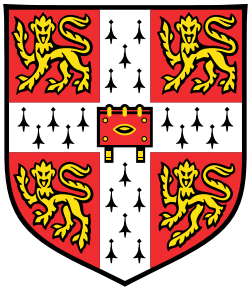
Statistical methods for the integrative analysis of single-cell multi-omics data



Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

April 2020

Chapter 1

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 [2]. 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.

1.1 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 [18]. 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**, 42].

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**, 42, 41]. 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.

1.1.1 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**, 39, 9, 36]. 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**, 28]. 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 [9]. By constructing a densely sampled reference data set, both works have let the ground for understanding transcriptomic variation during development.

1.1.2 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 [20].

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**, 22]. DNA methylation is globally removed and the chromatin attains its highest levels of accessibility [50]. Consistently, Hi-C experiments have

suggested a flexible chromatin landscape, with lack of topologically associating domains (TADs) or chromatin compartments [19, 12, 43], 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 [15, 43]. 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 [55]. 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 [6, 43].

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**, 53, 11, 37]. 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 [4].
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 [5, 31]. 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 [47].

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 [16]. H3K4me3 is thought to facilitate transcription by inducing a more efficient assembly of the transcriptional machinery [4, 48]. 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 [4, 34].
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 [8, 4]. Interestingly, inactivating PRC2 components in mouse embryos does not affect pre-implantation development, but the embryos become unviable after gastrulation[40]. This suggests that H3K27me3 has a critical role in regulating gene expression during cell fate commitment after germ layer specification.

Gastrulation: germ layer specification

The post-implantation blastocyst is 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 [54]. 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 [54, 23]. 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.[11].

I envision that the recent development of single-cell multi-modal technologies (described in **??**), 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 [13, 49, 26] and early post-implantation development [37], but gastrulation has remained elusive.

1.2 Results

1.2.1 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:

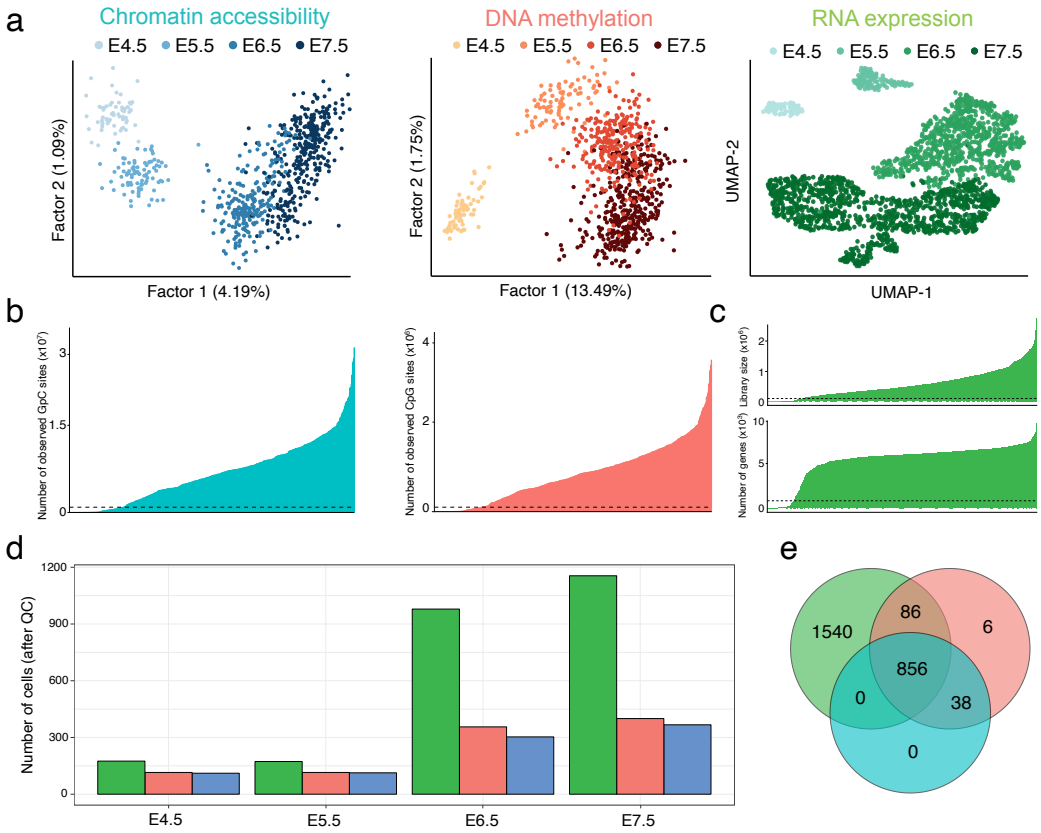


Figure 1.1: scNMT-seq gastrulation atlas. Data set overview.
(a) Dimensionality reduction for chromatin accessibility data (left, in blue), DNA methylation (middle, in red) and RNA expression (right, in green). For the gene expression data we applied UMAP[27]. For chromatin accessibility and DNA methylation data we applied Bayesian Factor Analysis[3].
(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).

1.2.2 Cell type assignment using the RNA expression data

To define cell type annotations we followed two independent strategies. For the E6.5 and E7.5 stages, we mapped the RNA expression profiles to the single-cell gastrulation atlas [33] (stages E6.5 to E8.0) using a matching mutual nearest neighbours algorithm [14]. 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 [21], as no transcriptomic atlas was available for these stages:

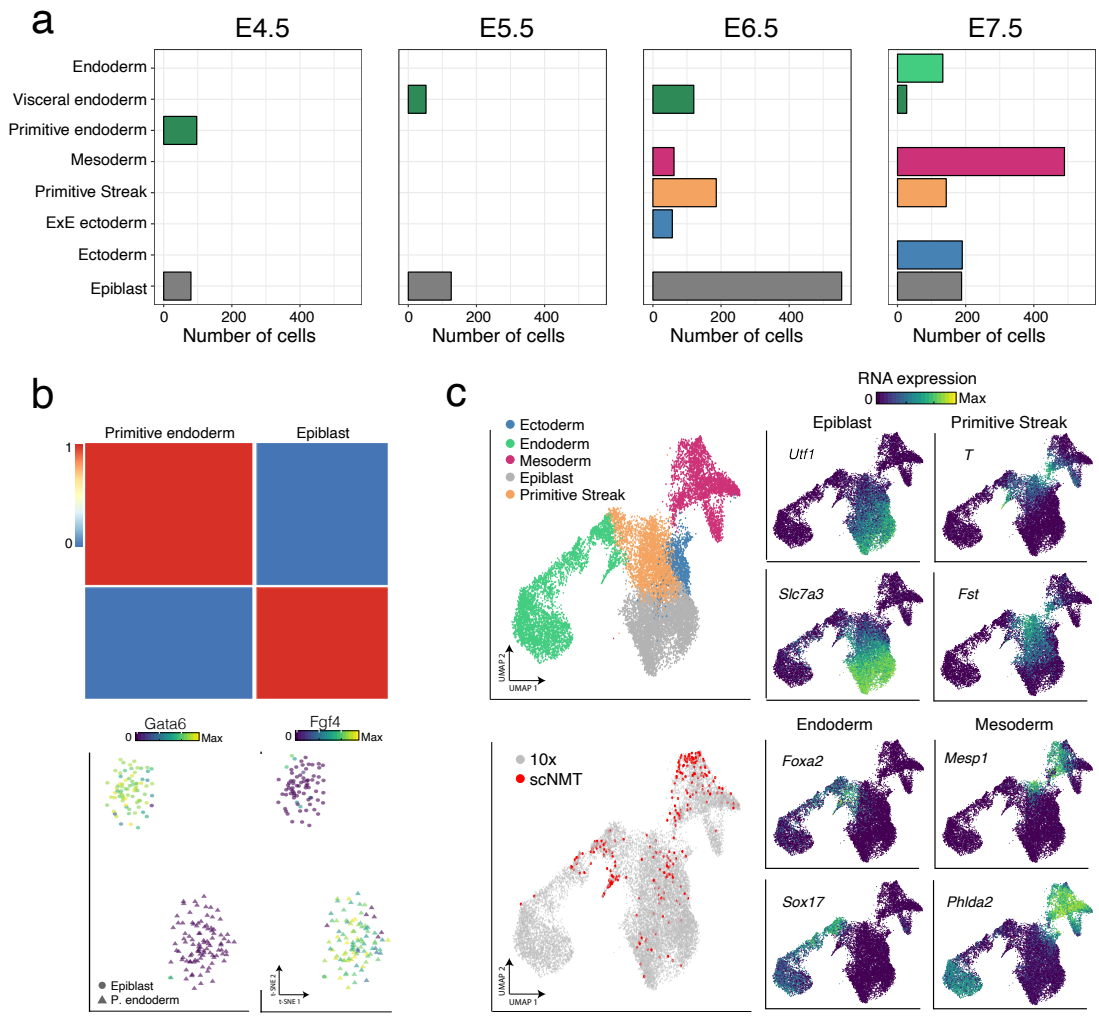


Figure 1.2: Cell type assignments using the RNA expression data.
(a) For each stage, the bar plots display the number of cells assigned to each lineage.
(b) Cell type assignment for E4.5 cells. The heatmap displays the consensus plot, representing the similarity between cells based on the averaging of clustering results from multiple combinations of clustering parameters[21]. 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 neighbours that were used to transfer labels to the scNMT-seq data set. The right plots display the RNA expression levels of marker genes for different cell types.

1.2.3 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 1.1). When fitting a separate model for stages E4.5, E5.5 and E6.5, the largest source of variation (Factor 1) separates cells by embryonic versus extra-embryonic origin, as expected (Figure 1.3). At E7.5 extra-embryonic cells were manually removed during the dissection and the first two latent factors discriminate the three germ layers (Figure 1.3).

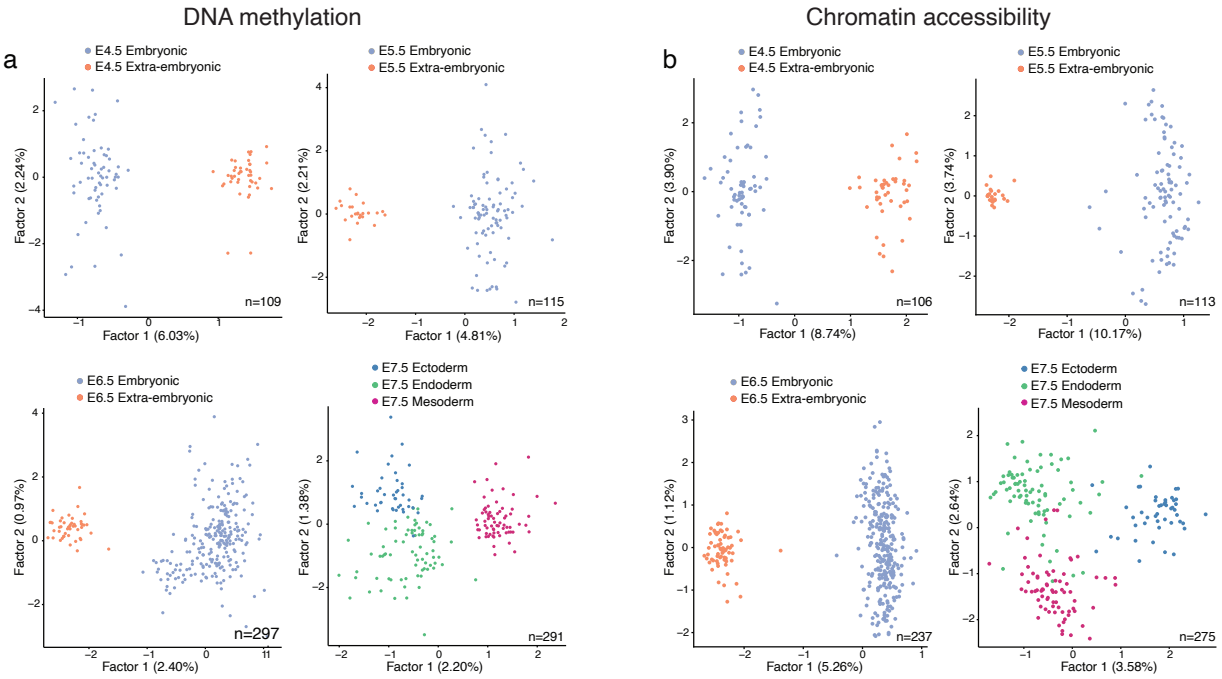


Figure 1.3: Dimensionality reduction of (a) DNA methylation and (b) chromatin accessibility data. Shown are scatter plots of the first two latent factors (sorted by variance explained) for models trained with cells from the indicated stages. From E4.5 to E6.5 cells are coloured by embryonic and extra-embryonic origin. At E7.5, cells are coloured by the primary germ layer.

1.2.4 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 Figure 1.4, mainly driven by a *de novo* methylation wave from E4.5 to E5.5 that preferentially targets CpG-poor genomic loci [5, 53] (Figure 1.4).
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 extra-embryonic tissues (t-test, Figure 1.5). Consistent with the DNA methylation changes, CpG-rich regions remain more accessible than CpG-poor regions of the genome.

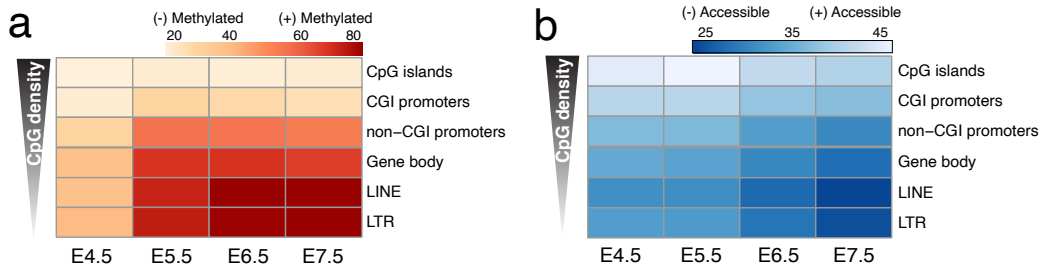


Figure 1.4: DNA methylation and chromatin accessibility levels per stage and genomic context.
Heatmaps display the mean levels across cells within a particular stage and across all loci within a particular genomic context.

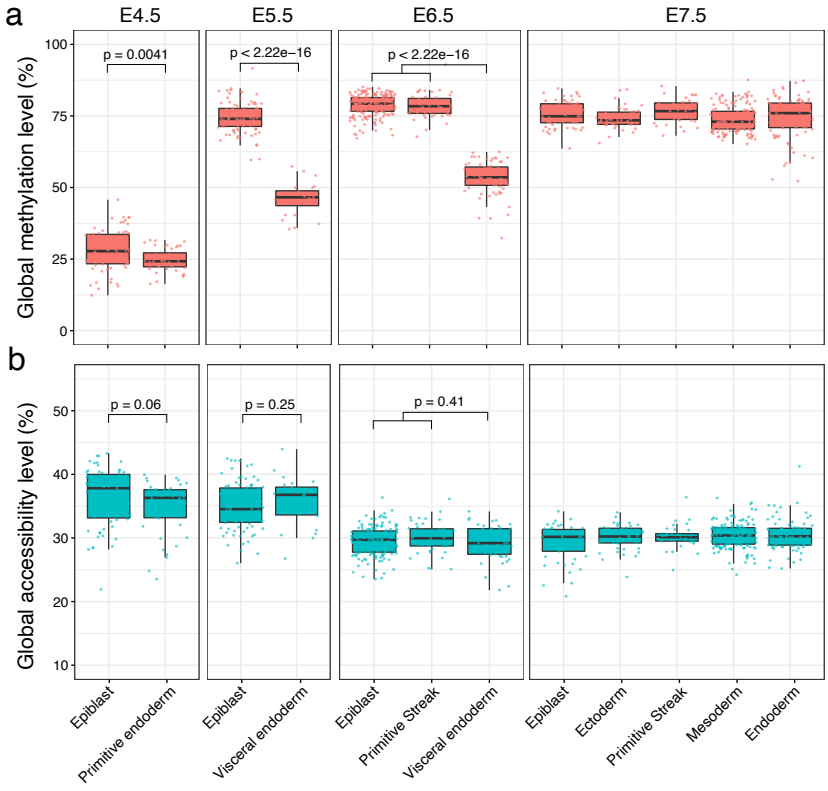


Figure 1.5: Global DNA methylation and chromatin accessibility levels per stage and lineage.
Box plots showing the distribution of genome-wide (a) CpG methylation levels or (b) GpC accessibility levels per stage and lineage. Each dot represents a single cell.

Next, we attempted to characterise the relationship between the transcriptome and the epigenome along differentiation. For simplicity we focused on gene promoters (defined as 2kb up and downstream from the transcription start site), as RNA expression and epigenetic readouts can be matched unambiguously. We calculated, for each gene, the correlation coefficient between RNA expression and the corresponding DNA methylation or chromatin accessibility levels. As a filtering criterion, we required, a minimum number of 1 CpG (methylation) or 3 GpC (accessibility) measurements in at least 50 cells for each genomic feature. In addition, we restricted the analysis to the top 5,000 most variable genes, according to the rationale of independent filtering [7].

We identified 125 genes whose expression shows significant correlation with promoter DNA methylation and 52 that show a significant correlation with chromatin accessibility Figure 1.6. Among the top hits we identified early pluripotency and germ cell markers, including *Dppa4*, *Dppa5a*, *Rex1*, *Tex19.1* and *Pou3f1* (Figure 1.6). Notably, all of them have a negative association between RNA expression and DNA methylation and a positive association between RNA expression and chromatin accessibility. Inspection of the transcriptomic and epigenetic dynamics reveals that the repression of these early pluripotency 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.

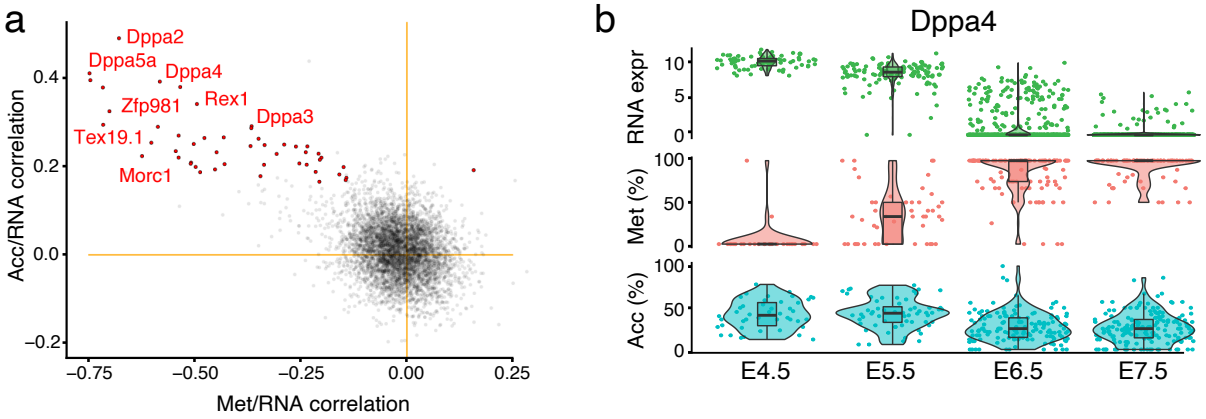


Figure 1.6: Genome-wide association analysis between RNA expression and the corresponding epigenetic status in gene promoters.
(a) Scatter plot of Pearson correlation coefficients between promoter DNA methylation versus RNA expression (x-axis); and promoter accessibility versus RNA expression (y-axis). Significant associations for both correlation 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.

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

In the previous section we demonstrated that exit from pluripotency is concomitant with the establishment of a repressive epigenetic landscape that is characterised by increasing levels of DNA methylation and decreasing levels of chromatin accessibility.

Next, we sought to investigate the coordinated changes between RNA expression and epigenetic status that define germ layer commitment. Instead of following a supervised approach, 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 1.7). 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. 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.

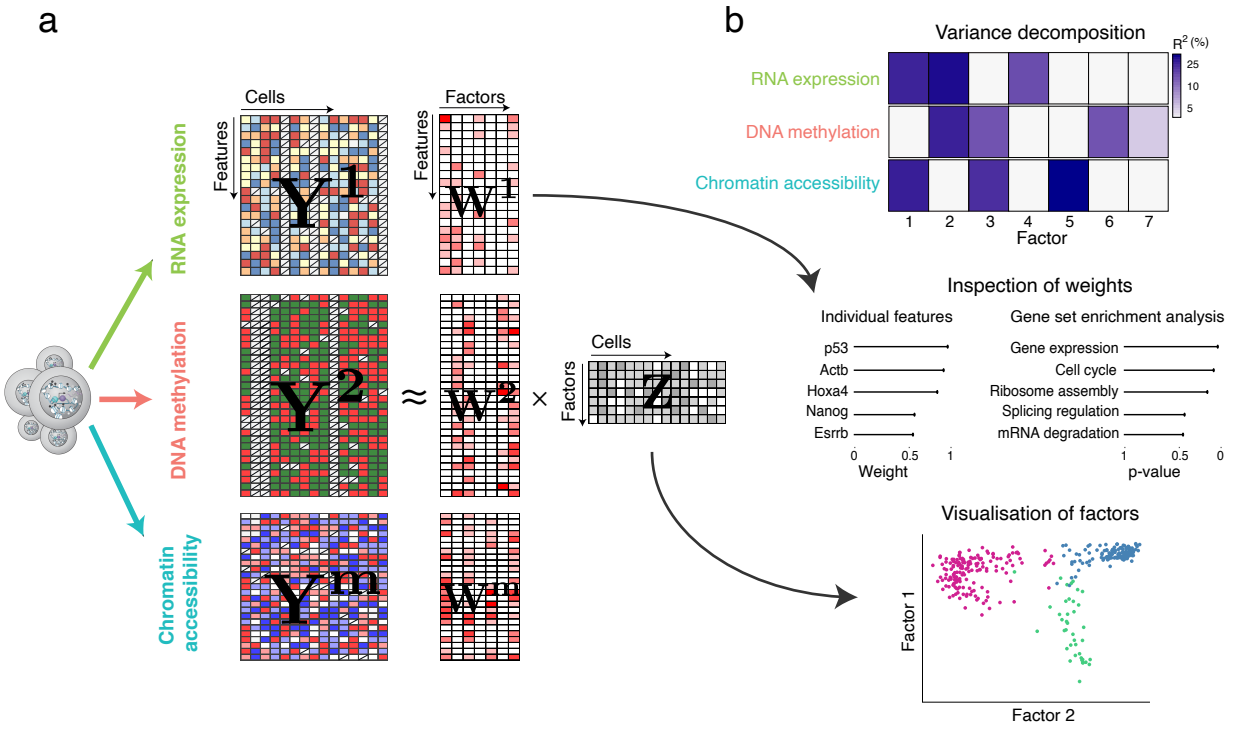


Figure 1.7: Multi-Omics Factor Analysis (MOFA): model overview and illustration of downstream analysis.
(a) Model overview: MOFA takes as input one or more data modalities (**Y**), extracted from the same cells. MOFA decomposes these matrices into a matrix of factors (Z) and a set of feature weight matrices (**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 weight matrices relate the low-dimensional space to the high-dimensional space by inferring a weight for each feature on each factor.
(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 cells in the factor space.

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

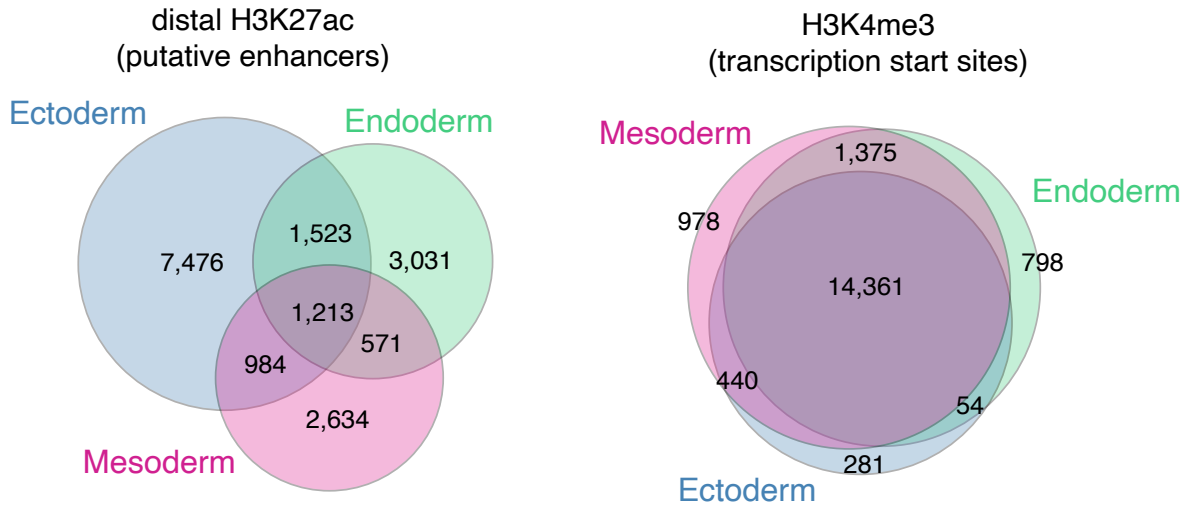


Figure 1.8: Venn diagrams showing overlap of peak calls for each lineage-specific histone mark, for distal H3K27ac (left) and all H3K4me3 (right). The figure shows that distal H3K27ac peaks (putative enhancer [10]) have moderate levels of overlap between the three germ layers. In contrast, H3K4me3 peaks (active transcription start sites [25]) are similar between the three germ layers.

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

MOFA model overview

MOFA identified 6 Factors capturing at least 1% of variance in the RNA expression data (Figure 1.9). The first two Factors (sorted by variance explained) captured the the emergence of the three germ layers, indicating that germ layer commitment is the largest source of variation across all molecular layers at E7.5. Notably, for these two Factors, MOFA links the variation at the gene expression level to concerted DNA methylation and chromatin accessibility changes at lineage-specific enhancer marks. Surprisingly, these two Factors capture very small amounts of the variation in DNA methylation and chromatin accessibility at promoters. This suggests that epigenetic changes in promoters may not be linked to germ layer commitment, with distal regulatory elements (i.e. enhancers) playing a more prominent role. Yet, we cannot rule out important variation in other epigenetic layers such as histone marks or chromatin conformation.

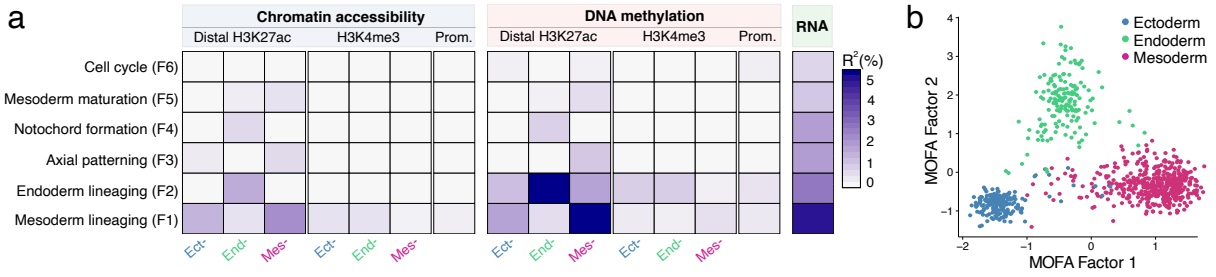


Figure 1.9: MOFA reveals coordinated epigenetic and transcriptomic variation at enhancer elements during germ layer commitment.
(a) Percentage of variance explained by each MOFA factor (rows) across data modalities (columns). Considered data modalities were RNA expression quantified over protein-coding genes (green); DNA methylation (red) and chromatin accessibility (blue) quantified on promoters, lineage-specific H3K4me3-marked sites and distal H3K27ac-marked sites (putative 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.

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

1.2.6 Differential DNA methylation and chromatin accessibility analysis

The MOFA analysis in the previous section reveals interesting genome-wide trends. We next attempted to pinpoint individual genomic elements that could be representative of the global patterns. This could be done by inspecting the feature weights in the MOFA model, but given that we can accurately classify cells into the three (discrete) germ layers, here we decided to adopt a more intuitive supervised approach. For each genomic element (with sufficient coverage), we calculated differential DNA methylation and chromatin accessibility between each germ layer versus the other two using a Fisher exact test for binomial proportions (Figures 1.10 and 1.11).

In general we observe that, consistent with the MOFA results, only enhancers display substantial amounts of epigenetic variation between the germ layers (Figure 1.10). As expected, endoderm enhancers seem to be more associated with endoderm commitment (more open and unmethylated in the endoderm cells) whereas mesoderm enhancers are more associated with mesoderm commitment (again, more open and unmethylated in the mesoderm cells). Notably, for both endoderm and mesoderm commitment events, the effect sizes associated with regions that display differential demethylation and chromatin accessibility are moderate (less than ≈30% change in levels, Figure 1.11) but coordinated across multiple enhancers (between ≈10% and ≈25% of the distal H3K27ac peaks, Figure 1.10).

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

below.

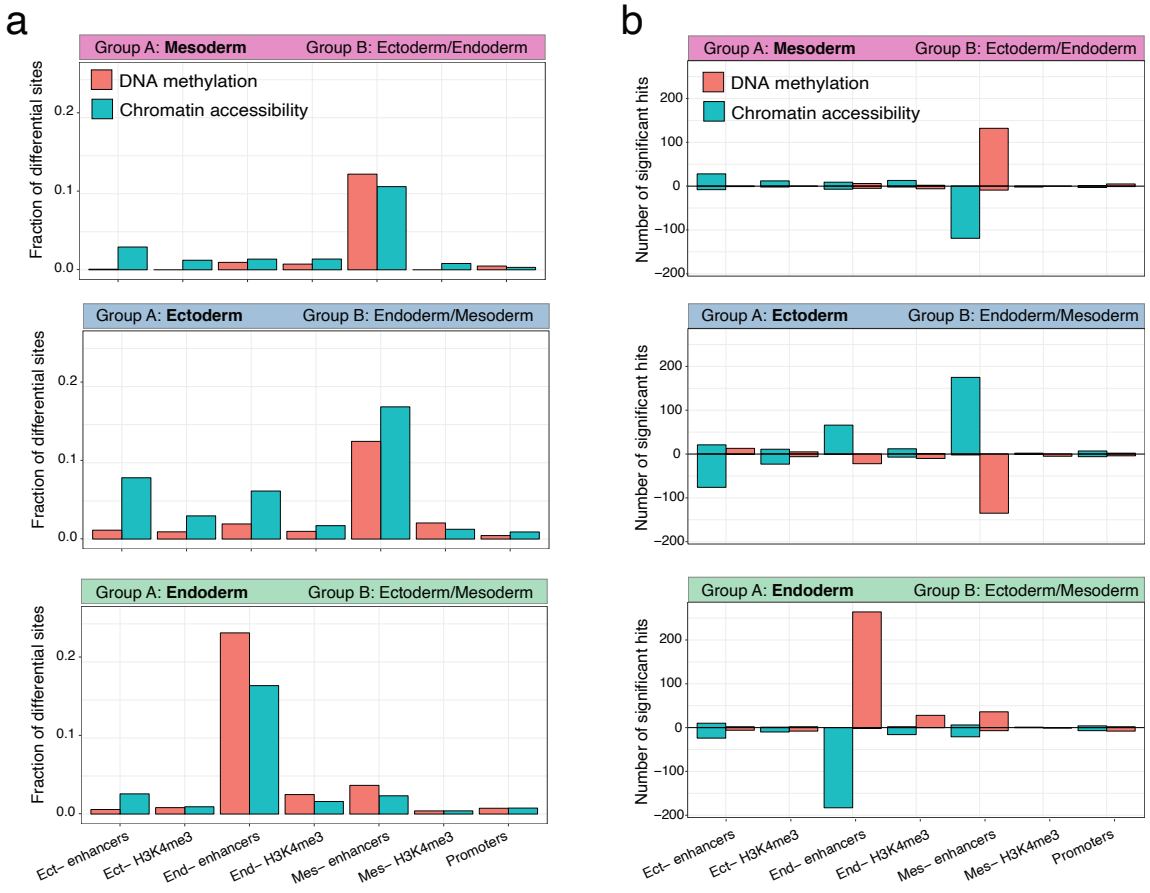


Figure 1.10: Differential DNA methylation and chromatin accessibility analysis between germ layers at E7.5.
Bar plots display (a) the fraction and (b) the total number of differentially methylated (red) or accessible (blue) loci (FDR<10%, Fisher exact test for binomial proportions, y-axis) per genomic context (x-axis). Each panel corresponds to the comparison of cells from one germ layer (group A) against cells comprising the other two germ layers present at E7.5 (Group B). For (b), positive values indicate increase in DNA methylation or chromatin accessibility in group A, whereas negative values indicate decrease in DNA methylation or chromatin accessibility.

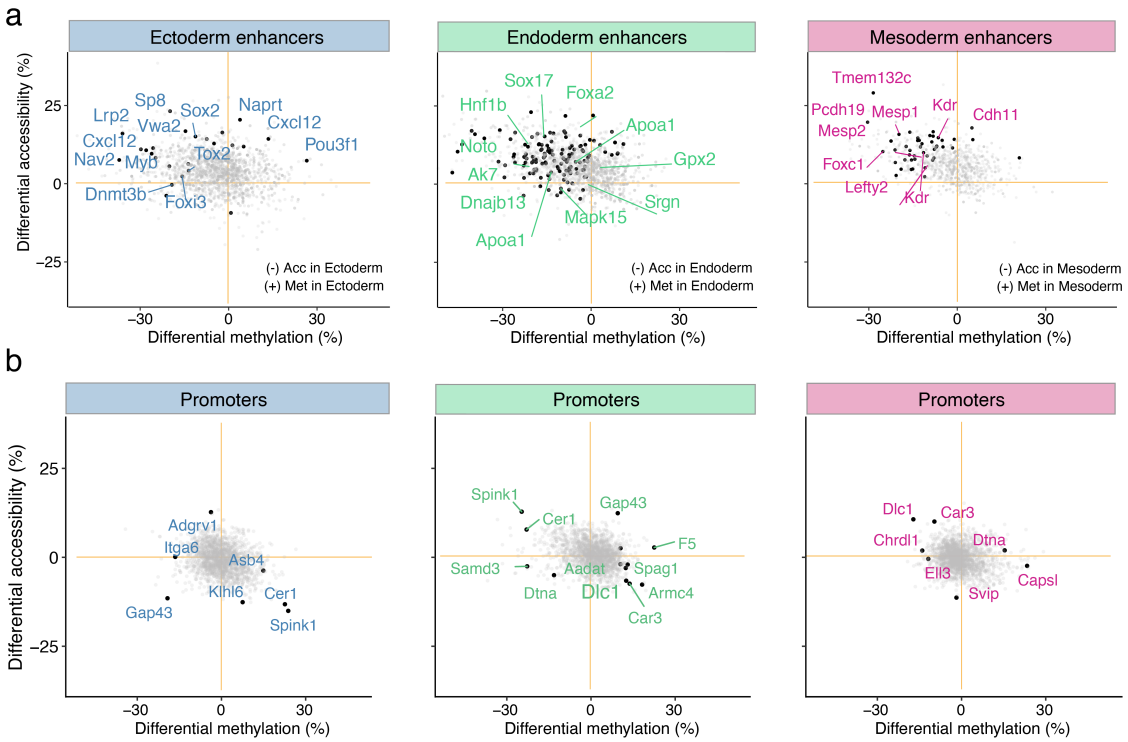


Figure 1.11: Differential DNA methylation and chromatin accessibility between germ layers at lineage-defining enhancers and promoters.
Scatter plots display differential DNA methylation (%, x-axis) and chromatin accessibility (%, y-axis) at (a) lineage-defining enhancers and (b) promoters. Comparisons are ectoderm versus non-ectoderm cells (left), endoderm versus non-endoderm cells (middle) and mesoderm versus non-mesoderm cells (right). Black dots depict gene-enhancer or gene-promoter pairs with significant changes in RNA expression and DNA methylation or chromatin accessibility (FDR<10%). Enhancers were associated with genes by a maximum genomic distance of 25kb. As a filtering criteria, we required 1 CpG (methylation) and 5 GpC (accessibility) observations in at least 10 cells per group Non-variable regions were filtered out prior to differential testing.

Characterisation of individual enhancers

The results above suggest that the establishment of lineage-specific epigenetic profiles results from the coordinated action of multiple elements located all across the genome, and hence the identification of individual putative regulatory elements is not trivial and probably requires a much larger data set than the one we profiled.

Nevertheless, when linking enhancers to genes by a maximum genomic distance of 25kb we identified some interesting gene-enhancer associations linked to key germ layer markers including *Snai1* and *Mesp2* for mesoderm, *Bmp2* and *Hnf1b* for endoderm, *Bcl11a* and *Sp8* for ectoderm (Figure 1.11). Their temporal dynamics for the three molecular layers are shown below:

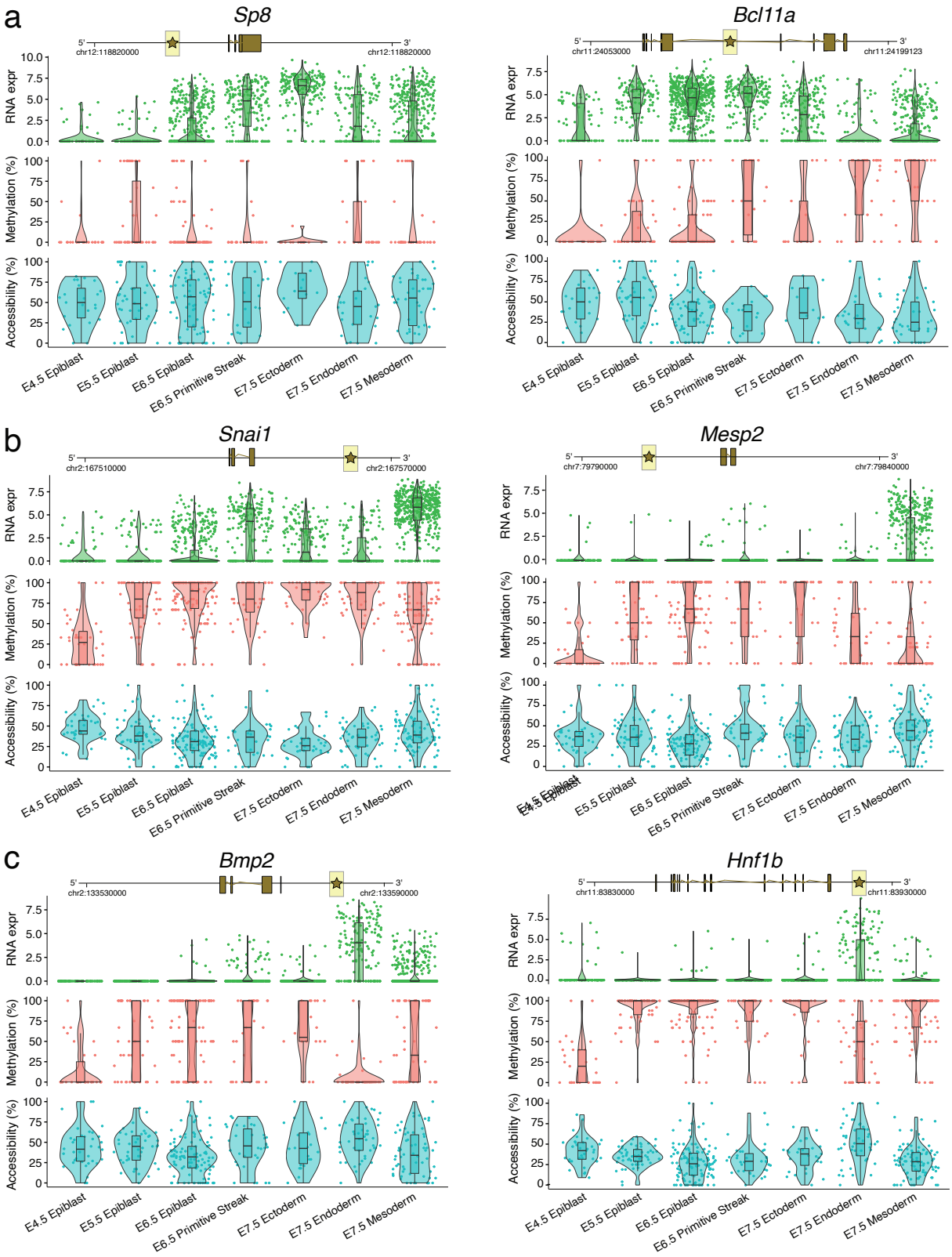


Figure 1.12: Illustrative examples of putative epigenetic regulation in enhancer elements during germ layer commitment.
Box and violin plots show the distribution of RNA expression (log normalised counts, green), DNA methylation (%, red) and chromatin accessibility (%, blue) levels per stage and lineage. Each dot corresponds to a cell. The enhancer region that is used to quantify DNA methylation and chromatin accessibility levels is represented with a star and highlighted in yellow in the genomic track above. Genes were linked to enhancers by overlapping genomic coordinates with a maximum distance of 50kb.

1.2.7 Transcription factor motif enrichment analysis

To identify transcription factors (TFs) that could drive the epigenetic variation in lineage-defining enhancers during germ layer commitment, we integrated the chromatin accessibility and RNA information as follows. For every TF with an associated motif in the Jaspar core 95 vertebrates data base we extracted its position-specific weight matrix and we tested for enrichment in differentially accessible distal H3K27ac sites using a background of all distal H3K27ac sites. To assess statistical significance we used a Fisher exact test, as implemented in the *meme suite* (v4.10.1). This information was then integrated with differential RNA expression between germ layers for the same TFs, quantified using the gene-wise negative binomial generalised linear model with quasi-likelihood test from edgeR. Reassuringly, 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 1.13).
Although this analysis serves as a good quality control for our results, it is important to keep in mind that using sequence information is only a proxy for true TF binding, and some essential TFs to not target specific motifs, including EOMES or T [45].

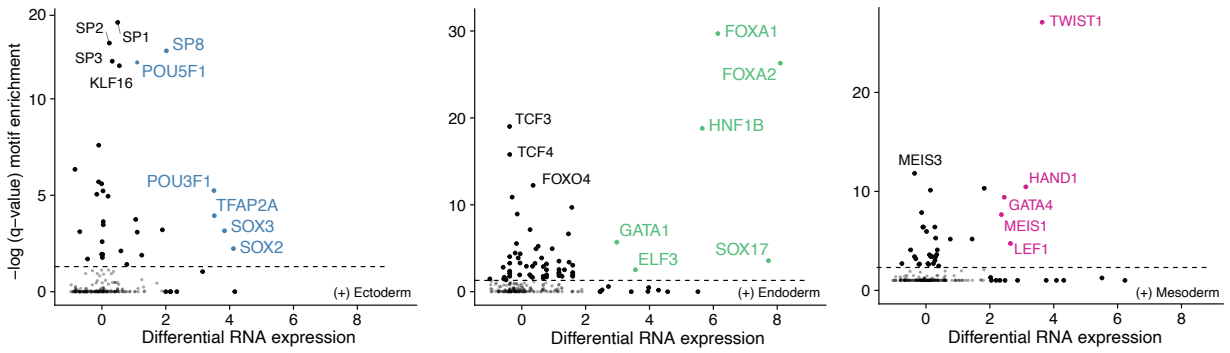


Figure 1.13: Transcription Factor motif enrichment analysis at lineage-defining distal H3K27ac sites. Shown is motif enrichment (-log10 q-value, y-axis) plotted against differential RNA expression (log fold change, x-axis) of the corresponding TF. The analysis is performed separately for each set of lineage-defining enhancers: ectoderm (left), endoderm (middle) and mesoderm (right). TFs with significant motif enrichment (FDR<1%) and differential RNA expression (FDR<1% and log-fold change higher than 2) are coloured and labelled.

1.2.8 Time resolution of the enhancer epigenome

In the previous section we have shown that distal regions marked with H3K27ac (i.e. putative enhancers) are the elements that drive or respond to germ layer specification at E7.5.
Next, we sought to explore how these epigenetic patterns are established. We visualised DNA methylation and chromatin accessibility levels at lineage-defining enhancers from E4.5 to E7.5 (Figure 1.14). Importantly, to interpret the visualisation, DNA methylation and chromatin accessibility values should be compared to the genome-wide background levels that are displayed as dashed lines.

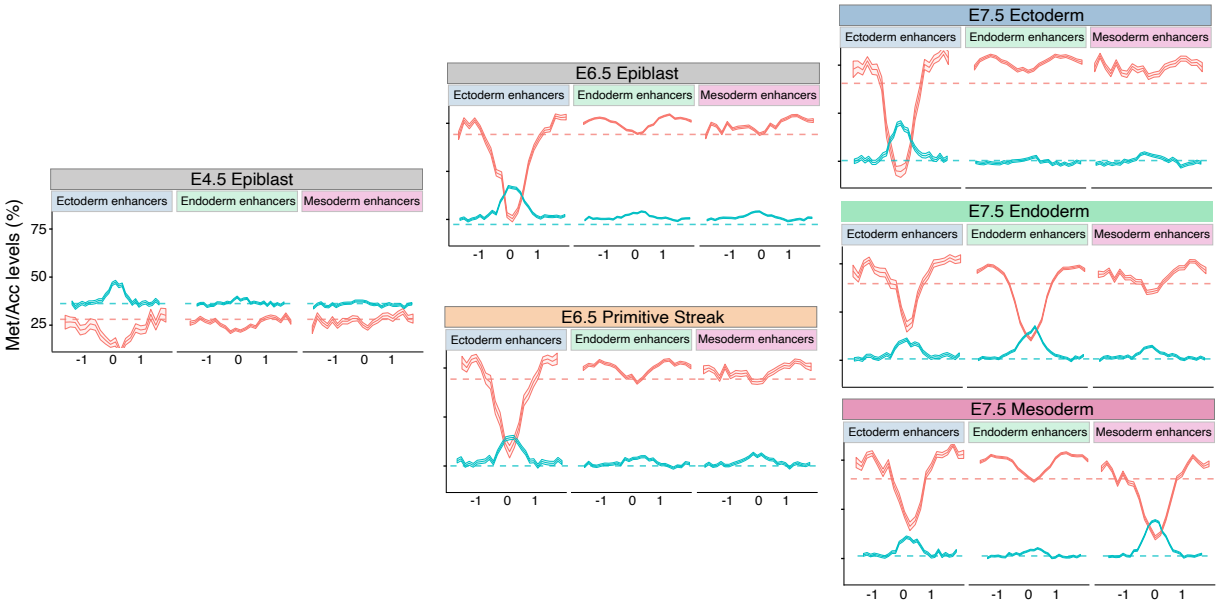


Figure 1.14: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers. Visualisation at pseudobulk resolution.
DNA methylation (red) and chromatin accessibility (blue) levels at lineage-defining enhancers quantified over different lineages across development. Shown are running 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 1.15).

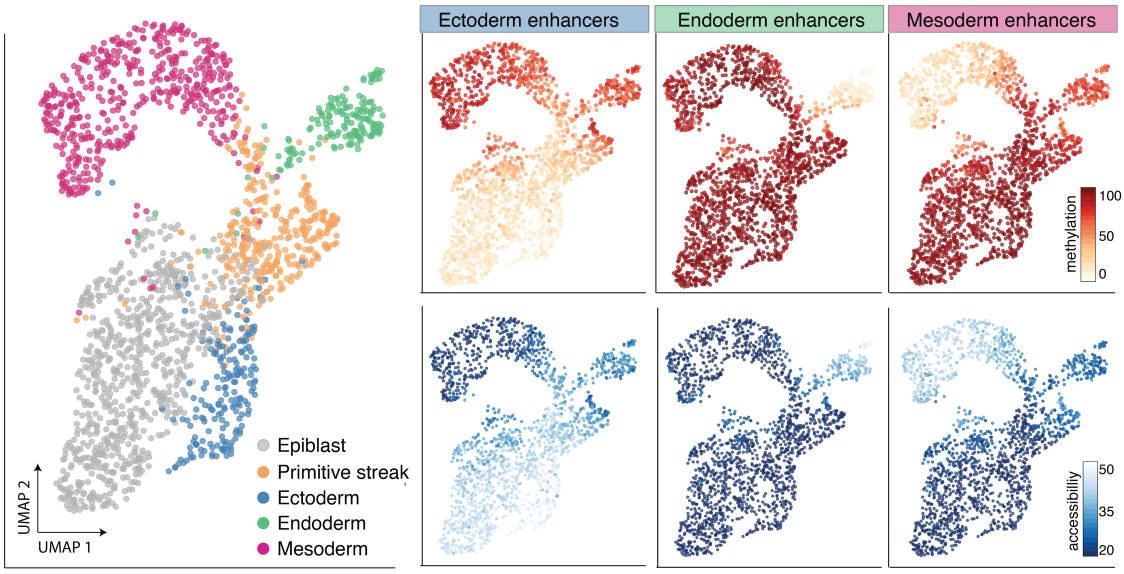


Figure 1.15: DNA methylation and chromatin accessibility dynamics at lineage-defining enhancers. Visualisation at single-cell resolution.
UMAP projection based on the MOFA factors inferred using all cells. In the left plot the cells are coloured according to their lineage. In the right plots cells are 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.

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 1.14 and Figure 1.15).
Upon germ layer specification at E7.5, mesoderm and endoderm enhancers undergo concerted demethylation from ≈ 80% to ≈50% in a lineage-specific manner (i.e. mesoderm enhancers demethylate in mesoderm cells, whereas endoderm enhancers demethylate in endoderm cells). Consistently, chromatin accessibility sharply increases from ≈ 25% to ≈45% upon lineage specification.

Ectoderm enhancers are primed in the early epiblast

In striking contrast to the mesoderm and endoderm enhancers, the ectoderm enhancers are open and demethylated as early as the E4.5 epiblast. Interestingly, the ectoderm cells share the same epigenetic profile (in enhancer elements) as the epiblast, characterised by demethylated and open ectoderm enhancers; and methylated and closed mesoderm and endoderm enhancers (Figure 1.14

and Figure 1.15). 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[52]
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 1.16). This result supports the first hypothesis, but does not rule out the second hypothesis.

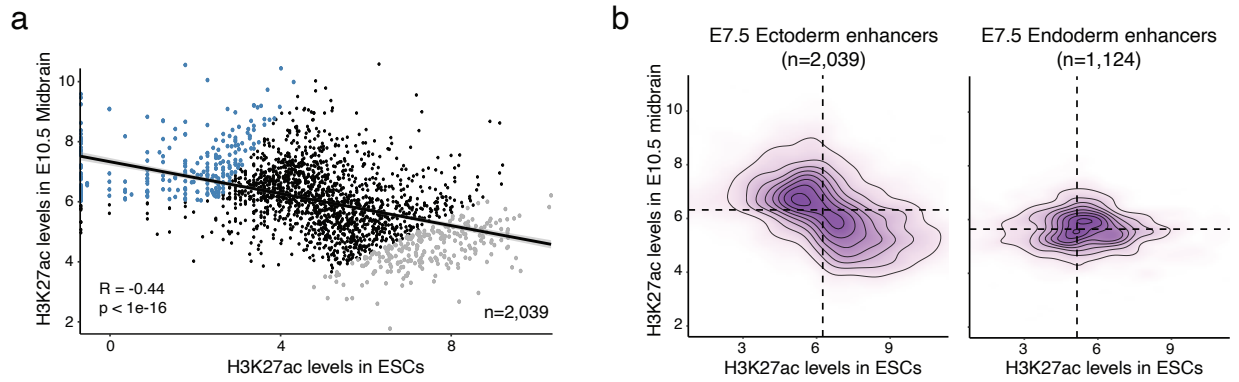


Figure 1.16: E7.5 ectoderm enhancers contain a mixture of pluripotency and neural signatures.
(a) Scatter plot of ectoderm enhancers' H3K27ac levels quantified in ESCs (pluripotency enhancers, x-axis) and E10.5 midbrain (neuroectoderm enhancers, y-axis). Each dot corresponds to an ectoderm enhancer (Figure 1.8). 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 1.16). Additionally, we also considered endoderm enhancers as a negative control.
For each class of enhancers, we quantified and visualised the DNA methylation and chromatin

accessibility dynamics along the epiblast-ectoderm trajectory Figure 1.17). We plotted absolute levels in (a) and normalised levels to the genome-wide background in (b). We remind the reader that to interpret the plot below, it is critical to compare the absolute levels to the genome-wide background levels.

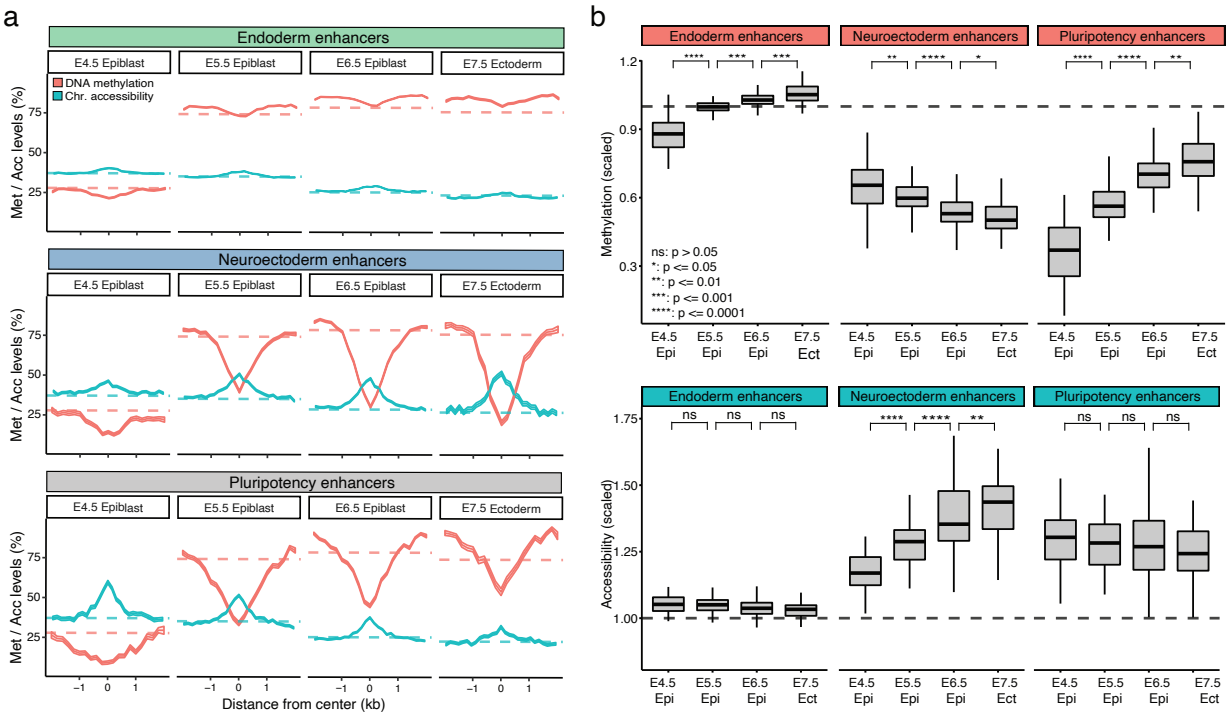


Figure 1.17: Pluripotency and neurectoderm enhancers display different DNA methylation and chromatin accessibility dynamics.
(a) Profiles of DNA methylation (red) and chromatin accessibility (blue) quantified along the epiblast-ectoderm trajectory. Each panel corresponds to a different genomic context. Profiles are quantified using running averages of 50-bp windows around the centre of the ChIP-seq peak for a total of 2 kb upstream and downstream. Solid lines display the mean across cells and shading displays the corresponding standard deviation. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue).
(b) Box plots of DNA methylation (top) and chromatin accessibility (bottom) levels quantified along the epiblast-ectoderm trajectory). Levels are scaled to the genome-wide background for each stage.

The three types of enhancers display very different epigenetic dynamics:

- Endoderm enhancers simply follow the genome-wide repressive dynamics, driven by a global increase in DNA methylation and a decrease in chromatin accessibility. Consistently, the relative levels for both measurements are close to ≈ 1.
- Pluripotency enhancers display an increase in DNA methylation from $\approx 15\%$ at E4.5 to $\approx 60\%$ at E7.5 and a decrease in chromatin accessibility from $\approx 50\%$ at E4.5 to $\approx 35\%$ at E7.5. This is similar to our previous result on the promoters dynamics of pluripotency genes (Figure 1.6). 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* [29, 17].

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

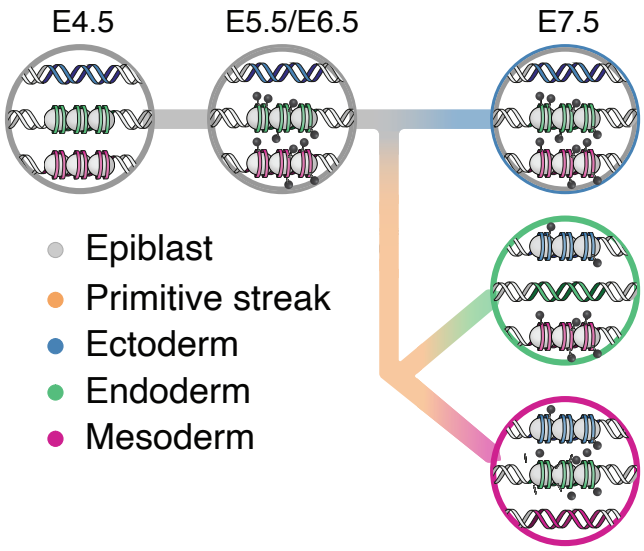


Figure 1.18: Schematic illustration of the hierarchical model for the epigenetic dynamics of germ layer commitment. Illustration designed by Veronique Juvin from SciArtWork.

1.2.9 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*[46]. 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 1.19). 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 1.19).
We find that, as cells begin to display a primitive streak phenotype, ectoderm-defining enhancers progressively decrease in accessibility and gain methylation, a process that continues as cells differentiate into the mesoderm and endoderm. In contrast, mesoderm and endoderm-defining enhancers simultaneously become hypomethylated and accessible only after commitment to these cell fates. In both cases, changes in DNA methylation and chromatin accessibility co-occur, suggesting a tight regulation of the two epigenetic layers.

In conclusion, we observe a sequential process where the inactivation of ectoderm enhancers precedes the activation of the mesendoderm enhancers. Interestingly, this resembles reprogramming of induced pluripotent stem cells, where the differentiated programme is repressed prior to the activation of the pluripotency programme[32]

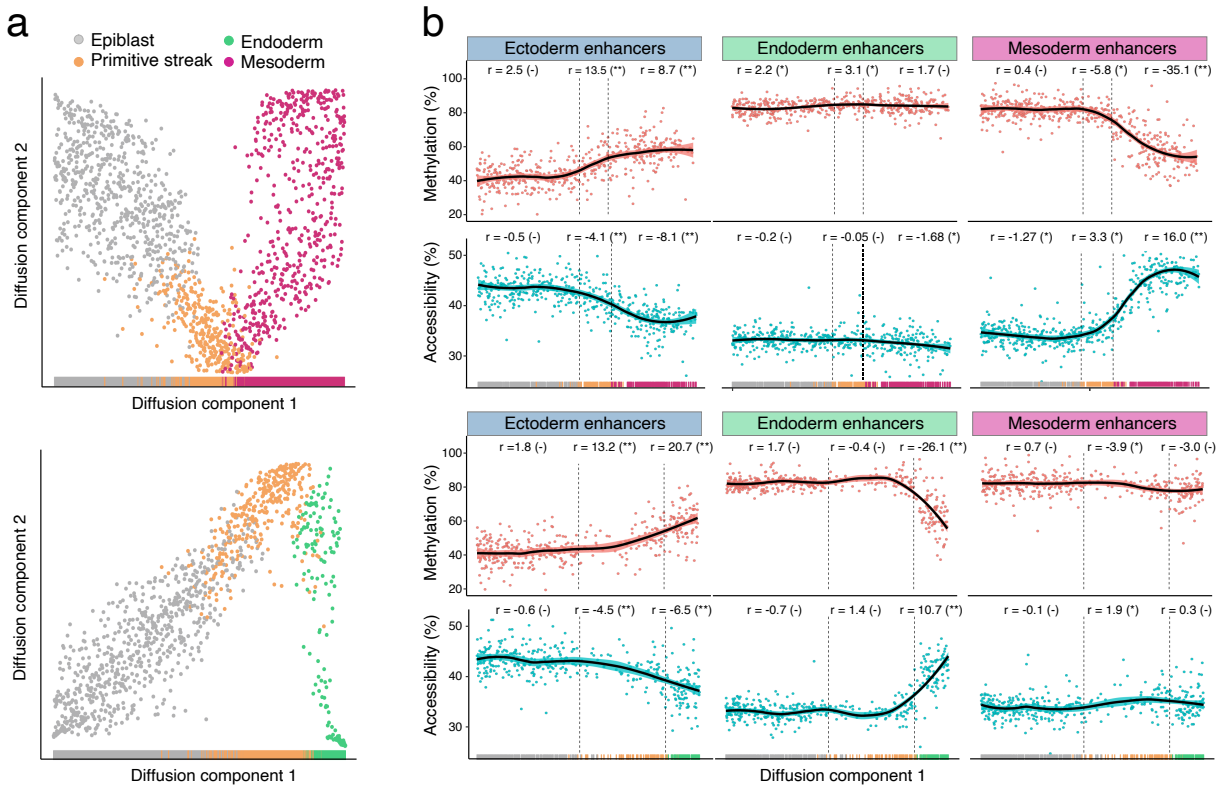


Figure 1.19: Silencing of ectoderm enhancers precedes activation of mesoderm and endoderm enhancers.
(a) Reconstructed mesoderm (top) and endoderm (bottom) commitment trajectories using a diffusion pseudotime method applied to the RNA expression data. Shown are scatter plots of the first two diffusion components, with cells coloured according to their lineage assignment. 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.

1.2.10 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 [35]. 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 [11, 38, 35, 24].

In our study, to test whether TET enzymes drive the lineage-specific demethylation events, we used an *in vitro* system where embryoid bodies were differentiated in serum conditions using both wild type (WT) mouse ESCs and cells that were deficient for all three TET enzymes (*Tet TKO*). The embryoid bodies were dissociated and subjected to scNMT-seq at days 2, 4-5, and 6-7 following the onset of differentiation.

Cell type assignment using the RNA expression

As in Figure 1.2, cell types were assigned by mapping the RNA expression profiles to the *in vivo* gastrulation atlas using a mutual nearest neighbours matching algorithm [14].
Notably, the WT cells from the EB differentiation protocol recapitulate the *in vivo* dynamics with remarkably accuracy (Figure 1.20). At day 2, most cells are in the pluripotent epiblast stage, which roughly corresponds to embryonic stages E4.5 to E5.5. At days 4-5, EBs begin the formation of primitive streak cells, as in embryonic stages E6.5 to E7.0. At days 6-7 of differentiation the primitive streak cells eventually commit to mesoderm (mostly) or endoderm fate, as in embryonic stages E7.0 to E8.0. In addition, at days 6-7 we observe the emergence of mature mesoderm structures including hematopoietic cell types.

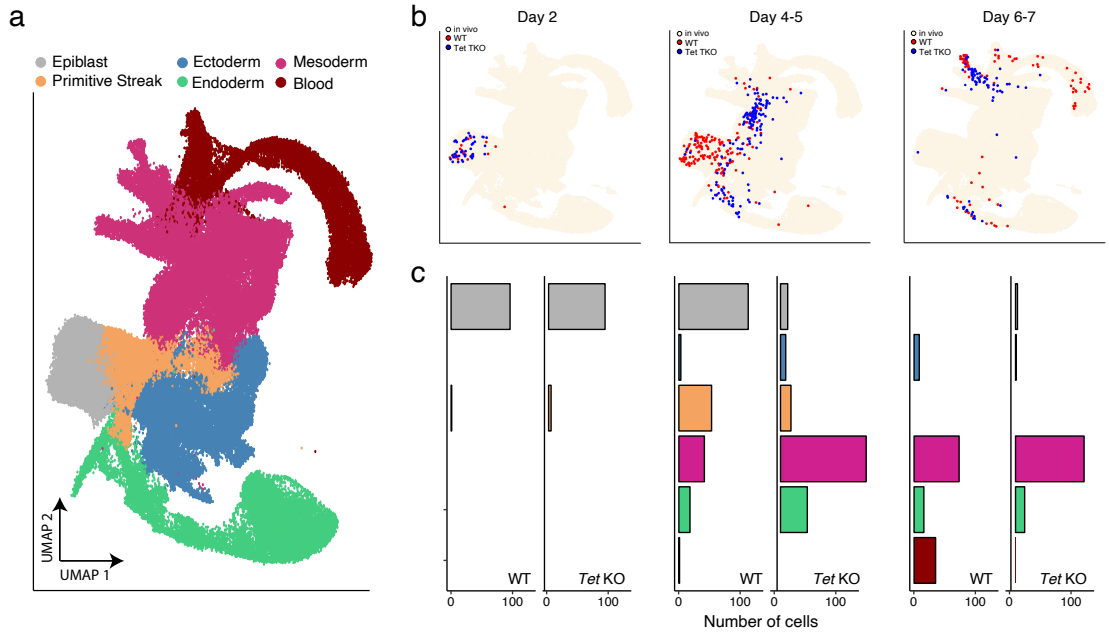


Figure 1.20: Cell type assignment for the Embryoid Body differentiation experiment.
(a) UMAP projection of the 10x atlas data set (stages E6.5 to E8.5, no extra-embryonic cells), where cells are coloured by lineage assignment.
(b) Same UMAP projection as in (a), but in this case, for each day of EB differentiation, cells are coloured by the the nearest neighbours that were used to assign cell type labels to the query cells. Cells from a WT genotype are shown in red and cells from a *Tet* TKO genotype are shown in blue.
(c) Bar plots display the cell type numbers for each day of EB differentiation, grouped by WT or *Tet* TKO genotype.

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

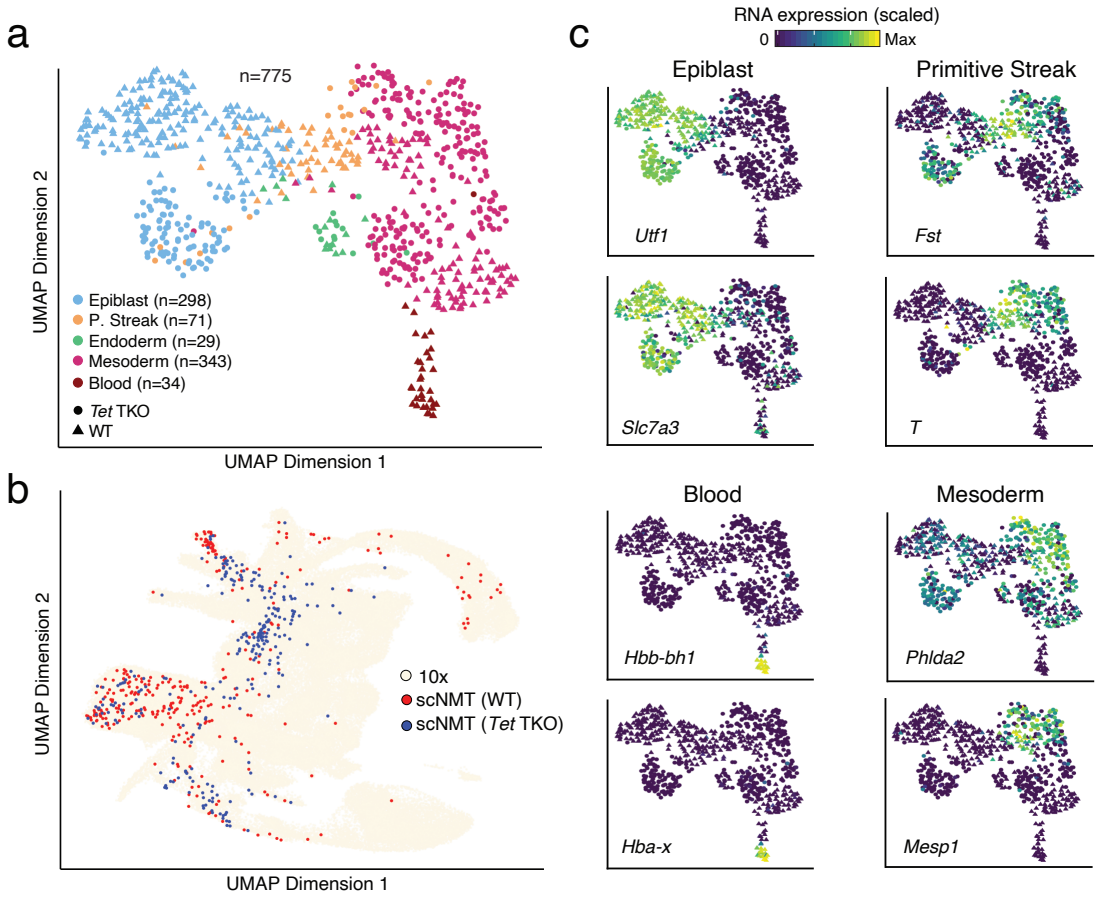


Figure 1.21: Embryoid bodies recapitulate the transcriptional heterogeneity of the mouse embryo.
(a) UMAP projection for the embryoid body dataset, where cells are coloured by lineage assignment and shaped by genotype (WT or *Tet* TKO).
(b) UMAP projection of the atlas data set (stages E6.5 to E8.5, no extra-embryonic cells). Cells coloured correspond to the nearest neighbours that were used to assign cell type labels to the EB dataset, red for WT and blue for *Tet* TKO.
(c) UMAP projection of embryoid body cells, as in (a), coloured by the relative RNA expression of marker genes.

Validation of epigenetic measurements

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

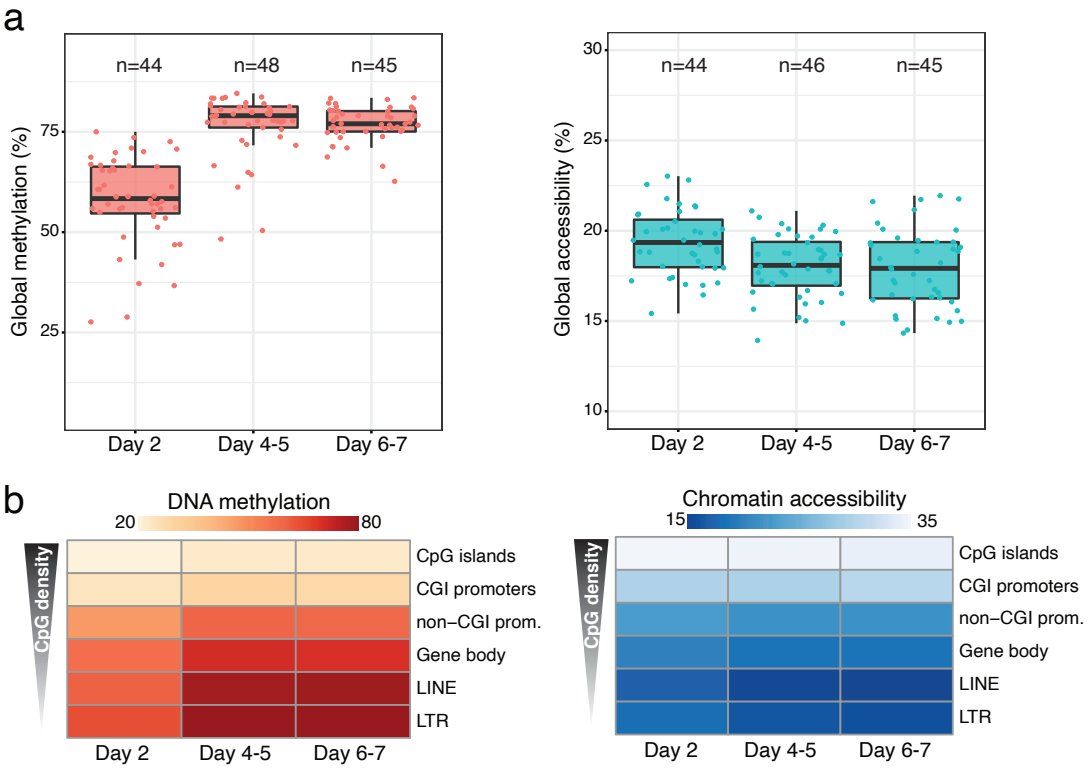


Figure 1.22: Global DNA methylation and chromatin accessibility levels during embryoid body differentiation (in WT cells).
(a) Box plots showing the distribution of genome-wide CpG methylation (left) or GpC accessibility levels (right) per stage and lineage. Each dot represents a single cell.
(b) Heatmap of DNA methylation (left) or chromatin accessibility (right) levels per stage and genomic context.

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

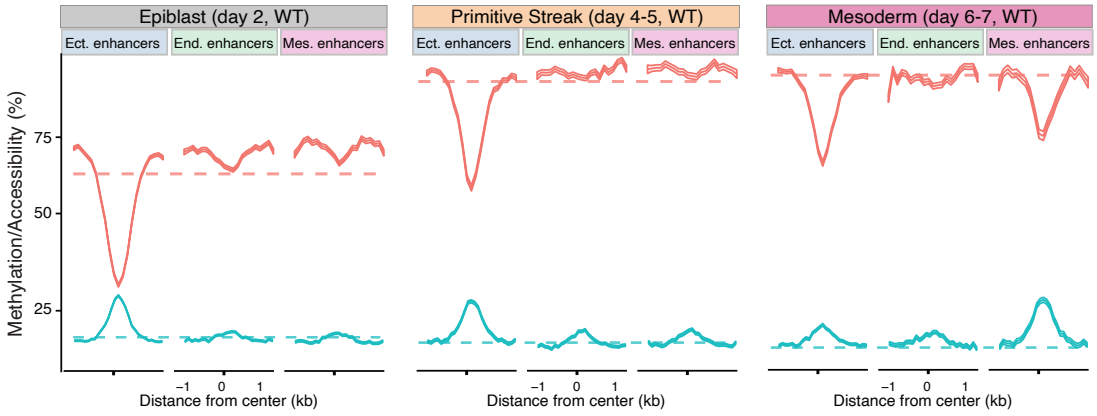


Figure 1.23: Profiles of DNA methylation (red) and chromatin accessibility (blue) at lineage-defining enhancers quantified along EB differentiation (only WT cells). Shown are running averages in consecutive 50bp windows around the centre of the ChIP-seq peaks (1kb upstream and downstream). Solid lines display the mean across cells and shading displays the corresponding standard deviation. Dashed horizontal lines represent genome-wide background levels for DNA methylation (red) and chromatin accessibility (blue).

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

Characterisation of the *TET* TKO phenotype

Having validated the EB system from a transcriptomic and epigenetic perspective, we proceed to compare the WT and the *TET* TKO cells.
At the epigenetic level, *TET* TKO epiblast-like cells (day 2) display higher levels of DNA methylation in ectoderm enhancers, but no differences in mesoderm or endoderm enhancers (Figure 1.24). No significant differences are observed between WT and *TET* TKO for chromatin accessibility. Interestingly, the *TET* TKO cells also display an increased proportion of cells undergoing mesendoderm transition (days 4-5, 95% versus 51% in the WT). This is suggestive of an early induction of gastrulation.

After the mesendoderm transition (days 4-5), mesoderm-committed *TET* TKO cells (days 6-7) failed to properly demethylate mesoderm-specific enhancers (Figure 1.24). This indicates that (1) enhancer demethylation is not required for early mesoderm commitment, and (2) demethylation of lineage-defining enhancers results from an active process that is at least partially driven by TET proteins.

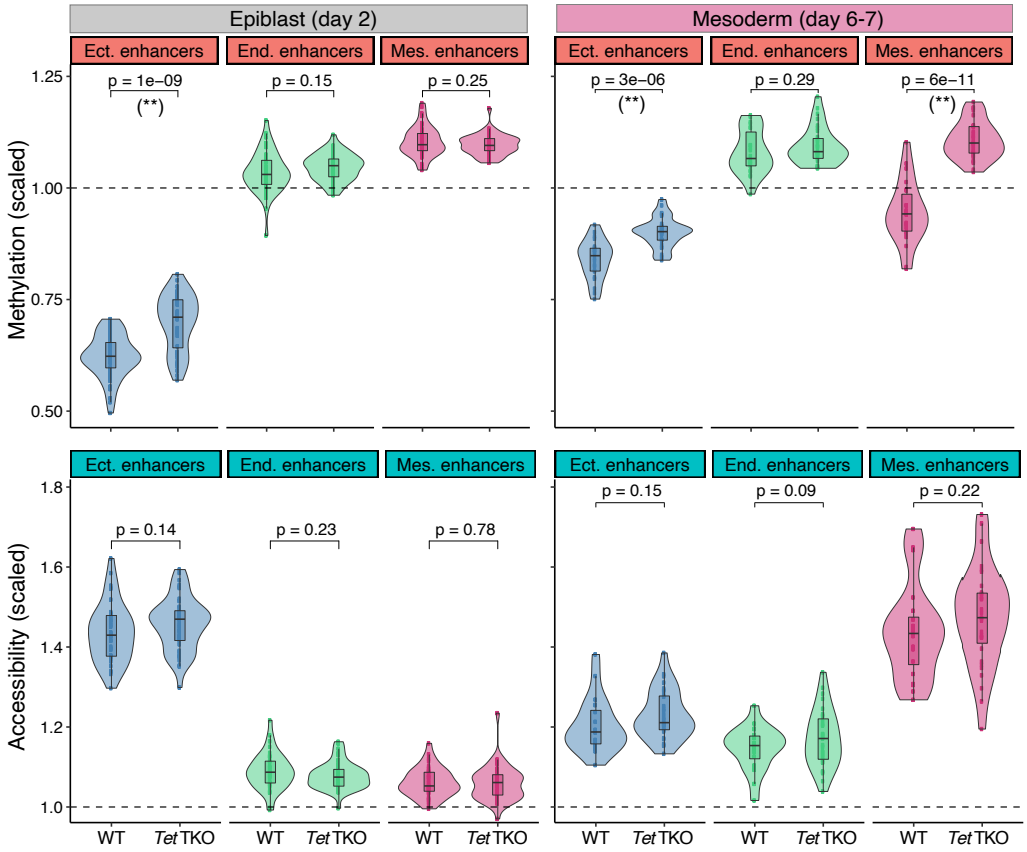


Figure 1.24: Overlayed box plots and violin plots display the distribution of DNA methylation (top) or chromatin accessibility values for lineage-defining enhancers in the epiblast-like cells at day 2 and the mesoderm-like cells at days 6-7.
The y-axis shows the DNA methylation or chromatin accessibility 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 1.20). This suggests that TET-mediated demethylation events, although not crucial for early mesendoderm commitment, seem to be important for subsequent cell fate decisions. Notably, our observations are concordant with findings from previous studies *in vivo* [11], 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.

1.2.11 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 unmethylated and accessible upon germ layer commitment. Most notably, when tracing back the epigenetic dynamics for the lineage-defining enhancers to the early epiblast stage, we observe that post-implantation cells display epigenetic priming for an ectoderm fate. This finding supports the existence of a default path in the Waddington landscape of development, with the ectoderm being the default germ layer in the embryo. In contrast, commitment to endoderm and mesoderm fates occurs by an active diversion from the default path driven by signalling cues in the primitive streak transient state.
Experimental evidence exist to support this hypothesis. Several groups have shown that, in the absence of external stimuli, ESCs differentiate to neurons [29, 17], 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.

1.2.12 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 laborious and expensive protocol, unsuitable for the profiling of large numbers of cells. In this study, we had to rely on pseudobulk approaches to obtain sufficient statistical power for some of our results. Also, it is likely that we have been underpowered to detect subtle yet important epigenetic variation. As discussed in Chapter 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 [1].

- 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 [33].

Appendix A

Mathematical derivations of MOFA

A.1 Deriving the variational inference algorithm

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

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

However, inspired by [44], 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 $\mathcal{L}(\mathbf{X})$[**Bishop**, 30]. The alternative and possibly easier approach is to define a parametric form for the distribution $q(\mathbf{X})$ with some parameters $\boldsymbol{\Theta}$ to be of the same form as the corresponding prior distribution $p(\mathbf{X})$. Then, one can find the gradients with respect to the parameters to obtain the coordinate ascent optimisation scheme. In our derivations we followed the first approach, but because we used conjugate priors the second one should converge to the same result.

Below we give the explicit update equations for every hidden variable in the MOFA+ model which are applied at each iteration of the variational inference algorithm.

A.2 Variational update equations

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

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

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

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

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

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

with

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

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

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

with

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

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

Prior distribution:

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

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

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

where:

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

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

Prior distribution:

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

Variational distribution:

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

where

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

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

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

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

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

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

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

with

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

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

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

with

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

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

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

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

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

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

where:

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

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

Prior distribution:

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

Variational distribution:

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

where

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

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

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

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

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

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

where:

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

A.3 Evidence Lower Bound

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

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

Log likelihood term Assuming a Gaussian likelihood:

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

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

Weights

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

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

Factors

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

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

ARD prior on the weights

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

Sparsity parameter of the weights

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

ARD prior on the Factors

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

Sparsity parameter of the Factors

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

Noise

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

Appendix B

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

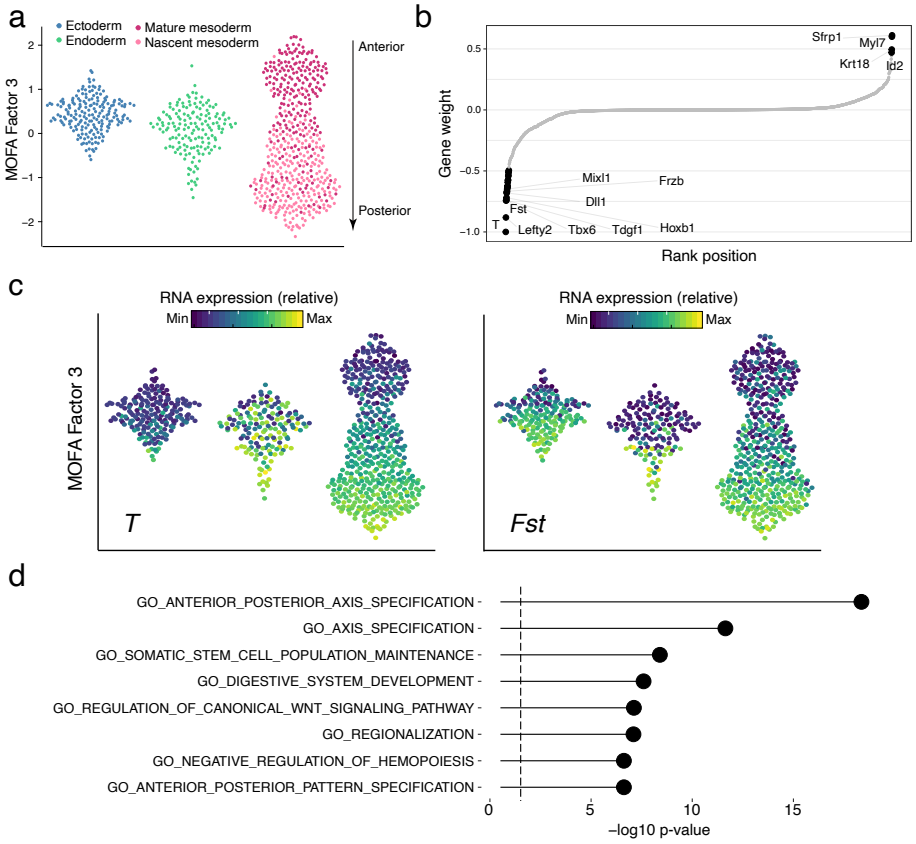


Figure B.1: Characterisation of MOFA Factor 3 as antero-posterior axial patterning and mesoderm maturation.
(a) Beeswarm plot of Factor 3 values, grouped and coloured by cell type. The mesoderm cells are subclassified into nascent and mature mesoderm (see Figure S2).
(b) RNA expression weights for Factor 3. Genes with large positive weights increase expression in the positive factor values (more anterior), whereas genes with negative weights increase expression in the negative factor values (more posterior).
(c) Same beeswarm plots as in (a), coloured by the relative RNA expression of genes with the highest positive (top) or negative (bottom) weight.
(d) Gene set enrichment analysis of the gene weights of Factor 3. Shown are the top most significant pathways from MSigDB C2 [**Subramanian2005**, **Ashburner2000**].

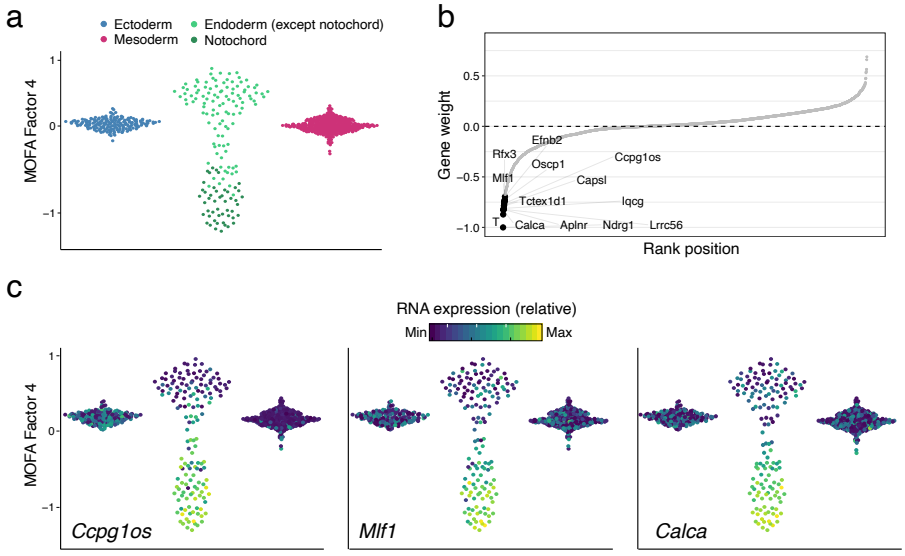


Figure B.2: Characterisation of MOFA Factor 4 as notochord formation.
(a) Beeswarm plot of Factor 4 values, grouped and coloured by cell type. The endoderm cells are subclassified into notochord (dark green) and not notochord (green) (see Figure S2).
(b) RNA expression weights for Factor 4. Genes with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

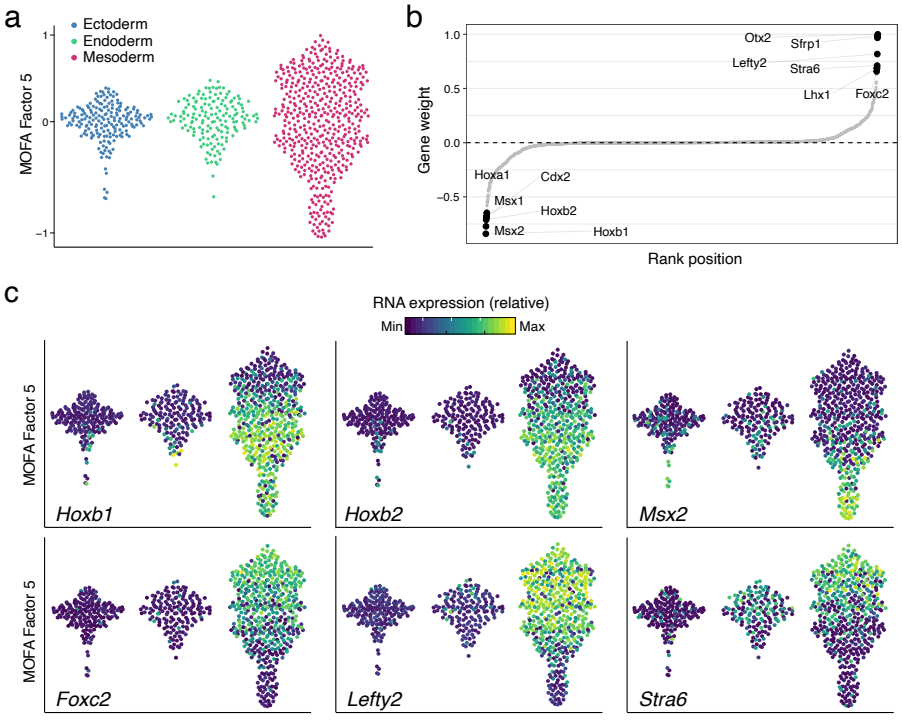


Figure B.3: Characterisation of MOFA Factor 5 as mesoderm patterning.
(a) Beeswarm plot of Factor 5 values, grouped and coloured by cell type.
(b) RNA expression weights for Factor 5. A higher absolute value indicates higher feature importance.
(c) Same beeswarm plots as in (a), coloured by the relative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

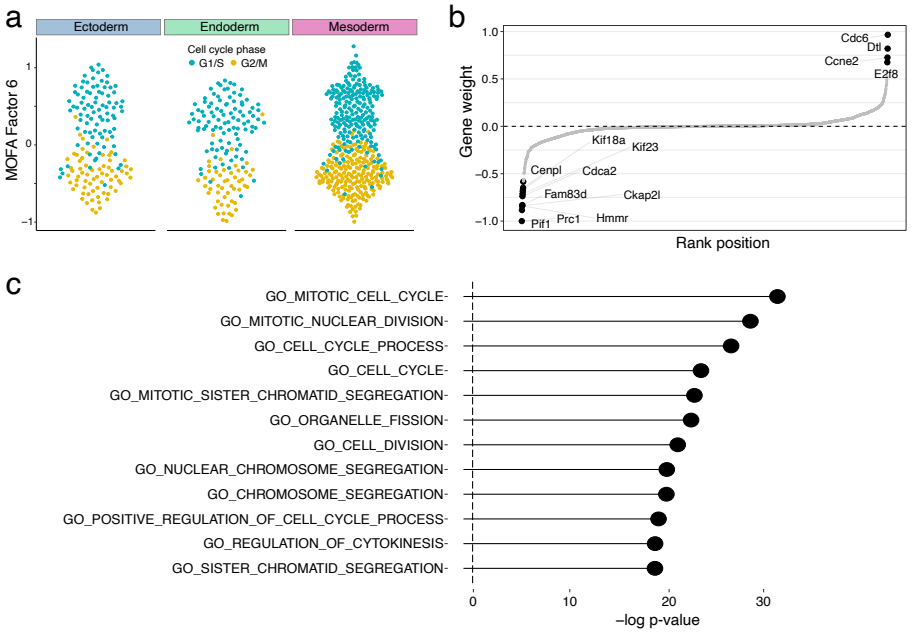


Figure B.4: Characterisation of MOFA Factor 6 as cell cycle.
(a) Beeswarm plot of Factor 6 values, grouped by cell type and coloured by infered cell cycle state using *cyclone*[**Scialdone2015**].
(b) RNA expression weights for Factor 6. Genes with large positive weights increase expression in the positive factor values (G1/S phase), whereas genes with negative weights increase expression in the negative factor values (G2/M phase).