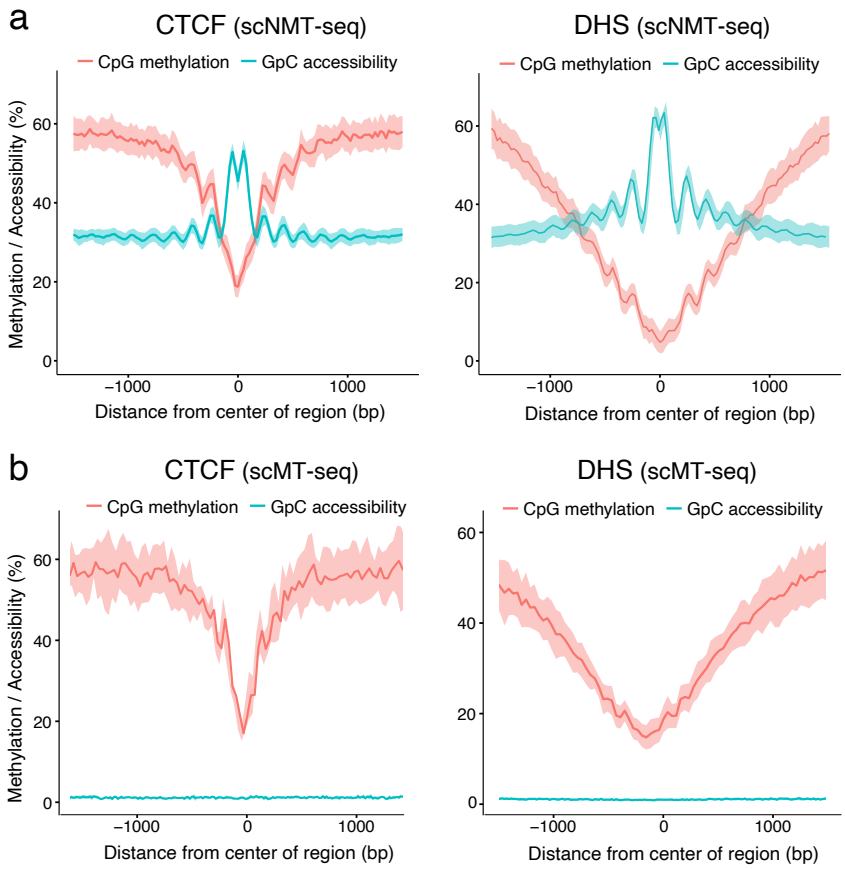


Figure 1.7: Accessibility and methylation profiles in regulatory genomic contexts. First, we pseudobulk the data set by pooling information across all cells. Next, we compute running averages of the CpG methylation (red) and the GpC accessibility (blue) in consecutive non-overlapping 50bp windows. Solid 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

Next, 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 [247, 8]. To get a measure of the 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 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.

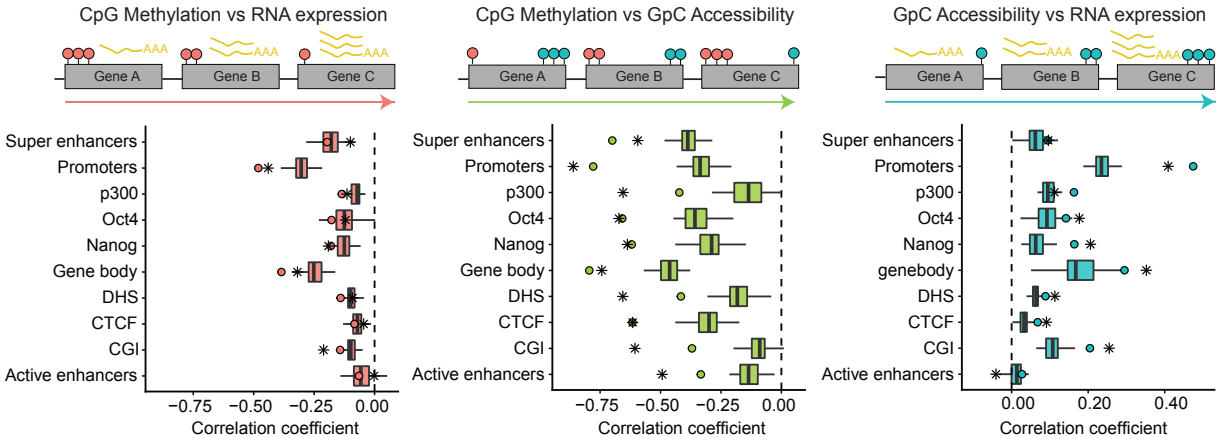


Figure 1.8: Quantification of linear associations between molecular layers.
The top diagram illustrates the computation of an association test per cell (across all loci in a given genomic context). The left panel shows DNA methylation versus RNA expression. The middle panel shows DNA methylation versus 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 [71, 50], respectively.

1.2.4 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*).

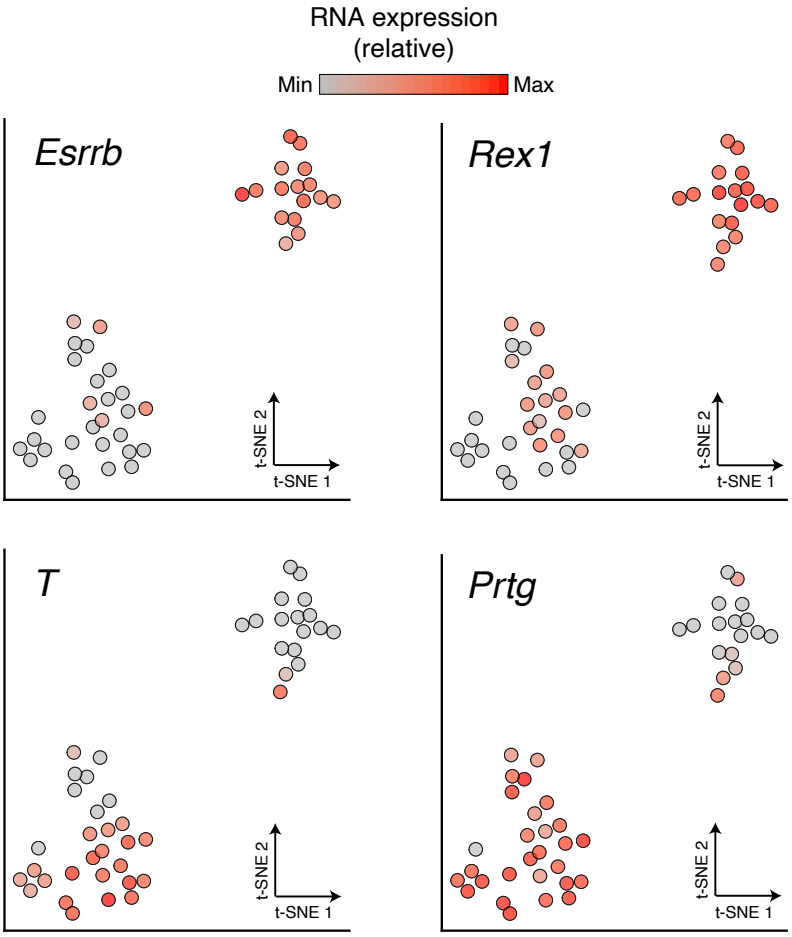


Figure 1.9: t-SNE Dimensionality reduction on the RNA expression profiles for the embryoid body cells.
The scatter plots show a t-SNE representation of the EB data. Cells are coloured based on expression of pluripotency factors (top) and differentiation markers (bottom).

Next, we tested for locus-specific associations between pairwise combinations of molecular layers (correlation across cells, Figure 1.10).

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 [185]. A previous study [8], identified a super enhancer near the gene that showed 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

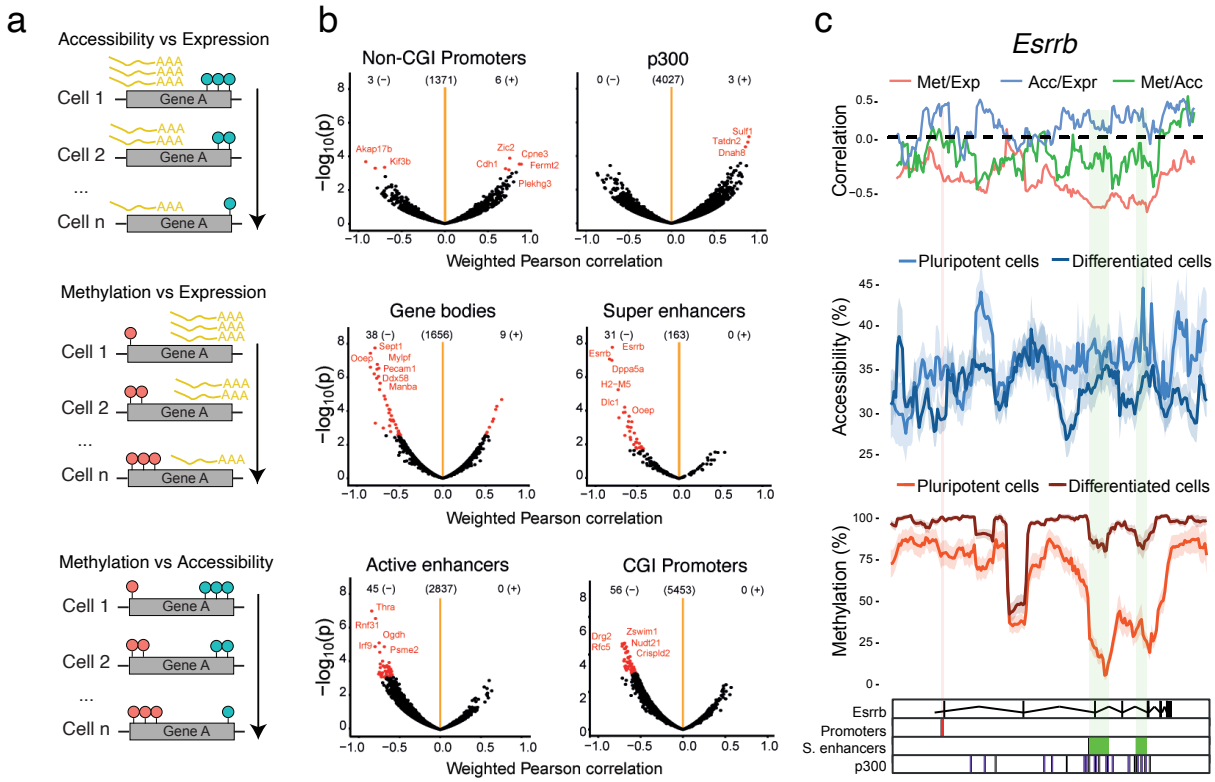


Figure 1.10: scNMT-seq enables the discovery of novel associations between transcriptomics and epigenetics at individual loci.
(a) Illustration for the correlation analysis, which results in one association test per locus (across cells).
(b) Pearson correlation coefficient (x-axis) and log10 p-value (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.

1.2.5 Inference of non-linear chromatin accessibility profiles at single nucleotide resolution

A clear advantage of scNMT-seq is the high resolution of the its chromatin accessibility readouts, namely a binary output for each observed GpC dinucleotide. As illustrated in Figure 1.7, GpC accessibility measurements rate extremely dynamic and display complex oscillatory patterns, likely due to presence of nucleosomes. This makes our approach of quantifying rates over a fixed genomic window not appropriate to capture the complexity of accessibility data. Therefore, we next attempted to exploit this high-resolution information to infer non-linear chromatin accessibility profiles at individual promoters.

The approach we followed is based on BPRMeth [112], a generalized linear regression model with Gaussian basis functions, coupled with a Bernoulli likelihood. A model was fit for every gene and

every cell, provided enough coverage (at least 10 GpC sites observed per gene across 40% of cells). Examples of inferred regression patterns are shown below:

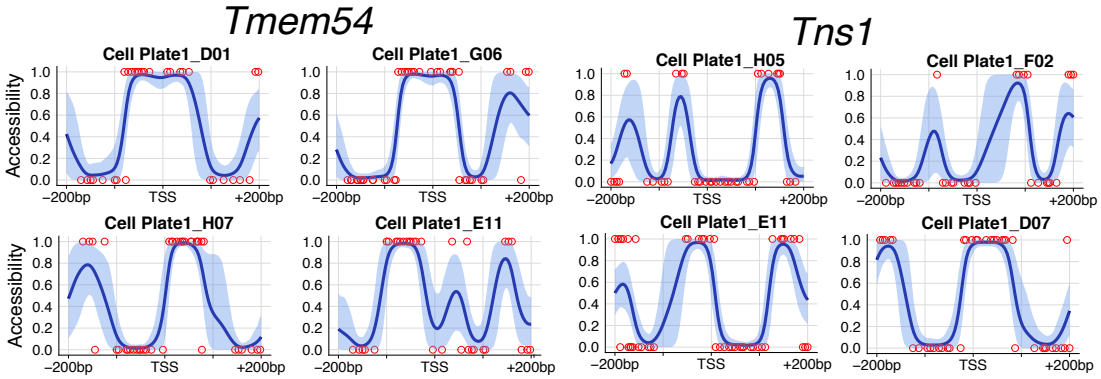


Figure 1.11: Illustrative examples of single-cell accessibility profiles around the transcription start site.
Shown are representative profiles for two genes, *Tmem54* and *Tns1*. Each panel corresponds to separate cell. The y-axis displays the binary GpC accessibility values, with 1 being accessibility and 0 inaccessible. The x-axis displays the genomic region around the TSS (200bp upstream and downstream). The blue area depicts the inferred (non-linear) accessibility profile using the BPRMeth model [112].

As a first validation step, we showed that the accessibility profiles inferred around the transcription start site (TSS) lead to a better prediction of RNA expression than using conventional accessibility rates:

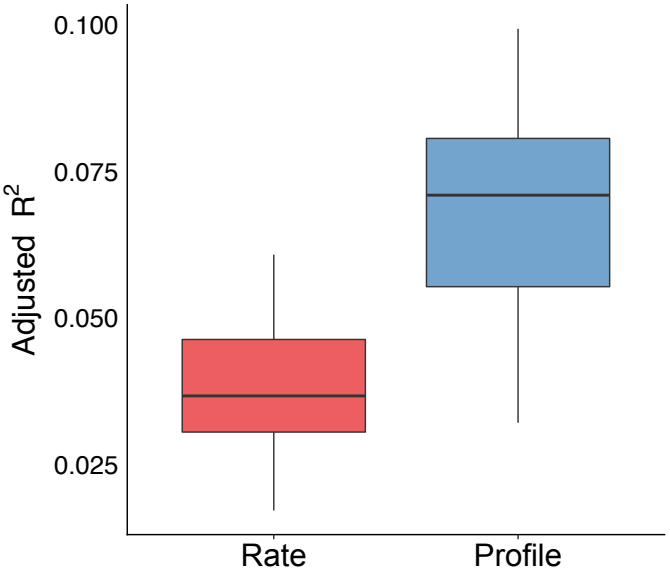


Figure 1.12: Prediction of RNA expression using conventional accessibility rates (red) and non-linear accessibility profiles (blue).
The y-axis displays the adjusted R^2 between observed RNA expression and predicted RNA expression. We fit a linear model per cell (across genes) where the response variable is RNA expression (log normalised counts) and the covariates are either the accessibility rate or the weights of the basis functions in BPRmeth.

Consistently, when inspecting individual genes we observe that highly expressed genes show characteristic patterns of nucleosome depleted regions around the TSS, whereas lowly expressed genes show low levels of chromatin accessibility:

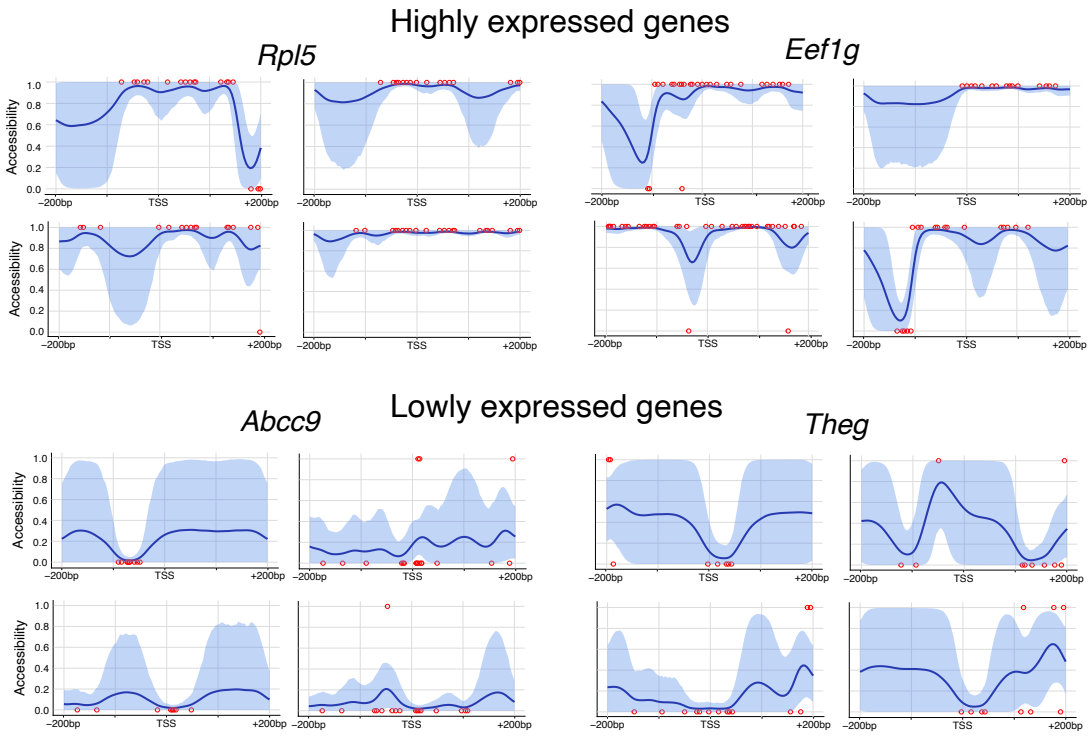


Figure 1.13: Single-cell accessibility profiles of representative genes with high and low expression levels.
Shown are *Rpl5* and *Eef1g* (high expression levels, top); *Abcc9* and *Theg* (low expression levels, bottom). Each panel corresponds to a separate cell. Axis are the same as in Figure 1.11.

Next, we attempted at linking the heterogeneity in chromatin accessibility profiles with the variability in RNA expression.

A challenge of this augmented representation is how to find a one-dimensional statistic that summarises the heterogeneity across cells (as the variance statistic in conventional rates), which can be in turn correlated with summary statistics from the RNA expression. The approach we followed here is to cluster cells (per gene) based on the similarity of the accessibility profiles, using a finite mixture model with an expectation-maximisation algorithm. The optimal number of clusters was estimated using a Bayesian Information Criterion.

After model fitting, we considered the number of clusters as a proxy for accessibility heterogeneity, the rationale being that homogeneous genes will be grouped in a single cluster, while heterogeneous genes will contain a higher number of clusters. The result of the analysis is shown below:

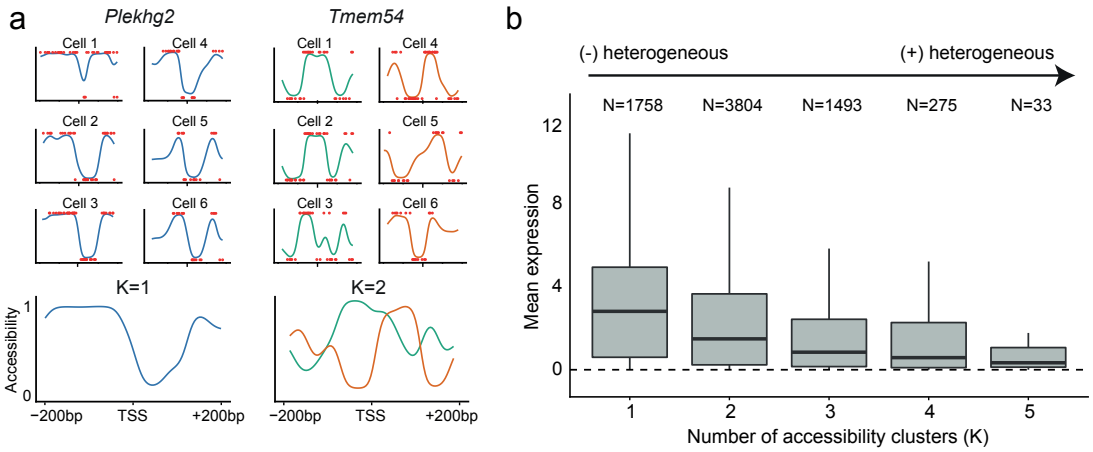


Figure 1.14: Clustering chromatin accessibility profiles at gene promoters.
(a) Accessibility profiles are fitted for each cell and gene using 200 bp windows around the TSS. Then, a clustering step is applied for each gene and the most likely number of clusters is estimated using a Bayesian Information Criterion. Genes with higher numbers of clusters correspond to genes with increased heterogeneity compared to genes with small numbers of clusters.
(b) Relationship between heterogeneity in the accessibility profile (x-axis) and average gene expression (across cells, y-axis).

Genes with homogeneous accessibility profiles (fewer clusters) are associated with higher average expression levels. This includes genes with housekeeping functions, which are known to display highly conserved epigenetic features [225]. In contrast, genes with heterogeneous accessibility (more clusters) are associated with lower expression levels. Interestingly, these genes are enriched for bivalent domains, containing both active H3K4me3 and repressive H3K27me3 histone marks (**??**). As reported in previous studies, bivalent chromatin is normally associated with lowly-expressed genes that are poised for activation upon cell differentiation, thus playing a fundamental role in pluripotency and development [258, 18]

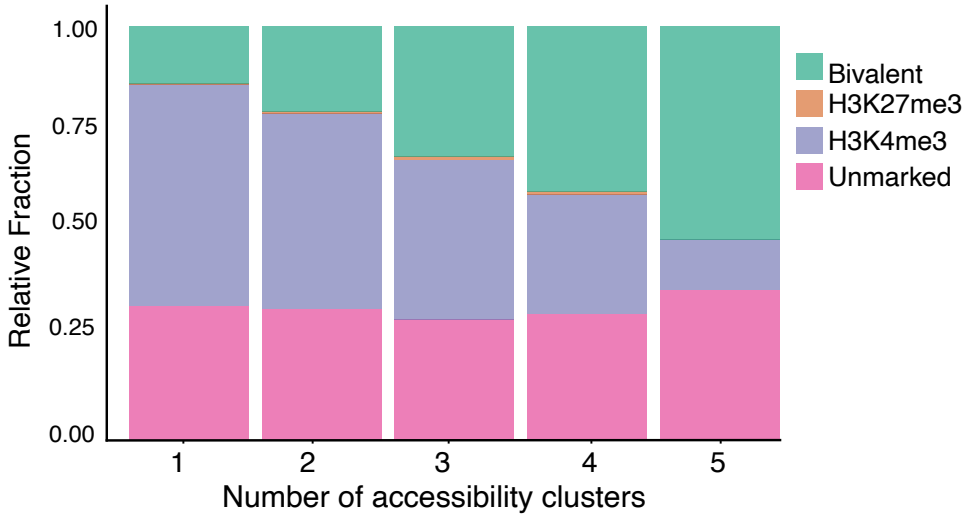


Figure 1.15: High levels of heterogeneity in chromatin accessibility are associated with the presence of bivalent histone marks (H3K4me3 and H3K27me3).
Proportion of gene promoters marked with different sets of histone mark combinations (y-axis), stratified by number of accessibility clusters (x-axis)

1.2.6 Exploration of epigenome connections along a developmental trajectory

The use of single-cell technologies has permitted the unbiased study of continuous trajectories by computationally reconstructing the *pseudotemporal* dynamics from the molecular profiles [253, 86, 218]. 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[86] to the EB data set, using the RNA expression of the 500 genes with highest biological overdispersion[**Lun2016**]. The first diffusion component was used to reconstruct a pseudotemporal ordering of cells from pluripotent to differentiated states:

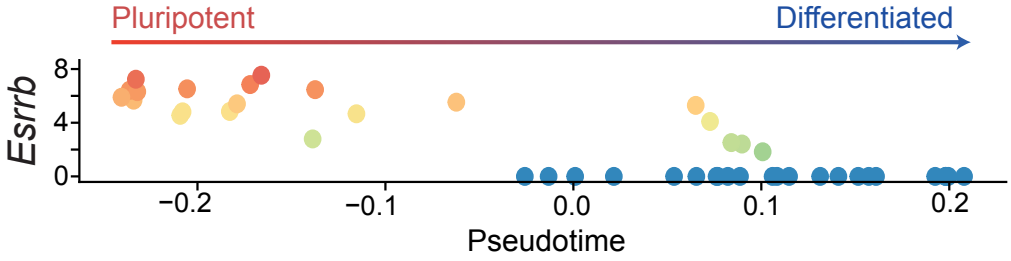


Figure 1.16: Reconstruction of developmental trajectory in embryoid body cells from the RNA expression data.
Each dot corresponds to one cell. The y-axis displays expression of *Esrrb*, a canonical pluripotency marker, and the x-axis shows the position of the cells in the first diffusion component.

Using the pseudotime reconstruction we investigated whether the strength of association between molecular layers (as calculated in Figure 1.8) are affected along the developmental trajectory. To do this, we simply correlated the correlation coefficient across genes between each pair of molecular layers (one value per cell) versus the pseudotime position (Figure 1.17). 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 1.17), 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.

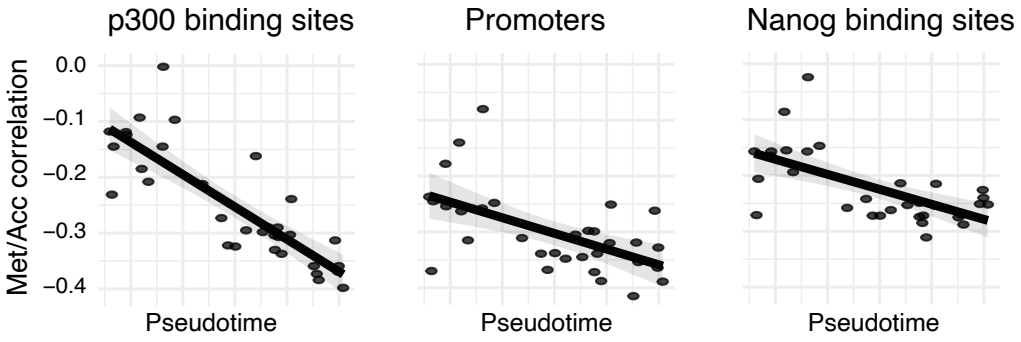


Figure 1.17: Developmental trajectory is associated changes in methylation-accessibility coupling.
Shown is the location of each cell in pseudotime (x-axis) and the corresponding Pearson correlation coefficients between methylation and accessibility (y-axis) in three different genomic contexts with regulatory roles.

1.3 Conclusions and open perspectives

In this Chapter I have 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. We have shown that this novel technology can be used to study the connection between the epigenome and the transcriptome in an unsupervised manner.

This protocol 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. These are the lines of research that I believe are important to improve scNMT-seq:

- **Scalability**: scRNA-seq protocols are reaching the astonishing numbers of millions of cells per experiment, compared to the limited cell numbers achieved in multi-modal experiments [39, 40, 81]. As in scRNA-seq, the maturation of multi-modal techniques will have a trade-off between sensitivity and scalability [42]. Under the assumption that most of the variance explained by epigenetic layers is driven by cis-regulation [17], we emphasise the importance of obtaining high-resolution measurements as provided by scNMT-seq. Hence, effort should be placed on making this comprehensive technology more scalable, which can be achieved in the short term by a series of technical improvements.
  First, barcodes are currently added at the end of the protocol, which limits a single experiment to the size of the plate (typically 96 cells). 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 cells [**DR-seq**, 174].
  Similarly, the physical separation of mRNA from genomic DNA is also carried out at the start of the protocol and individually for each cell. Given that it is a time-consuming and expensive process, this step should also be performed after pooling [**DR-seq**].

Finally, sequencing costs are substantially decreasing (even faster than predicted by Moore's law [241]). Yet, the generation of scNMT-seq libraries remains inexorably expensive. Hence, we 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 [83], an initial round of ATAC protocol to select open chromatin [232] 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 [7]. These becomes even more pronounced in scNMT-seq where almost $\approx 50\%$ of CpG dinucleotides are removed to avoid technical biases (see Section 1.2.3.1). Nonetheless, as discussed in Section 1.2.1, an important advantage of bisulfite approaches is that missing data can be easily discriminated from inaccessible chromatin. Therefore, the imputation of DNA methylation data will be a critical step to enable genome-wide analysis.
  Most of the methods developed for bulk data are unsuccesful because they do not account for cell-to-cell variability [7]. A successful single-cell strategy based on deep learning has been proposed (DeepCpG[7]), but is a complex model that is difficult to train and does not scale to large studies. Faster and accurate Bayesian approaches have also been considered (Melissa [113]), albeit the model is restricted to small genomic annotations. An interesting direction would be to extend DeepCpG and Melissa to exploit the richness of information in the GpC accessibility data to refine the imputation of CpG measurements.

- **Adding more molecular layers**: the scNMT-seq protocol 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 [194, 69, 162, 65]. 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[103], allele-specific fractions[58] and RNA velocity information [135].
  From the experimental side, NMT-seq can theoretically be combined with novel single-cell assays that quantify transcription factor binding [173] and histone modifications [114].

- **Denoising**: in scNMT-seq the CGC positions (27%) suffer from off-target effects of the GpC methylase [116]. In this work we have excluded those measurements to avoid undesired technical variation. Yet, no attempts have been carried to quantify this effect. If small enough, one could denoise the resulting CpG measurements by machine learning techniques that use sequence context information.

- **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 [141] 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.

1.4 Theoretical foundations

1.4.1 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 on 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 data sets 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 data sets. On the other hand, simplistic models will fit the data poorly, leading to poor explanatory power.

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

1.4.1.1 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[236, 22, 176]. However, in this setting there is no penalisation for model complexity, making maximum likelihood solutions prone to overfit in cases where the data is relatively sparse. Generalisations that account for model complexity have been proposed and include regularising terms that shrink parameters to small values. However, these are often particular cases of the more general framework of Bayesian statistics [91, 22, 176].

1.4.1.2 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 [15], 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 [22, 176].
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 [22, 176, 77].

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 information is available one could set set uninformative priors by following maximum entropy principles [106].

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 [200, 22, 176, 77]. The existence of conjugate priors is one of the major reasons that justify the widespread use of exponential family distributions in Bayesian models [77]. An example is the Automatic Relevance Determination prior discussed in **??**.

Again, the milestone 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 the 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 [22, 176, 77]. 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 dsitribution is chosen smartly, 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[22, 176, 77]. 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.

1.4.1.3 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 [77, 24].

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 data sets [24].
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 [22, 176, 24].

1.4.1.4 Laplace approximation

The Laplace approximation is probably the simplest of the deterministic tecniques, where the aim is to construct a Gaussian approximation around the mode of the true posterior distribution using a second-order Taylor expansion [22, 176].
Suppose $\mathbf{X}$ contains all unobserved variables. The true posterior distribution can be written as:

$$p(\mathbf{X}) = \frac{f(\mathbf{X})}{Z}$$

where $f(\mathbf{X})$ is a function that depends on the unobserved variables and Z is an unknown normalisation constant to ensure that $\int p(\mathbf{X})d\mathbf{X} = 1$.

The second-order Taylor expansion of $\log f(\mathbf{X})$ centered around its (known) mode $\hat{\mathbf{X}}$ is:

$$\log f(\mathbf{X}) \approx \log f(\hat{\mathbf{X}}) - \frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}})^T \mathbf{A} (\mathbf{X} - \hat{\mathbf{X}})$$

where $\mathbf{A} = \nabla^2 \log f(\hat{\mathbf{X}})$ is the Hessian matrix of $\log f(\mathbf{X})$ evaluated at $\hat{\mathbf{X}}$.
Notice three things. First, the first-order term of the Taylor expansion is zero because $\hat{\mathbf{X}}$ is a stationary point. Second, the log function is monotonically increasing and therefore a maximum of $\log f(\mathbf{X})$ is also a maximum of $f(\mathbf{X})$. Third, the mode of the posterior $p(\mathbf{X})$ must be known, which requires the use of (complex) optimisation algorithms.
Taking the exponential in both sides:

$$f(\mathbf{X}) \approx f(\hat{\mathbf{X}}) \exp\{-\frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}})^T \mathbf{A} (\mathbf{X} - \hat{\mathbf{X}})\}$$

which leads to the following multivariate Gaussian distribution approximation $q(\mathbf{X}) = \mathcal{N}\left(\mathbf{X} \mid \hat{\mathbf{X}}, \mathbf{A}\right)$:

$$q(\mathbf{X}) = \frac{\mid A \mid^{1/2}}{(2\pi^{d/2})} \exp\{-\frac{1}{2}(\mathbf{X} - \hat{\mathbf{X}})^T \mathbf{A} (\mathbf{X} - \hat{\mathbf{X}})\}$$

where d is the number of unobserved variables.
Despite its simplicity, the Laplace approximation is a useful strategy that has been successfully applied in practice. Nonetheless, this approximation has notable caveats: first, is limited by its own local definition, ignoring all the density beyond the mode of the posterior. Second, it does not

apply to discrete variables. Third, the inversion of the Hessian is very expensive in high-dimensional settings.

1.4.1.5 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 [24, 272]. 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:

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

Note that the KL divergence is not a proper distance metric, as it is not symmetric. In fact, using the reverse KL divergence $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$ defines a different inference framework called expectation propagation [167].

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 [22, 176]) 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):

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

Hence, minimising the KL divergence is equivalent to maximising $\mathcal{L}(\mathbf{X})$ Figure 1.18:

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y}) \Big) d\mathbf{X} \\ &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Y})] - \mathbb{E}_q[\log q(\mathbf{X})]\end{aligned} \tag{1.1}$$

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 (1.1) can be computed in closed form.
In some occasions (see section X), we will use the following form for the ELBO:

$$\mathcal{L}(\mathbf{X}) = \mathbb{E}_q[\log p(\mathbf{Y}|\mathbf{X})] + (\mathbb{E}_q[\log p(\mathbf{X})] - \mathbb{E}_q[\log q(\mathbf{X})]) \tag{1.2}$$

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 $q(\mathbf{X})$. The following image summarises the general picture of variational learning:

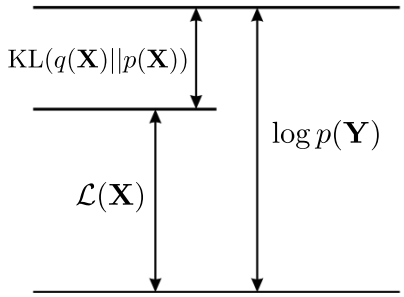


Figure 1.18: The quantity $\mathcal{L}(\mathbf{X})$ provides a lower bound on the true log marginal likelihood $\log p(\mathbf{Y})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(q||p)$ between the variational distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$

There are several approaches to define $q(\mathbf{X})$, the two most commonly used are called (unparametric) mean-field and (parametric) fixed-form [272, 24].

1.4.1.6 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[221]:

$$q(\mathbf{X}) = \prod_{i=1}^{M} q(\mathbf{x}_i) \tag{1.3}$$

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 1.19). Yet, this is a key assumption to obtain an analytical inference scheme that yields surprisingly accurate results [23, 68, 28].

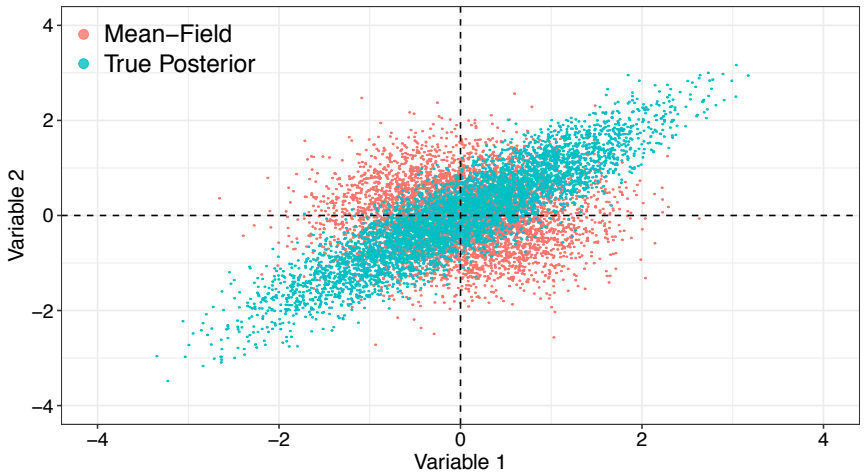


Figure 1.19: Illustrative example of sampling from a true posterior distribution (blue) versus a fitted mean-field varaitional distribution (red) in a model with two (correlated) unobserved variables. The mean-field approximation wrongly assumes that the unobserved variables are independent.

Using calculus of variations (derivations can be found in [22, 176]), it follows that the optimal distribution $q(\mathbf{X})$ that maximises the lower bound $\mathcal{L}(\mathbf{X})$ is

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const} \tag{1.4}$$

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 $\hat{p}(\mathbf{x}_i)$.

1.4.1.7 Fixed-form variational inference

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

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \tag{1.5}$$

$$= \mathbb{E}[\log(q(\mathbf{X})) - \log(p(\mathbf{X}, \mathbf{Y}))] \tag{1.6}$$

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 [28, 204, 28].

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 [272, 24].

However, while the parametric assumption certainly limits the flexibility of variational distributions, the advantage of this formulation is that it unveils the possibility to use fast gradient-based methods for the inference procedure [98, 204].

1.4.1.8 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 stringkly different properties. This can be understood by inspecting the differences between the two KL divergence formulas:

Variational inference:

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = -\int_z q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \tag{1.7}$$

Expectation propagation:

$$\mathrm{KL}(p(\mathbf{X}|\mathbf{Y})||q(\mathbf{X})) = -\int_z p(\mathbf{X}|\mathbf{Y}) \log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \tag{1.8}$$

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 (1.8) than in Equation (1.7), because of the true posterior density being on the denominator. 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 (1.7) there is a strong penalty for having low-density $q(\mathbf{X})$ values.
As discussed in [22], the practical consequences of this duality can be observed when the posterior is multi-modal, as in any sufficiently complex model. 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 parameter itself.
A detailed mathematical treatment of EP, including derivations for specific examples, can be found in [22, 176, 167]

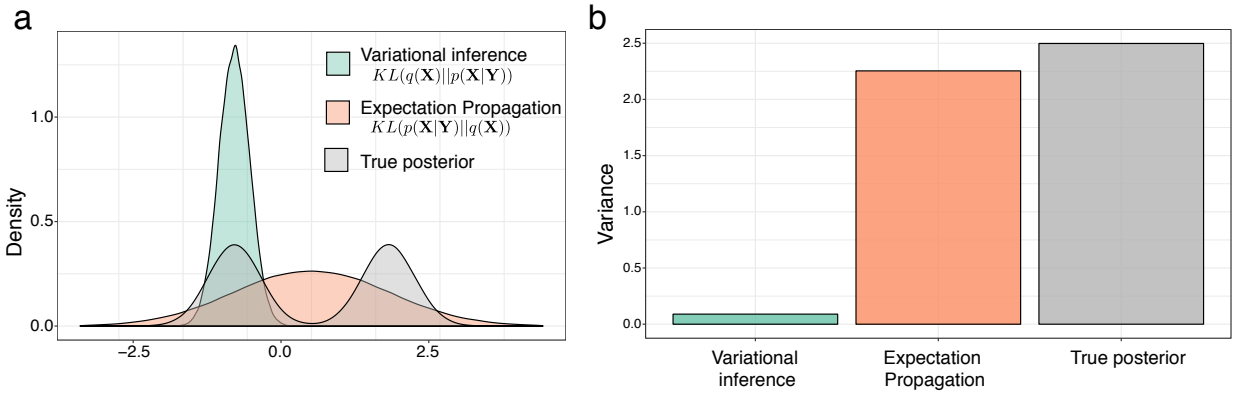


Figure 1.20: Illustrative comparison of Variational inference and Expectation Propagation. Shown is the (a) Density and (b) Variance of the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$ (grey), the variational distribution (orange) and the expectation propagation distribution (green).

Following the rationale above, it is easy to predict that variational inference tends to be underestimate the variance of the posterior density. Yet, empirical research have shown that this is acceptable, provided that a good model selection is performed [23].

1.4.1.9 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 data sets and it has been applied to a myriad of different problems, including genome-wide association studies [41], population genetics, [201], network analysis [219] and natural language processing [25].

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[24, 272, 178]. As an example, the mean-field setting makes strong independence assumptions about the parameters. Although it tends to besurprisingly effective, it is not clear in which applications the dependencies between the parameters are important enough than 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 [24, 272].

As we shall demonstrate later, alternative strategies have been considered to allow some dependencies between the variables, resulting in *structured* mean-field approximations[97, 249]. However, they often lead to very complex (if not intractable) inference frameworks.

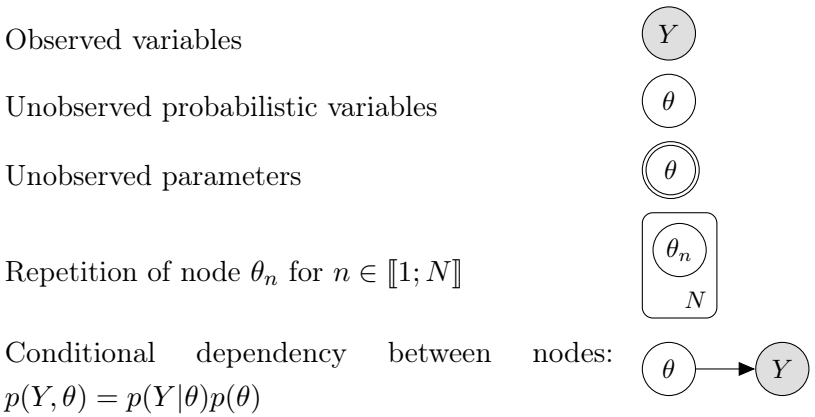
Finally, another area of extensive research is how to extend the applicability of VI to non-conjugate models. As discussed in Section 1.4.1.3, the ELBO of non-conjugate models contains intractable integrals, and setting up an inference scheme requires the use of either stochastic Monte Carlo approximations or deterministic approximations that introduce additional lower bounds [272, 223,

64]. In this thesis we follow this rationale to derive an inference framework for a model with non-gaussian likelihoods.

1.4.2 Graphical notations 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 [22]. 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 [22]. 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 [60].



1.4.3 Latent variable models for genomics

With the exponential growth in the use of high-throughput genomics, biological data sets are increasingly high dimensional, both in terms of samples and features. A key principle of biological data sets 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 (LVMs) [129, 234, 142, 196, 138, 235, 166].

1.4.3.1 General mathematical formulation

Given a dataset $\mathbf{Y}$ of N samples and D features, LVMs 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 a considerable challenges in interpretation [275]. Hence, for most applications where interpretability is important, $f(\mathbf{X}|\mathbf{\Theta})$ is assumed to be linear [**XX**]:

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T \tag{1.9}$$

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 or loadings, which provide the linear mapping between the features and the factors.
Note that the aim in dimensionality reduction is to exploit the coordinated heterogeneity between features, and hence features are 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 [207].

1.4.3.2 Principal component Analysis

Principal Component Analysis (PCA) is the most popular technique for dimensionality reduction [99, 211]. Starting from Equation (1.9), two formulations of PCA exist [22]. 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:

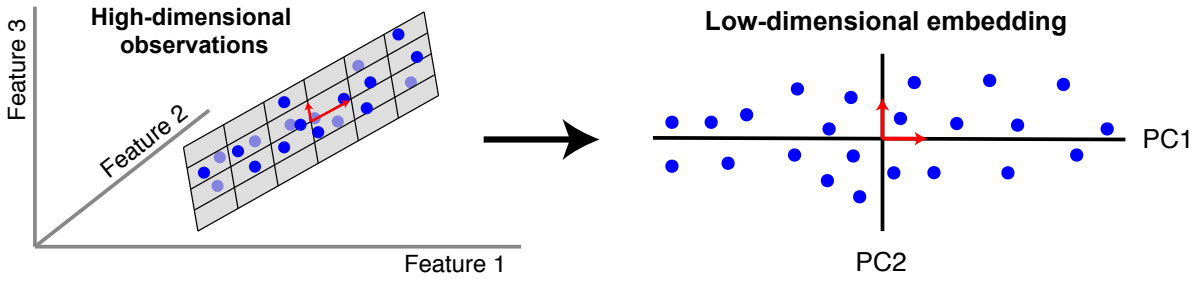


Figure 1.21

For a single principal component, the optimisation problem is:

$$\underset{\|\mathbf{w}\|=1}{\arg\max} = \mathbf{w}_1^T\mathbf{Y}^T\mathbf{Y}\mathbf{w} \tag{1.10}$$

where $\mathbf{Y}^T\mathbf{Y} = \mathbf{S} \in \mathbb{R}^{D\times D}$ is the data covariance matrix and $\mathbf{w}_1^T$ is the vector of loadings.
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 (1.10) 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 all 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.

In both cases, solving the optimisation problems via Lagrange multipliers leads, remarkably, to the same solution:

$$\mathbf{S}\mathbf{w}_k = \lambda_k \mathbf{w}_k \tag{1.11}$$

Hence, the loading vectors $\mathbf{w}_k$ are the eigenvectors of $\mathbf{S}$, which can be computed via singular value decomposition [22].

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 loading vector $\mathbf{w}$ (Figure 1.22). 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 $\|\mathbf{y}_n - \mathbf{z}_n \mathbf{w}^T\|$:

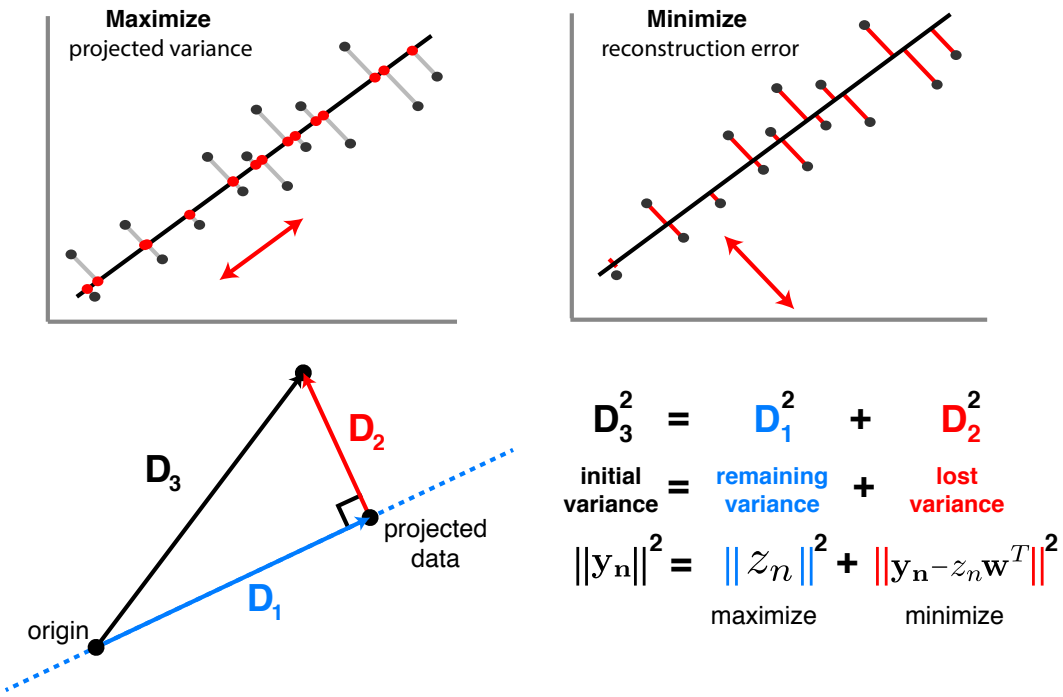


Figure 1.22: In the maximum variance formulation we aim at maximising the variance of the projected data (blue line), whereas in the minimum error formulation we are aimed at minimising the residual variance (red line). Given a fixed total variance (black line), both strategies are equivalent

The main strength of PCA relies on its simplicity and closed form solution. Additionally, the linear mapping has the advantage of yielding interpretable loadings, 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 data sets [145]. 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 [104]. Third, there is no rationality on how to evaluate the fit and perform model selection. Finally, it does not offer a principled way of modelling prior information about the data.

1.4.3.3 Probabilistic Principal Component Analysis and Factor Analysis

A probabilistic version of PCA was initially proposed in [248]. It can be formulated by converting some (or all) fixed parameters into random variables and adding an explicit noise term to Equation (1.9):

$$\mathbf{Y} = \mathbf{WZ} + \boldsymbol{\epsilon} \tag{1.12}$$

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:

$$p(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right)$$
$$p(\epsilon) = \mathcal{N}\left(\epsilon \,|\, 0, \sigma^2\right)$$

All together, this leads to a normally-distributed likelihood:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \sigma) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathcal{N}\left(y_{n,d} \,|\, \mathbf{w}_{:,k}^{T} \mathbf{z}_{n,:}, \sigma^2 \mathbf{I}\right) \tag{1.13}$$

The corresponding graphical model is:

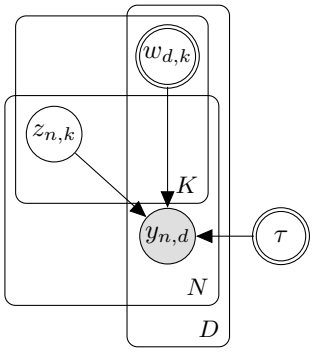


Figure 1.23: Graphical model for probabilistic PCA. The latent variables are modelled as random variables, whereas the loadings and the noise are modelled as deterministic parameters.

Importantly, the choice of the distribution for ϵ implies that the noise of each feature is independent but restricted to have the same variance σ. 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 [216, 22].

The inference procedures involves learning the parameters $\mathbf{W}$, and σ^2 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 [216, 22]. 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 σ^2. 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 [248].
Interestingly, the EM solution of probabilistic PCA lies in the same subspace than the traditional PCA solution [248], 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 data sets.

1.4.3.4 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:

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

where τ is the precision (inverse of the variance) of the noise term. A generalisation to Bayesian Factor Analysis follows by allowing a separate noise term per feature:

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

where a_0 and b_0 are fixed hyperparameters. As in Equation (1.13), 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:

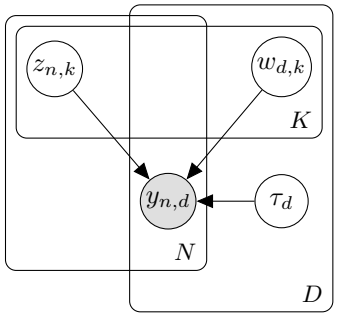


Figure 1.24: Graphical model for Bayesian Factor Analysis. All unobserved variables are modelled as random variables.

1.4.3.5 Hierarchical priors: Automatic relevance determination

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 [21, 20]. 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 [123, 260, 125, 35, 118, 277].

As an example, a major challenge in PCA is how to determine the dimensionality of the latent space (i.e. the number of principle components). As we will show, the use of hierarchical prior distributions allows the model to introduce sparsity assumptions on the loadings in such a way that the model automatically learns the number of factors.
In the context of Factor Analysis, one the first sparsity priors to be proposed was the Automatic Relevance determination (ARD) prior [179, 157, 21, 20].

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

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 1.25).
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 1.26.

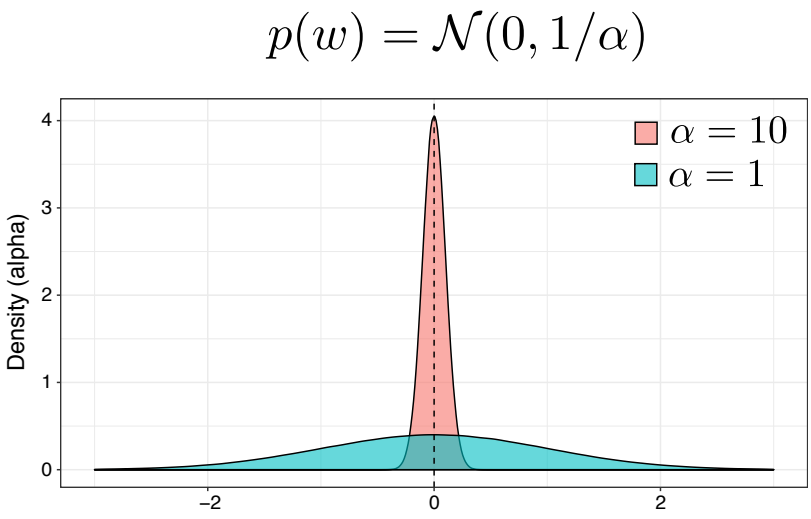


Figure 1.25: Visualisation of the sparsity-inducing Automatic Relevance Determination prior

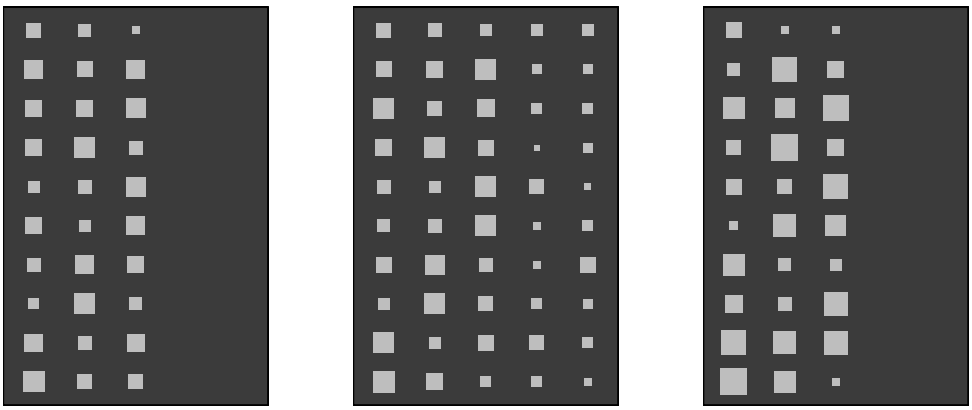


Figure 1.26: Hinton plots display the values of the loading matrix, similar to a heatmap, where bigger squares depict larger loadings. Shown are the Hinton plots for (a) the true weights, (b) the infered weights by a Factor Analysis model with no ARD prior (middle), and (c) the infered weights by a Factor Analysis model with ARD prior per factor. This figure was generated using simulated data with $N = 100$ samples, $D = 10$ features and $K = 3$ factors.

1.4.3.6 Hierarchical priors: 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 [234, 76].
The variability observed in biological data is driven both by technical factors and biological factors. Thechnical 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 [76]. Hence, a practical factor analysis model should be able to learn factors with different degrees of sparsity.
The ARD prior proposed in Section 1.4.3.5 allows entire factors to be dropped out from the model, but it provides a weak degree of regularisation when it comes to inactivating individual loadings 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 [168, 249]:

$$p(w_{d,k} \mid \alpha_k, \theta_k) = (1 - \theta_k)\mathbb{1}_0(w_{d,k}) + \theta_k \mathcal{N}\left(w_{d,k} \mid 0, \alpha_k^{-1}\right) \tag{1.14}$$

$$p(\theta_k) = \text{Beta}\left(\theta_k \mid a_0^\theta, b_0^\theta\right) \tag{1.15}$$

$$p(\alpha_k) = \mathcal{G}\left(\alpha_k \mid a_0^\alpha, b_0^\alpha\right) \tag{1.16}$$

The corresponding graphical model is:

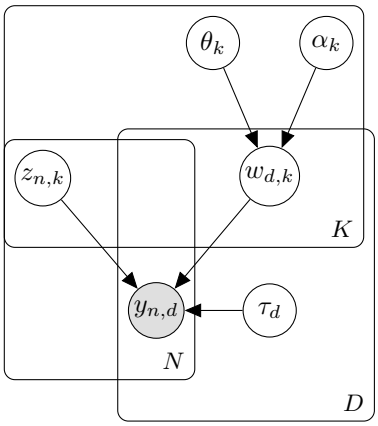


Figure 1.27: Graphical model for Bayesian sparse Factor Analysis. A double sparsity-inducing prior is used on the loadings: an ARD prior to prune inactive factors and a spike-and-slab prior to inactive individual features within the active factors.

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 shrinked 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.

1.4.4 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 [268, 144, 271]. 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 refered to as the multi-view learning problem [268, 144].

A tempting approach to circumvent the multi-view learning problem is to simply concatenate all different data sets before applying conventional (single-view) latent variable models [213]. 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 [193]. 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 of which 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 [268] and a more genomics-oriented perspective can be found in [213]. For the purpose of this thesis, we will describe only the use of latent variable models for multi-view data integration.

1.4.4.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between two datasets [100, 88].
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, Therefore, the second pair of canonical variables $\mathbf{u}_2$ and $\mathbf{v}_2$ is found out of 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 [100, 88].
Because of its simplicity and efficient computation, CCA has widespread use as a dimensionality reduction technique [88]. 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 $D >> N$ [161, 84]. Hence, probabilistic versions with sparsity assumptions that reduce overfitting and improve interpretability followed.

1.4.4.2 Probabilistic Canonical Correlation Analysis

Following the derivation of probabilistic PCA [248], a similar effort enabled a probabilistic formulation of CCA as a generative model [13].
In this model, the two matrix of observations $\mathbf{Y}^1$ and $\mathbf{Y}^2$ are decomposed in terms of two loading matrices $\mathbf{W}^1$ and $\mathbf{W}^2$ but a joint latent matrix $\mathbf{Z}$:

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

With the following prior probability distributions:

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

As in [248], the loadings 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:

$$p(\mathbf{Y}^1|\mathbf{W}^1, \mathbf{Z}, \tau_1) = \prod_{n=1}^{N}\prod_{d=1}^{D_1} \mathcal{N}\left(y_{n,d}^1 \,|\, (\mathbf{w}_{:,k}^1)^T \mathbf{z}_n, \tau_1^{-1}\right) \tag{1.17}$$
$$p(\mathbf{Y}^2|\mathbf{W}^2, \mathbf{Z}, \tau_2) = \prod_{n=1}^{N}\prod_{d=1}^{D_2} \mathcal{N}\left(y_{n,d}^2 \,|\, (\mathbf{w}_{:,k}^2)^T \mathbf{z}_n, \tau_2^{-1}\right)$$

The corresponding graphical model is:

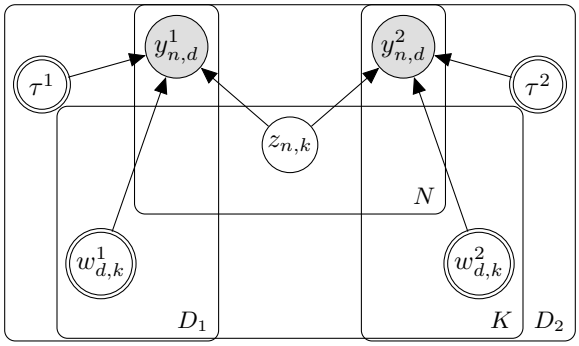


Figure 1.28: Graphical model for probabilistic Canonical Correlation Analysis

Notice that the observations for both data sets 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 [13]. Nonetheless, one of the many advantage of a probabilistic formulation is that it enables a broad range of principled extensions into larger graphical models.

1.4.4.3 Bayesian Canonical Correlation Analysis

A fully Bayesian treatment of CCA followed based on exactly the same principle presented in Section 1.4.3.4 by introducing prior distributions to all unobserved variables [262, 124]:

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

Resulting in the same likelihood model as in Equation (1.17). 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. An independent noise term can be defined per view or per per feature. One could also model correlated noise by generalising the distribution to a multivariate gaussian with full-rank covariance. [262, 124].

The corresponding graphical model is:

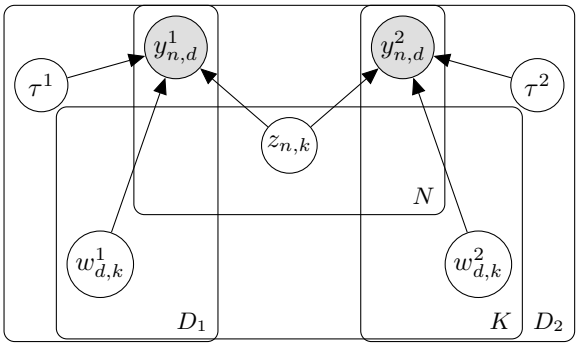


Figure 1.29: Graphical model for Bayesian Canonical Correlation Analysis

As expected, in practice this yields a more sparse solution than traditional CCA (Figure 1.30):

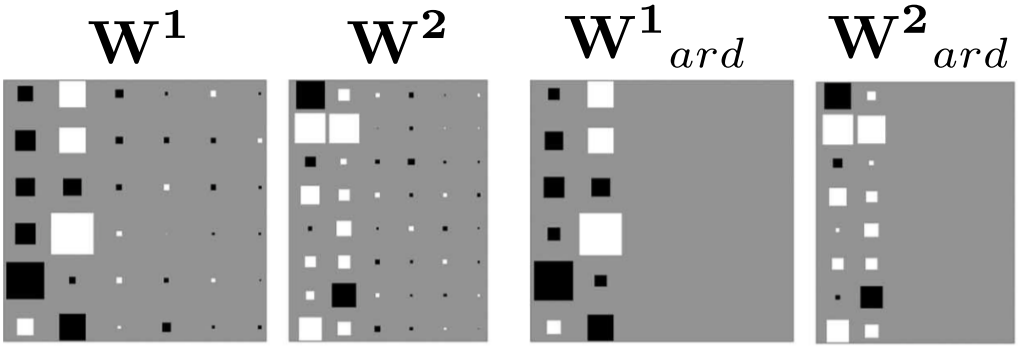


Figure 1.30: Comparison of the Hinton's diagram of $\mathbf{W}^1$ and $\mathbf{W}^2$ for the maximum likelihood CCA model (two left plots) and the variational bayes CCA model (two right plots). Reprinted from [262] with modifications.

1.4.4.4 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 [260] and a series of generalisations followed, tailored with specific assumptions for different applications [125, 143, 35, 118, 277, 209]. 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 1.4.4.3) 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 facto 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:

$$p(\mathbf{W}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k}^{m} \,|\, 0, \frac{1}{\alpha_k^m}\right) \tag{1.18}$$

$$p(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{1.19}$$

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:

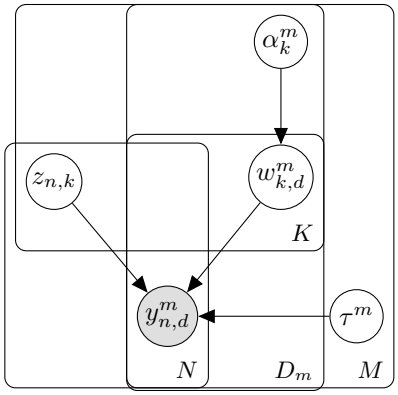


Figure 1.31: Graphical model for Bayesian Group Factor Analysis

Finally, notice that if $M = 1$ the model reduces to Bayesian PCA (Section 1.4.3.4), 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) 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.

1.5 Multi-Omics Factor Analysis

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 [10].
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 lead 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.

1.5.1 Model description

MOFA is a multi-view generalisation of traditional Factor Analysis to M input matrices (or views) based on the framework of Group Factor Analysis (discussed in **??**).
The input data consists on M views $\mathbf{Y}^m \in \mathbb{R}^{N \times D_m}$ with non-overlapping features that often represent different assays. However, there is flexibility in the definition of views.
Formally, the input data is factorised as:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m \tag{1.20}$$

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 weights that relate the high-dimensional space to the low-dimensional latent representation. Finally, $\boldsymbol{\epsilon}^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic:

$$p(\epsilon_d^m) = \mathcal{N}\left(\epsilon_d^m \,|\, 0, 1/\tau_d^m\right) \tag{1.21}$$

Non-gaussian noise models can also be defined (see Section 1.5.6), but unless otherwise stated, we will always assume Gaussian residuals.
Altogether, this results in the following likelihood:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \mathbf{T}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{n=1}^{N} \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_n^T \mathbf{w}_d^m, 1/\tau_d^m\right) \tag{1.22}$$

Notice that the mathematical formulation so far is equivalent to the Group Factor Analysis described in **??**.

1.5.1.1 Prior distributions for the factors

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

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{1.23}$$

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

1.5.1.2 Prior distributions for the weights

The key determinant of the model is the regularization used on the prior distributions for the weights. Here we encode two levels of sparsity, a (1) view- and factor-wise sparsity and (2) an individual feature-wise sparsity. The aim of the factor- and view-wise sparsity is to disentangle the activity of factors to the different views, such that the weight vector $\mathbf{w}^m_{:,k}$ is shrunk to zero if the factor k does not explain any variation in view m.
In addition, we place a second layer of sparsity which encourages inactive weights on each individual feature. Mathematically, we express this as a combination of an Automatic Relevance Determination (ARD) prior [157] for the view- and factor-wise sparsity and a spike-and-slab prior [168] for the feature-wise sparsity: However, this formulation of the spike-and-slab prior contains a Dirac delta function, which makes the inference procedure troublesome. To solve this we introduce a re-parametrization of the weights w as a product of a Gaussian random variable $\hat{w}$ and a Bernoulli random variable s, [249] resulting in the following prior:

$$p(\hat{w}^m_{dk}, s^m_{dk}) = \mathcal{N}\left(\hat{w}^m_{dk} \,|\, 0, \frac{1}{\alpha^m_k}\right) \mathrm{Ber}(s^m_{dk} \,|\, \theta^m_k) \tag{1.24}$$

In this formulation α^m_k controls the activity of factor k in view m and θ^m_k controls the corresponding fraction of active weights (i.e. the sparsity levels).

Finally, we define conjugate priors for θ and α:

$$p(\theta^m_k) = \mathrm{Beta}\left(\theta^m_k \,|\, a^\theta_0, b^\theta_0\right) \tag{1.25}$$

$$p(\alpha^m_k) = \mathcal{G}\left(\alpha^m_k \,|\, a^\alpha_0, b^\alpha_0\right) \tag{1.26}$$

with hyper-parameters $a^\theta_0, b^\theta_0 = 1$ and $a^\alpha_0, b^\alpha_0 = 1e^{-5}$ to get uninformative priors. Posterior values of θ^m_k close to 0 implies that most of the weights of factor k in view m are shrinked to 0 (sparse factor). In contrast, a value of θ^m_k close to 1 implies that most of the weights are non-zero (non-sparse factor). A small value of α^m_k implies that factor k is active in view m. In contrast, a large value of α^m_k implies that factor k is inactive in view m.

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

$$
\begin{aligned}
p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = & \prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{d,k}^m | \theta_k^m) \\
& \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \\
& \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \\
& \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right).
\end{aligned}
\tag{1.27}
$$

and the corresponding graphical model is shown in Figure 1.33. This completes the definition of the MOFA model.

1.5.1.3 Interpretation of the factors

Each factor ordinates cells along a one-dimensional axis centered at zero. Samples with different signs indicate opposite phenotypes, with higher absolute value indicating a stronger effect. Intuitively, their interpretation is similar to that of principal components in PCA.
For example, if the k-th factor captures the variability associated with cell cycle, we could expect cells in the Mitosis state to be at one end of the factor (irrespective of the sign, only the relative positioning being of importance). In contrast, cells in G1 phase are expected to be at the other end of the factor. Cells with intermediate phenotype, or with no clear phenotype (for example if no cell cycle genes are profiled), are expected to be located around zero.

1.5.1.4 Interpretation of the weights

The weights provide a score for each gene on each factor. Genes with no association with the factor are expected to have values close to zero, as specified by the prior. In contrast, genes with strong association with the factor are expected to have large absolute values. The sign of the loading indicates the direction of the effect: a positive loading indicates that the feature is more active in the cells with positive factor values, and viceversa.
Following the cell cycle example from above, genes that are upregulated in the M phase are expected to have large positive weights, whereas genes that are downregulated in the M phase (or, equivalently, upregulated in the G1 phase) are expected to have large negative weights.

1.5.1.5 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.

1.5.2 Downstream analysis

Once trained, the MOFA model can be queried for a set of downstream analysis (Figure 1.32):

- **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.
- **Ordination of the samples in the latent space**: as in any latent variable model, the samples can be visualised in the latent space using scatterplots or beeswarm plots. As we will demonstrate, simply by colouring or shaping the samples in the factor space using external covariates one can easily characterise the etiology of some of the 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 help to reveals the molecular signatures that underlie each factor.
- **Association analysis between factors and external covariates**: multi-omic data sets typically consist on 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 that are not used for model training 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 data sets.
- **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.

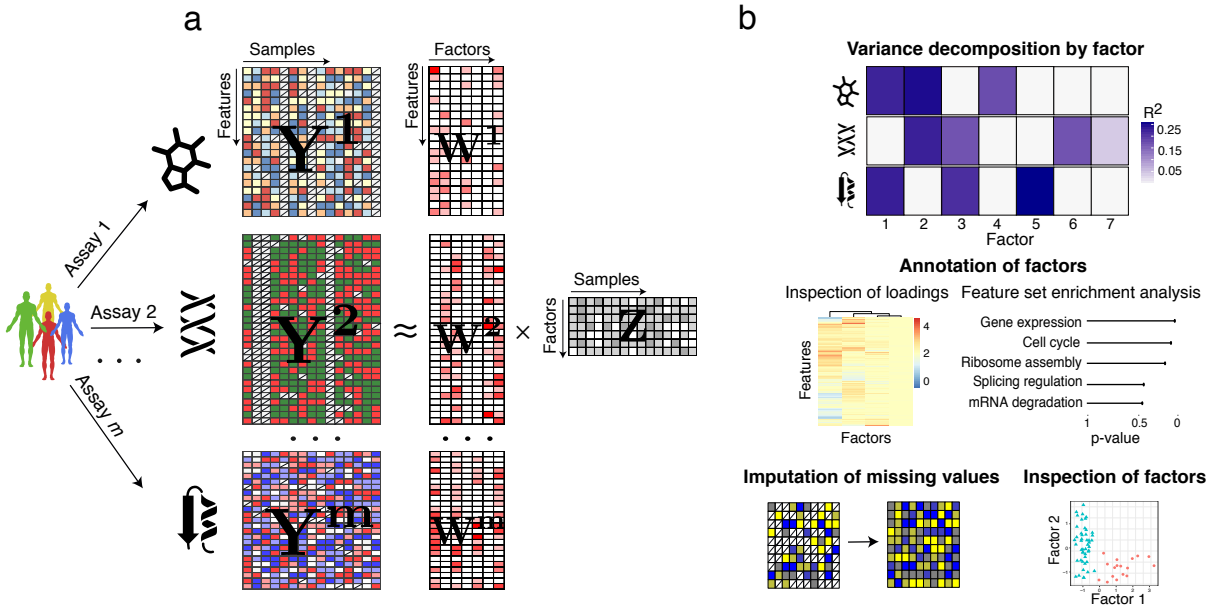


Figure 1.32: MOFA overview. The model takes M data matrices as input $(\mathbf{Y}^1, \cdots, \mathbf{Y}^M)$, one or more from each data modality, with co-occurrent samples but features that are not necessarily related and can differ in numbers. MOFA decomposes these matrices into a matrix of factors $(\mathbf{Z})$ and M weight matrices, one for each data modality $(\mathbf{W}^1, \cdots, \mathbf{W}^M)$. White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denotes missing values. The fitted MOFA model can be queried for different downstream analyses, including a variance decomposition to assess the proportion of variance explained by each factor in each data modality.

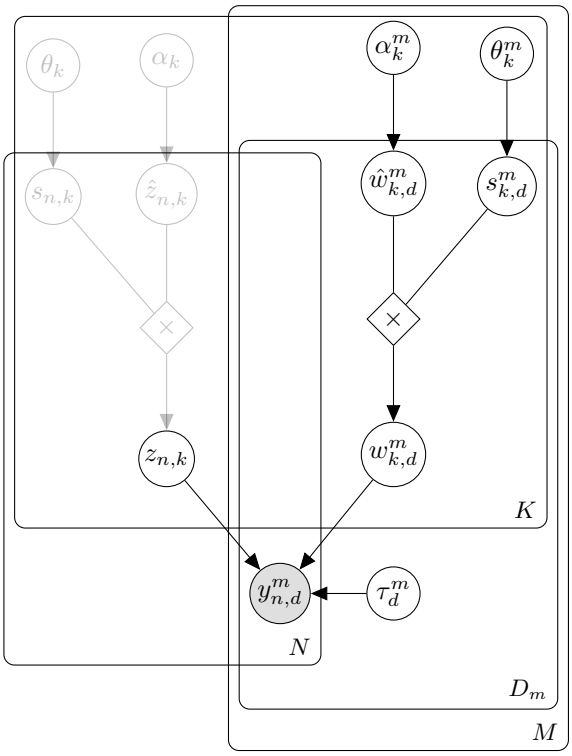


Figure 1.33: Graphical model for MOFA. The white circles represent hidden variables that are infered by the model, whereas the grey circles represent the observed variables. There are a total of four plates, each one representing a dimension of the model: M for the number of views, N for the number of 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 4 where we implement a spike-and-slab prior on the factors.

1.5.2.1 Inference

To make the model scalable to large data sets we adopt a Variational inference framework with a structured mean field approximation. A detailed overview is given in Section .1.0.1, and details on the variational updates for the MOFA model are given in **??**. To enable efficient inference for non-Gaussian likelihoods we employ local bounds [105, 223]. This is described in detail in Section 1.5.6.

1.5.3 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 instances 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:

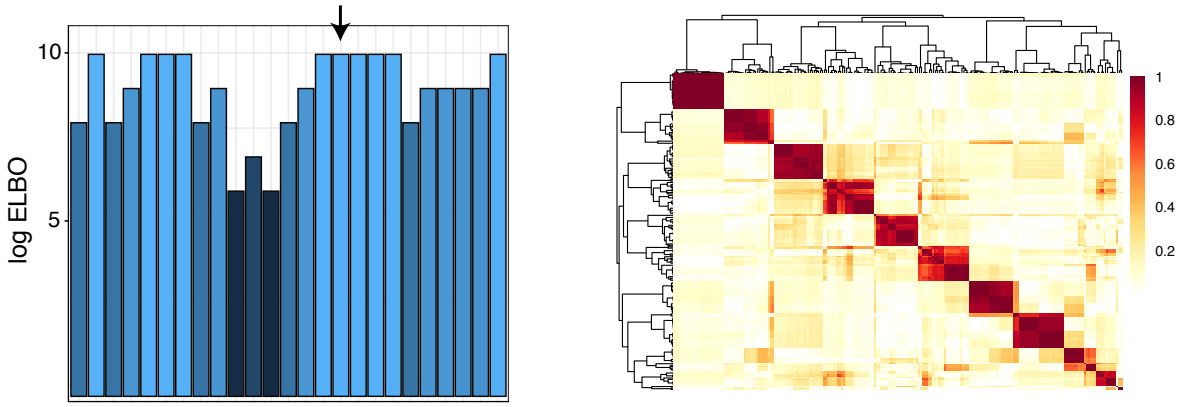


Figure 1.34: Model selection and robustness analysis in MOFA.
The left plot the log ELBO (y-axis) for 25 model instances (x-axis). The arrow indicates the model with the highest ELBO that would be selected for downstream analysis. The right plot displays the absolute value of the Pearson correlation 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.

1.5.4 Learning the number of factors

As described in **??**, the use of an ARD prior allows factors to be actively prunned by the model if their variance explained is negligible. In the implementation we control the prunning 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 (Figure 1.35). Thus, the optimal number of factors can be selected by the model selection strategy outlined in Section 1.5.3.

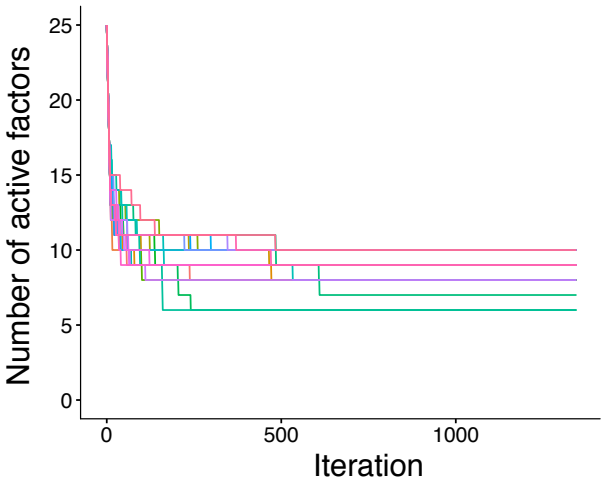


Figure 1.35: Training curve for the number of active factors across 25 different model instances.
The y-axis displays the number of active factors. The x-axis displays the iteration number. Different lines denote different model instances.

1.5.5 Monitoring convergence

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

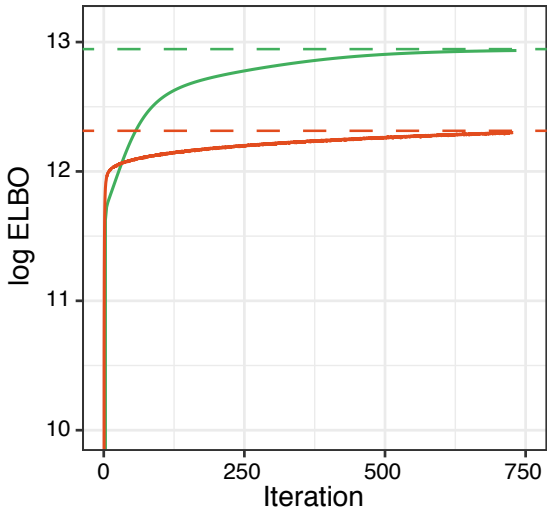


Figure 1.36: Training curve for two diffent initialisations of MOFA. The y-axis displays the log of the ELBO, with higher values indicating a better fit. The x-axis displays the iteration number. The horizontal dash lines mark the value of the ELBO upon convergence.

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

1.5.6 Modelling and inference with non-Gaussian data

To implement efficient variational inference in conjunction with a non-Gaussian likelihood we adapt prior work from [223] using local variational bounds. The key idea is to dynamically approximate non-Gaussian data by Gaussian pseudo-data based on a second-order Taylor expansion. To make the approximation justifiable we need to introduce variational parameters that are adjusted alongside the updates to improve the fit.

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

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

Expanding the MOFA model to non-Gaussian likelihoods we now assume a general likelihood of the form $p(\mathbf{Y}|\mathbf{X}) = p(\mathbf{Y}|\mathbf{C})$ with $\mathbf{C} = \mathbf{Z}\mathbf{W}^T$, that can write as

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

with $f_{nd}(c_{nd}) = -\log p(y_{nd}|c_{nd})$. We dropped the view index m to keep notation uncluttered.
Extending [223] to our heteroscedastic noise model, we require $f_{nd}(c_{nd})$ to be twice differentiable and bounded by κ_d, such that $f''_{nd}(c_{nd}) \leq \kappa_d \, \forall n, d$. This holds true in many important models as for example the Bernoulli and Poisson case. Under this assumption a lower bound on the log likelihood can be constructed 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 $\boldsymbol{\zeta} = \zeta_{nd}$ 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 above optimization problem, 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 algorithm propsed in [223] then alternates between updates of $\boldsymbol{\zeta}$ and q($\boldsymbol{\Theta}$). The update for $\boldsymbol{\zeta}$ is given by

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

where the expctations are taken with respect to the corresponding q distributions.
On the other hand, the updates for $q(\mathbf{X})$ can be shown to be identical to the variational Bayesian updates with a conjugate Gaussian likelihood when replacing the observed data $\mathbf{Y}$ by a pseudo-data $\hat{\mathbf{Y}}$ and the precisions τ_{nd} (which were treated as random variables) by the constant terms κ_d introduced above.
The pseudodata is given by

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

Depending on the log likelihoods $f(\cdot)$ different κ_d are used resulting in different pseudo-data updates. Two special cases implemented in MOFA are the Poisson and Bernoulli likelihood described in the following.

Bernoulli likelihood for binary data

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

$$\mathbf{Y}|\mathbf{Z}, \mathbf{W} \sim \mathrm{Ber}(\sigma(\mathbf{Z}\mathbf{W}^T)),$$

where $\sigma(a) = (1 + e^{-a})^{-1}$ is the logistic link function and $\mathbf{Z}$ and $\mathbf{W}$ are the latent factors and weights in our model, respectively.
In order to make the variational inference efficient and explicit as in the Gaussian case, we aim to approximate the Bernoulli data by a Gaussian pseudo-data as proposed in [223] and described above which allows to recycle all the updates from the model with Gaussian views. While [223]

assumes a homoscedastic approximation with a spherical Gaussian, we adopt an approach following [105], which allows for heteroscedaticity and provides a tighter bound on the Bernoulli likelihood. Denoting $c_{nd} = (\mathbf{Z}\mathbf{W}^T)_{nd}$ the Jaakkola upper bound [105] on the negative log-likelihood is given by

$$\begin{aligned} -\log\left(p(y_{nd}|c_{nd})\right) &= -\log\left(\sigma\left((2y_{nd}-1)c_{nd}\right)\right) \\ &\leq -\log(\zeta_{nd}) - \frac{(2y_{nd}-1)c_{nd} - \zeta_{nd})}{2} + \lambda(\zeta_{nd})\left(c_{nd}^2 - \zeta_{nd}^2\right) \\ &=: b_J(\zeta_{nd}, c_{nd}, y_{nd}) \end{aligned}$$

with λ given by $\lambda(\zeta) = \frac{1}{4\zeta}\tanh\left(\frac{\zeta}{2}\right)$.
This can easily be derived from a first-order Taylor expansion on the function $f(x) = -\log(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = \frac{x}{2} - \log(\sigma(x))$ in x^2 and by the convexity of f in x^2 this bound is global as discussed in [105].
In order to make use of this tighter bound but still be able to re-use the variational updates from the Gaussian case we re-formulate the bound as a Gaussian likelihood on pseudo-data $\hat{\mathbf{Y}}$.
As above we can plug this bound on the negative log-likelihood into the variational optimization problem to obtain

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

This is minimized iteratively in the variational parameter ζ_{nd} and the variational distribution of Z,W:
Minimizing in the variational parameter ζ this leads to the updates given by

$$\zeta_{nd}^2 = \mathbb{E}[c_{nd}^2]$$

as described in [105], [22].
For the variational distribution $q(\mathbf{Z}, \mathbf{W})$ we observe that the Jaakkola bound can be re-written as

$$b_J(\zeta_{nd}, c_{nd}, y_{nd}) = -\log\left(\varphi\left(\hat{y}_{nd}; c_{nd}, \frac{1}{2\lambda(\zeta_{nd})}\right)\right) + \gamma(\zeta_{nd}),$$

where $\varphi(\cdot; \mu, \sigma^2)$ denotes the density function of a normal distribution with mean μ and variance σ^2 and γ is a term only depending on $\boldsymbol{\zeta}$. This allows us to re-use the updates for $\mathbf{Z}$ and $\mathbf{W}$ from a setting with Gaussian likelihood by considering the Gaussian pseudo-data

$$\hat{y}_{nd} = \frac{2y_{nd} - 1}{4\lambda(\zeta_{nd})}$$

updating the data precision as $\tau_{nd} = 2\lambda(\zeta_{nd})$ using updates generalized for sample- and feature-wise precision parameters on the data.

Poisson likelihood for count data

When observations are a 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 the rate function and has to be convex and log-concave in order to ensure that the likelihood is log-concave.
As done in [223], here we choose the following rate function: $\lambda(c) = \log(1 + e^c)$.

Then an upper bound of the second derivative of the log-likelihood is given by

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

The pseudodata updates are given by

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

1.5.7 Model validation with simulated data

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

1.5.7.1 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, ten model instances were trained for every simulation scenario.
We observe that in most settings the model accurately recovers the correct number of factors (Figure 1.37). 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.

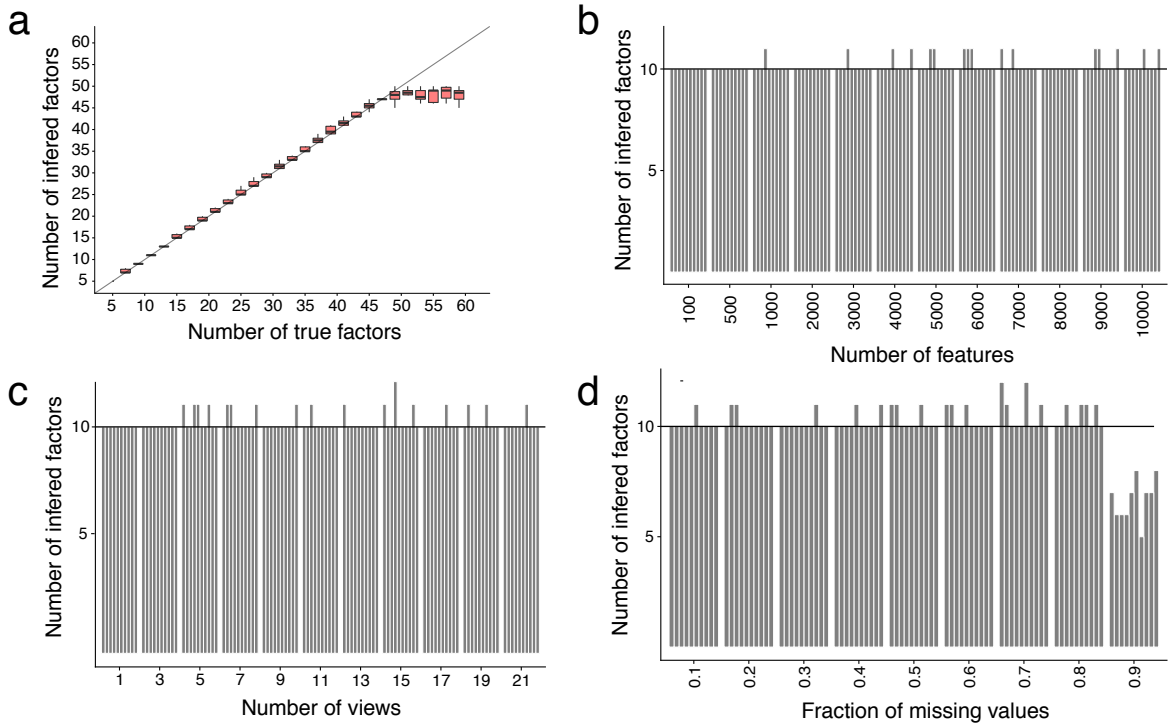


Figure 1.37: Assessing the ability to recover simulated factors.
In all plots the y-axis displays the number of infered factors. (a) x-axis displays the number of true factors, and boxplots summarise the distribution across 10 model instances. For (c-d) the true number of factors was set to $K = 10$ 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.

1.5.7.2 View-wise sparsity on the weights

One of the most important statistical assumptions underlying MOFA is the ARD prior aimed at disentangling the activity of factors across views (see Section 1.4.3.5 and Section 1.5.1.
We simulated data from the generative model were 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 [169]. In iCluster each factor shares the same sparsity constraint across all views, and hence the model is less accurate when it comes to the detection of factors that show differential activity across different views.:

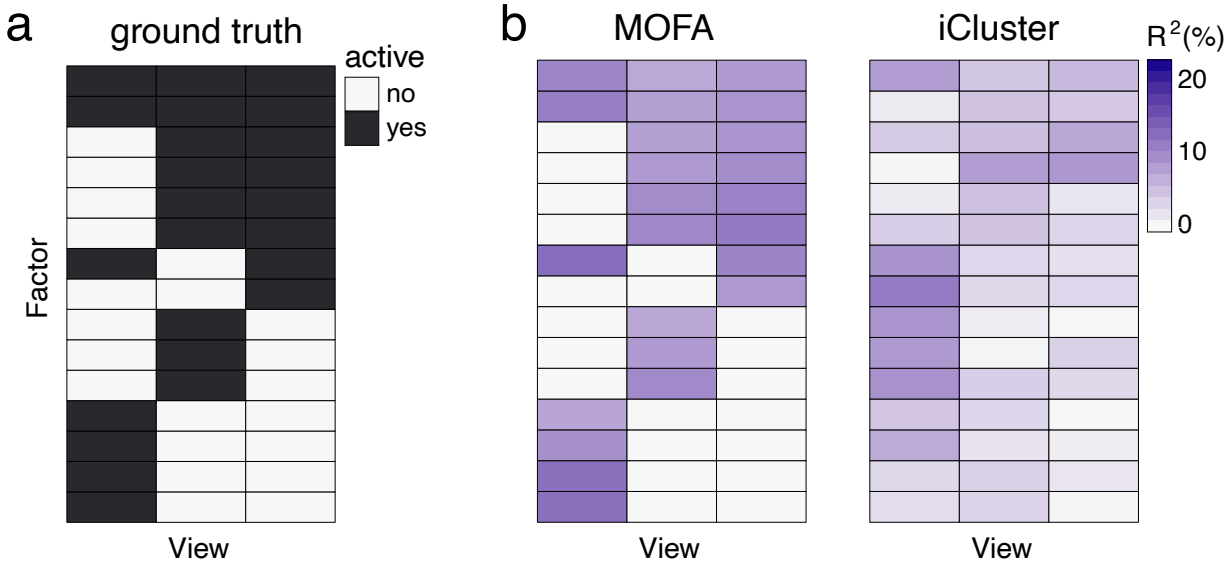


Figure 1.38: Evaluating the ability to recover differential factor activity across views. (a) The true activity pattern, with factors sampled to display differential activity across views. (b) Percentage of variance explained for each factor in each view, for MOFA and iCluster[169].

1.5.7.3 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 1.5.1.2).
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, the model without spike-and-slab priors contains the ARD prior, which should provide some degree of regularisation. To compare both options to a non-spare 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 1.39).

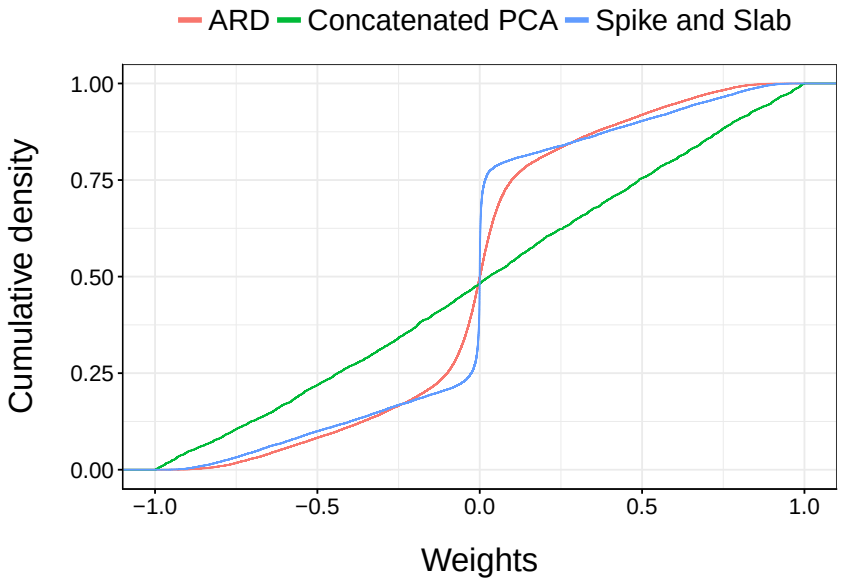


Figure 1.39: Assessing the sparsity priors on the weights.
The plot shows the empirical cumulative density function of the weights for an arbitrary factor in a single view. The weights were simulated with a sparsity level of $\theta_k^m = 0.5$ (50% of active features.)

1.5.7.4 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 1.5.6, 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:

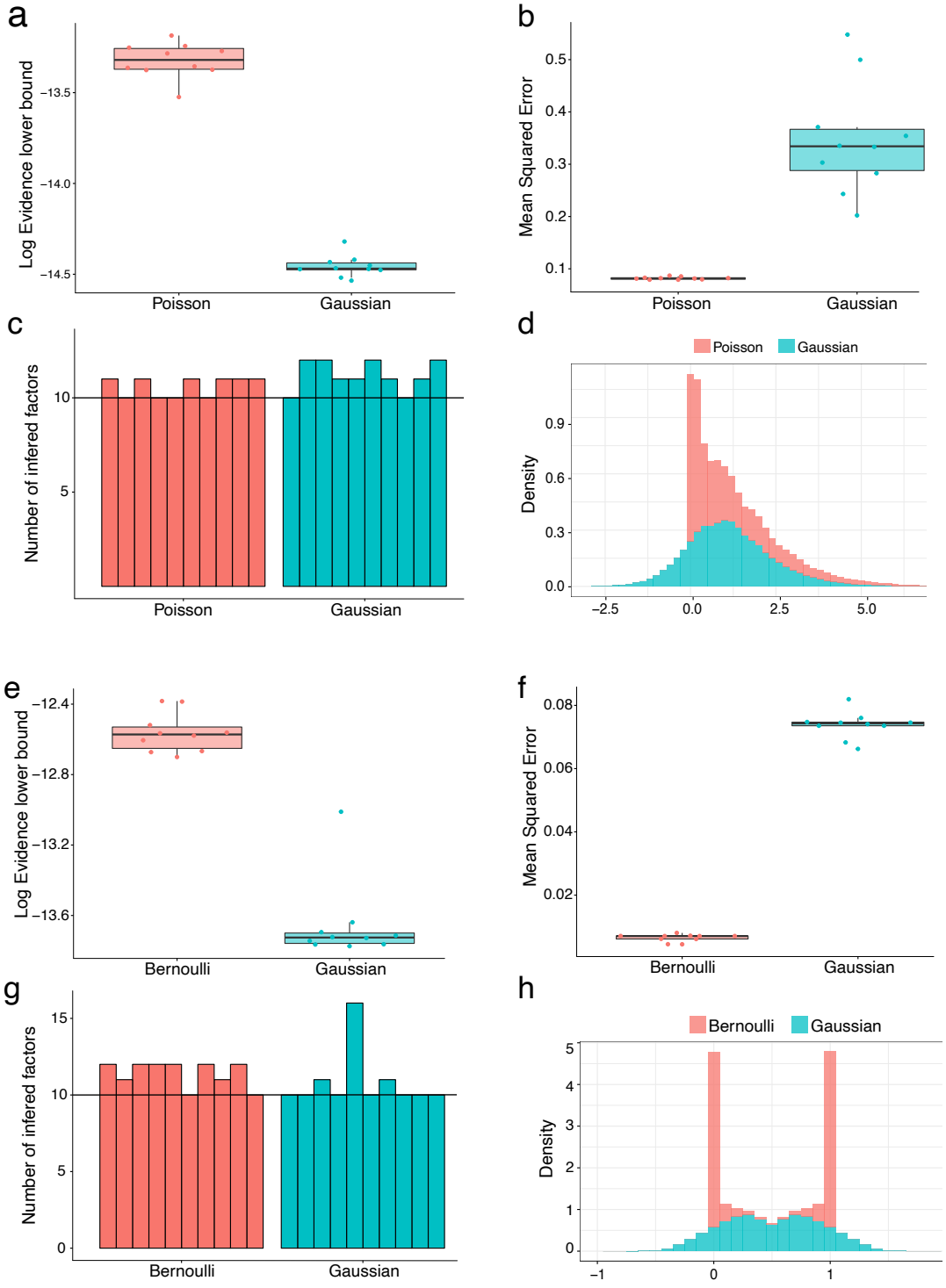


Figure 1.40: Validation of the non-gaussian likelihood models using simulated data. (a-d) Comparison of Poisson and Gaussian likelihood models applied to count data. (e-h) Comparison of Bernoulli and Gaussian likelihood models applied to binary data. (a,e) The y-axis displays the ELBO for each model 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.

1.5.8 Application to chronic lymphocytic leukaemia

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 [44, 51, 4].

To demonstrate the potential of the method, we applied MOFA 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[59]. We selected this data set for three main reasons: (1) The complex missing data structure, with nearly 40% samples having incomplete assays (see **??**). As described in Section 1.5.1.5, the inference framework implemented in MOFA should cope with large amounts of missing values, including missing assays. (2) The different data modalities: after data processing, three assays had continuous observations whereas for the somatic mutations the observations were binary. As described in Section 1.5.6, MOFA can combine different likelihood models. (3) The existence of clinical covariates: this provides an excellent test to evaluate whether the MOFA factors can capture the variation underlying clinically-relevant phenotypes.

1.5.8.1 Data overview and processing

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 [**Love2014**]. 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 [62]. 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.
- **TO FINISHSomatic 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.

For more details on the data generation steps we refer the reader to [59].

1.5.8.2 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 detects Factors which are shared across several

data modalities (Factors 1 and 2, sorted by variance explained). Some factors capture sources of covariation between two data modalities (Factor 3 and 5, active in the RNA expression and drug response). In addition, some factors capture 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.

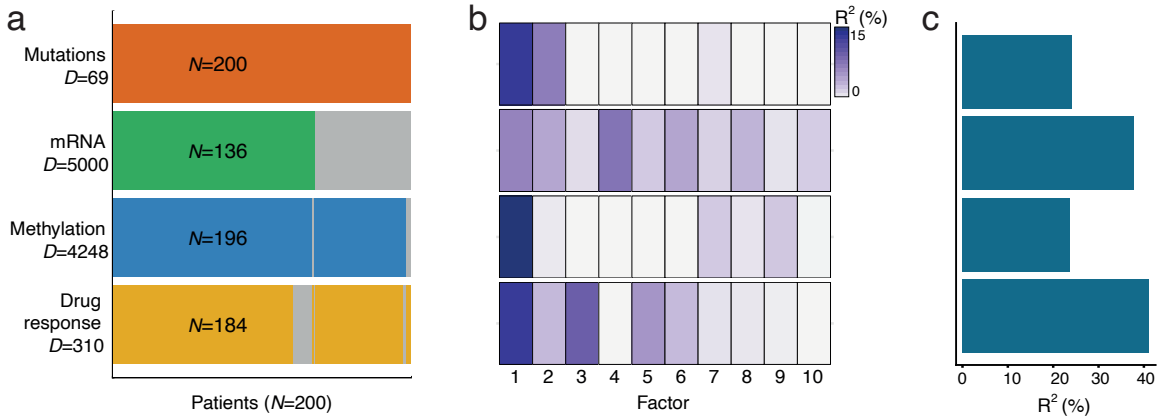


Figure 1.41: Application of MOFA to a study of chronic lymphocytic leukaemia. Model overview.
(a) Data overview. Assays are shown in different rows (D = number of features) and samples (N) in columns, with missing samples shown using grey bars. Notice that some samples are missing entire assays.
(b) 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 all molecular layers, from the genome to the transcriptome and ultimately in the drug response assay.
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 1.42). 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 [67, 34, 54].

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 1.42). Note that this latent representation enables simple patient stratification into molecular subgroups (see dashed lines), a first step towards personalised medicine.
Interestingly, 24 patients lacked IGHV status measurements (see grey crosses) due to quality control filtering in the DNA sequencing assay. Nonetheless, MOFA is 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.

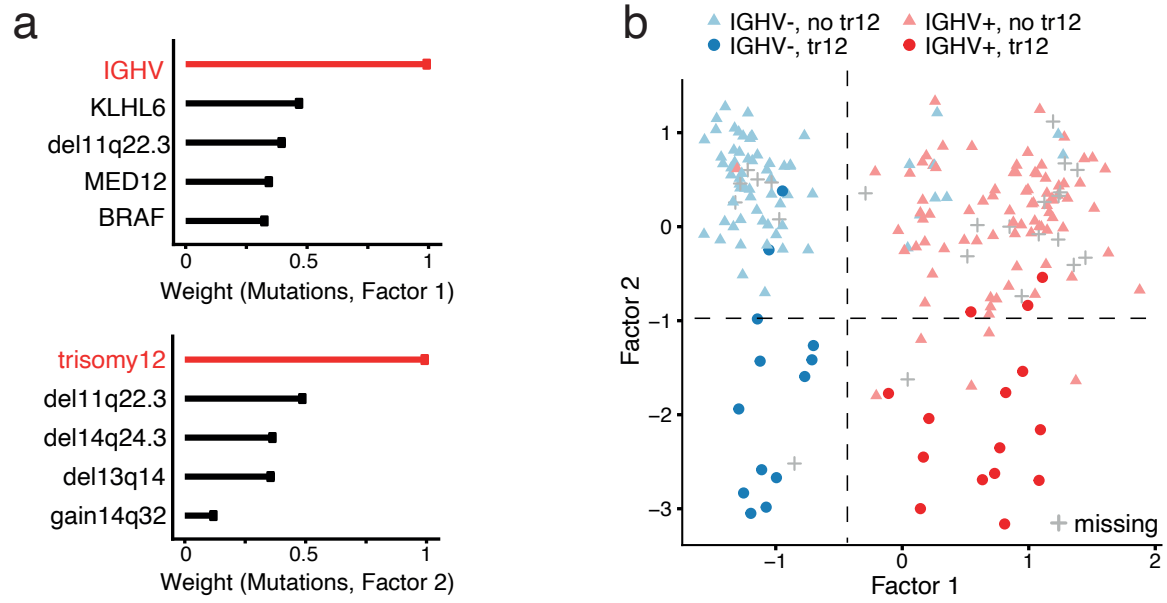


Figure 1.42: Visualisation of the genetic signature underlying Factor 1 and 2
(a) Absolute loadings of the top features of Factors 1 and 2 in the Mutations data. (b) Visualization of samples using Factors 1 and 2. The colours denote the IGHV status of the tumours; symbol shape and colour tone indicate 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[67]. 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 resposne to chemoimmunotherapy, whereas unmutated IGHV patients have a worse prognosis [67, 34, 54].
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 (Figures 1.42 and 1.43), as also suggested in [198].

1.5.8.3 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[257, 171]. Heatmaps of the RNA expression levels for these genes reveals clear differences between samples when ordinated 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.

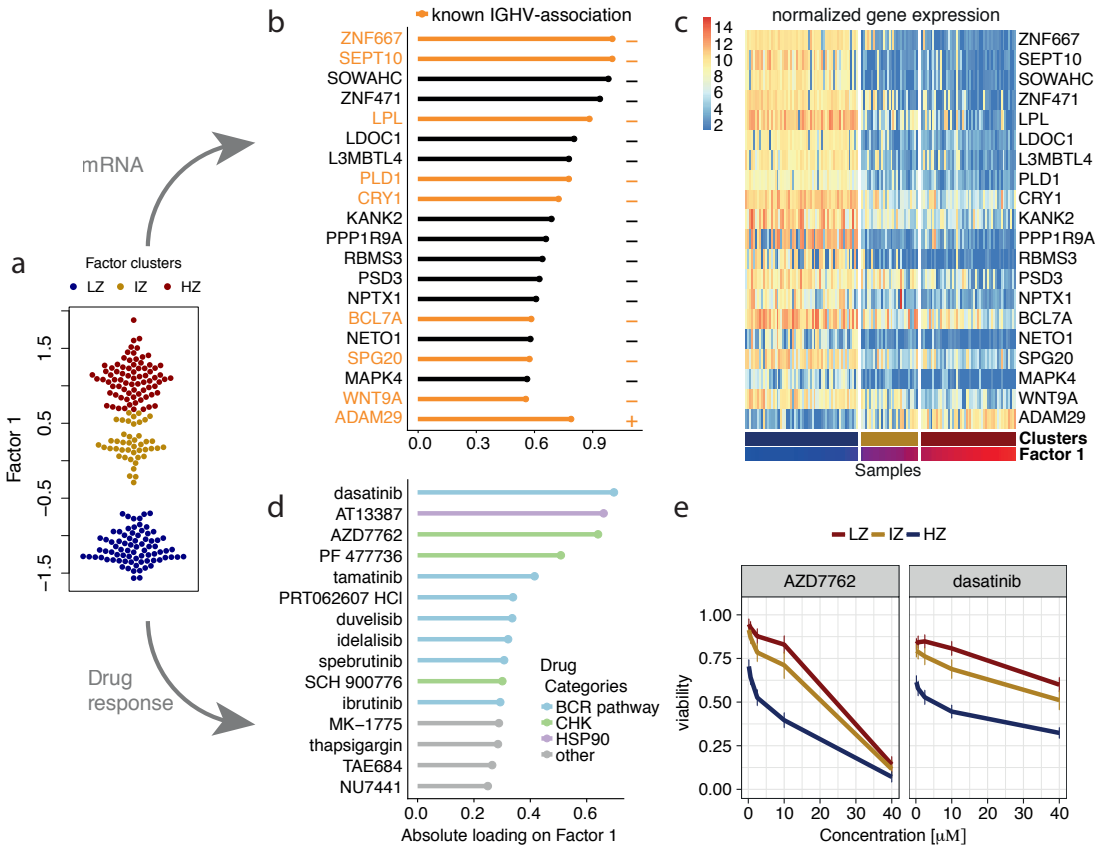


Figure 1.43: Characterization of MOFA Factor 1 as IGHV status.
(a) Beeswarm plot with Factor 1 values for each sample with colours corresponding to three groups found by 3-means clustering with low factor values (LZ), intermediate factor values (IZ) and high factor values (HZ).
(b) Absolute weights for the genes with the largest absolute weights in the mRNA data. Plus or minus symbols on the right indicate the sign of the loading. Genes highlighted in orange were previously described as prognostic markers in CLL and associated with IGHV status (Vasconcelos et al, 2005; Maloum et al, 2009; Trojani et al, 2012; Morabito et al, 2015; Plesingerova et al, 2017).
(c) Heatmap of gene expression values for genes with the largest weights as in (b).
(d) Absolute weights of the drugs with the largest weights, annotated by target category.
(e) Drug response curves for two of the drugs with top weights, stratified by the clusters as in (a).

1.5.8.4 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 1.5.2, 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 [**Fabregat2015**, **Ashburner2000**].

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 [74],

in MOFA we implemented several scoring schemes and a variety of parametric and unparametric statistical tests. We refer the reader to [74] for details.

Appling Gene Set Enrichment Analysis on the RNA weights using the Reactome annotations [**Fabregat2015**] 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[57].
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 (SD07, MIS-43, SD51) and DNA damage response (fludarabine, nutlin-3, doxorubicine) (Figure S12c-d).

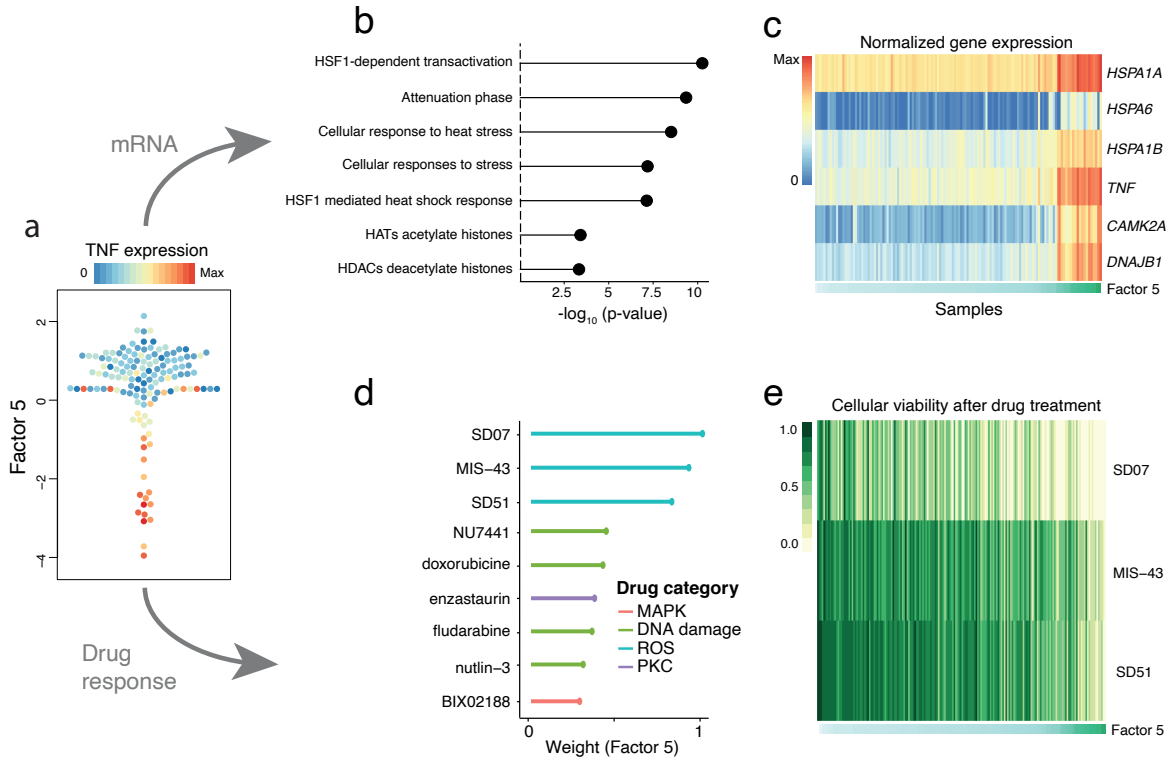


Figure 1.44: Characterization of Factor 5 in the CLL data as oxidative stress response.
(a) Beeswarm plot of Factor 5. Colours denote the expression of TNF, an inflammatory stress marker that is present among the top RNA weights.
(b) Gene set enrichment analysis for the top Reactome pathways. Displayed are the top pathways with the strongest enrichment in the RNA weights. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.
(c) Heatmap of mRNA expression values for representative genes with the largest weights. Samples are ordered by their factor values.
(d) Scaled weights for the top drugs with the largest loading, annotated by target category.
(e) Heatmap of drug response values for the top three drugs with largest weight.

1.5.8.5 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 [52] 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 1.45).

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).

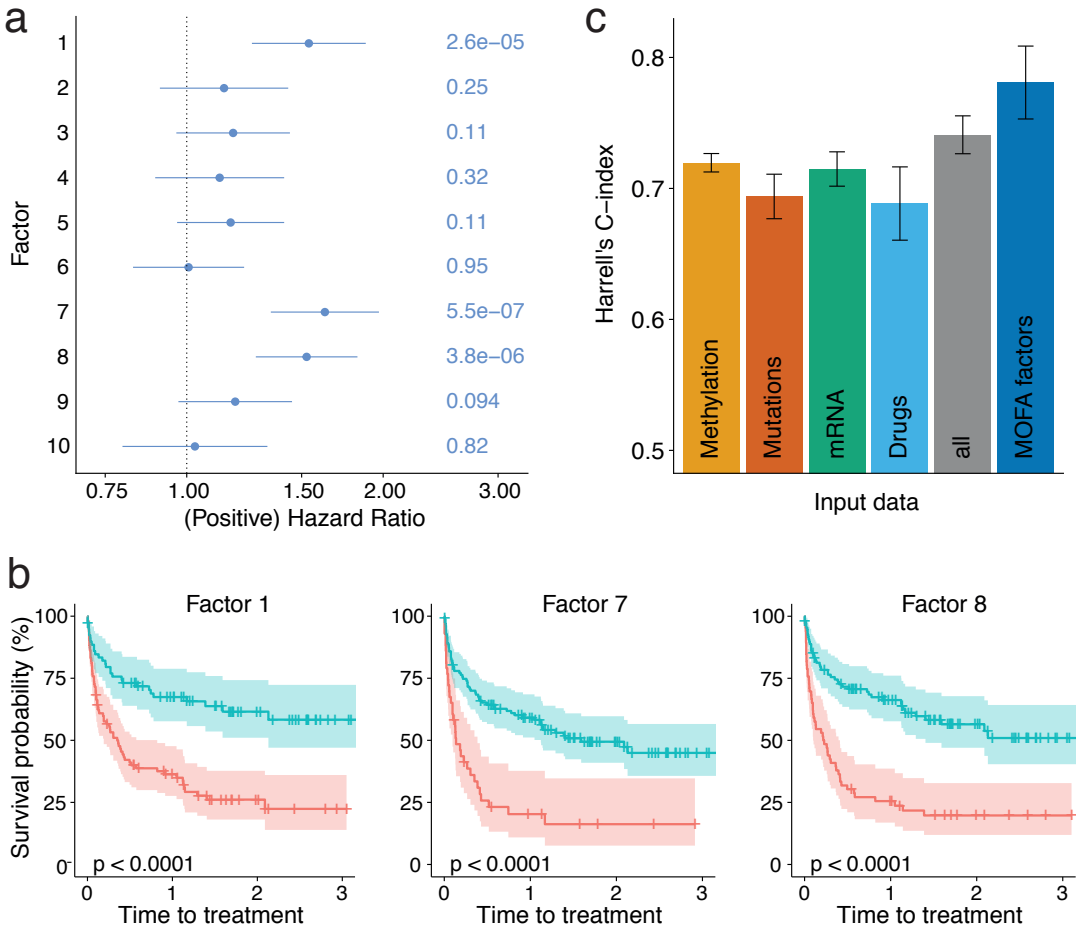


Figure 1.45: Association analysis between MOFA Factors and clinical putcome.
(a) Association of MOFA factors to time to next treatment using a univariate Cox regression with N = 174 samples (96 of which are uncensored cases) and p-values based on the Wald statistic. Error bars denote 95% confidence intervals. Numbers on the right denote p-values for each predictor.
(b) Kaplan-Meier plots measuring time to next treatment for the individual MOFA factors. The cut-points on each factor were chosen using maximally selected rank statistics, and p-values were calculated using a log-rank test on the resulting groups.
(c) Prediction accuracy of time to treatment for N = 174 patients using multivariate Cox regression trained with the 10 MOFA factors, as well using the first 10 principal components applied to single data modalities and the full data set. Shown are average values of Harrell's C-index from fivefold cross-validation. Error bars denote standard error of the mean.

1.5.8.6 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 (3.17)) has a mean of zero and does not contribute to the predictions.
The equation above computes a point estimate for every sample n and feature d, 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 [77]. Nonetheless, due to the nature of the optimisation problem in variational inference, the variance of the posterior distributions can be underestimated (see Section 1.4.1.8). In addition, this would more complex to implement and would result in a significant increase in computational complexity. Hence, and also because of the additional computational complexity, we did not attempt this approach.

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 [160], a k-nearest neighbour method [254].
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.

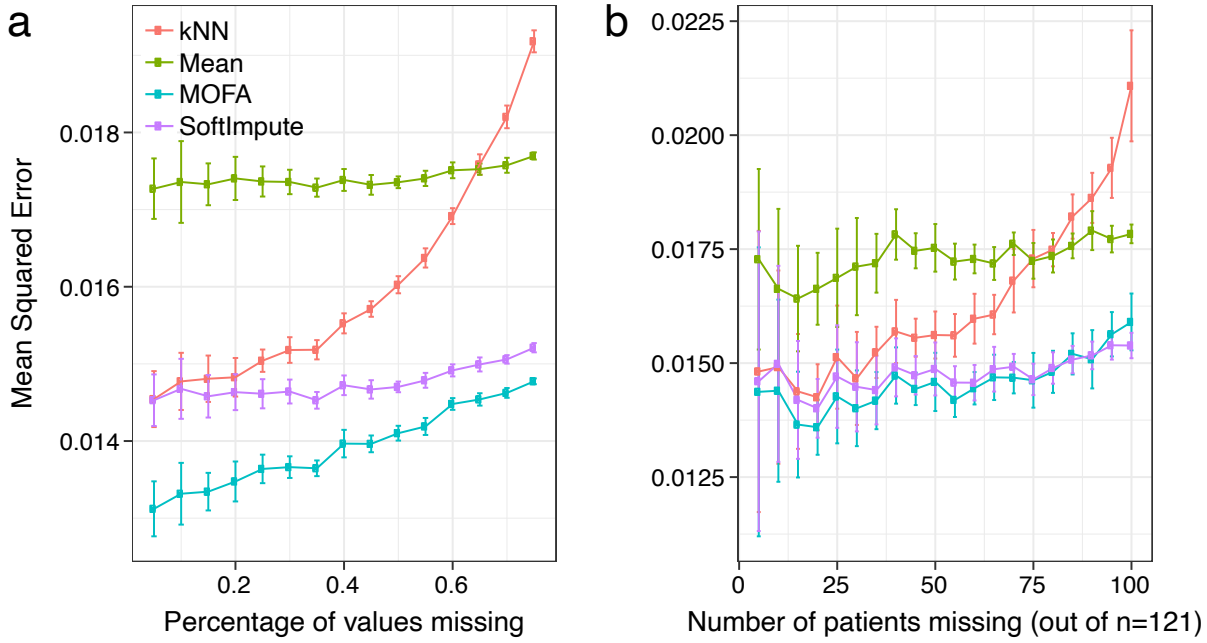


Figure 1.46: Evaluation of imputation performance in the drug response assay.
The y-axis shows the mean-squared error (MSE) across 15 trials for increasing fractions of missing data (x-axis). Two experiments were considered: (a) values missing at random and (b) entire assays missing at random. Each point displays the mean across all trials and the error bars depict the corresponding standard deviations.

1.5.9 Application to single-cell multi-omics

The emergence of single-cell multi-modal techniques has created open opportunities for the development of novel computational strategies [239, 49, 42].
To show case how MOFA can be used to integrate single-cell multi-omics data, we considered a simple data set that consists on 87 ESCs where RNA expression and DNA methylation were simultaneously measured using single-cell Methylation and Transcriptome sequencing (scM&T-seq)[8]. Two populations of ESCs were profiled: the first one contains 16 cells grown in 2i media, which is known to induce a native pluripotency state associated with genome-wide DNA hypomethylation [71]. The second population contains 71 cells grown in serum media, which contain stimuli that trigger a primed pluripotency state poised for differentiation [251].

1.5.10 Data processing

The RNA expression data was processed using *scran*[154] to obtain log normalised counts adjusted by library size. Feature selection was performed by selecting the top 5,000 most overdispersed genes[134]. A Gaussian likelihod was used for this data modality.
The DNA methylation data was processed as described in Chapter 1. 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.

1.5.10.1 Model overview

In this data set, MOFA inferred 3 factors with a minimum explained variance of 1% (Figure 1.47). 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* [170]. As previously described both *in vitro* [8] and *in vivo* [12], 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 [75].
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.

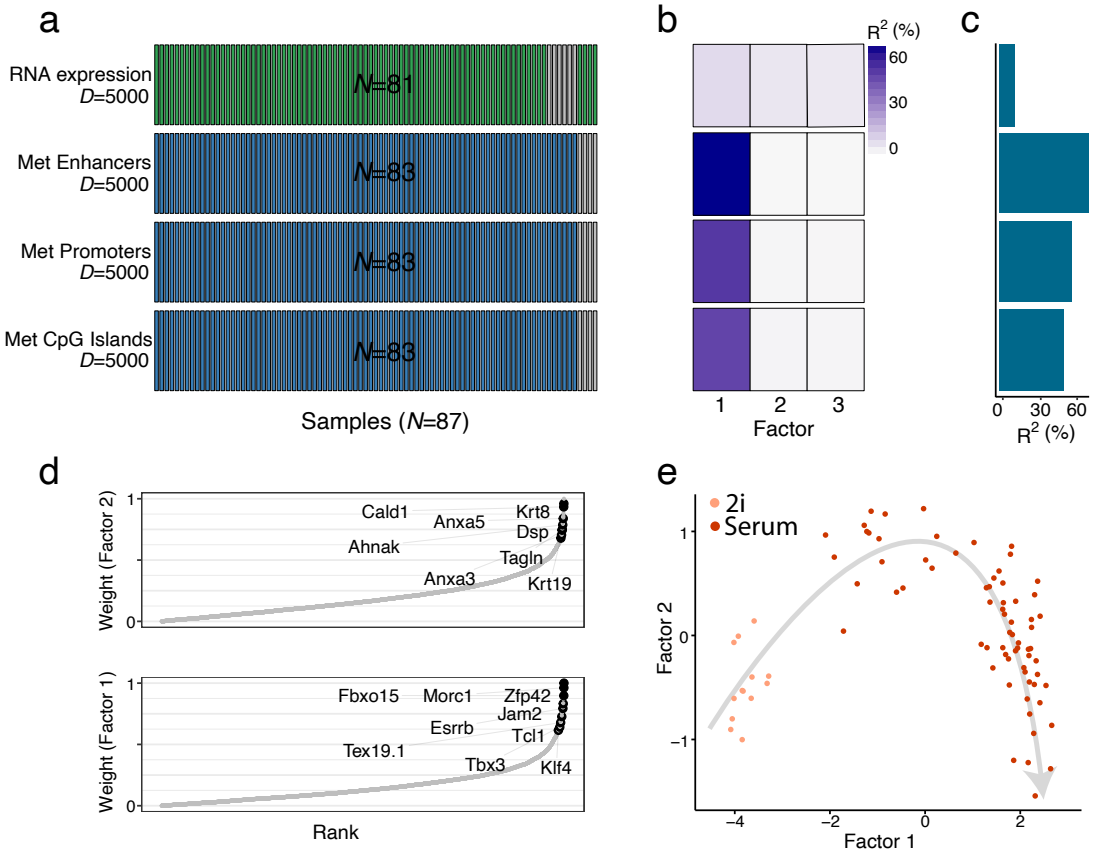


Figure 1.47: MOFA recovers a differentiation process from a single-cell multi-omics data set.
(a) Overview of the data modalities. Rows indicate number of features (D) and columns indicate number of samples (N). Grey bars denote missing samples.
(b) Fraction of variance explained per factor (column) 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.

1.5.11 Limitations and open perspectives

MOFA solves important challenges for the integrative analysis of (single-cell) multi-omics data sets. 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[133]. Non-linear approaches, including deep neural networks or variational autoencoders have shown promising results when it comes to dimensionality reduction [147, 61, 153], batch correction[153], denoising [66] or imputation [148]. Interestngly, 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 [241, 39].
  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 [96, 24, 97]. 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.

- **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 1.32). 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.

- **Tailored likelihoods for single-cell analysis**: MOFA enables the modular extension to arbitrary non-gaussian likelihoods, provided that they can be locally bounded and integrated into the variational framework (see Section 1.5.6). New likelihood models such as zero-inflated negative binomial distributions [212] could make MOFA more suited to the analysis of single-cell data.

- **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 [77].

- **Incorporation of prior information**: an unsupevised 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 [33].

Chapter 2

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

In this chapter I will describe a study where we combined scNMT-seq (Chapter 1) and MOFA (Chapter 2) to explore the relationship between the transcriptome and the epigenome during mouse gastrulation.

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 [9]. 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 all 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, Oliver Stegle and Wolf Reik supervised the project. The article was jointly written by Stephen Clark, Carine Stapel and me, with input from all authors.

2.1 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 [110]. 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 [**Solnica-Krezel2012**, **Tam2007**, 244].

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 [**Arnold2009**, **Tam2007**, 244, 243]. Although differences exist between species, the morphogenic process of gastrulation is evolutionary conserved throughout the animal kingdom [**Solnica-Krezel2012**]. 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.

2.1.1 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 genome-wide studies. Only after the breakthrough made by the introduction of single-cell sequencing technologies it has been possible to generate comprehensive molecular roadmaps of embryonic development [**PijuanSala2019**, 222, 39, 208]. In a pioneer study, [**PijuanSala2019**] 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 [**Scialdone2016**, **Ibarra-Soria2018**, **Wen2017**, 170]. 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 [39]. By constructing a densely sampled reference data set, both works have let the ground for understanding transcriptomic variation during development.

2.1.2 Epigenetic studies

RNA expression is a big and central piece in the puzzle of understanding embryonic development, but still a single piece. The next step is to connect this information to the accompanying epigenetic changes, which are becoming more accessible to profile with single-cell technologies. 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 [117].

2.1.2.1 Pre-implantation: establishment of the pluripotent state

The first efforts to interrogate the epigenetic dynamics 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 [**Smith2012**, 139]. DNA methylation is globally removed and the chromatin attains its highest levels of accessibility [266]. Consistently, Hi-C experiments have suggested a flexible chromatin landscape, with lack of topologically associating domains (TADs) or chromatin compartments [115, 63, 245], providing a plausible explanation for the remarkably 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 [87, 245]. 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 [278]. 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 [18, 245].

2.1.2.2 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 the myriad of 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, both at the bulk level and at the single cell level [**Auclar2014**, **Dai2016**, 272, 217]. All studies found that the hypomethylated state in E3.5 blastocysts is followed by a *de novo* DNA methylation wave upon implantation (between E4.5 and E5.5) 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 [11].
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 [12, 181]. Interestingly, mouse ESCs remain viable despite complete loss of DNA methylation, but they are uncapable of undergoing cell fate commitment and remain in the pluripotent state [256].

The interplay of histone marks during post-implantation development is complex and remains poorly understood. H3K4me3 is detected at transcription start sites after the zygotic genome activation, and remains remarkably stable across different pluripotency stages as well as in differentiated cell types [94]. H3K4me3 is thought to facilitate transcription by inducing a more efficient assembly of the transcriptional machinery [11, 259]. 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 [11, 199].
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 [29, 11]. Interestingly, inactivating PRC2 components in mouse embryos does not affect pre-implantation development, but the embryos become unviable after gastrulation[224]. This suggests that H3K27me3 has a critical role in regulating gene expression during cell fate commitment after germ layer specification.

2.1.2.3 Gastrulation: germ layer specification

The post-implantation blastocyst is relatively homogeneous and can be characterised to 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 have been reported where the authors attempted to manually dissect each germ layer, followed by bulk sequencing [273]. This revealed that the relatively homogeneous epigenetic landscape at the E6.5 epiblast is succedeed by a more dynamic landscape, driven by the emergence of regulatory elements that become activated in a lineage-specific manner, as suggested by extensive demethylation events [273, 140]. 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.[**Dai2016**].

I envision that the recent development of single-cell multi-modal technologies (described in Section 1.1.3), where epigenomes can be unequivocally assigned to transcriptomes at single-cell resolution, will unveil novel opportunities to study the cell fate commitment events during gastrulation. These methodologies have been successful in pre-implantation stages [81, 263, 151] and early post-implantation development [217], but gastrulation has remained elusive.

2.2 Results

2.2.1 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 1) 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:

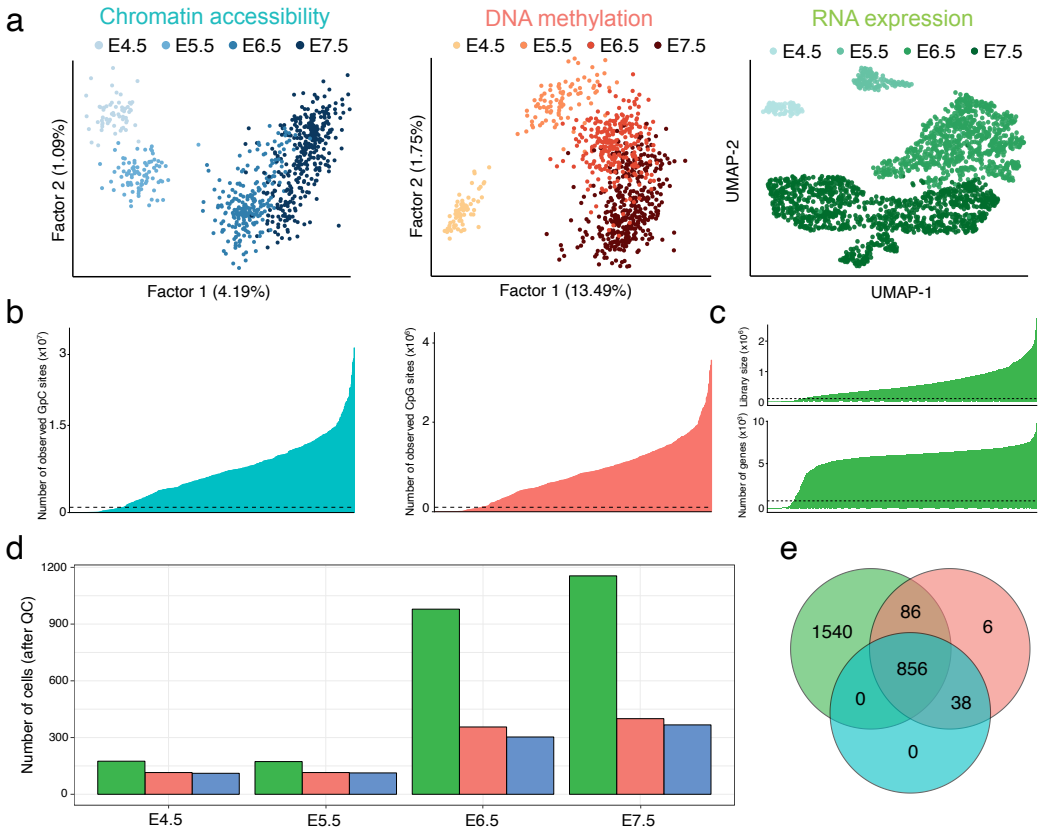


Figure 2.1: scNMT-seq gastrulation atlas. Data set overview.
(a) Dimensionality reduction for chromatin accessibility data (left, in blue), DNA methylation (middle, in red) and RNA expression (right, in green). For the gene expression data we applied UMAP[164]. For chromatin accessibility and DNA methylation data we applied Bayesian Factor Analysis[10].
(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 one cell, and 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 discarded 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 discarded on the basis of poor coverage.
(d) Number of cells that pass quality control for each molecular layer, grouped by stage. Note that for 1,419 out of 2,524 total cells only the RNA expression was sequenced.
(e) Venn Diagram displaying the number of cells that pass quality control for RNA expression (green), DNA methylation (red), chromatin accessibility (blue).

2.2.1.1 Validation of DNA methylation data and chromatin accessibility data

To validate the DNA methylation and chromatin accessibility data, we performed dimensionality reduction across 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 2).

Reassuringly, we observe that for both modalities the model with all cells captures a developmental progression from E4.5 to E7.5 (Figure 2.1). 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 extraembryonic origin, as expected (Figure 2.2). At E7.5 extra-embryonic cells were manually removed during the dissection and the first two latent factors discriminate the three germ layers (Figure 2.2).

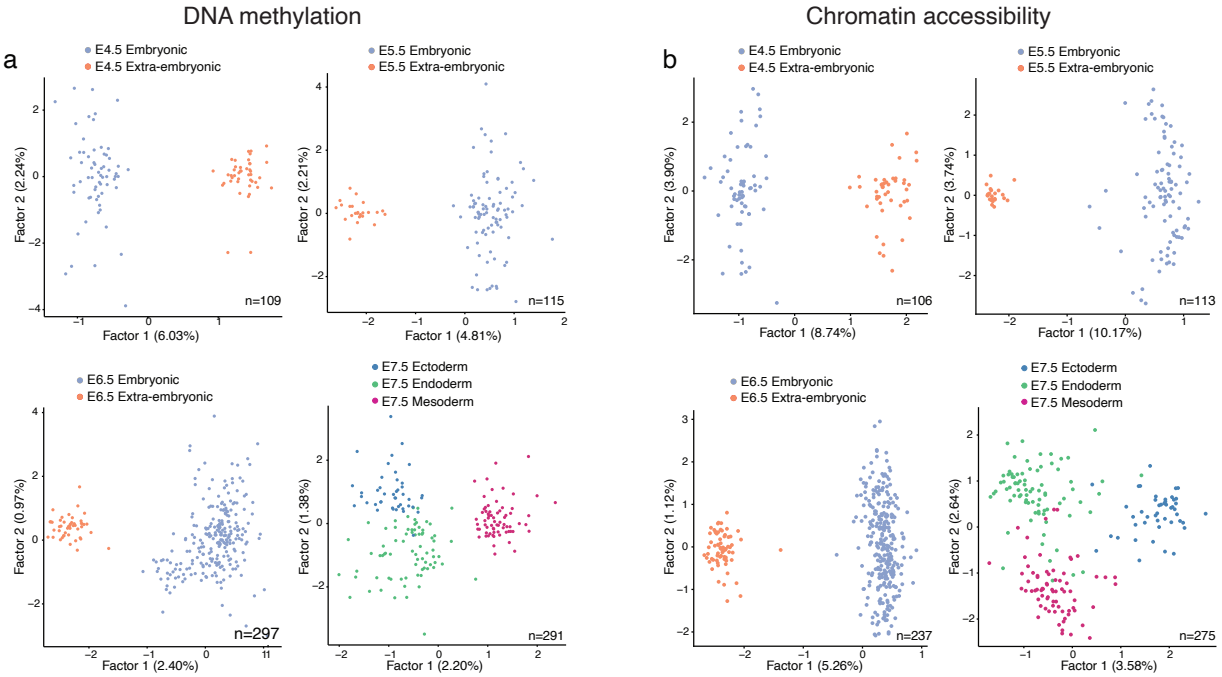


Figure 2.2: Dimensionality reduction of (a) DNA methylation and (b) chromatin accessibility data. Shown are scatter plots of the first two latent factors (sorted by variance explained) for models trained with cells from the indicated stages. From E4.5 to E6.5 cells are coloured by embryonic and extra-embryonic origin. At E7.5, cells are coloured by the primary germ layer.

2.2.2 Cell type assignment using the RNA expression data

To define cell type annotations we followed two independnet stategies. For the E6.5 and E7.5 stages, we mapped the RNA expression profiles to the single-cell gastrulation atlas [192] (stages E6.5 to E8.0) using a matching mutual nearest neighbours algorithm [85]. In short, the count matrices for both data sets were concatenated and normalised together. Then, Principal Component Analysis was applied, followed by batch correction in the atlas to remove the technical variability between experiments. 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.
For the E4.5 and E5.5 stages, we used a consensus clustering method [122], as no transcriptomic atlas was available for these stages:

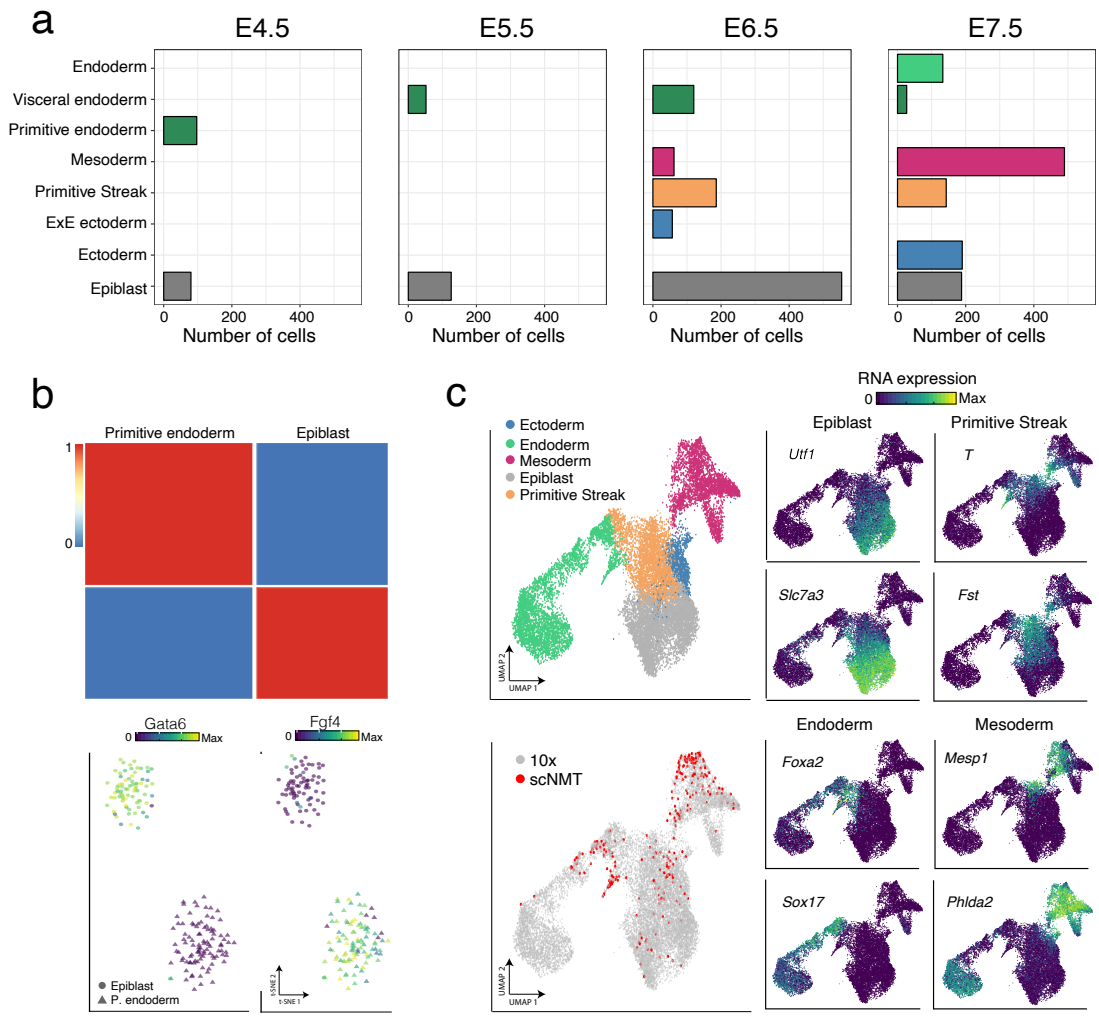


Figure 2.3: Cell type assignments using the RNA expression data.
(a) For each stage, the bar plots display the number of cells assigned to each lineage.
(b) Cell type assignment for E4.5 cells. The heatmap displays the consensus plot, representing the similarity between cells based on the averaging of clustering results from multiple combinations of clustering parameters[122]. A similarity of 0 (blue) indicates that the two cells are always assigned to different clusters, whereas a similarity of 1 (red) means that the two cells are always assigned to the same cluster.
The scatter plot displays a t-SNE representation of the RNA expression data coloured by the expression of *Fgf4*, a known E4.5 epiblast marker and *Gata6*, a known E4.5 primitive endoderm marker.
(c) 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 neighbors 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.

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

First, we explored the changes in DNA methylation and chromatin accessibility along each stage transition. Globally, CpG methylation levels rise from ≈ 25% to ≈ 75% in the embryonic tissue and ≈ 50% in the extra-embryonic tissue **??**, mainly driven by a *de novo* methylation wave from E4.5 to E5.5 that preferentially targets CpG-poor genomic loci [12, 272] (Figure 2.4).
In contrast to the sharp increase in DNA methylation between E4.5 and E5.5, we observed a more gradual decline in global chromatin accessibility from ≈ 38% at E4.5 to ≈ 29% at E7.5, with no significant differences between embryonic and extraembryonic tissues (t-test, Figure 2.5). Consistent with the DNA methylation changes, CpG-rich regions remain more accessible than CpG-poor regions of the genome.

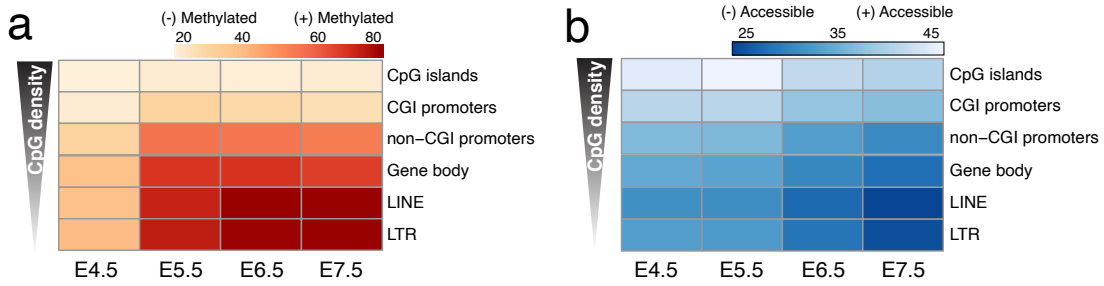


Figure 2.4: DNA methylation and chromatin accessibility levels per stage and genomic context.
Heatmaps display the mean levels across cells within a particular stage and across all loci within a particular genomic context.

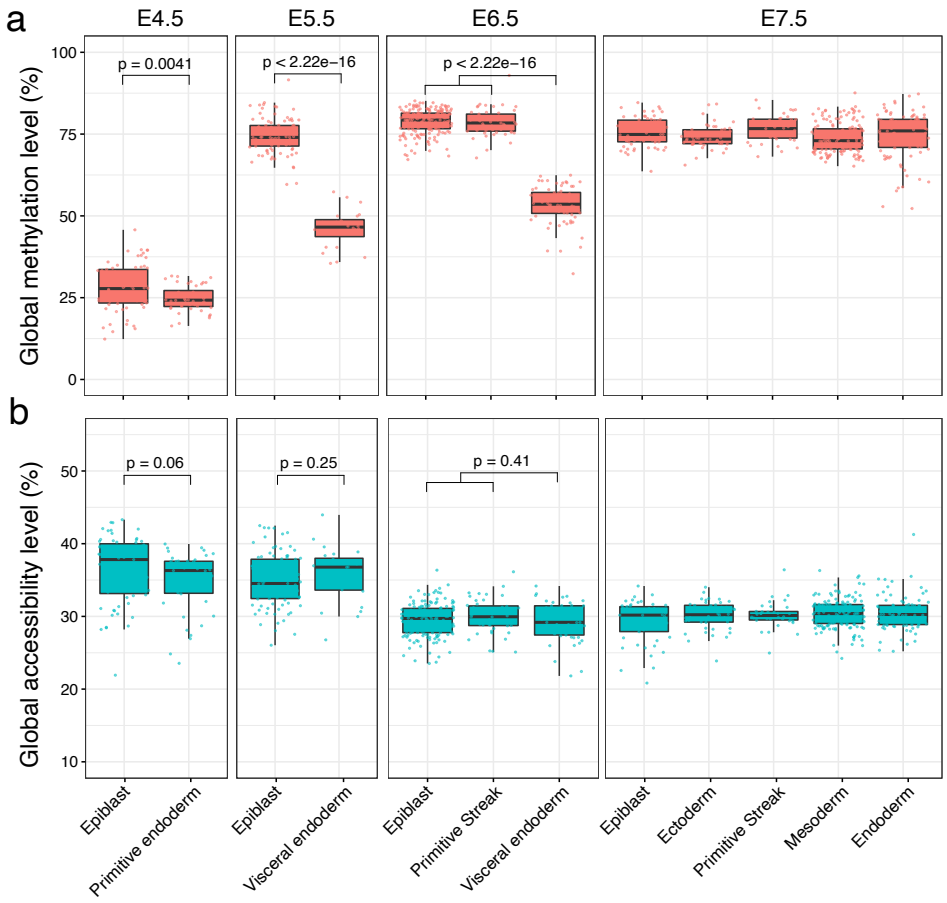


Figure 2.5: Global DNA methylation and chromatin accessibility levels per stage and lineage.
Box plots showing the distribution of genome-wide (a) CpG methylation levels or (b) GpC accessibility levels per stage and lineage. Each dot represents a single cell.

Next, we attempted to characterise the relationship between the transcriptome and the epigenome along differentiation. For simplicity we focused on gene promoters, as RNA expression and epigenetic readouts can be unambiguously matched. We calculated for each gene, the correlation coefficient between RNA expression and the corresponding DNA methylation or chromatin accessibility levels at its promoter (defined as 2kb up and downstream from the transcription start site). 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 [27].
We identified 125 genes whose expression shows significant correlation with promoter DNA methylation and 52 that show a significant correlation with chromatin accessibility Figure 2.6. Among the top hits we identify early pluripotency and germ cell markers, including *Dppa4*, *Dppa5a*, *Rex1*, *Tex19.1* and *Pou3f1* (Figure 2.6). Notably, all of them have a negative association between RNA expression and DNA methylation and a positive associatin between RNA expression and chromatin accessibility. Inspection of the transcriptomic and epigenetic dynamics reveals that the repression of these early pluripotency markets are concomitant with the genome-wide trend of DNA methylation gain and chromatin closure.

In addition, this analysis identifies novel genes, including *Trap1a*, *Zfp981*, *Zfp985*, as well as a number of metabolism genes (e.g. *Apoc1*, *Pla2g1b*, *Pla2g10*) that may have yet unknown roles in pluripotency or germ cell development.

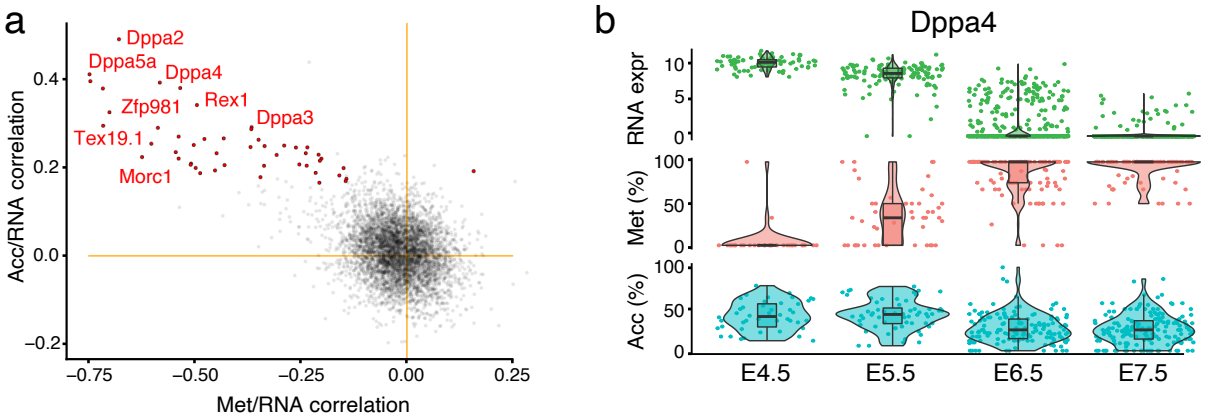


Figure 2.6: Genome-wide associacion analysis between RNA expression and the corresponding epigenetic status in gene promoters.
(a) Scatter plot of Pearson correlation coefficients between promoter DNA methylation versus RNA expression (x-axis); and promoter accessibility versus RNA expression (y-axis). Significant associations for both correlation types (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 (left) display the distribution of chromatin accessibility (% levels, blue), RNA expression (log2 counts, green) and DNA methylation (% levels, red) values per stage and lineage. Each dot corresponds to one cell.

2.2.4 Multi-omics factor analysis reveals coordinated variability between the transcriptome and the epigenome during germ layer formation

In the previous section we have 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. Instead of following a supervised approach, here we performed an unsupervised integrative analysis using Multi-Omics Factor Anaysis (MOFA, presented in Chapter 2). 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 2.7). 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, hence enabling the interpretation of the underlying molecular variation. 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.

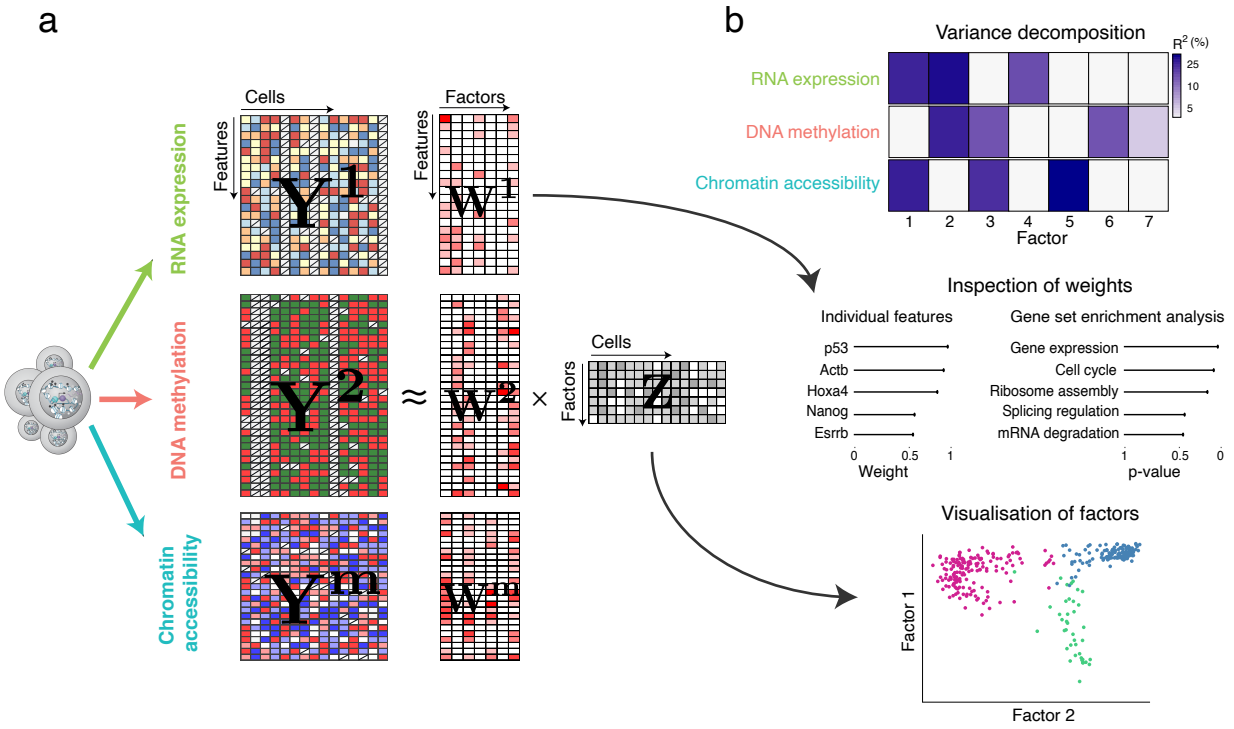


Figure 2.7: Multi-Omics Factor Analysis (MOFA): model overview and illustration of downstream analysis.
(a) Model overview: MOFA takes as input one or more data modalities (Y), extracted from the same samples (individual cells in this case). MOFA decomposes these matrices into a matrix of factors (Z) and a set of feature weight matrices (W), one for each data modality. The Z matrix contains the low dimensional representation of cells in terms of a few number of latent factors. The W matrices relate the low-dimensional space to the high-dimensional space by inferring a weight for each feature on each factor. When interpreting a factor, the absolute value of the loading is used as a measure of feature importance.
(b) Downstream analysis: the fitted MOFA model can be queried for different downstream analyses, including (i) variance decomposition, assessing the proportion of variance (R^2) explained by each factor in each data modality, (ii) semi-automated factor annotation based on the inspection of weights and gene set enrichment analysis, (iii) visualization of the samples in the factor space.

2.2.4.1 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.[267]. An overview on the numbers and the overlap of the lineage-specific histone marks is given in the following figure:

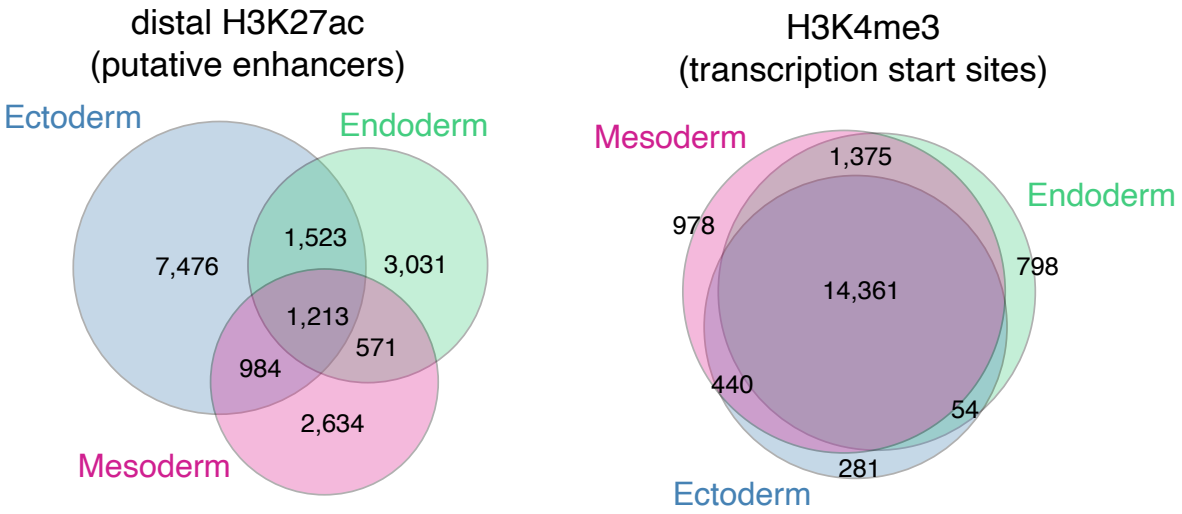


Figure 2.8: Venn diagrams showing overlap of peak calls for each lineage-specific histone mark, for distal H3K27ac (left) and all H3K4me3 (right). The figure shows that distal H3K27ac peaks (putative enhancer [53]) have moderate levels of overlap between the three germ layers. In contrast, H3K4me3 peaks (active transcription start sites [146]) are similar between the three germ layers.

Additionally, we quantified DNA methylation and chromatin accessibilty in gene promoters, again defined as 2kb upstream and downstream of the transcription start sites.

2.2.4.2 Model overview

MOFA identified 6 Factors capturing at least 1% of variance in the RNA expression data. The first two Factors (sorted by variance explained) captured the the emergence of the three germ layers. 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.

This supports other studies that identified distal elements as lineage-driving regulatory regions.

Interestingly, the effect sizes associated with regions that display differential demethylation and chromatin accessibility are moderate (less than 30% change) but coordinated across multiple enhancers (between 10% and 25% of the H3K27ac peaks).

Inspection of gene-enhancer associations identified enhancers linked to key germ layer markers including Lefty2, Mesp1, Mesp2 (mesoderm), Foxa2, Noto, Sox17 (endoderm), and Cxcl12, Sox2, Sp8 (ectoderm).

Intriguingly, ectoderm-specific enhancers show fewer associations than their meso- and endoderm counterparts, a finding that is explored further below.

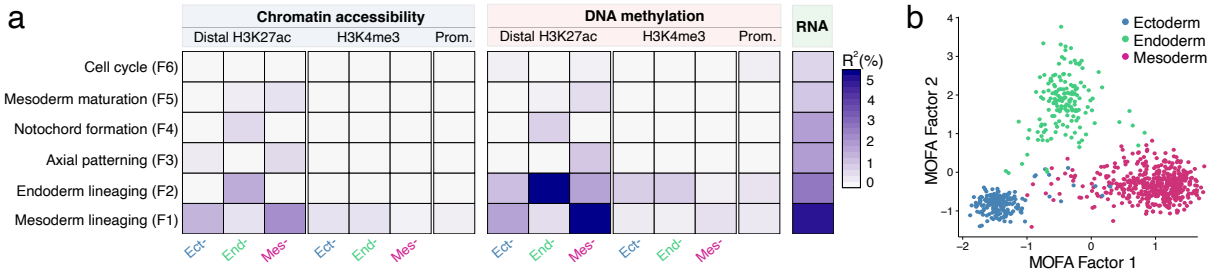


Figure 2.9: Multi-omics factor analysis reveals coordinated epigenetic and transcriptomic variation at enhancer elements during germ layer commitment.
(a) Percentage of variance explained by each MOFA factor (rows) across data modalities (columns). Considered data modalities were RNA expression quantified over protein-coding genes (green); DNA methylation (red) and chromatin accessibility (blue) quantified on promoters, lineage-specific H3K4me3-marked sites and distal H3K27ac-marked sites (enhancers). Factors are sorted by the total variance explained across all data modalities.
(b) Scatter plot of MOFA Factor 1 (x-axis) and MOFA Factor 2 (y-axis). Cells are coloured according to their lineage assignment (see Figure S2).

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

2.2.5 Characterisation of individual enhancers

The MOFA analysis in the previous section reveals interesting genome-wide trends. Next, we attempted to pinpoint individual enhancers that are representative of the global patterns.

SCATTERPLOT OF ENHANCERS

SCATTERPLOT OF PROMOTERS

EXAMPLE GENE

2.2.6 Transcription factor motif enrichment analysis

To identify transcription factors (TF)s 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). This information was then integrated with differential RNA expression between germ layers for the same TFs, quantified using the genewise negative binomial generalised linear model with quasi-likelihood test from edgeR. Not unexpectedly, this analysed 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 2.10).
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 T [**Tosic2019**].

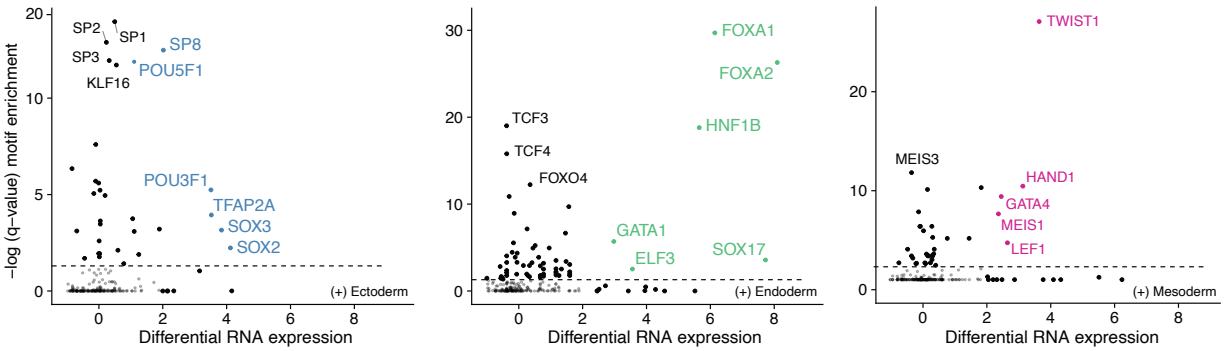


Figure 2.10: Transcription Factor motif enrichment analysis at lineage-defining distal H3K27ac sites. Shown is motif enrichment (-log10 q-value, y-axis) plotted against differential RNA expression (log fold change, x-axis) of the corresponding TF. The analysis is performed separately for each set of 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.

2.2.7 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 2.11). 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.

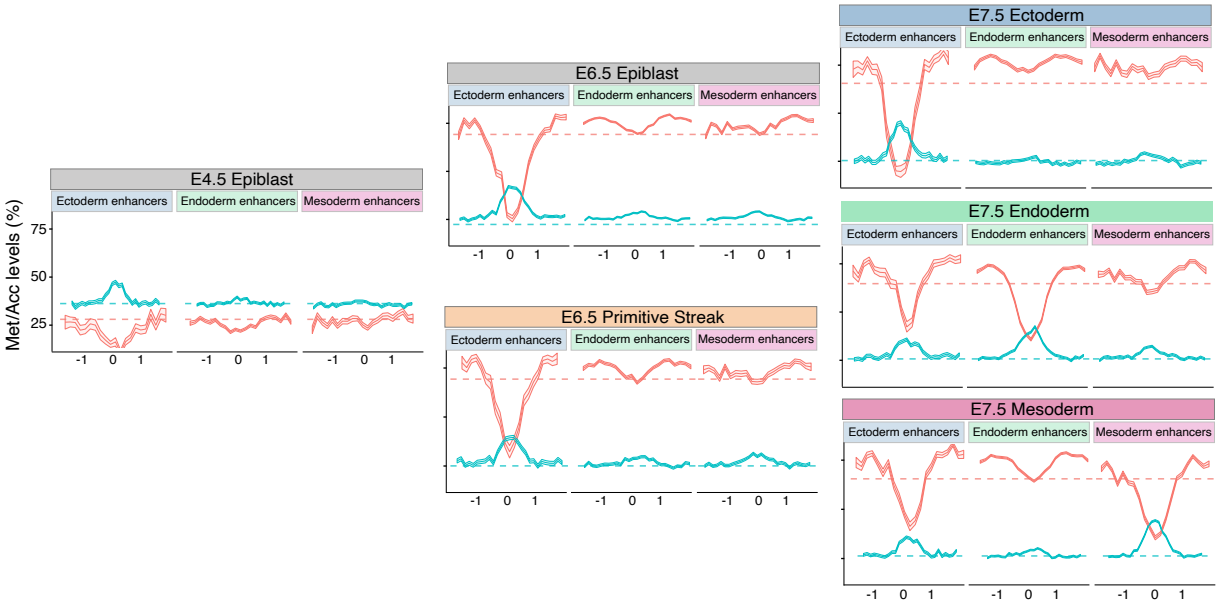


Figure 2.11: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers. Visualisation at pseudobulk resolution.
DNA methylation (red) and chromatin accessibility (blue) levels at lineage-defining enhancers quantified over different lineages across development. Shown are running 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:

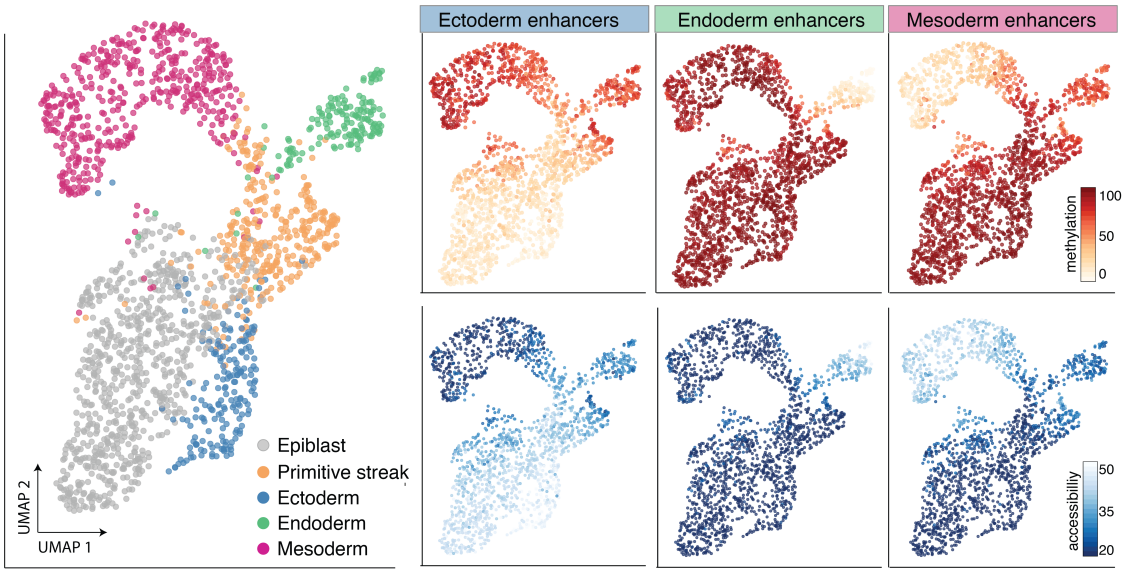


Figure 2.12: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers. Visualisation at single-cell resolution.
UMAP projection based on the MOFA factors inferred using all cells. In the left plot the cells are coloured according to their lineage. In the right plots cells are 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.

2.2.7.1 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 2.11 and Figure 2.12).
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.

2.2.7.2 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 2.11

and Figure 2.12).
Upon commitment to mesoderm and endoderm, ectoderm enhancers become partially repressed.

Two hypothesis 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[**Feng2014**]
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 2.13. This result supports the first hypothesis, but does not rule out the second hypothesis.

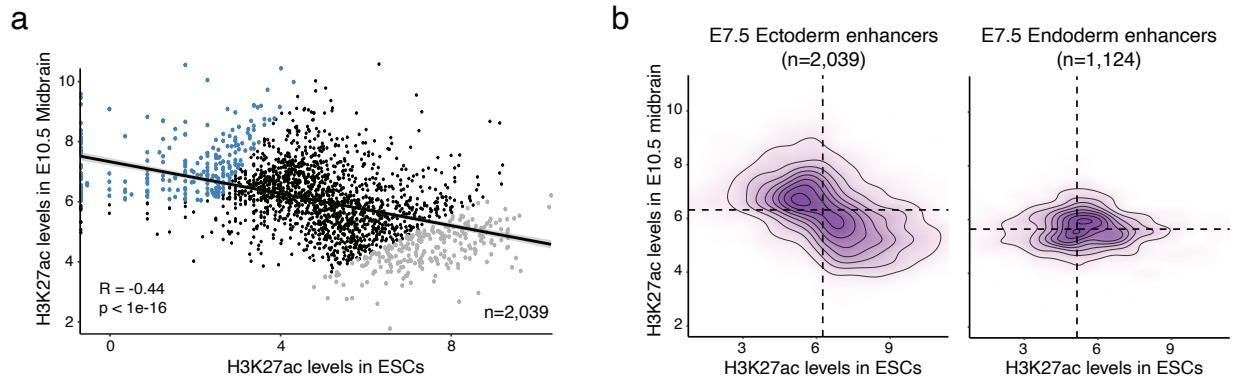


Figure 2.13: E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures.
(a) Scatter plot of ectoderm enhancers' H3K27ac levels quantified in ESCs (pluripotency enhancers, x-axis) and E10.5 midbrain (neuroectoderm enhancers, y-axis). Each dot corresponds to an ectoderm enhancer (Figure 2.8). 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 2.13). 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 2.14). 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.

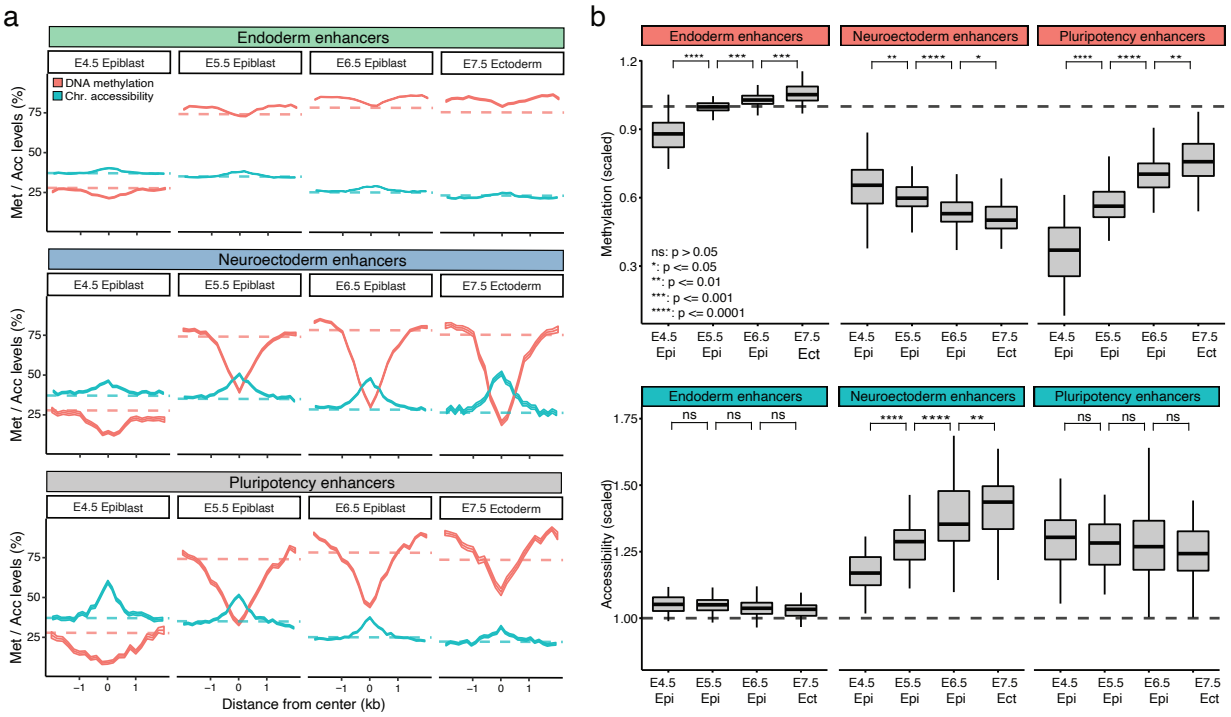


Figure 2.14: Pluripotency and neurectoderm enhancers display different DNA methylation and chromatin accessibility dynamics.
(a) Profiles of DNA methylation (red) and chromatin accessibility (blue) quantified along the epiblast-ectoderm trajectory. Each panel corresponds to a different genomic context. 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 2.6). 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 hypothesis are correct. Ectoderm enhancers at E7.5 contain a mixture of pluripotency and neuroectoderm signatures. However, 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* [175, 95].

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

2.2.8 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*[250]. 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 2.16). 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 2.16).
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.

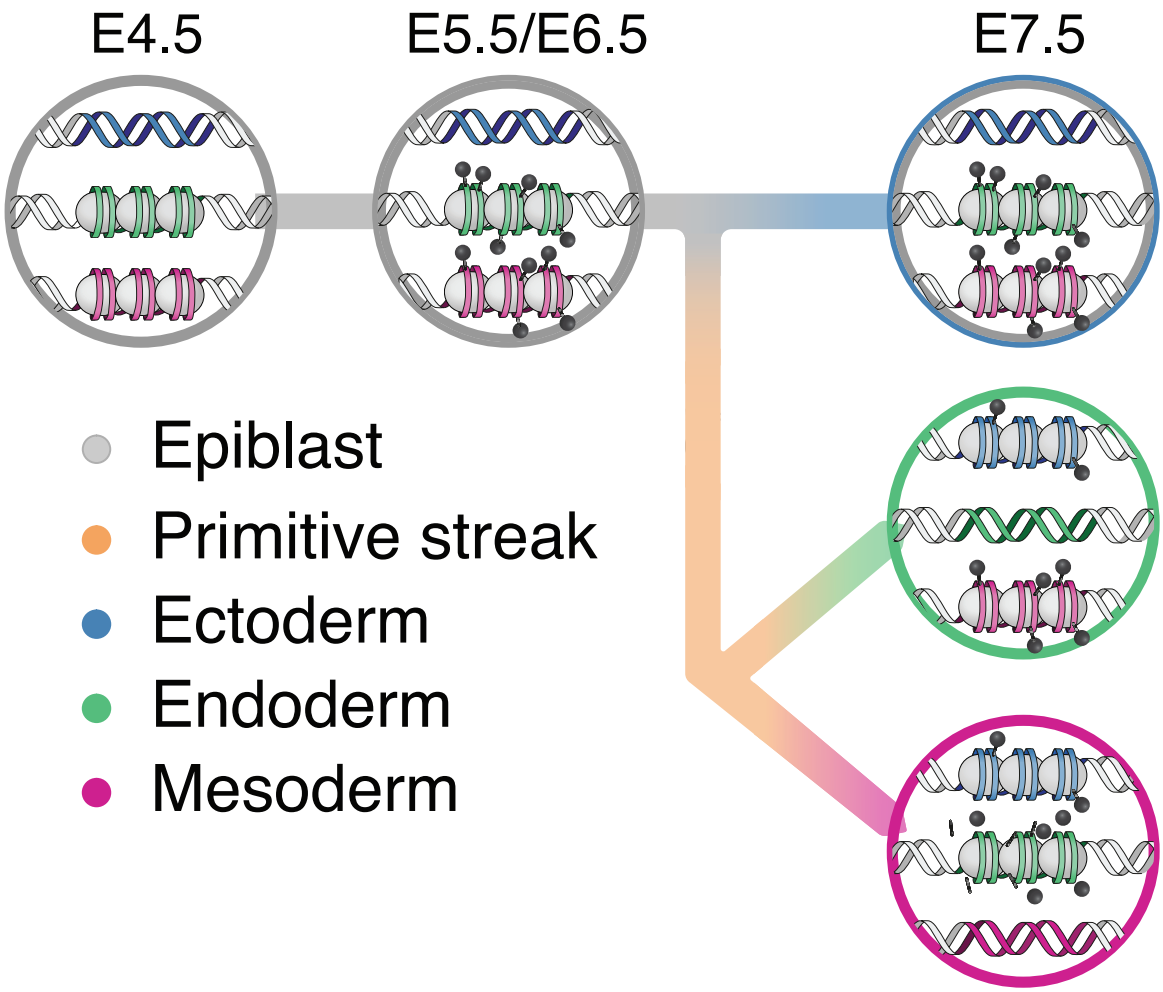


Figure 2.15: Schematic illustration of the hierarchical model for the epigenetic dynamics of germ layer commitment.
Illustration designed by Veronique Juvin from SciArtWork.

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[184]

2.2.9 TET enzymes are required for efficient demethylation of lineage-defining enhancers

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 [206]. This discovery fundamentally changed our understanding of DNA methylation, suggesting that it is not as statistic 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 [**Dai2016**, 220, 206, 144].

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.

2.2.9.1 Cell type assignment using the RNA expression

As in Figure 2.3, cell types were assigned by mapping the RNA expression profiles to the *in vivo* gastrulation atlas using a mutual nearest neighbours matching algorithm [85].
Notably, the WT cells from the EB differentiation protocol recapitulate the *in vivo* dynamids with remarkably accuracy. 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.

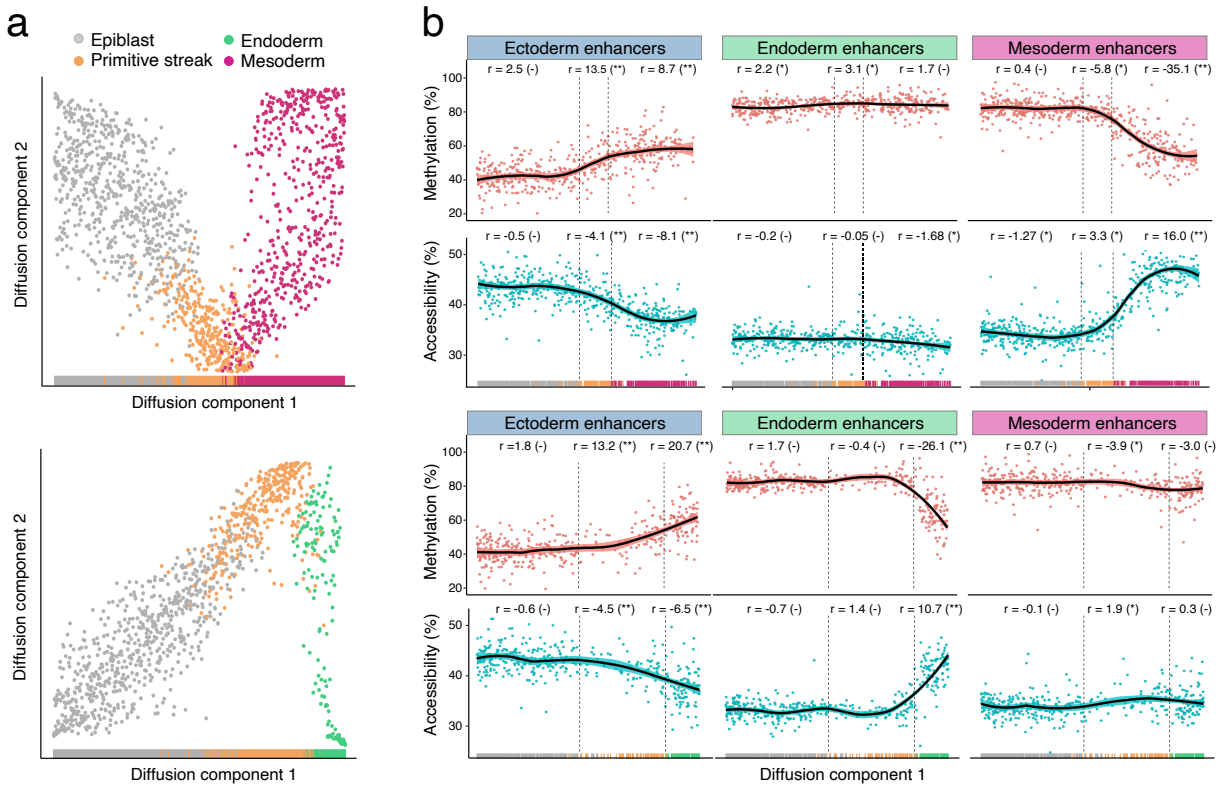


Figure 2.16: Silencing of ectoderm enhancers precedes activation of mesoderm and endoderm enhancers.

(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. For both cases, ranks along the first diffusion component are selected to order cells according to their differentiation state.

(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 estimates. In addition, for each scenario we fit a piece-wise linear regression model for epiblast, primitive streak and mesoderm or endoderm cells (vertical lines indicate the discretised lineage transitions). 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 $p < 0.01$).

(c) Density plots showing differential DNA methylation (%, x-axis) and chromatin accessibility (%, y-axis) at lineage-defining enhancers calculated for each of the lineage transitions.

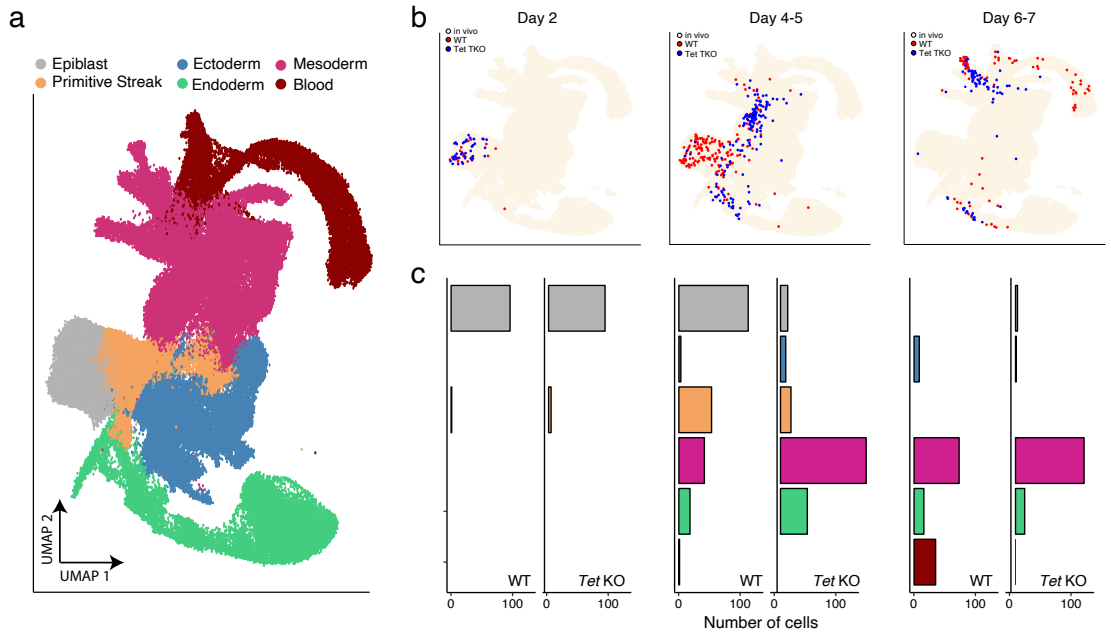


Figure 2.17: Cell type assignment for the Embryoid Body differentiation experiment.
(a) UMAP projection of the 10x atlas data set (stages E6.5 to E8.5, no extraembryonic cells), where cells are coloured by lineage assignment.
(b) Same UMAP projection as in (a), but in this case, for each day of EB differentiation, 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:

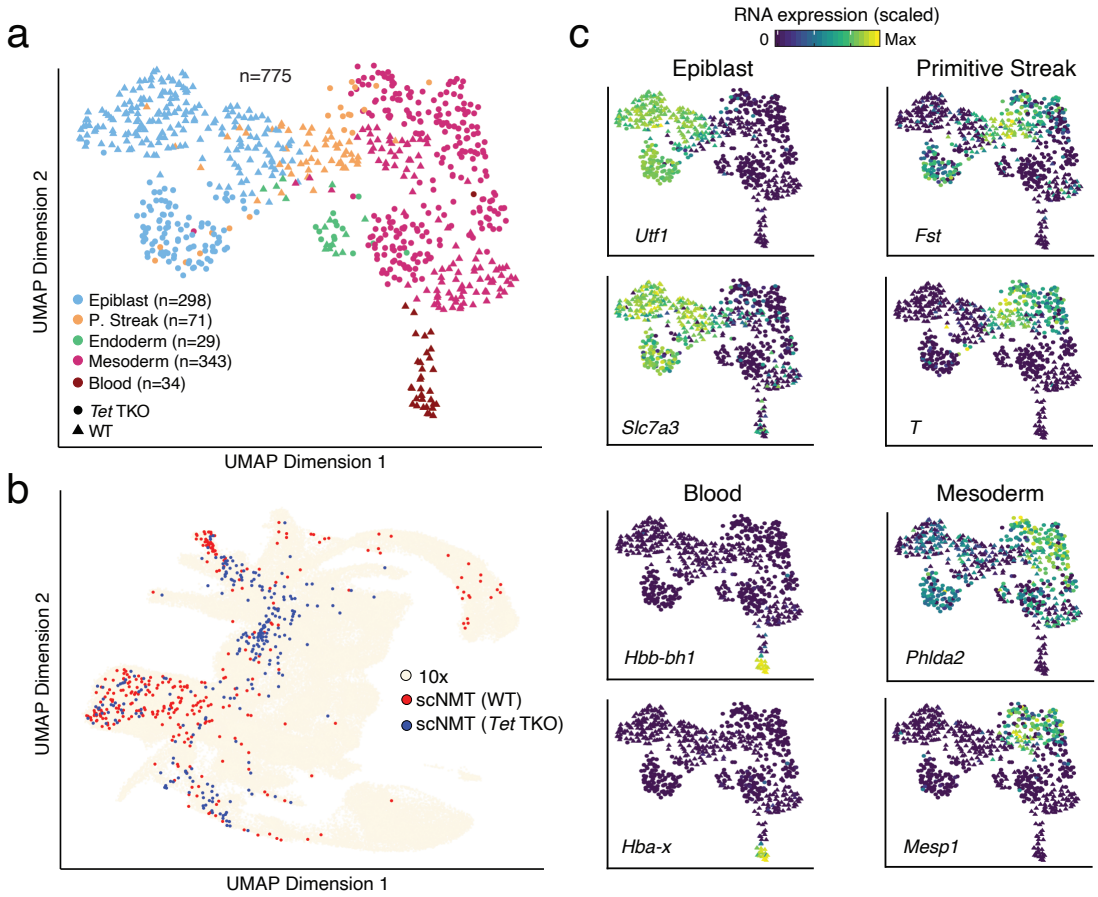


Figure 2.18: Embryoid bodies recapitulate the transcriptional heterogeneity of the mouse embryo.
(a) UMAP projection for the embryoid body dataset, where cells are coloured by lineage assignment and shaped by genotype (WT or *Tet* TKO).
(b) UMAP projection of the atlas data set (stages E6.5 to E8.5, 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.

2.2.9.2 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:

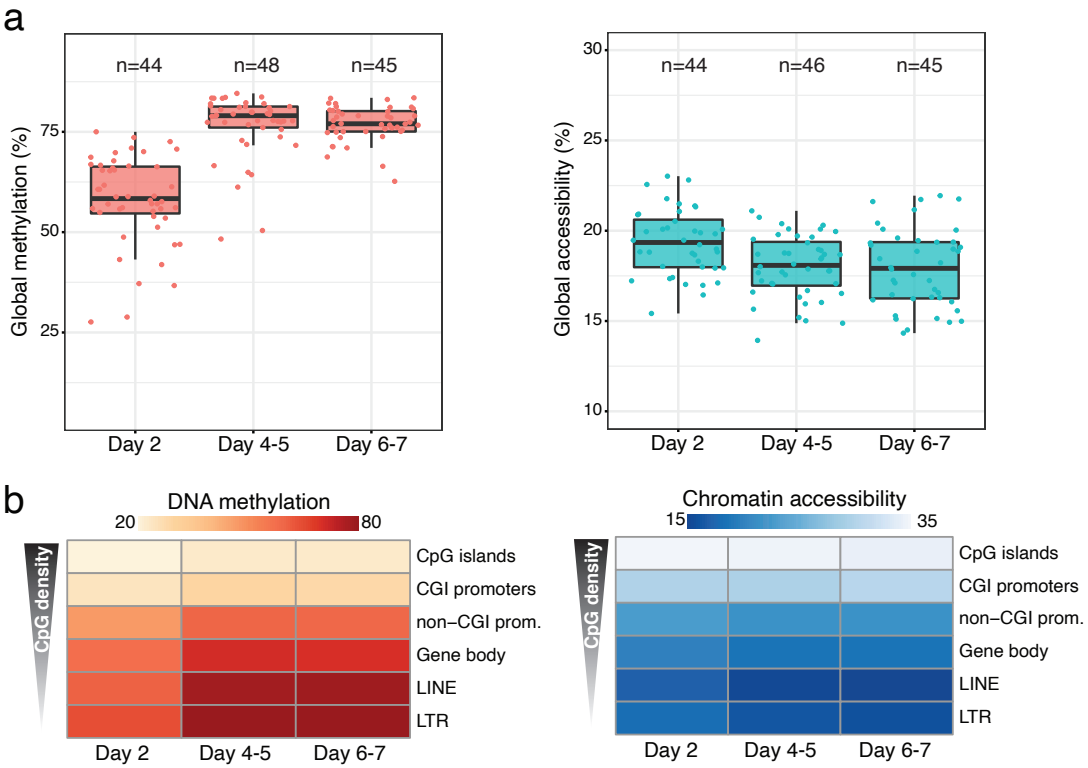


Figure 2.19: Global DNA methylation and chromatin accessibility levels during embryoid body differentiation (WT).
(a) Box plots showing the distribution of genome-wide CpG methylation (left) or GpC accessibility levels (right) per stage and lineage. Each dot represents a single cell.
(b) Heatmap of DNA 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%.

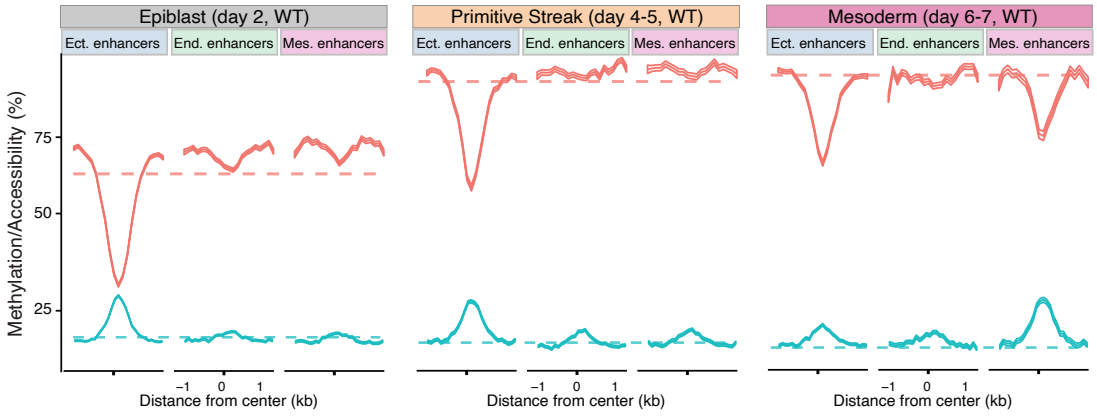


Figure 2.20: Profiles of DNA methylation (red) and chromatin accessibility (blue) at lineage-defining enhancers quantified along EB differentiation (only WT cells). Shown are running averages in consecutive 50bp windows around the center of the ChIP-seq peaks (1kb upstream and downstream). Solid lines 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).

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.

2.2.9.3 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 2.21). 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 2.21. 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.

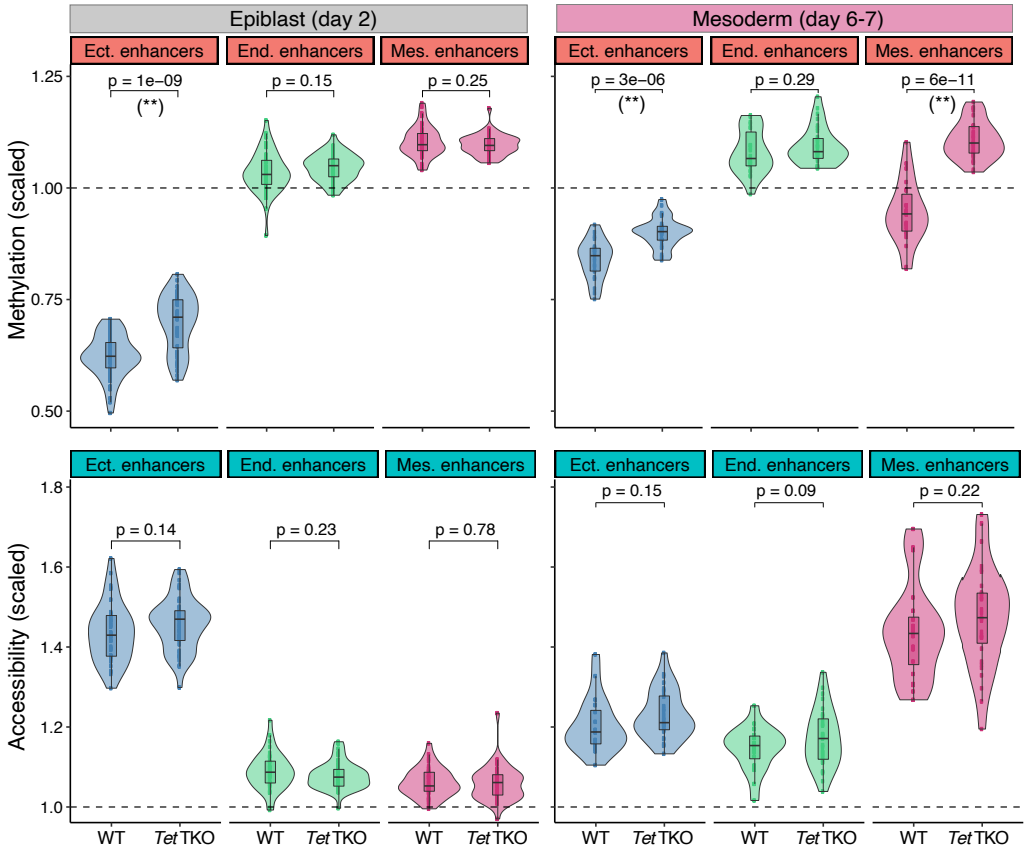


Figure 2.21: Overlayed box plots and violin plots display the distribution of DNA methylation (top) or chromatin accessibility values for lineage-defining enhancers in the epiblast-like cells at day 2 and the mesoderm-like cells at days 6-7.
The y-axis shows the DNA methylation or chromatin accessibility 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 2.17). 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* **[Dai2016]**, 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.

All together, this *in vitro* part of our study confirms that EBs are a suitable model to study the epigenetics of germ layer specification. We hope this provides a valuable resource for other researchers looking to study lineage specification in light of the 3Rs of the ethical use of animals in research.

2.2.10 Conclusions

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 demethylated and accessible upon germ layer commitment. Most notably, when tracing back the epigenetic dynamics for the lineage-defining enhaners 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 driven by signalling cues in the primitive streak transient state.
Interestingly, experimental evidence exist to support this hypothesis. Several groups have shown that, in the absence of external stimuli, ESCs differentiate to neurons [175, 95], 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.

More generally, we speculate that asymmetric epigenetic priming, where early progenitors are epigenetically primed for a default cell type, may be a more general and poorly understood feature of lineage commitment.

2.2.11 Limitations and future perspectives

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

- Scalability: in its current form, scNMT-seq is a labourious 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 1 we are taking steps to make it more high-throughput in order to eventually apply it to post-gastrulation and early 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. One could try repeat the analysis after attempting imputation of the DNA methylation and chromatin accessibility measurements [7].

- 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)? Also, how could we target the default pathway? Is there a way to artifically methylate all ectoderm enhancers (assuming we are able to accurately identify them) 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, dissecting mechanistic roles of important genes using knock out mice is challenging and time-consuming. More importantly, the phenotypic effects of the mutation can be masked by gross devleopmental defects. For this reasons, we are going to explore the usage of chimeric embryos where *TET* TKO tdTomato+ ESCs are injected into wild-type blastocysts. If the procedure is successful, the adult will contain a mixture of WT and *TET* TKO cells that can be separated upon embryo collection using FACS [**PijuanSala2019**].

Chapter 3

MOFA+: an improved framework for the comprehensive integration of structured single-cell data

In Chapter 2 we developed Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of multi-modal 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) data sets, the inference scheme implemented in MOFA is still limited in scalability. In addition, the increased experimental throughput has facilitated the simultaneous study of multiple experimental conditions, even in a combinatorial fashion[210].
MOFA makes strong assumptions about the dependencies across samples and it hence has no principled way of modelling data sets where the samples are structured into multiple groups, where groups can correspond to batches, donors or independent studies. By pooling and contrasting information across experimental conditions, it would be possible to obtain more comprehensive insights into the complexity underlying biological systems.

In this new Chapter we improve the model formulation in MOFA with the aim of performing integrative analysis of large-scale datasets that are *structured* into multiple data modalities and/or multiple groups.

The work discussed in this chapter has been peer-reviewed and published in [**Argelaguet2020**]. The project was conceived by Damien Arnol, Britta Velten and me. The mathematical derivations and the implementation of the stochastic variational inference scheme was done by Damien Arnol, Yonatan Deloro and me. The downstream analysis package was implemented by Danila Bredikhin and me. I generated all figures and I wrote the manuscript with feedback from all authors. John C. Marioni and Oliver Stegle supervised the project.

3.1 Theoretical foundations

3.1.1 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 commonly used probability distributions belong to the exponential family, including the normal or Gaussian, Gamma, Poisson, Bernoulli, Exponential, etc. Exponential family distributions can be represented in the following form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\{\eta(\boldsymbol{\theta})T(\mathbf{x}) - A(\boldsymbol{\theta})\} \tag{3.1}$$

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 frequently used in this thesis are the following:

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:

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

In the context of Bayesian inference, the main property that make exponential family distributions indispensable is that they have conjugate priors (i.e. a combination of likelihood and prior distributions which ensure a closed-form posterior distribution which is of the same form as the prior). As we have discussed in Chapter 2, this property is crucial for enabling efficient statistical inference, otherwise posterior distributions must be computed using expensive and approximate numerical methods.

3.1.2 Gradient ascent

Gradient ascent is a first-order optimization algorithm for finding the maximum of a function [**Bishop2006**, 176]. Formally, for a differentiable function $F(x)$, the iterative scheme of gradient ascent is:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \nabla F(\mathbf{x}^{(t)}) \tag{3.2}$$

In short, it 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 $\sum_t (\rho^{(t)})^2 < \infty$. Then F is guaranteed to converge to the global maximum [214] if the objective function is convex. If F is not convex, the algorithm is sensible to the initialisation $\mathbf{x}^{t=0}$ and can converge to local maxima instead of the global maximum.

3.1.2.1 Stochastic gradient ascent

Gradient ascent becomes prohibitively slow with large datasets, mainly because of the computational cost involved in the iterative calculation of gradients [230].
A simple strategy to speed up gradient ascent is to replace the actual gradient ∇F by an estimate $\hat{\nabla} F$ using a randomly selected subset of the data (minibatch). The iterative scheme is then defined in the same way as in standard gradient ascent:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \hat{\nabla} F(\mathbf{x}^{(t)}) \tag{3.3}$$

3.1.2.2 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 parameters 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 an Euclidean norm, and the direction of steepest ascent is hence dependent on the Euclidean geometry of the θ space. Why is this problematic when working with probability distributions? Because it does not consider the uncertainity that underlies probability distributions. 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

$$\begin{aligned} \psi_1 &\sim \mathcal{N}(0\,|\,5) \qquad & \psi_3 &\sim \mathcal{N}(0\,|\,1) \\ \psi_2 &\sim \mathcal{N}(10\,|\,5) \qquad & \psi_4 &\sim \mathcal{N}(10\,|\,1) \end{aligned} \tag{3.4}$$

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_1 and ψ_2 than between ψ_3 and ψ_4:

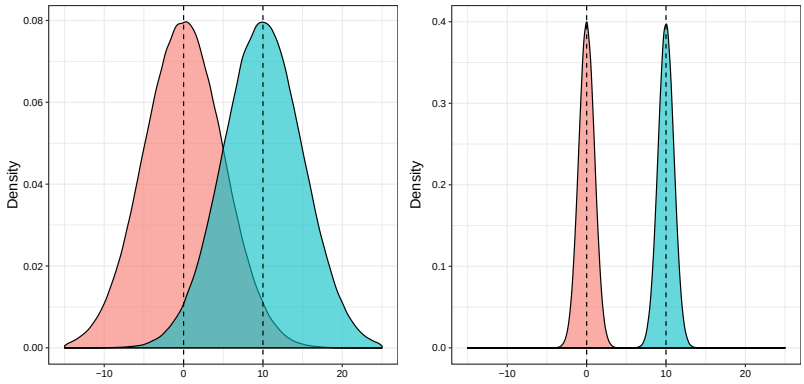


Figure 3.1: Illustration of the problem of using Euclidean distances to measure distances between parameters of distributions.
In both plots, the red and blue distributions are separated by the same Euclidean distance of 100. Yet, the distance in probability space between the two distributions is higher 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 [5, 159].
To find the optimal $\hat{d}_{KL}$, one needs to solve the following optimisation problem:

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

where c is an arbitrary constant. Previous works have shown that this can be solved by introducing Lagrange multipliers and Taylor expansions (see [5, 130]). The solution corresponds to the standard (Euclidean) gradient pre-multiplied by the inverse of the Fisher Information Matrix of $q(x|\theta)$:

$$\hat{d}_{KL} \propto \mathbf{F}^{-1}(\theta)\nabla_\theta\mathcal{L}(\theta) \tag{3.5}$$

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 [130, 5, 98].

3.1.3 Derivation of a stochastic variational inference algorithm

In this section I will show how to derive a stochastic variational inference algorithm for general Bayesian models. This work is inspired from [98] which we adapted and implemented in the MOFA+ model. A comprehensive mathematical derivation of the algorithm is not sought in this chapter. Instead, I will describe a modified and simplified derivation to gist the essential. For a complete mathematical derivation we refer the reader to [98].
Also, this section builds upon three theoretical foundations that have been introduced before: Variational inference (Section .1.0.1), exponential family distributions (Section 3.1.1) and (natural) gradient ascent (Section 3.1.2).

3.1.3.1 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 ($\boldsymbol{\alpha}_z$).

This leads to the following factorisation of the joint distribution:

$$p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}_\beta, \boldsymbol{\alpha}_z) = p(\mathbf{Z}|\boldsymbol{\alpha}_z)p(\boldsymbol{\beta}|\boldsymbol{\alpha}_\beta) \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\beta}) \tag{3.6}$$

and the corresponding graphical model representation:

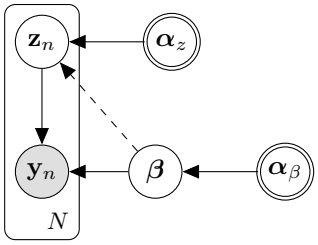


Figure 3.2: Graphical model for a general probabilistic model where unobserved variables are classified as global and local.
The dashed line indicates that the connection between global and local variables is optional and it is not used in the MOFA model.

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 β 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 (3.1))

$$\begin{aligned} p(\beta|\alpha_\beta) &= h(\beta)\exp\{\eta_g(\alpha_\beta)t(\beta) - a_g(\alpha_\beta)\} \\ p(z_{nk}|\alpha_z) &= h(z_{nk})\exp\{\eta_l(\alpha_z)t(z_{nk}) - a_l(\alpha_z)\} \end{aligned} \tag{3.7}$$

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

$$\begin{aligned} p(\beta|\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha}) &= h(\beta)\exp\{\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})^T t(\beta) - a_g(\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha}))\} \\ p(\mathbf{z}_n|\mathbf{y}_{nj},\mathbf{z}_{nj},\beta) &= h(\mathbf{z}_n)\exp\{\eta_l(\mathbf{y}_{nj},\mathbf{z}_{nj},\beta)^T t(\mathbf{z}_n) - a_l(\eta_l(\mathbf{y}_{nj},\mathbf{z}_{nj},\beta))\} \end{aligned} \tag{3.8}$$

3.1.3.2 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:

$$q(\beta|\lambda) = h(\beta)\exp\{\eta_g(\lambda)t(\beta) - a_g(\lambda)\} \tag{3.9}$$

$$q(z_{nk}|\phi_{nk}) = h(z_{nk})\exp\{\eta_l(\phi_n)t(z_{nk}) - a_l(z_{nk})\} \tag{3.10}$$

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, see **??**) factorises as:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_{q(\mathbf{Z},\beta)}[\log p(\mathbf{Y},\mathbf{Z},\beta)] - \mathbb{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] - \mathbb{E}_{q(\beta)}[\log q(\beta)] \\ &= \sum_{n=1}^{N}\mathbb{E}_{q(\mathbf{z}_n,\beta)}[\log p(\mathbf{y}_n,\mathbf{z}_n,\beta)] - \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{q(z_{nk})}[\log q(z_{nk})] - \mathbb{E}_{q(\beta)}[\log q(\beta)] \end{aligned} \tag{3.11}$$

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. Assumign 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.

3.1.3.3 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 β. Then, from the assumption that the complete conditionals and the variational distributions belong to the exponential family (Equations (3.8) to (3.9)):

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

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:

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_{q(z)}[\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})] - \lambda) \tag{3.12}$$

and setting it to zero leads to the solution:

$$\lambda = \mathbb{E}_{q(z)}[\eta_g(\mathbf{Y},\mathbf{Z},\boldsymbol{\alpha})] \tag{3.13}$$

3.1.3.4 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 (3.8) to (3.9)):

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

Taking the gradient with respect to ϕ_{nk}:

$$\nabla_\phi \mathcal{L}(\phi_{nk}) = \nabla^2_\phi a_l(\phi_{nk})(\mathbb{E}_{q(\beta,\mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n,\mathbf{z}_{nj},\beta)] - \phi_{nk}) \tag{3.14}$$

and setting it to zero leads to the following solution:

$$\phi_{nk} = \mathbb{E}_{q(\beta,\mathbf{z}_{nj})}[\eta_l(\mathbf{y}_n,\mathbf{z}_{nj},\beta)] \tag{3.15}$$

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

However, as discussed in Section 3.1.2.2, 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.

3.1.3.6 Deriving the natural gradients for the global variational parameters

From Equation (3.12), 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:

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

3.1.3.7 Deriving the natural gradients for the local variational parameters

From Equation (3.14), 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.

3.1.3.8 Stochastic variational inference algorithm using natural gradients

After replacing the euclidean gradient with the natural gradients, the model can be optimised using the following algorithm:

Algorithm 2 Stochastic variational inference algorithm using natural gradients

1: Initialise the global parameters $\boldsymbol{\lambda}^{(t=0)}$.
2: Initialise step size $\rho^{(t=0)}$
3: **repeat**
4: sample $\mathcal{B}$ a mini-batch of samples of size S
5: **for** each local variational parameter ϕ_{nk} such that n is in batch $\mathcal{B}$ **do**
6:

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

7: **end for**
8: **for** each global variational parameter λ **do**
9:

$$\begin{aligned} \lambda^{(t+1)} &= \lambda^{(t)} + \rho^{(t)} \hat{\nabla}_\lambda \mathcal{L}^S(\lambda) \\ &= (1-\rho^{(t)})\lambda^{(t)} + \rho^{(t)} \mathbb{E}_{q^{(t+1)}(z)} \left[\frac{N}{S} \eta_g(\mathbf{Y}_{[n\in\mathcal{B}],:}, \mathbf{Z}_{[n\in\mathcal{B}],:}, \boldsymbol{\alpha}) \right] \end{aligned} \tag{3.16}$$

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

Notice that the stochastic nature of the 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 high 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[205], 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.

3.2 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 data sets.

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:

$$\mathbf{Y}_g^m = \mathbf{Z}_g \mathbf{W}^{mT} + \boldsymbol{\epsilon}_g^m \tag{3.17}$$

where $\mathbf{Z}_g \in \mathbb{R}^{N_g \times K}$ are a set of G matrices that contains 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 MOFA v1.

It is important to get the intuition for the multi-group formulation right. The aim of the multi-group framework is not to capture differential changes in **mean** levels between the groups (as for example when doing differential RNA expression) but rather to exploit the covariation of features. The aim is to find out which sources of variability (i.e. which latent Factors) are present in the different groups and which ones are exclusive to 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 to be compared across views, only the covariation patterns are of interest. To achieve this, the features are centred per view and also per group (i.e. all intercept effects are regressed out) before fitting the model. The following figure summarises the MOFA+ pipeline:

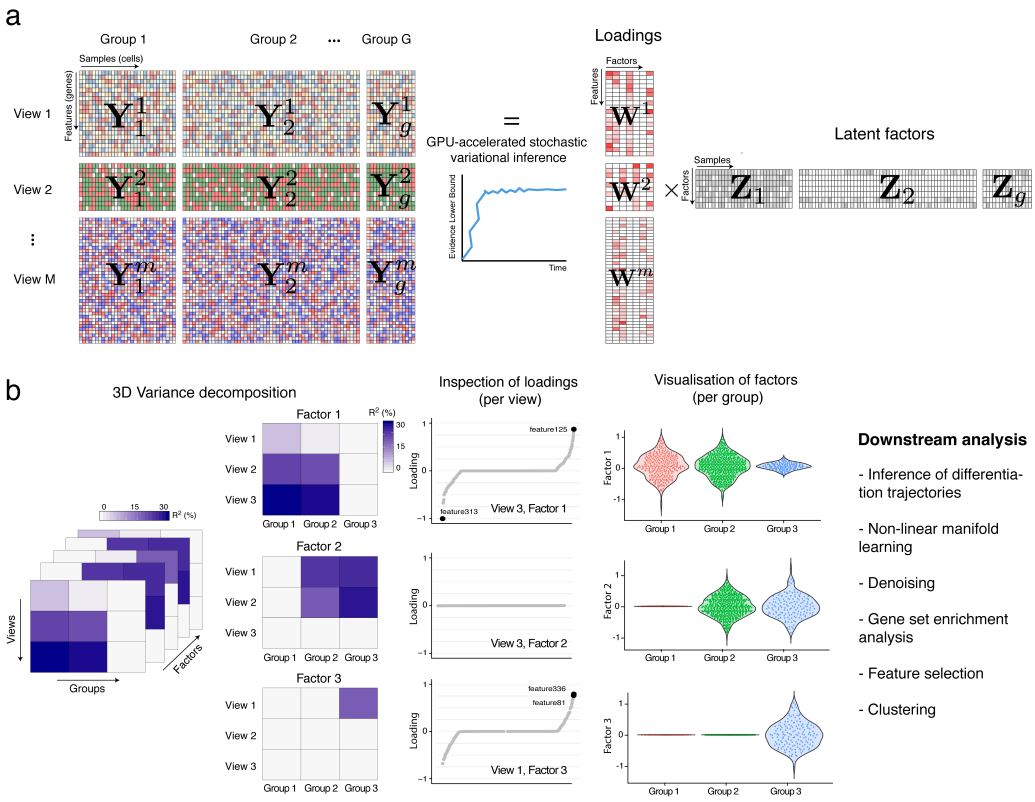


Figure 3.3: Multi-Omics Factor Analysis v2 (MOFA+) provides an unsupervised framework for the integration of multi-group and multi-view single-cell data.
(a) Model overview: the input consists of multiple data sets structured into M views and G groups. Views consist of non-overlapping sets of features that can represent different assays. Analogously, groups consist of non-overlapping sets of samples that can represent different conditions or experiments. Missing values are allowed in the input data. MOFA+ exploits the dependencies between the features to learn a low-dimensional representation of the data (Z) defined by K latent factors that capture the global sources of molecular variability. For each factor, the weights (W) link the high-dimensional space with the low-dimensional manifold and provide a measure of feature importance. The sparsity-inducing priors on both the factors and the weights enable the model to disentangle variation that is unique to or shared across the different groups and views. Model inference is performed using GPU-accelerated stochastic variational inference.
(b) The trained MOFA+ model can be queried for a range of downstream analyses: 3D variance decomposition, quantifying the amount of variance explained by each factor in each group and view, inspection of feature weights, visualisation of factors and other applications such as clustering, inference of non-linear differentiation trajectories, denoising and feature selection.

3.2.1 Model priors and likelihood

3.2.1.1 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 feature-wise spike-and-slab prior (reparametrised[249]):

$$p(\hat{w}_{dk}^m, s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \text{Ber}(s_{dk}^m \,|\, \theta_k^m) \tag{3.18}$$

with the corresponding conjugate priors for θ and α:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{3.19}$$

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{3.20}$$

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 aim of the spike-and-slab prior is to push individual weights to zero to yield a more interpretable solution.
For more details, we refer the reader to Chapter 2.

3.2.1.2 Prior on the factors

In MOFA v1 we adopted an isotropic Gaussian prior:

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{3.21}$$

which assumes *a priori* an unstructured latent space. This is the assumption that we want to break. Following the same logic as in the factor and view-wise ARD prior, the integration of multiple groups of samples requires introducing a *structured* prior that captures the existence of different groups, such that some factors are allowed to be active in different subsets of groups.

To formalise the intuition above we simply need to copy the double sparsity prior from the weights to the factors:

$$p(\hat{z}_{nk}^g, s_{nk}^g) = \mathcal{N}(\hat{z}_{nk}^g \,|\, 0, 1/\alpha_k^g)\,\text{Ber}(s_{nk}^g \,|\, \theta_k^g) \tag{3.22}$$

$$p(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, a_0^\theta, b_0^\theta\right) \tag{3.23}$$

$$p(\alpha_k^g) = \mathcal{G}\left(\alpha_k^g \,|\, a_0^\alpha, b_0^\alpha\right), \tag{3.24}$$

where g is the index of the sample groups.
Notice that the spike-and-slab prior is introduced for completeness but is not necessarily required, and can be disabled by fixing $\mathbb{E}[\theta_k^g] = 1$.

3.2.1.3 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 individual feature and per group:

$$p(\epsilon_g^m) = \mathcal{N}\left(\epsilon_g^m \,|\, 0, /\tau_g^m \mathbf{I}_{Dm}\right) \tag{3.25}$$

$$p(\tau_g^m) = \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_g^m \,|\, a_0^\tau, b_0^\tau\right) \tag{3.26}$$

This formulation is important to capture the (realistic) events where a specific feature may be highly variable in one group but non-variable in another group.
In addition, as in MOFA v1, non-Gaussian noise models can also be defined, but unless otherwise stated, we will always assume Gaussian residuals.

3.2.1.4 Likelihood

Altogether, this results in the following likelihood:

$$p(\mathbf{Y}|\mathbf{W},\mathbf{Z},\mathbf{T}) = \prod_{m=1}^{M}\prod_{g=1}^{G}\mathcal{N}\left(\mathbf{Y}_g^m \,|\, \mathbf{Z}_g\mathbf{W}^{mT}, 1/\boldsymbol{\tau}_g^m\right) \tag{3.27}$$

3.2.1.5 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:

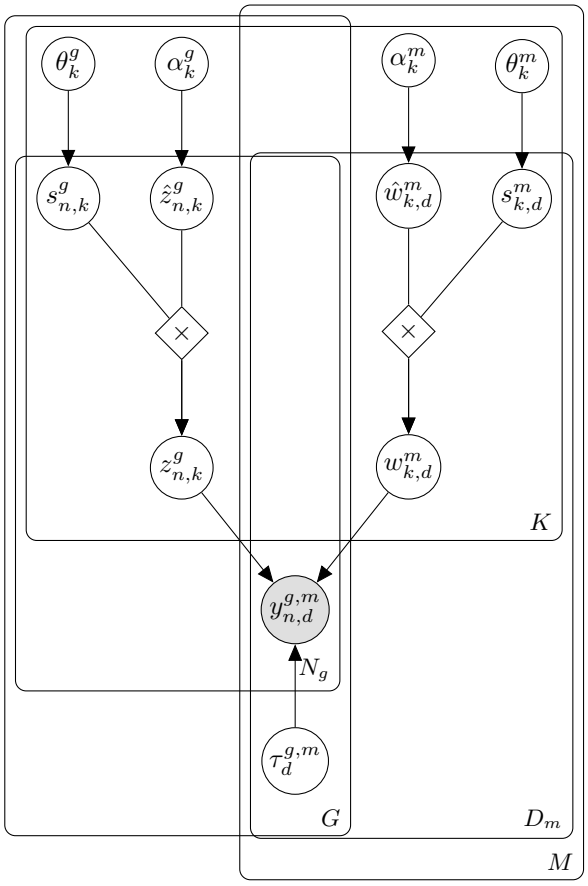


Figure 3.4: Graphical model for MOFA+.
The white circles represent hidden variables that are inferred by the model, whereas the grey circles represent the observed variables. There are a total of five plates, each one representing a dimension of the model: M for the number of views, G for the number 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.

3.2.2 Solving the rotational invariance problem

Conventional Factor Analysis is invariant to rotation in the latent space[276]. To demonstrate this property, let us apply an arbitrary rotation to the weights and the factors, specified by the rotation matrix $\mathbf{R} \in \mathbb{R}^{K \times K}$:

$$\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{R}^{-1}$$
$$\tilde{\mathbf{W}} = \mathbf{R}\mathbf{W}$$

First, note that the model likelihood is unchanged by this rotation, irrespective of the prior distribution used.

$$p(\mathbf{Y}|\tilde{\mathbf{Z}}\tilde{\mathbf{W}}, \tau) = p(\mathbf{Y}|\mathbf{Z}\mathbf{R}^{-1}\mathbf{R}\mathbf{W}, \tau) = p(\mathbf{Y}|\mathbf{Z}\mathbf{W}, \tau)$$

However, the prior distributions of the factors and the weights are only invariant to rotations when using isotropic Normal priors:

$$\ln p(\mathbf{W}) \propto \sum_{k=1}^{K}\sum_{d=1}^{D} w_{d,k}^2 = \mathrm{Tr}(\mathbf{W}^T\mathbf{W}) = \mathrm{Tr}(\mathbf{W}^T\mathbf{R}^{-1}\mathbf{R}\mathbf{W}) = \mathrm{Tr}(\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})$$

where we have used the property $\mathbf{R}^T = \mathbf{R}^{-1}$ that applies to rotation matrices. The same derivation follows for the factors $\mathbf{Z}$.
In practice, this property renders conventional Factor Analysis unidentifiable, hence limiting its interpretation and applicability. Sparsity assumptions, however, partially address the rotational invariance problem [**Hore2015**].

It is important to remark that the factors are nonetheless invariant to permutations. This implies that under different initial conditions, the order of the factors is not necessarily the same in independent model fittings. To address this we manually sort factors *a posteriori* based on total variance explained.

3.2.3 Stochastic variational inference algorithm

In Section 3.1.3 I have explained how to derive a stochastic variational inference (SVI) algorithm for a general Bayesian model using an adapted version of the formulation introduced in [98].
To apply the SVI algorithm to MOFA the first step is to choose the *local* and *global* dimensions. Just as a reminder, the local dimension will be factorised in the ELBO and thus the one where the stochastic gradients apply.
In single-cell studies we expect increasingly large data sets (more cells) but the number of features to remain roughly constant and the natural dimension to define as *local* is the axis of the samples and the global dimension to be axis of features.

In the case of the MOFA+ model the variables that are classified as *local* are the Factors $\mathbf{Z_g} = \{z_{nk}^g\}$, which due to the reparametrisation of the spike-and-slab prior consists on the element-wise product of two matrices: $\hat{\mathbf{Z}}_\mathbf{g}$ and $\mathbf{S_g}$. All other hidden variables are global: $\mathbf{tau^{gm}}$, $\mathbf{W^m}$ and $\mathbf{S^m}$ (whose

term-wise product gives $\mathbf{W}^m$), $\alpha^{\mathbf{m}}, \theta^{\mathbf{m}}$, as well as $\alpha^{\mathbf{g}}$ and $\theta^{\mathbf{g}}$ when adding the spike-and-slab prior over the Factors.

3.3 Model validation

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

3.3.1 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 matched those obtained from conventional inference across a range of batch sizes, learning rates and forgetting rates.
In terms of speed, GPU-accelerated SVI inference was up to ≈ 20x faster than VI, with speed differences becoming more pronounced with increasing number of cells. For completeness, we also compared the 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.

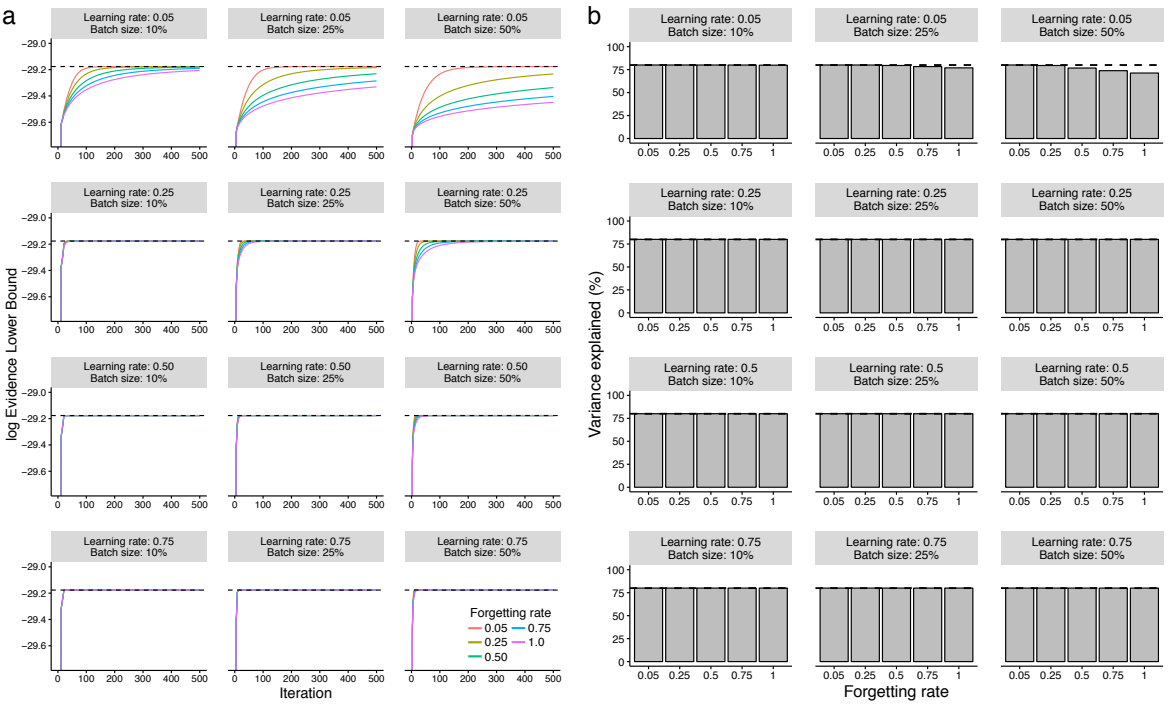


Figure 3.5: Validation of stochastic variational inference using simulated data.
(a) Line plots display the iteration number of the inference (x-axis) and the log- Evidence Lower Bound (ELBO) on the y-axis. Panels correspond to different values of batch sizes (10%, 25%, 50% of the data) and initial learning rates (0.05, 0.25, 0.5, 0.75). Colors correspond to different forgetting rates (0.05, 0.25, 0.5, 0.75, 1.0). The dashed horizontal line indicates the ELBO achieved using standard 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.

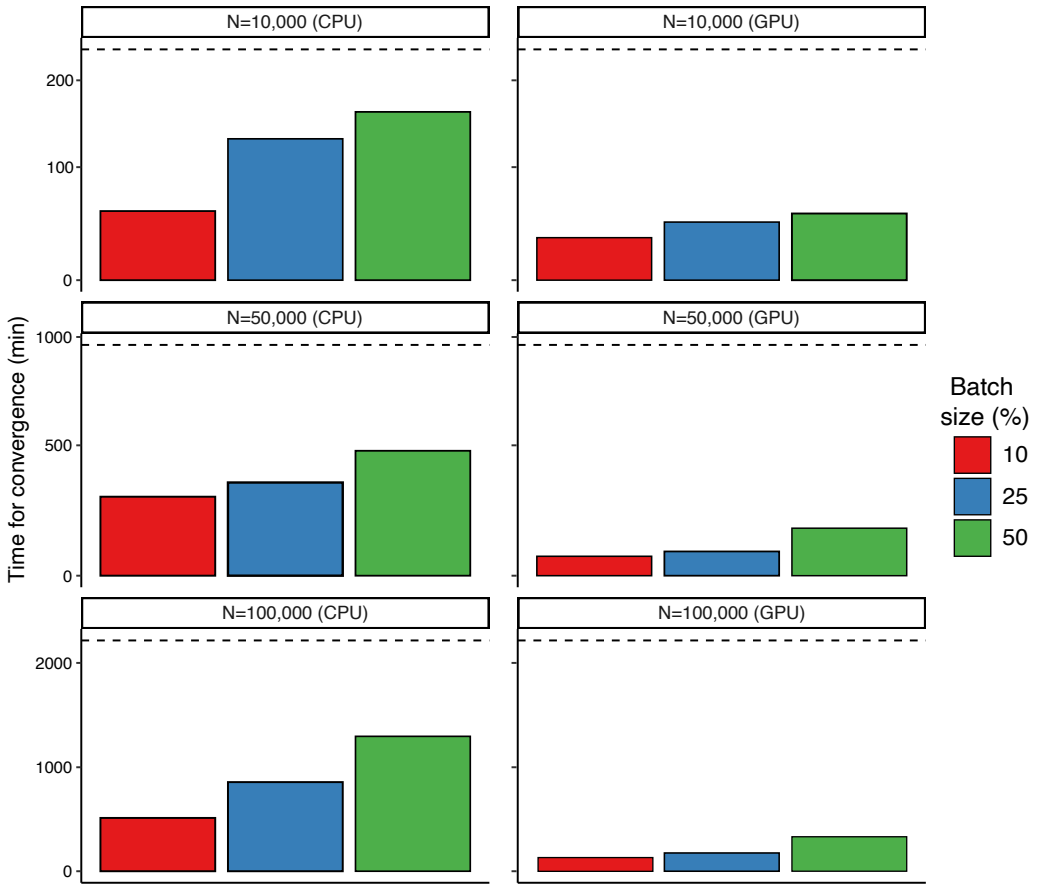


Figure 3.6: Evaluation of convergence speed for stochastic variational inference using simulated data.
Bar plots show the time elapsed for training MOFA+ models with stochastic variational inference (SVI). Colors represent different batch sizes (10%, 25% or 50%). The dashed line indicates the training time for standard VI.
VI models were trained using a single E5-2680v3 CPU. SVI models were trained either using a single E5-2680v3 CPU (first column) or using an Nvidia GTX 1080Ti GPU (second column).

3.3.2 Multi-group structure

Finally, we evaluated whether the double view and group-wise sparsity prior enables the detection of factors with simultaneous differential activity between groups and views.
We simulated data with the following parameters: $M = 2$ modalities, $G = 2$ 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 modalities and groups (Figure 3.7, see ground truth). The task is to recover the true factor activity structure given a random initialisation.
We fit 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 differing amounts of variance across groups and across views, MOFA+ was able to more accurately reconstruct the true factor activity patterns:

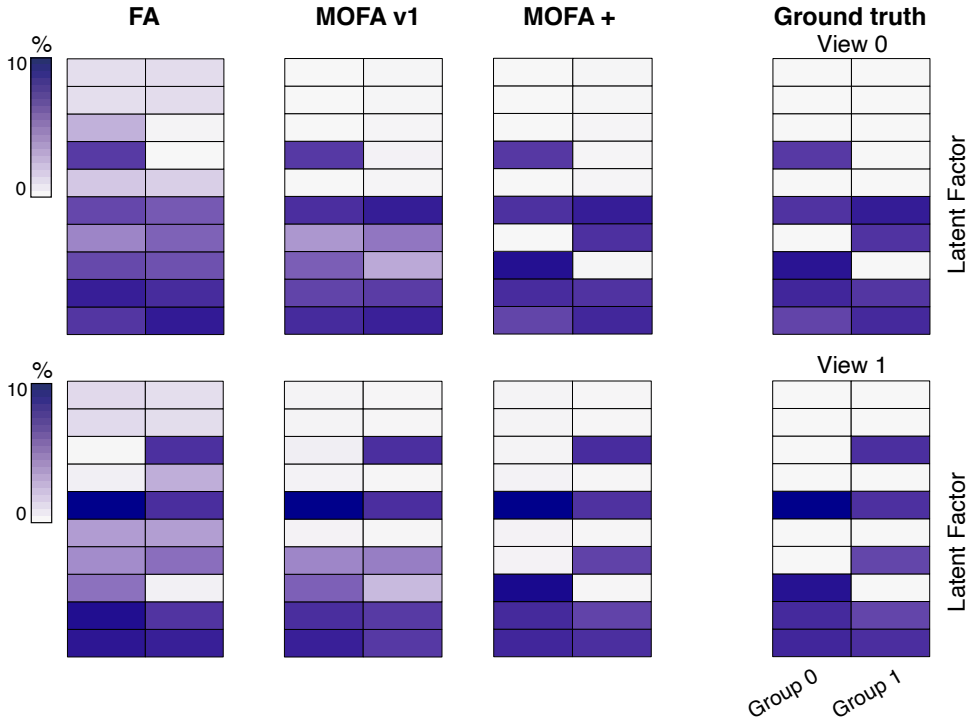


Figure 3.7: Validation of group-wise ARD prior in the factors using simulated data. Representative example of the resulting variance explained patterns. The first row of heatmaps correspond to modality 0 and the second row to modality 1. In each heatmap, the first column corresponds to group 0 and the 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 modality 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.

3.4 Applications

3.4.1 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[192]. This data set, which has been introduced in Chapter 3, consists on a single view (RNA expression) but with a clear group structure where cells belongs to different biological replicates at different 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.

MOFA+ identified 7 Factors that explain at least 1% of variance (across all groups). Notably, this latent representation captures between 35% and 55% of the total transcriptional heterogeneity per embryo:

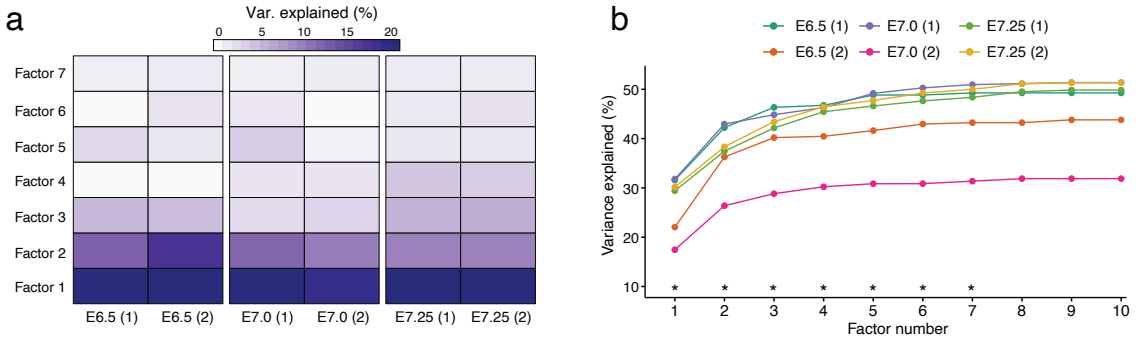


Figure 3.8: MOFA+ variance explained estimates in the gastrulation scRNA-seq atlas. (a) Heatmap displays the variance explained (%) for each factor (rows) in each group (pool of mouse embryos at a specific developmental stage, columns). The bar plots show the variance explained per group with all factors.
(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).

3.4.1.1 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 3.9; *Rhox5* and *Bex3* for ExE ectoderm (not shown); *Mesp1* and *Phlda2* for nascent mesoderm Figure 3.10. Other factors captured technical variation due to metabolic stress that affects all batches in a similar fashion (Factor 3, Figure 3.11).
The characterisation of other factors is described in[**Argelaguet2020**] and is not reproduced here for simplicity.

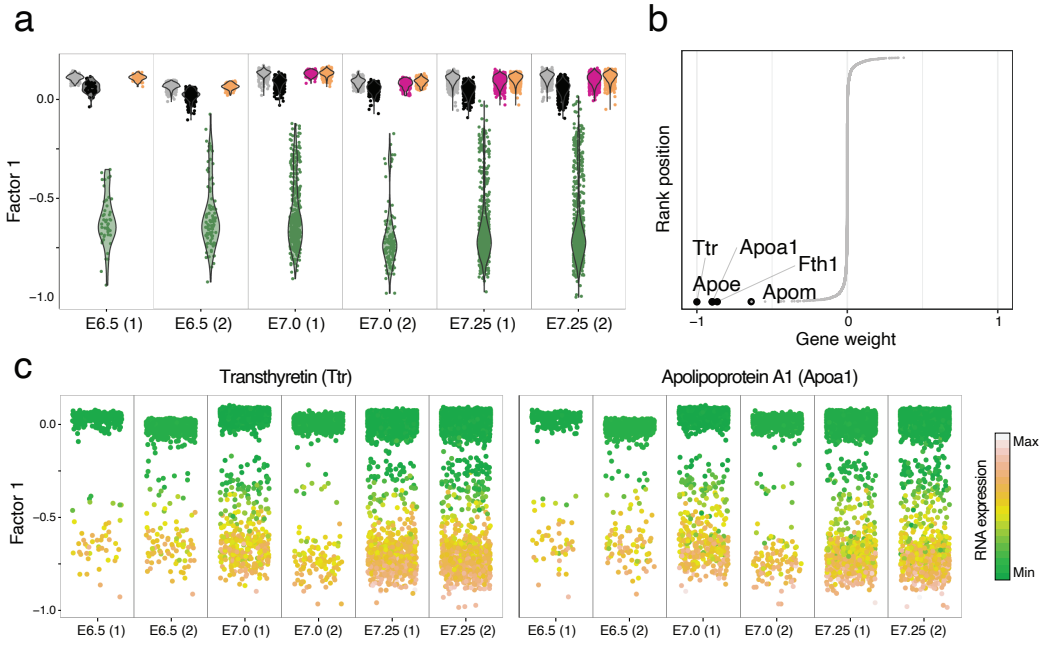


Figure 3.9: Characterisation of Factor 1 as extra-embryonic (ExE) endoderm formation.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

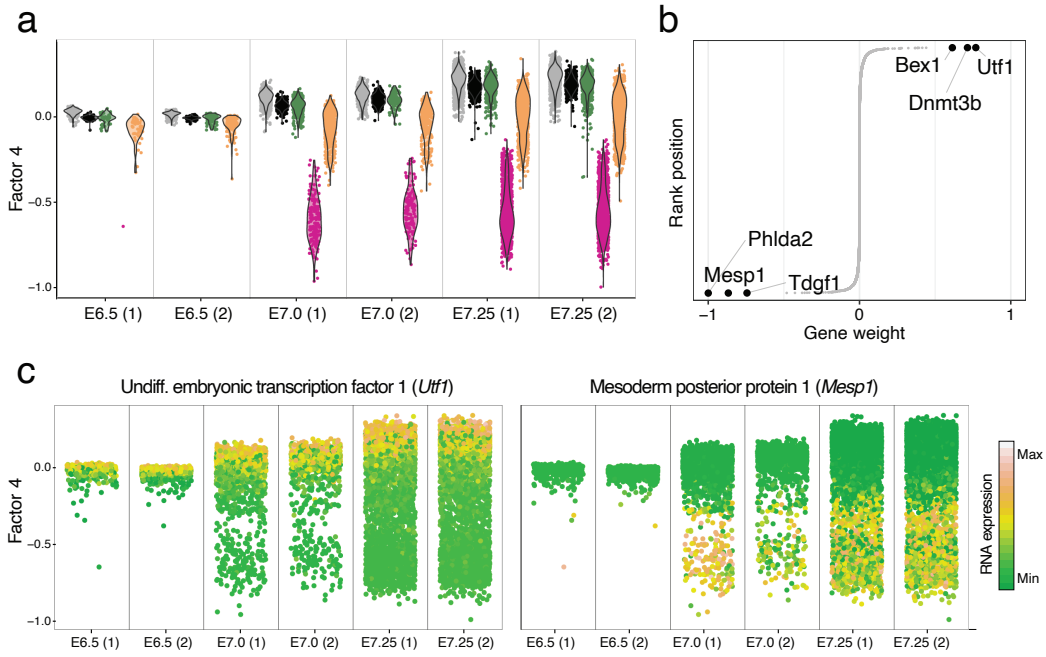


Figure 3.10:
Characterisation of Factor 4 as mesoderm commitment.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

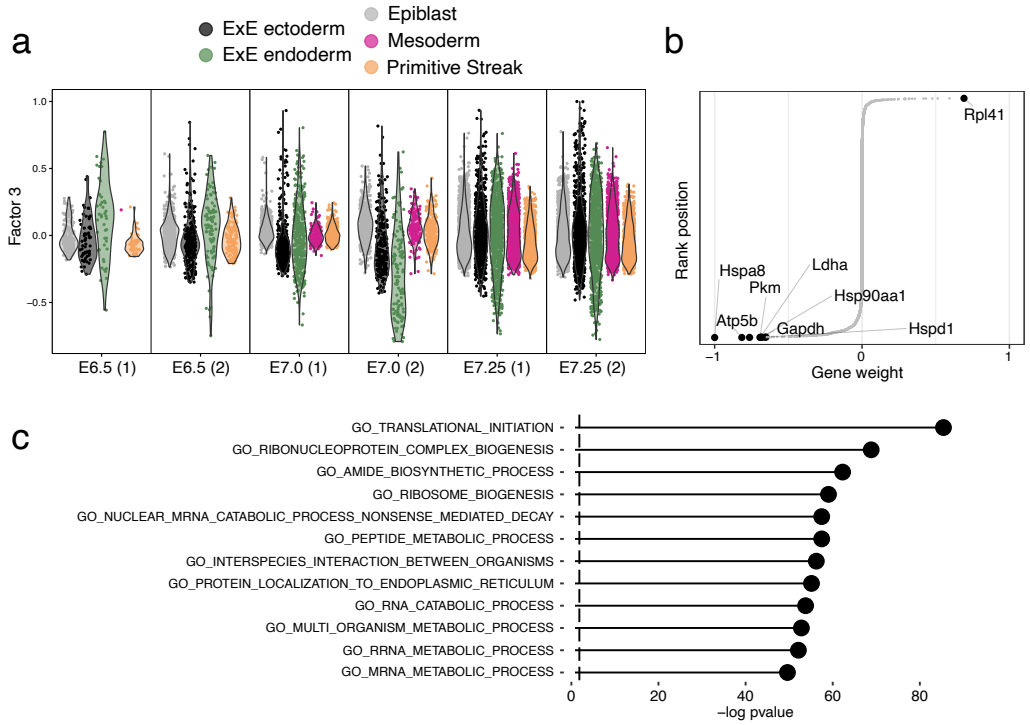


Figure 3.11:
Characterisation of Factor 3 as cell-to-cell differences in metabolic activity.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top seven genes with largest weight (in absolute values)
(c) Gene set enrichment analysis applied to the gene weights using the Reactome gene sets[**Fabregat2015**]. Significance is assessed via a parametric. Resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

Interestingly, Factors display differeent signatures of activity (variance explained) across developmental stages. For example, the variance explained by Factor 1 remains constant across developmental progression (Figure 3.8), indicating that commitment to 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 *structured* scRNA-seq datasets.

3.4.2 Identification of context-dependent methylation signatures associated with cellular diversity in the mammalian cortex

As a second use case, we considered how MOFA+ can be used to investigate cellular heterogeneity in epigenetic signatures between populations of neurons. This application illustrates how a multi-group and multi-view structure can be defined from seemingly uni-modal data.

We considered a data set of of 3,069 cells isolated from the frontal cortex of young adult mouse, where DNA methylation was profiled using single-cell bisulfite sequencing[155].
Some background to motivate our experimental design: in mammalian genomes, DNA methylation predominantly occurs at CpG dinucleotides (mCG), with more than 75% of CpG sites being methylated in differentiated cell types. By contrast non-CpG methylation (mCH) has been historically dismissed as methodological artifact of incomplete bisulfite conversion, until recent works have confirmed their existence in restricted cell types. Yet, evidence for a potential functional role remains controversial[93].

Here we used MOFA+ to dissect the cellular heterogeneity associated with mCH and mCG in the mouse frontal cortex. As input data we quantified mCH and mCG levels at gene bodies, promoters and putative enhancer elements. Each combination of genomic and sequence context was defined as a separate view.
As described in Chapters 1 and 3, methylation levels were calculated per cell and genomic feature using a binomial model where the number of successes correspond to the number of reads that support methylation (or accessibility) and the number of trials the total number of reads.
Finally, to explore the influence of the neuron's location we grouped cells according to their cortical layer: Deep, Middle or Superficial (**??**). Notably, the resulting data set is extremely sparse, which hampers the use of conventional dimensionality reduction techniques.The probabilistic framework underlying MOFA+ naturally enables the handling of missing values by ignoring the corresponding terms in the likelihood function.

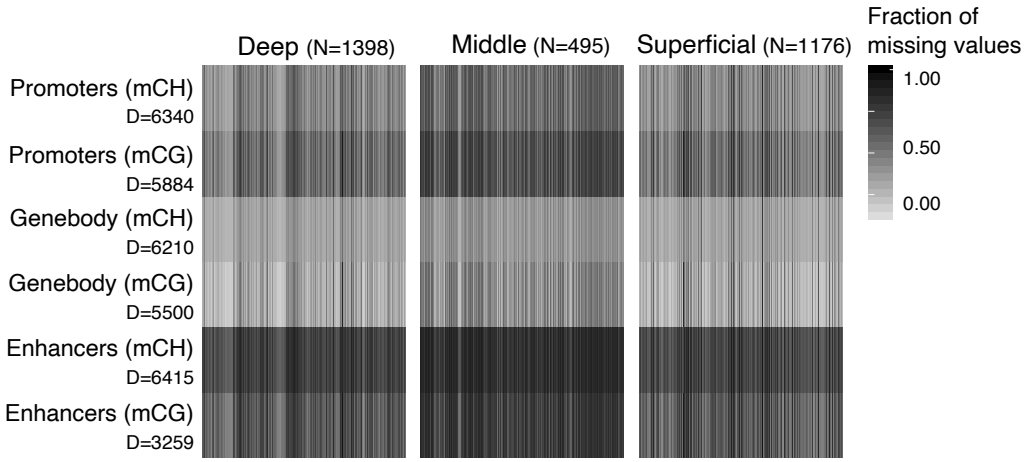


Figure 3.12

MOFA+ identifies 10 factors with a minimum variance explained of 1% in at least one data modality.

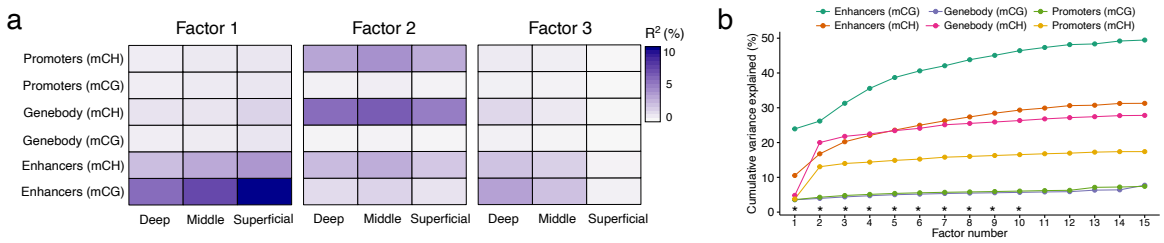


Figure 3.13: MOFA+ variance explained estimates in the frontal cortex DNA methylation data set.
(a) Percentage of variance explained for each factor across the different groups (cortical layer, x-axis) and views (genomic context, y-axis). For simplicity, only the first three factors are shown.
(b) Cumulative 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 data modality).

Factor 1, the major source of variation, is linked to the existence of inhibitory and excitatory neurons, the two major classes of neurons (Figure 3.15). This factor shows significant mCG activity across all cortical layers, mostly driven by coordinated changes in enhancer elements, but to some extent also gene bodies.

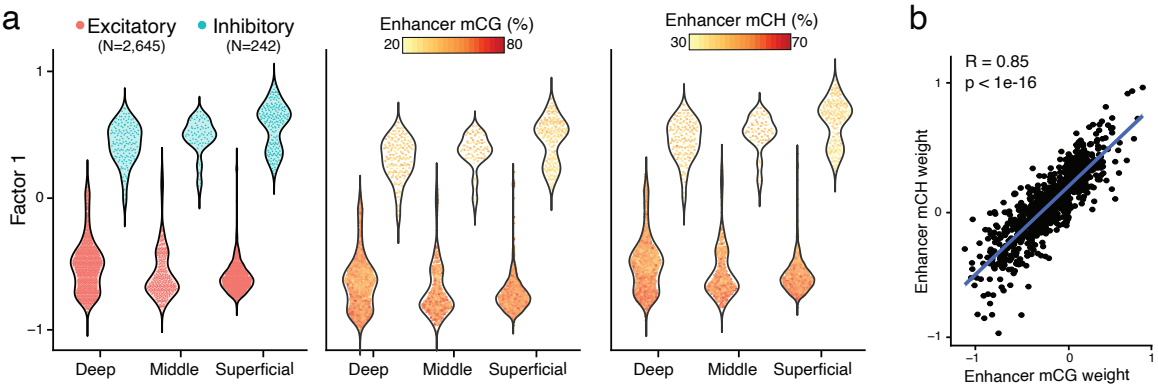


Figure 3.14: Characterisation of Factor 1 as DNA methylation signatures distinguishing inhibitory versus excitatory cell types[Luo2016].
(a) Beeswarm plots of Factor values per group (cortical layer). In the left plot, cells are coloured by neuron class. In the middle and right plots the cells are coloured by average mCG and mCH levels (%), respectively, of the top 100 enhancers with the largest weights.
(b) Correlation of enhancer mCG weights (x-axis) and mCH weights (y-axis)

Factor 2 captures genome-wide differences in global mCH levels (R=0.99, not shown), most likely to be a technical source of variation.

Factor 3 captures heterogeneity linked to the increased cellular diversity along cortical depth, with the Deep layer displaying significantly more diversity of excitatory cell types than the Superficial layer (**??**).

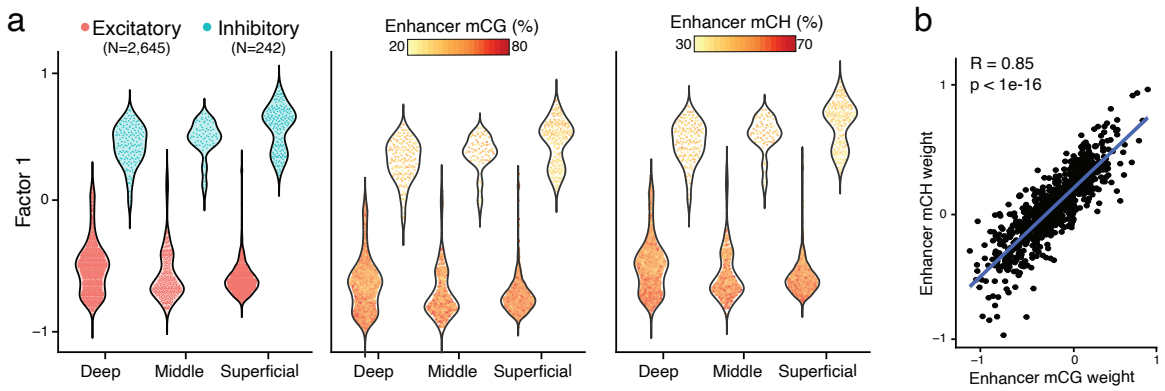


Figure 3.15: Characterisation of Factor 3 as increased cellular diversity along cortical depth.
(a) Beeswarm plots of Factor values per group (cortical layer). In the left plot, cells are coloured by neuron class. In the middle and right plots the cells are coloured by average mCG and mCH levels (%), respectively, of the top 100 enhancers with the largest weights.
(b) Correlation of enhancer mCG weights (x-axis) and mCH weights (y-axis).

The (linear) MOFA factors can be combined by further non-linear dimensionality reduction algorithms such as UMAP or t-SNE. In this case, we show that the t-SNE projections reveals the existence of multiple subpopulations of both excitatory and inhibitory cell types. Notably, the MOFA+ factors are significantly better at identifying these subpopulations than the conventional approach of using Principal Component Analysis with imputed measurements:

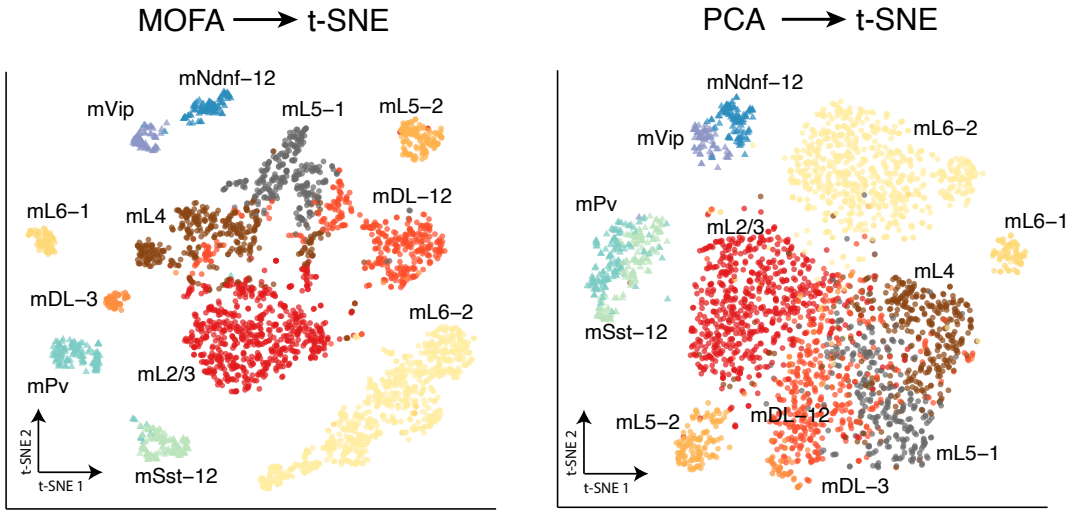


Figure 3.16: Comparison of MOFA+ Factors and Principal Components as input to t-SNE.
The scatterplots display t-SNE projections when using as input MOFA+ factors (left) or principal components (right). Each dot represents a cell, coloured by cell type[**Luo2016**]. To ensure a fair comparison we used the same number of PCs and MOFA+ Factors ($K = 15$). Feature-wise imputation of missing values is applied for the PCA.

Interestingly, in addition to the dominant mCG signal, MOFA+ connects Factor 1 and Factor 3 to variation in mCH, which suggest a role of mCH in cellular diversity. We hypothesise that this could be supported if the genomic regions that show mCH signatures are different than the ones marked

the conventional mCG signatures. To investigate this, we correlated the mCH and mCG feature loadings for each factor and genomic context (Figure 3e and Figure S14). In all cases we observe a strong positive dependency, indicating that mCH and mCG signatures are spatially correlated and target similar loci. Taken together, these results supports the hypothesis that mCH and mCG tag the same genomic loci and are associated with the same sources of variation, suggesting that the presence of mCH may be the result of non-specific *de novo* methylation as a by-product of the establishment of mCG.

3.4.3 Identification of molecular signatures of lineage commitment during mammalian embryogenesis

As a last application, we considered a substantially more complex dataset with multiple sample groups and data modalities. The dataset consists of a multi-omic atlas of mouse gastrulation where scNMT-seq was used to simultaneously profile RNA expression, DNA methylation and chromatin accessibility in 1,828 cells at multiple stages of development[9]. This is the data set that I introduced in Chapter 3.

In this dataset MOFA+ can be used to deline the coordinated variation between the transcriptome and the epigenome and detect at which stage(s) of development it occurs.
The main difference with respect to the MOFA analysis presented in Chapter 3 is that MOFA+ enables a multi-stage characterisation of the data set's variation. In FIGURE XXX we only considered stage E7.5, whereas here we can employ the multi-group functionality to perform a simultaneous analysis across multiple stages.

3.4.3.1 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 are described in Chapter 3.

As in the MOFA analysis presented in Chapter 3, here 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[9] (Figure 3.17). As discussed in Chapter 3, the CpG methylation (endogenous DNA methylation) and GpC methylation (proxy for chromatin accessibility) result in very sparse readouts that are challenging to analyse with standard statistical approaches.

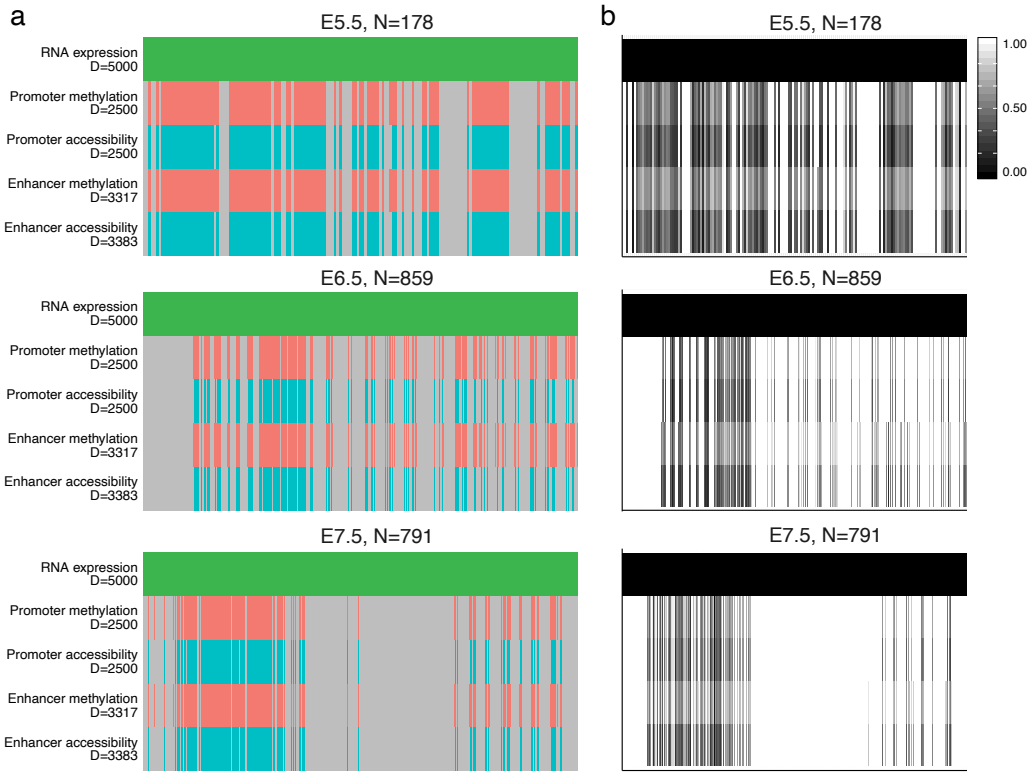


Figure 3.17: Overview of the scNMT-seq mouse gastrulation data set used as input for MOFA+.
(a) Structure of the input data in terms of modalities (x-axis) versus samples (y-axis). Each panel corresponds to a different group (embryonic stage). Grey bars represent missing modalities.
(b) Structure of the missing values in the data. For each cell and modality, the colour displays the fraction of missing values.

3.4.3.2 Model overview

In this data set MOFA+ identifies 8 factors with a minimum variance explained of 1% in the RNA expression (Figure 3.18). Interestingly, this plot indicates important differences in the heterogeneity of the considered data modalities. The MOFA+ Factors explain 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, suggesting that the sources of variation captured beyond Factor 2 are largely driven by RNA expression alone.

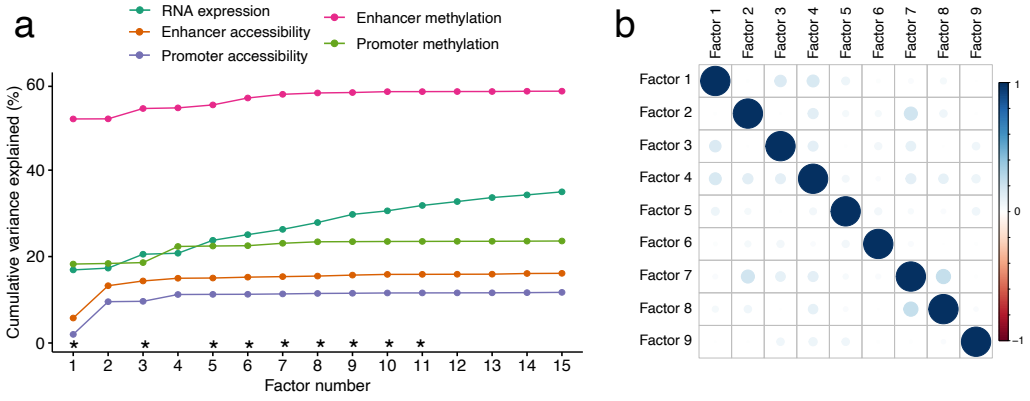


Figure 3.18: Unsupervised characterisation of MOFA+ factors from the scNMT-seq gastrulation data set.
(a) Cumulative variance explained (per view, y-axis) versus factor number (x-axis). Asterisks indicate the factors that are selected for downstream analysis (minimum of 1% variance explained in the RNA 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.

3.4.3.3 Characterisation of Factors 1 and 2

The first factor captured the formation of ExE endoderm, a cell type that is present across all stages (Figure 4a), in agreement with our previous results using the independently generated transcriptomic atlas of mouse gastrulation (Figure 2). MOFA+ links Factor 1 to changes across all molecular layers. Notably, the distribution of weights for DNA methylation are skewed towards negative values (at both enhancers and promoters), indicating that ExE endoderm cells are characterised by a state of global demethylation, consistent with previous studies 44.

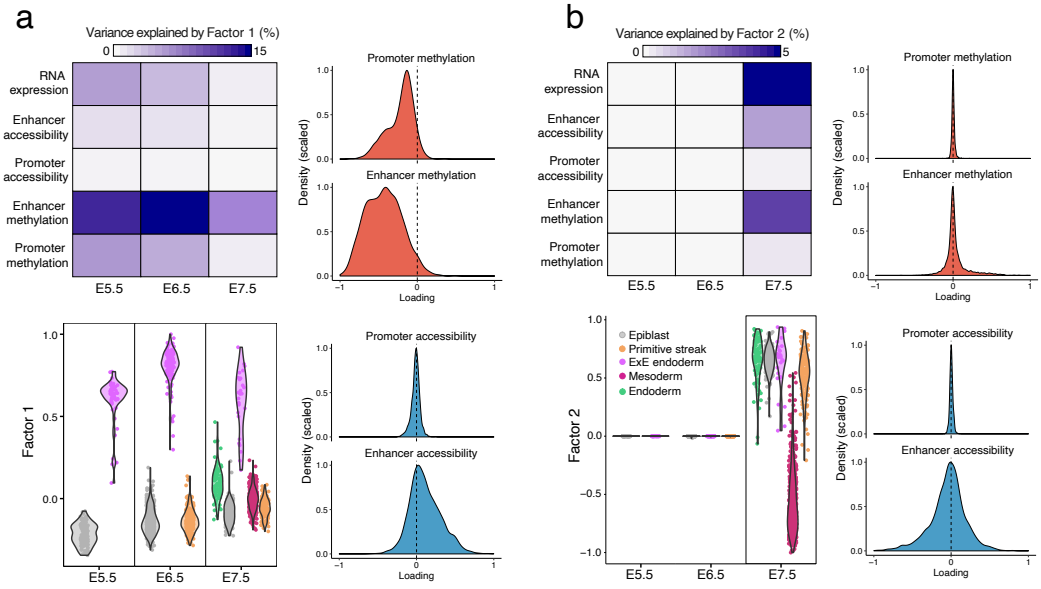


Figure 3.19

3.4.3.4 Characterisation of other Factors

As discussed above, the rest of the MOFA+ Factors explain significantly less variance than Factors 1 and 2, and they are mostly driven by the RNA expression (Figure 3.18). Their etiolgoy can be identified by the inspection of gene weights and by gene set enrichment analysis. For simplicity, I will only show Factor 6, which captures cell-cycle variation that is consistenly found across all three embryonic stages.

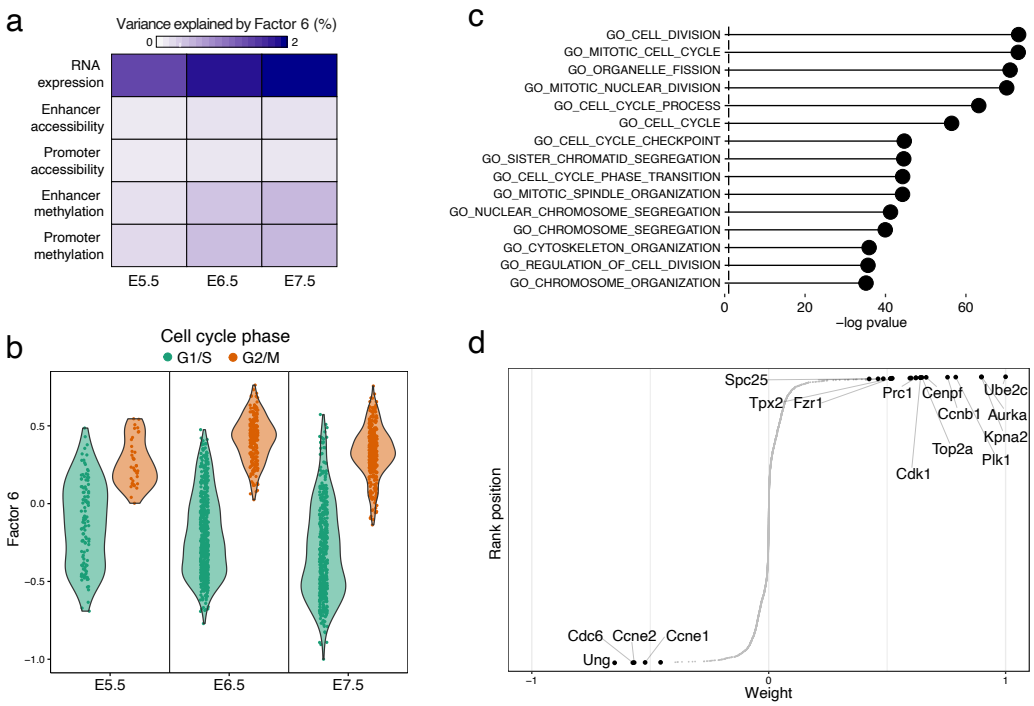


Figure 3.20: Characterisation of Factor 6 as cell cycle variation.
(a) Variance explained by Factor 6 in each group (embryonic stage, columns) and data modality (rows).
(b) Distribution of Factor 6 values per group (embryonic stage, x-axis), with cells coloured by the inferred cell cycle state using *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 labeled.

3.4.3.5 Conclusion

3.5 A note on the implementation

The inference framework in MOFA+ is implemented in Python, whereas the downstream analysis and visualisations are implemented in R. GPU acceleration is implemented using CuPy[182], an open-source matrix library accelerated with NVIDIA CUDA.
To facilitate adoption of the method, we deploy MOFA+ as open-source software (https://github.com/bioFAM/MOF with multiple tutorials and a web-based analysis workbench, hopefully enabling a user-friendly characterisation of structured single-cell data.

3.6 Limitations and open perspectives

In this Chapter we proposed a generalisation of the MOFA model for the principled analysis of large-scale *structured* data sets. This solves some of the limitations of the MOFA model presented in Chapter 2, but a significant amount 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 rigurous and scalable non-linear generalisation of factor models. [2].

- **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 [270, 197]. 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 [**XX**]. A clear application for this is spatial transcriptomics data, where one could build a (spatial) covariance matrix using cell-to-cell distances which can then be imposed in the prior distribution of the latent factors (recall that in MOFA the prior distribution for the factors assumes independence between samples). This would improve the detection of sources of variation with a spatial component.s

.1 Appendix: mathematical derivations of MOFA

.1.0.1 Deriving the variational inference algorithm

The theoretical foundations for the variational inference scheme are described in Section .1.0.1. 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:

$$\begin{aligned} q(\mathbf{X}) &= q\left(\{\widehat{\mathbf{Z}^{\mathbf{g}}}, \mathbf{S}^{\mathbf{g}}, \alpha^{\mathbf{g}}, \theta^{\mathbf{g}}\}, \{\widehat{\mathbf{W}^{\mathbf{m}}}, \mathbf{S}^{\mathbf{m}}, \alpha^{\mathbf{m}}, \theta^{\mathbf{m}}\}, \{\tau^{\mathbf{gm}}\}\right) \\ &= \prod_{g=1}^{G}\prod_{n=1}^{N_g}\prod_{k=1}^{K} q(\hat{z}_{nk}^{g}, s_{nk}^{g}) \prod_{g=1}^{G}\prod_{k=1}^{K} q(\alpha_k^g) \prod_{g=1}^{G}\prod_{k=1}^{K} q(\theta_k^g) \\ &\times \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{kd}^{m}, s_{kd}^{m}) \prod_{m=1}^{M}\prod_{k=1}^{K} q(\alpha_k^m) \prod_{m=1}^{M}\prod_{k=1}^{K} q(\theta_k^m) \\ &\times \prod_{g=1}^{G}\prod_{m=1}^{M}\prod_{d=1}^{D_m} q(\tau_d^{gm}) \end{aligned} \tag{28}$$

However, inspired by [249], 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 . 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 $\mathcal{L}(\mathbf{X})$[22, 176]. 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.

.1.1 Variational update equations

Factors

For every group g, sample n and factor k:

Prior distribution $p(\hat{z}_{nk}^g, s_{nk}^g)$:

$$p(\hat{z}_{nk}^g, s_{nk}^g) = \mathcal{N}(\hat{z}_{nk}^g \,|\, 0,\ 1/\alpha_k^g)\,\mathrm{Ber}(s_{nk}^g \,|\, \theta_k^g) \tag{29}$$

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

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

$$q(s^g_{nk}) = \text{Ber}(s^g_{nk}|\gamma^g_{nk}) \tag{30}$$

with

$$\begin{aligned}
\gamma^g_{nk} &= \frac{1}{1+\exp(-\lambda^g_{nk})} \\
\lambda^g_{nk} &= \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle} - 0.5 \ln \left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle} \right) \\
&+ \frac{\langle \tau^{gm}_d \rangle}{2} \frac{\left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} y^{gm}_{nd} \langle w^m_{kd} \rangle - \sum_{j \neq k} \langle s^g_{nj} \hat{z}^g_{nj} \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w^m_{kd} \rangle \langle w^m_{jd} \rangle \right)^2}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}}
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
q(\hat{z}^g_{nk}|s^g_{nk} = 0) &= \mathcal{N}\left(\hat{z}^g_{nk}|0, 1/\alpha^g_k\right) \\
q(\hat{z}^g_{nk}|s^g_{nk} = 1) &= \mathcal{N}\left(\hat{z}^g_{nk}\middle|\mu_{z^g_{nk}}, \sigma^2_{z^g_{nk}}\right)
\end{aligned} \tag{32}$$

with

$$\begin{aligned}
\mu_{z^g_{nk}} &= \frac{\sum_{m=1}^{M} \sum_{d=1}^{D_m} y^{m,g}_{nd} \langle w^m_{kd} \rangle - \sum_{j \neq k} \langle s^g_{nj} \hat{z}^g_{nj} \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w^m_{kd} \rangle \langle w^m_{jd} \rangle}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}} \\
\sigma^2_{z^g_{nk}} &= \frac{\langle \tau^{gm}_d \rangle^{-1}}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w^m_{kd})^2 \rangle + \frac{\langle \alpha^g_k \rangle}{\langle \tau^{gm}_d \rangle}}
\end{aligned} \tag{33}$$

ARD prior on the factors

For every group g and factor k:

Prior distribution:

$$p(\alpha^g_k) = \mathcal{G}\left(\alpha^g_k \,|\, a^\alpha_0, b^\alpha_0\right) \tag{34}$$

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

$$q(\alpha^g_k) = \mathcal{G}\left(\alpha^g_k \,|\, \hat{a}^\alpha_{gk}, \hat{b}^\alpha_{gk}\right) \tag{35}$$

where:

$$\begin{aligned}
\hat{a}^\alpha_{gk} &= a^\alpha_0 + \frac{N_g}{2} \\
\hat{b}^\alpha_{gk} &= b^\alpha_0 + \frac{\sum_{n=1}^{N_g} \langle (\hat{z}^g_{nk})^2 \rangle}{2}
\end{aligned} \tag{36}$$

Sparsity parameter of the Factors

For every group g and factor k:

Prior distribution:

$$p(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, a_0^\theta, b_0^\theta\right) \tag{37}$$

Variational distribution:

$$q(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, \hat{a}_{gk}^\theta, \hat{b}_{gk}^\theta\right) \tag{38}$$

where

$$\begin{aligned} \hat{a}_{gk}^\theta &= \sum_{n=1}^{N_g} \langle s_{nk}^g \rangle + a_0^\theta \\ \hat{b}_{gk}^\theta &= b_0^\theta - \sum_{n=1}^{N_g} \langle s_{nk}^g \rangle + N_g \end{aligned} \tag{39}$$

Feature weights

For every view m, feature d and factor k:

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

$$p(\hat{w}_{kd}^m, s_{kd}^m) = \mathcal{N}(\hat{w}_{kd}^m \,|\, 0,\ 1/\alpha_k^m)\, \text{Ber}(s_{kd}^m \,|\, \theta_k^m) \tag{40}$$

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

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

$$q(s_{kd}^m) = \text{Ber}(s_{kd}^m | \gamma_{kd}^m) \tag{41}$$

with

$$\begin{aligned} \gamma_{kd}^m &= \frac{1}{1+\exp(-\lambda_{kd}^m)} \\ \lambda_{kd}^m &= \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} - 0.5 \ln \left(\sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} \right) \\ &+ \frac{\langle \tau_d^{gm} \rangle}{2} \frac{\left(\sum_{g=1}^{G} \sum_{n=1}^{N_g} y_{nd}^{gm} \langle z_{nk}^g \rangle - \sum_{j \neq k} \langle s_{jd}^m \hat{w}_{jd}^m \rangle \sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle z_{nk}^g \rangle \langle z_{nj}^g \rangle \right)^2}{\sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle}} \end{aligned} \tag{42}$$

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

$$\begin{aligned} q(\hat{w}_{kd}^m | s_{kd}^m = 0) &= \mathcal{N}\left(\hat{w}_{kd}^m | 0, 1/\alpha_k^m\right) \\ q(\hat{w}_{kd}^m | s_{kd}^m = 1) &= \mathcal{N}\left(\hat{w}_{kd}^m \middle| \mu_{w_{kd}^m}, \sigma_{w_{kd}^m}^2\right) \end{aligned} \tag{43}$$

with

$$\mu_{w_{kd}^m} = \frac{\sum_{g=1}^{G}\sum_{n=1}^{N_g} y_{nd}^{gm}\langle z_{nk}^g\rangle - \sum_{j\neq k}\langle s_{jd}^m \hat{w}_{jd}^m\rangle \sum_{g=1}^{G}\sum_{n=1}^{N_g}\langle z_{nk}^g\rangle\langle z_{nj}^g\rangle}{\sum_{g=1}^{G}\sum_{n=1}^{N_g}\langle (z_{nk}^g)^2\rangle + \frac{\langle \alpha_k^m\rangle}{\langle \tau_d^{gm}\rangle}}$$
$$\sigma^2_{w_{kd}^m} = \frac{\langle \tau_d^{gm}\rangle^{-1}}{\sum_{g=1}^{G}\sum_{n=1}^{N_g}\langle (z_{nk}^g)^2\rangle + \frac{\langle \alpha_k^m\rangle}{\langle \tau_d^{gm}\rangle}} \tag{44}$$

ARD prior on the weights

For every view m and factor k:

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

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

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

$$q(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \mid \hat{a}_{mk}^\alpha, \hat{b}_{mk}^\alpha\right) \tag{45}$$

where:

$$\hat{a}_{mk}^\alpha = a_0^\alpha + \frac{D_m}{2}$$
$$\hat{b}_{mk}^\alpha = b_0^\alpha + \frac{\sum_{d=1}^{D_m}\langle (\hat{w}_{kd}^m)^2\rangle}{2} \tag{46}$$

Sparsity parameter of the weights

For every view m and factor k:

Prior distribution:

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

Variational distribution:

$$q(\theta_k^m) = \text{Beta}\left(\theta_k^m \mid \hat{a}_{mk}^\theta, \hat{b}_{mk}^\theta\right) \tag{47}$$

where

$$\hat{a}_{mk}^\theta = \sum_{d=1}^{D_m}\langle s_{kd}^m\rangle + a_0^\theta$$
$$\hat{b}_{mk}^\theta = b_0^\theta - \sum_{d=1}^{D_m}\langle s_{kd}^m\rangle + D_m \tag{48}$$

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

$$q(\tau_d^{gm}) = \mathcal{G}\left(\tau_d^{gm} \,|\, \hat{a}_d^{gm}, \hat{b}_d^{gm}\right) \tag{49}$$

where:

$$\begin{aligned} \hat{a}_d^{gm} &= a_0^\tau + \frac{N_g}{2} \\ \hat{b}_d^{gm} &= b_0^\tau + \frac{1}{2}\sum_{n=1}^{N_g}\langle\left(y_{nd}^{gm} - \sum_k^K w_{kd}^m z_{nk}^g\right)^2\rangle \end{aligned} \tag{50}$$

.1.2 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 (1.2), 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:

$$\mathcal{L} = \mathbb{E}_{q(X)} \ln p(Y|X) - \mathrm{KL}\left(q(X)||p(X)\right) \tag{51}$$

Log likelihood term Assuming a Gaussian likelihood:

$$\begin{aligned} \mathbb{E}_{q(X)} \ln p(Y|X) = &- \sum_{m=1}^M \frac{N D_m}{2}\ln(2\pi) + \sum_{g=1}^G \frac{N_g}{2}\sum_{m=1}^M\sum_{d=1}^{D_m}\langle\ln(\tau_d^{gm})\rangle \\ &- \sum_{g=1}^G\sum_{m=1}^M\sum_{d=1}^{D_m}\frac{\langle\tau_d^{gm}\rangle}{2}\sum_{n=1}^{N_g}\left(y_{nd}^{m,g} - \sum_{k=1}^K\langle s_{kd}^m \hat{w}_{kd}^m\rangle\langle z_{nk}^g\rangle\right)^2 \end{aligned} \tag{52}$$

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

$$\begin{aligned} \mathbb{E}_q[\ln p(\hat{W}, S)] = &- \sum_{m=1}^M \frac{K D_m}{2}\ln(2\pi) + \sum_{m=1}^M \frac{D_m}{2}\sum_{k=1}^K \ln(\alpha_k^m) - \sum_{m=1}^M \frac{\alpha_k^m}{2}\sum_{d=1}^{D_m}\sum_{k=1}^K\langle(\hat{w}_{kd}^m)^2\rangle \\ &+ \langle\ln(\theta)\rangle\sum_{m=1}^M\sum_{d=1}^{D_m}\sum_{k=1}^K\langle s_{kd}^m\rangle + \langle\ln(1-\theta)\rangle\sum_{m=1}^M\sum_{d=1}^{D_m}\sum_{k=1}^K(1-\langle s_{kd}^m\rangle) \end{aligned} \tag{53}$$

$$\begin{aligned}\mathbb{E}_q[\ln q(\hat{W},S)] = &-\sum_{m=1}^{M}\frac{KD_m}{2}\ln(2\pi)+\frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\ln(\langle s^m_{kd}\rangle\sigma^2_{w^m_{kd}}+(1-\langle s^m_{kd}\rangle)/\alpha^m_k)\\ &+\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}(1-\langle s^m_{kd}\rangle)\ln(1-\langle s^m_{kd}\rangle)-\langle s^m_{kd}\rangle\ln\langle s^m_{kd}\rangle\end{aligned} \tag{54}$$

Factors

$$\begin{aligned}\mathbb{E}_q[\ln p(\hat{Z},S)] = &-\sum_{g=1}^{G}\frac{N_gK}{2}\ln(2\pi)+\sum_{g=1}^{G}\frac{N_g}{2}\sum_{k=1}^{K}\ln(\alpha^g_k)-\sum_{g=1}^{G}\frac{\alpha^g_k}{2}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\langle(\hat{z}^g_{nk})^2\rangle\\ &+\langle\ln(\theta)\rangle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\langle s^g_{nk}\rangle+\langle\ln(1-\theta)\rangle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}(1-\langle s^g_{nk}\rangle)\end{aligned} \tag{55}$$

$$\begin{aligned}\mathbb{E}_q[\ln q(\hat{Z},S)] = &-\sum_{g=1}^{G}\frac{N_gK}{2}\ln(2\pi)+\frac{1}{2}\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\ln(\langle s^g_{nk}\rangle\sigma^2_{z^g_{nk}}+(1-\langle s^g_{nk}\rangle)/\alpha^g_k)\\ &+\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}(1-\langle s^g_{nk}\rangle)\ln(1-\langle s^g_{nk}\rangle)-\langle s^g_{nk}\rangle\ln\langle s^g_{nk}\rangle\end{aligned} \tag{56}$$

ARD prior on the weights

$$\begin{aligned}\mathbb{E}_q[\ln p(\boldsymbol{\alpha})] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(a^\alpha_0\ln b^\alpha_0+(a^\alpha_0-1)\langle\ln\alpha_k\rangle-b^\alpha_0\langle\alpha_k\rangle-\ln\Gamma(a^\alpha_0)\Big)\\ \mathbb{E}_q[\ln q(\boldsymbol{\alpha})] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\Big(\hat{a}^\alpha_k\ln\hat{b}^\alpha_k+(\hat{a}^\alpha_k-1)\langle\ln\alpha_k\rangle-\hat{b}^\alpha_k\langle\alpha_k\rangle-\ln\Gamma(\hat{a}^\alpha_k)\Big)\end{aligned} \tag{57}$$

Sparsity parameter of the weights

$$\begin{aligned}\mathbb{E}_q\left[\ln p(\boldsymbol{\theta})\right] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\big((a_0-1)\times\langle\ln(\pi^m_{d,k})\rangle+(b_0-1)\langle\ln(1-\pi^m_{d,k})\rangle-\ln(\mathrm{B}(a_0,b_0))\big)\\ \mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right] &= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}\big((a^m_{k,d}-1)\times\langle\ln(\pi^m_{d,k})\rangle+(b^m_{k,d}-1)\langle\ln(1-\pi^m_{d,k})\rangle-\ln(\mathrm{B}(a^m_{k,d},b^m_{k,d}))\big)\end{aligned} \tag{58}$$

ARD prior on the Factors

$$\begin{aligned}\mathbb{E}_q[\ln p(\boldsymbol{\alpha})] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\Big(a^\alpha_0\ln b^\alpha_0+(a^\alpha_0-1)\langle\ln\alpha_k\rangle-b^\alpha_0\langle\alpha_k\rangle-\ln\Gamma(a^\alpha_0)\Big)\\ \mathbb{E}_q[\ln q(\boldsymbol{\alpha})] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\Big(\hat{a}^\alpha_k\ln\hat{b}^\alpha_k+(\hat{a}^\alpha_k-1)\langle\ln\alpha_k\rangle-\hat{b}^\alpha_k\langle\alpha_k\rangle-\ln\Gamma(\hat{a}^\alpha_k)\Big)\end{aligned} \tag{59}$$

Sparsity parameter of the Factors

$$\begin{aligned}
\mathbb{E}_q\left[\ln p(\boldsymbol{\theta})\right] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\Big((a_0-1)\times\langle\ln(\pi_{n,k}^g)\rangle + (b_0-1)\langle\ln(1-\pi_{n,k}^g)\rangle - \ln(\mathrm{B}(a_0,b_0))\Big) \\
\mathbb{E}_q\left[\ln q(\boldsymbol{\theta})\right] &= \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\Big((a_{k,n}^g-1)\times\langle\ln(\pi_{n,k}^g)\rangle + (b_{k,n}^g-1)\langle\ln(1-\pi_{n,k}^g)\rangle - \ln(\mathrm{B}(a_{k,n}^g,b_{k,n}^g))\Big)
\end{aligned} \tag{60}$$

Noise

$$\begin{aligned}
\mathbb{E}_q[\ln p(\boldsymbol{\tau})] &= \sum_{m=1}^{M} D_m a_0^\tau \ln b_0^\tau + \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm}(a_0^\tau-1)\langle\ln\tau_d^{gm}\rangle - \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm} b_0^\tau\langle\tau_d^{gm}\rangle - \sum_{m=1}^{M} D_m\ln\Gamma(a_0^\tau) \\
\mathbb{E}_q[\ln q(\boldsymbol{\tau})] &= \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\Big(\hat{a}_{dgm}^\tau\ln\hat{b}_{dgm}^\tau + (\hat{a}_{dgm}^\tau-1)\langle\ln\tau_d^{gm}\rangle - \hat{b}_{dgm}^\tau\langle\tau_d^{gm}\rangle - \ln\Gamma(\hat{a}_{dgm}^\tau)\Big)
\end{aligned} \tag{61}$$

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

[3] U. D. Akavia et al. "An Integrated Approach to Uncover Drivers of Cancer". In: *Cell* 143.6 (2010), pp. 1005–1017. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2010.11.013.

[4] 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.

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

[6] 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 (Oct. 1999).

[7] C. Angermueller et al. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning". In: *Genome Biology* 18.1 (2017), p. 67.

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

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

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] S. B. Baylin and P. A. Jones. “A decade of exploring the cancer epigenome —biological and translational implications”. In: *Nature Reviews Cancer* 11 (Sept. 2011).

[17] J. T. Bell et al. “DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines”. In: *Genome Biology* 12.1 (2011).

[18] 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.

[19] 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.

[20] C. Bishop. “Variational Principal Components”. In: *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN’99*. Vol. 1. 1999, pp. 509–514.

[21] 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.

[22] C. M. Bishop. “Pattern recognition”. In: *Machine Learning* 128 (2006), pp. 1–58.

[23] 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.

[24] 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].

[25] 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.

[26] M. J. Bonder et al. “Disease variants alter transcription factor levels and methylation of their binding sites”. In: *Nature Genetics* 49 (Dec. 2016).

[27] 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).

[28] 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].

[29] 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.

[30] J. D. Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current Protocols in Molecular Biology* 109.1 (2015).

[31] J. D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523 (June 2015).

[32] 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 (Oct. 2013).

[33] F. Buettner et al. "f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq". In: *Genome Biol.* 18.1 (2017), p. 212.

[34] 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.

[35] K. Bunte et al. "Sparse group factor analysis for biclustering of multiple data sources". In: *Bioinformatics* 32.16 (2016), pp. 2457–2463.

[36] A. Butler et al. "Integrating single-cell transcriptomic data across different conditions, technologies, and species". In: *Nature Biotechnology* 36.5 (2018), pp. 411–420.

[37] 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.

[38] J. Cao et al. "Joint profiling of chromatin accessibility and gene expression in thousands of single cells". In: *Science* 361.6409 (Sept. 2018), p. 1380.

[39] 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.

[40] J. Cao et al. "The single-cell transcriptional landscape of mammalian organogenesis". In: *Nature* 566.7745 (2019), pp. 496–502.

[41] 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.

[42] 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.

[43] L. Chen et al. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells". In: *Cell* 167.5 (2016), 1398–1414.e24.

[44] 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.

[45] X. Chen et al. "A rapid and robust method for single cell chromatin accessibility profiling". In: *Nature Communications* 9.1 (2018), p. 5345.

[46] 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 (Feb. 2017).

[47] 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.

[48] 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.

[49] 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.

[50] T. E. P. Consortium et al. "An integrated encyclopedia of DNA elements in the human genome". In: *Nature* 489 (Sept. 2012).

[51] J. C. Costello et al. "A community effort to assess and improve drug sensitivity prediction algorithms". In: *Nature Biotechnology* 32 (June 2014).

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] N. C. Dempsey et al. "Differential heat shock protein localization in chronic lymphocytic leukemia". In: *Journal of Leukocyte Biology* 87.3 (2020/03/23 2010), pp. 467–476.

[58] 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.

[59] S. Dietrich et al. "Drug-perturbation-based stratification of blood cancer". In: *J. Clin. Invest.* 128.1 (2018), pp. 427–445.

[60] L. Dietz. *Directed Factor Graph Notation for Generative Models*. 2010.

[61] 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.

[62] 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.

[63] Z. Du et al. “Allelic reprogramming of 3D chromatin architecture during early mammalian development”. In: *Nature* 547 (2017).

[64] 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.

[65] 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.

[66] G. Eraslan et al. “Single-cell RNA-seq denoising using a deep count autoencoder”. In: *Nature Communications* 10.1 (2019), p. 390.

[67] G. Fabbri and R. Dalla-Favera. “The molecular pathogenesis of chronic lymphocytic leukaemia”. In: *Nat. Rev. Cancer* 16.3 (2016), pp. 145–162.

[68] 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.

[69] J. Fan et al. “Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data”. In: *Genome Research* 28.8 (Aug. 2018), pp. 1217–1227.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] H. R. Frost, Z. Li, and J. H. Moore. “Principal component gene set enrichment (PCGSE)”. In: *BioData mining* 8.1 (2015).

[75] 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.

[76] 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].

[77] A. Gelman et al. *Bayesian Data Analysis, Third Edition*. Hardcover. 2013.

[78] M. Gerstung et al. "Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes". In: *Nature Communications* 6 (Jan. 2015).

[79] 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.

[80] 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.

[81] F. Guo et al. "Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells". In: *Cell Research* 27 (June 2017).

[82] H. Guo et al. "Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing". In: *Nature Protocols* 10 (Apr. 2015).

[83] 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 (Dec. 2013), pp. 2126–2135.

[84] 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.

[85] L. Haghverdi et al. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". In: *Nature Biotechnology* 36 (2018).

[86] L. Haghverdi et al. "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nature Methods* 13 (Aug. 2016).

[87] 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.

[88] W. Hardle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2007, pp. 321–30. ISBN: 978-3-540-72243-4.

[89] T. Hashimshony et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (2012), pp. 666–673.

[90] 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.

[91] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman, 2015. ISBN: 9781498712163.

[92] H. H. He et al. "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification". In: *Nature Methods* 11 (Dec. 2013).

[93] Y. He and J. R. Ecker. "Non-CG Methylation in the Human Genome". In: *Annual Review of Genomics and Human Genetics* 16.1 (2015). PMID: 26077819, pp. 55–77. DOI: 10.1146/annurev-genom-090413-025437.

[94] N. D. Heintzman et al. "Histone modifications at human enhancers reflect global cell-type-specific gene expression". In: *Nature* 459.7243 (2009).

[95] A. Hemmati-Brivanlou and D. Melton. "Vertebrate Embryonic Cells Will Become Nerve Cells Unless Told Otherwise". In: *Cell* 88.1 (1997).

[96] M. D. Hoffman et al. "Stochastic variational inference". In: *The Journal of Machine* (2013).

[97] 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.

[98] M. Hoffman et al. "Stochastic Variational Inference". In: *arXiv e-prints*, arXiv:1206.7051 (2012), arXiv:1206.7051. eprint: 1206.7051.

[99] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.

[100] 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.

[101] Y. Hou et al. "Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas". In: *Cell Research* 26 (Feb. 2016).

[102] Y. Hu et al. "Simultaneous profiling of transcriptome and DNA methylome from a single cell". In: *Genome Biology* 17.1 (2016), p. 88.

[103] 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.

[104] 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.

[105] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1 (2000), pp. 25–37.

[106] E. Jaynes. "Prior Probabilities". In: *IEEE Transactions on Systems Science and Cybernetics* 4.3 (1968), pp. 227–241.

[107] C. Jiang and B. F. Pugh. "Nucleosome positioning and gene regulation: advances through genomics". In: *Nature Reviews Genetics* 10 (Mar. 2009).

[108] W. Jin et al. "Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples". In: *Nature* 528 (Nov. 2015).

[109] 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.

[110] 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.

[111] P. A. Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13 (May 2012).

[112] C.-A. Kapourani and G. Sanguinetti. "BPRMeth: a flexible Bioconductor package for modelling methylation profiles". In: *Bioinformatics* 34.14 (2018), pp. 2485–2486. DOI: 10.1093/bioinformatics/bty129.

[113] C.-A. Kapourani and G. Sanguinetti. "Melissa: Bayesian clustering and imputation of single cell methylomes". In: *bioRxiv* (2018). DOI: 10.1101/312025.

[114] H. S. Kaya-Okur et al. "CUT&Tag for efficient epigenomic profiling of small samples and single cells". In: *bioRxiv* (Jan. 2019), p. 568915.

[115] Y. Ke et al. "3D Chromatin Structures of Mature Gametes and Structural Reprogramming during Mammalian Embryogenesis". In: *Cell* 170.2 (2017).

[116] T. K. Kelly et al. "Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules". In: *Genome Research* 22.12 (Dec. 2012), pp. 2497–2506.

[117] 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.

[118] 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.

[119] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. "Bayesian approach to single-cell differential expression analysis". In: *Nature Methods* 11 (May 2014).

[120] 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.

[121] M. Kim et al. "Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli". In: *Nature Communications* 7 (Oct. 2016).

[122] V. Y. Kiselev et al. "SC3: consensus clustering of single-cell RNA-seq data". In: *Nature Methods* 14.5 (2017).

[123] A. Klami and S. Kaski. "Probabilistic approach to detecting dependencies between data sets". In: *Neurocomputing* 72.1 (2008), pp. 39–46.

[124] 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.

[125] A. Klami et al. "Group factor analysis". In: *IEEE transactions on neural networks and learning systems* 26.9 (2015), pp. 2136–2147.

[126] A. M. Klein et al. "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5 (2015), pp. 1187–1201.

[127] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* (2019).

[128] 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.

[129] 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.

[130] A. Kristiadi. *Natural Gradient Descent*. https://wiseodd.github.io/techblog/2018/03/14/natural-gradient/. Blog. 2019.

[131] 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.

[132] 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.

[133] M. Kuhn and J. Kjell. "Applied Predictive Modeling". In: (2013).

[134] 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.

[135] G. La Manno et al. "RNA velocity of single cells". In: *Nature* 560.7719 (2018), pp. 494–498.

[136] 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.

[137] G. R. G. Lanckriet et al. "A statistical framework for genomic data fusion". In: *Bioinformatics* 20.16 (2004).

[138] 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.

[139] H. J. Lee, T. A. Hore, and W. Reik. "Reprogramming the Methylome: Erasing Memory and Creating Diversity". In: *Cell Stem Cell* 14.6 (2014), pp. 710–719. ISSN: 1934-5909. DOI: https://doi.org/10.1016/j.stem.2014.05.008.

[140] H. J. Lee et al. "Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos". In: *Nature Communications* 6.1 (2015).

[141] I. Lee et al. "Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing". In: *bioRxiv* (Jan. 2018), p. 504993.

[142] J. T. Leek and J. D. Storey. "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis". In: *PLoS Genet.* 3.9 (2007), e161.

[143] 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.

[144] Y. Li, F.-X. Wu, and A. Ngom. "A review on machine learning principles for multi-view biological data integration". In: *Briefings in Bioinformatics* 19.2 (2016), pp. 325–340. ISSN: 1477-4054. DOI: 10.1093/bib/bbw113.

[145] 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.

[146] 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.

[147] 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.

[148] D. Lin et al. "An integrative imputation method based on multi-omics datasets". In: *BMC Bioinformatics* 17.1 (2016), p. 247.

[149] 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.

[150] R. Lister et al. "Human DNA methylomes at base resolution show widespread epigenomic differences". In: *Nature* 462 (Oct. 2009).

[151] L. Liu et al. "Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity". In: *Nature Communications* 10.1 (2019).

[152] Y. Liu et al. "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis". In: *Nature Biotechnology* 31 (Jan. 2013).

[153] R. Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (2018), pp. 1053–1058.

[154] 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.

[155] C. Luo et al. "Robust single-cell DNA methylome profiling with snmC-seq2". In: *Nature Communications* 9.1 (2018), p. 3824. DOI: 10.1038/s41467-018-06355-2.

[156] I. C. Macaulay et al. "G&T-seq: parallel sequencing of single-cell genomes and transcriptomes". In: *Nature Methods* 12 (Apr. 2015).

[157] D. J. MacKay. "Bayesian methods for backpropagation networks". In: *Models of neural networks III*. Springer, 1996, pp. 211–254.

[158] 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.

[159] 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].

[160] 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.

[161] 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.

[162] D. J. McCarthy et al. "Cardelino: Integrating whole exomes and single-cell transcriptomes to reveal phenotypic impact of somatic variants". In: *bioRxiv* (Jan. 2018), p. 413047.

[163] D. J. McCarthy et al. "Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R". In: *Bioinformatics* 33.8 (2017), pp. 1179–1186. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw777.

[164] L. McInnes, J. Healy, and J. Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018. arXiv: 1802.03426 [stat.ML].

[165] A. Meissner et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis". In: *Nucleic Acids Research* 33.18 (2005), pp. 5868–5877. ISSN: 0305-1048. DOI: 10.1093/nar/gki901.

[166] C. Meng et al. "Dimension reduction techniques for the integrative analysis of multi-omics data". In: *Brief. Bioinform.* 17.4 (2016), pp. 628–641.

[167] T. P. Minka. "Expectation Propagation for approximate Bayesian inference". In: *arXiv e-prints*, arXiv:1301.2294 (2013), arXiv:1301.2294. arXiv: 1301.2294.

[168] 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.

[169] 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.

[170] 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.

[171] 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.

[172] A. Mortazavi et al. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". In: *Nature Methods* 5 (May 2008).

[173] A. Moudgil et al. "Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells". In: *bioRxiv* (Jan. 2019), p. 538553.

[174] R. M. Mulqueen et al. "Highly scalable generation of DNA methylation profiles in single cells". In: *Nature Biotechnology* 36 (Apr. 2018).

[175] I. Munoz-Sanjuan and A. H. Brivanlou. "Neural induction, the default model and embryonic stem cells". In: *Nature Reviews Neuroscience* 3.4 (2002).

[176] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 9780262018029.

[177] 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.

[178] 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.

[179] R. M. Neal. *Bayesian learning for neural networks*. 1995.

[180] K. Nordstrom et al. "Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data". In: *bioRxiv* (2019). DOI: 10.1101/547596.

[181] 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.

[182] 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.

[183] E. Papalexi and R. Satija. “Single-cell RNA sequencing to explore immune cell heterogeneity”. In: *Nature Reviews Immunology* 18 (Aug. 2017).

[184] B. Papp and K. Plath. “Epigenetics of Reprogramming to Induced Pluripotency”. In: *Cell* 152.6 (2013).

[185] 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.

[186] 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.

[187] 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.

[188] F. Paul et al. “Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors”. In: *Cell* 163.7 (2015), pp. 1663–1677.

[189] V. M. Peterson et al. “Multiplexed quantification of proteins and transcripts in single cells”. In: *Nature Biotechnology* 35 (Aug. 2017).

[190] S. Picelli et al. “Full-length RNA-seq from single cells using Smart-seq2”. In: *Nature Protocols* 9 (Jan. 2014).

[191] B. L. Pierce et al. “Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms”. In: *Nature Communications* 9.1 (2018), p. 804.

[192] B. Pijuan-Sala et al. “A single-cell molecular map of mouse gastrulation and early organogenesis”. In: *Nature* 566.7745 (2019), pp. 490–495.

[193] 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.

[194] 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.

[195] S. Pott. “Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells”. In: *eLife* 6 (2017). Ed. by B. Ren, e23203. ISSN: 2050-084X. DOI: 10.7554/eLife.23203.

[196] I. Pournara and L. Wernisch. "Factor analysis for gene regulatory networks and transcription factor activity profiles". In: *BMC Bioinformatics* 8.1 (2007), p. 61.

[197] N. Qian. "On the momentum term in gradient descent learning algorithms". In: *Neural Networks* 12.1 (1999), pp. 145–151.

[198] 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.

[199] A. Rada-Iglesias et al. "A unique chromatin signature uncovers early developmental enhancers in humans". In: *Nature* 470.7333 (2011).

[200] H. Raiffa and R. Schlaifer. *Applied statistical decision theory*. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961. ISBN: 9780875840178.

[201] 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.

[202] B. H. Ramsahoye et al. "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a". In: *Proceedings of the National Academy of Sciences* 97.10 (2000), pp. 5237–5242. ISSN: 0027-8424. DOI: 10.1073/pnas.97.10.5237.

[203] D. Ramskold et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature Biotechnology* 30 (July 2012).

[204] 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].

[205] 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*. ICML'13. JMLR.org, 2013, pp. II-298–II-306.

[206] K. D. Rasmussen and K. Helin. "Role of TET enzymes in DNA methylation, development, and cancer". In: *Genes & Development* 30.7 (2016).

[207] 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.

[208] A. Regev et al. "Science Forum: The Human Cell Atlas". In: *eLife* 6 (2017), e27041. DOI: 10.7554/eLife.27041.

[209] 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).

[210] J. M. Replogle et al. "Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing". In: *Nature Biotechnology* (2020).

[211] M. Ringnér. "What is principal component analysis?" In: *Nat. Biotechnol.* 26 (2008), p. 303.

[212] D. Risso et al. "A general and flexible method for signal extraction from single-cell RNA-seq data". In: *Nature Communications* 9.1 (2018), p. 284.

[213] M. D. Ritchie et al. “Methods of integrating data to uncover genotype–phenotype interactions”. In: *Nature Reviews Genetics* 16 (Jan. 2015).

[214] H. Robbins and S. Monro. “A stochastic approximation method”. In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[215] 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.

[216] D. B. Rubin and D. T. Thayer. “EM algorithms for ML factor analysis”. In: *Psychometrika* 47.1 (1982), pp. 69–76.

[217] S. Rulands et al. “Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency”. In: *Cell Systems* 7.1 (2018).

[218] W. Saelens et al. “A comparison of single-cell trajectory inference methods: towards more accurate and robust tools”. In: *bioRxiv* (Jan. 2018), p. 276907.

[219] 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.

[220] J. L. Sardina et al. “Transcription Factors Drive Tet2-Mediated Enhancer Demethylation to Reprogram Cell Fate”. In: *Cell Stem Cell* 23.5 (2018).

[221] 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.

[222] N. Schaum et al. “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris”. In: *Nature* 562.7727 (2018).

[223] M. Seeger and G. Bouchard. “Fast variational Bayesian inference for non-conjugate matrix factorization models”. In: *Artificial Intelligence and Statistics*. 2012, pp. 1012–1018.

[224] Y. Shan et al. “PRC2 specifies ectoderm lineages and maintains pluripotency in primed but not naive ESCs”. In: *Nature Communications* 8.1 (2017).

[225] X. She et al. “Definition, conservation and epigenetics of housekeeping and tissue-enriched genes”. In: *BMC Genomics* 10.1 (2009), p. 269.

[226] S. A. Smallwood et al. “Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity”. In: *Nature Methods* 11 (July 2014).

[227] Z. D. Smith and A. Meissner. “DNA methylation: roles in mammalian development”. In: *Nature Reviews Genetics* 14 (Feb. 2013).

[228] S. Soderholm et al. “Multi-Omics Studies towards Novel Modulators of Influenza A Virus-Host Interaction”. en. In: *Viruses* 8.10 (2016).

[229] 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 (Feb. 2010), pdb.prot5384.

[230] J. C. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. Hoboken, N.J: J. Wiley, 2003.

[231] H. Specht et al. "Automated sample preparation for high-throughput single-cell proteomics". In: *bioRxiv* (2018), p. 399774.

[232] R. Spektor et al. "methyl-ATAC-seq measures DNA methylation at accessible chromatin". In: *bioRxiv* (Jan. 2018), p. 445486.

[233] 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.

[234] 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.

[235] 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.

[236] 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].

[237] M. Stoeckius et al. "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14 (July 2017).

[238] K. Struhl and E. Segal. "Determinants of nucleosome positioning". In: *Nature Structural &Amp; Molecular Biology* 20 (Mar. 2013).

[239] T. Stuart and R. Satija. "Integrative single-cell analysis". In: *Nature Reviews Genetics* (2019).

[240] J. H. Sul et al. "Accurate and Fast Multiple-Testing Correction in eQTL Studies". In: *The American Journal of Human Genetics* 96.6 (2015), pp. 857–868.

[241] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nature Protocols* 13 (Mar. 2018).

[242] V. Svensson et al. "Power analysis of single-cell RNA-sequencing experiments". In: *Nature Methods* 14 (Mar. 2017).

[243] 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.

[244] 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.

[245] 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.

[246] C. Thiele, K. Wunderling, and P. Leyendecker. "Multiplexed and single cell tracing of lipid metabolism". In: *Nature Methods* 16.11 (2019), pp. 1123–1130.

[247] R. E. Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489 (Sept. 2012).

[248] M. Tipping and C. Bishop. "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society* 61(3) (1999), pp. 611–22.

[249] 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.

[250] 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).

[251] 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.

[252] 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.

[253] C. Trapnell et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32 (Mar. 2014).

[254] O. Troyanskaya et al. "Missing value estimation methods for DNA microarrays". In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[255] M. Tsompana and M. J. Buck. "Chromatin accessibility: a window into the genome". In: *Epigenetics & Chromatin* 7.1 (2014), p. 33.

[256] 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).

[257] 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.

[258] N. L. Vastenhouw and A. F. Schier. "Bivalent histone modifications in early embryogenesis". In: *Current Opinion in Cell Biology* 24.3 (2012), pp. 374–386.

[259] N. L. Vastenhouw et al. "Chromatin signature of embryonic pluripotency is established during genome activation". In: *Nature* 464.7290 (2010).

[260] S. Virtanen et al. "Bayesian group factor analysis". In: *Artificial Intelligence and Statistics*. 2012, pp. 1269–1277.

[261] B. Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature Methods* 11 (Jan. 2014).

[262] C. Wang. "Variational Bayesian Approach to Canonical Correlation Analysis". In: *IEEE Trans Neural Netw* 3.18 (2007).

[263] Y. J. Wang et al. "Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues". In: *bioRxiv* (Jan. 2019), p. 541433.

[264] 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.

[265] 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.

[266] J. Wu et al. "The landscape of accessible chromatin in mammalian preimplantation embryos". In: *Nature* 534 (2016).

[267] Y. Xiang et al. "Epigenomic analysis of gastrulation identifies a unique chromatin state for primed pluripotency". In: *Nature Genetics* 52.1 (2020).

[268] 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].

[269] W.-S. Yong, F.-M. Hsu, and P.-Y. Chen. "Profiling genome-wide DNA methylation". In: *Epigenetics & Chromatin* 9.1 (2016), p. 26. ISSN: 1756-8935. DOI: 10.1186/s13072-016-0075-3.

[270] M. D. Zeiler. *ADADELTA: An Adaptive Learning Rate Method*. 2012. arXiv: 1212.5701 [cs.LG].

[271] 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).

[272] C. Zhang et al. "Advances in Variational Inference". In: *arXiv e-prints*, arXiv:1711.05597 (2017), arXiv:1711.05597. arXiv: 1711.05597.

[273] 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.

[274] 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.

[275] 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.

[276] J.-h. Zhao and P. L. Yu. "A note on variational Bayesian factor analysis". In: *Neural Networks* 22.7 (2009), pp. 988–997. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2008.11.002.

[277] S. Zhao et al. "Bayesian group factor analysis with structured sparsity". In: *Journal of Machine Learning Research* 17.196 (2016), pp. 1–47.

[278] Y. Zhao and B. A. Garcia. "Comprehensive Catalog of Currently Documented Histone Modifications". In: *Cold Spring Harbor Perspectives in Biology* 7.9 (Sept. 2015).

[279] G. X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8 (Jan. 2017).

[280] C. Ziegenhain et al. "Comparative Analysis of Single-Cell RNA Sequencing Methods". In: *Molecular Cell* 65.4 (2017), 631–643.e4.

[281] R. Zilionis et al. "Single-cell barcoding and sequencing using droplet microfluidics". In: *Nature Protocols* 12 (Dec. 2016).

[282] I. Zvetkova et al. "Global hypomethylation of the genome in XX embryonic stem cells". In: *Nature Genetics* 37 (Oct. 2005).