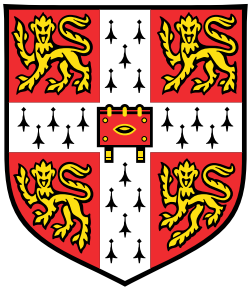
# Statistical methods for the integrative analysis of single-cell multi-omics data



**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

# 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 [121, 75, 171, 174]. 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 [222, 111, 75].

### 1.1.1 Single-cell RNA sequencing

single-cell RNA sequencing (scRNA-seq) protocols differ extensively in terms of scalability, costs and sensitivity [223, 128]. Broadly speaking, they can be classified into plate-based and droplet-based methods. In plate-based methods such as CEL-seq [83] and Smart-seq[188, 177], 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[97], allele-specific fractions[54] and RNA velocity information [127].

Droplet-based methods are based on the use of droplet microfluidics technology [250]. 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 [119, 257], Drop-seq[149] and the comercial 10x Genomics Chromium [255]. As a trade-off, the increased high throughput of droplet-based approaches comes at the expense of reduced sensitivity[256, 242, 224].

More recently, a third type of scRNA-seq methodology emerged based on a combinatorial cellular indexing strategy [33, 199, 36], 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 [197]. 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 [75, 111, 17].

#### 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 [103, 173, 130, 211]. Its classical roles include the silencing of repeating elements, inactivation of the X chromosome, gene imprinting, and repression of gene expression [105]. Consistently, the disruption of the DNA methylation machinery is associated with multiple dysfunctions, including cancer [14], autoimmune diseases [143] and neurological disorders [4].
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 [87, 187, 141].

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) [210, 77, 74, 65]. 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 [68, 44, 42]. 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, [210] 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, [210] 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.

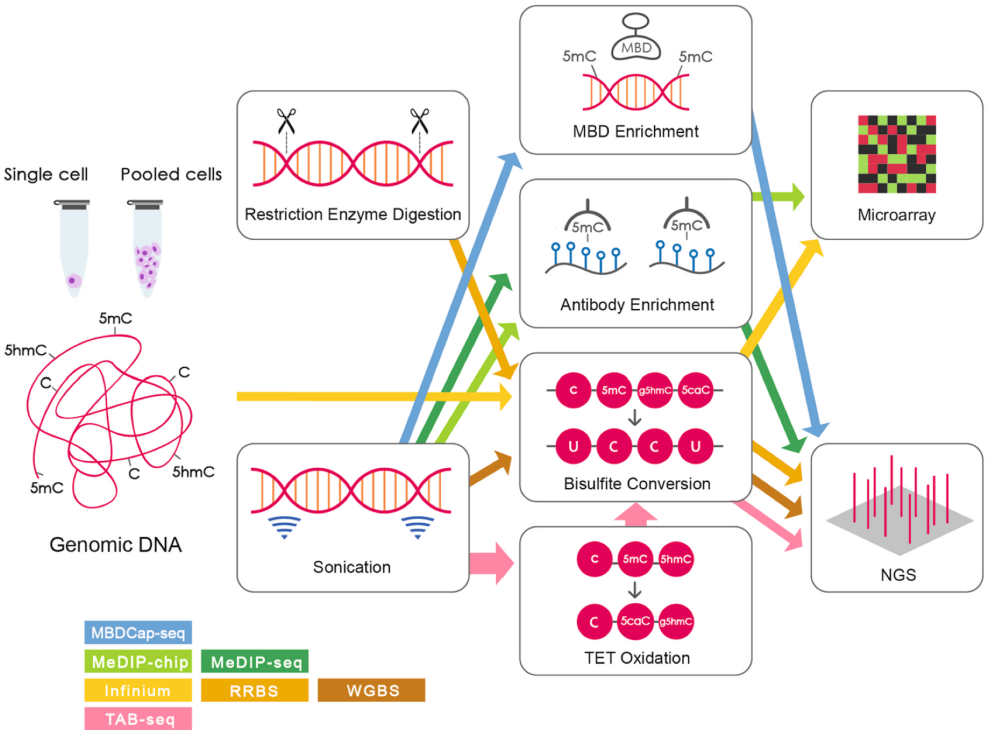



**Figure 1.1:** Workflow of DNA methylation profiling protocols. Reprinted from [246]

Alongside scBS, other bulk sequencing methods were also adapted to the single cell resolution, with different trade-offs between coverage and costs. For instance, [**Guo2015**] 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 [**Meissner2005**]. 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 [120, 235]. The positioning of the nucleosomes in the nucleus provide an important layer of gene regulation, mostly by exposing or sheltering transcription factors binding sites [101]. In general, active regulatory regions tend to have low occupancy of nucleosomes, whereas inactive regions show a high density of nucleosomes [221]. 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)

[213], transposase-accessible chromatin followed by sequencing (ATAC-seq) [29] and Nucleosome Occupancy and Methylome-sequencing (NOMe-seq) [110]. A systematic comparison with a controlled experimental design can be found in [169].

- **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 [213]. Although this methodology became one of the gold standards to map chromatin accessibility by the ENCODE consortium [46, 228], it has now been reported that DNase I introduces significant cleavage biases, thus affecting its reliability to infer transcription factor footprints [86].

- **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 [27, 235, 169].

- **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 [113]. Hence, after M.CviPI treatment followed by bisulfite sequencing, GpC methylation marks can be interpreted as direct read outs for chromatin accessibility. [110]. 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 [110, 181, 169]. 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.

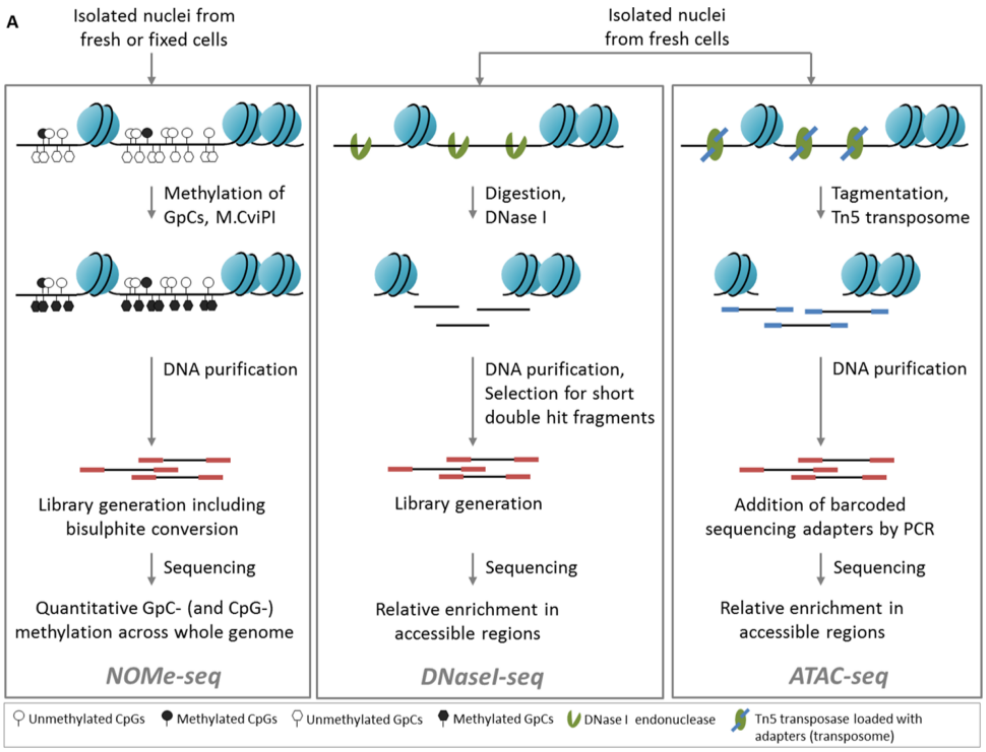



**Figure 1.2:** High-level overview of the workflows for the three main chromatin accessibility assays: NOMe-seq, DNase-seq and ATAC-seq. Reprinted from [169].

As with DNA methylation, single-cell profiling methods for chromatin accessibility also emerged from its bulk counterparts, including ATAC-seq[28], NOMe-seq [181] and DNase-seq [102]. Due to its cost-effective strategy, single-cell ATAC-seq (scATAC-seq) has become the most popular technique to map open chromatin [52, 35, 41]. Compared to bulk ATAC-seq, scATAC-seq libraries are notably sparse. In a saturated library, [52] 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, [51] 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 [84, 197]. 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 [222, 38], 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 [175]. 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) [147]. 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 [147, 95, 6, 96].
- **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[220]) and RNA expression and protein sequencing assay (REAP-seq[176]). 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 [34], 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 [43].
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) [6] and Nucleosome Occupancy and Methylation sequencing (NOMe-seq) [110, 181]. 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[113, 110]. 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[113, 110].

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[210], whereas the RNA fraction undergoes Smart-seq2 [177].

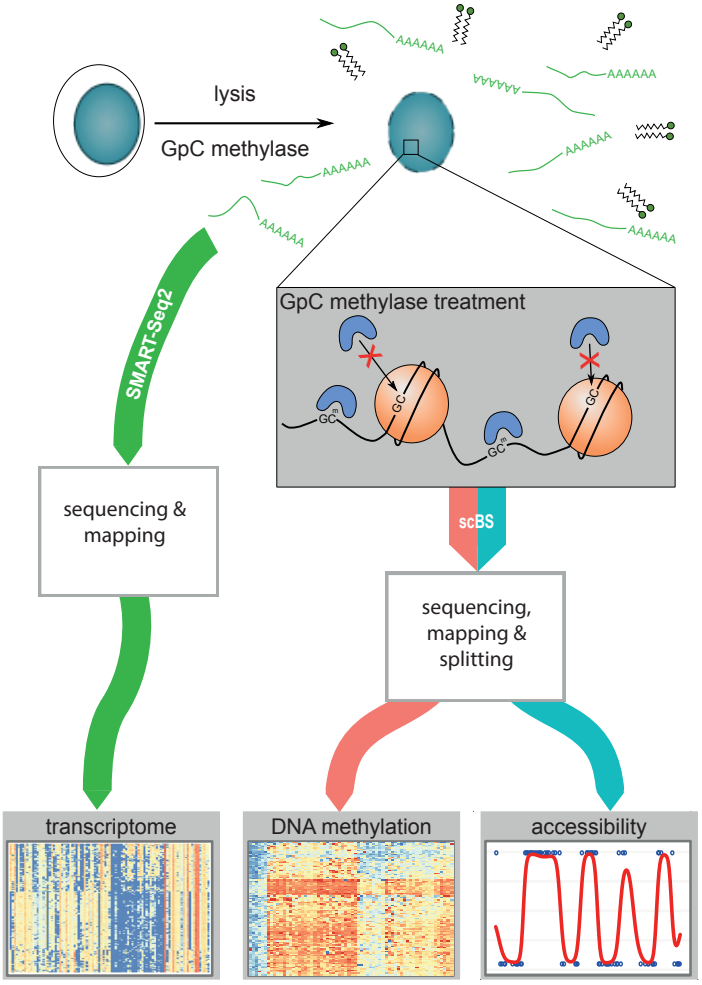



**Figure 1.3: scNMT-seq protocol overview.**
In the first step, cells are isolated and lysed. Second, cells are incubated with a GpC methyltransferase. Third, the RNA fraction is separated using oligo-dT probes and sequenced using Smart-seq2. The DNA fraction undergoes scBS-seq library preparation and 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 [110, 181, 169]. 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 [169].

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 [110]. 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 [124] 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 [210], 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 [210, 6, 66].

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.

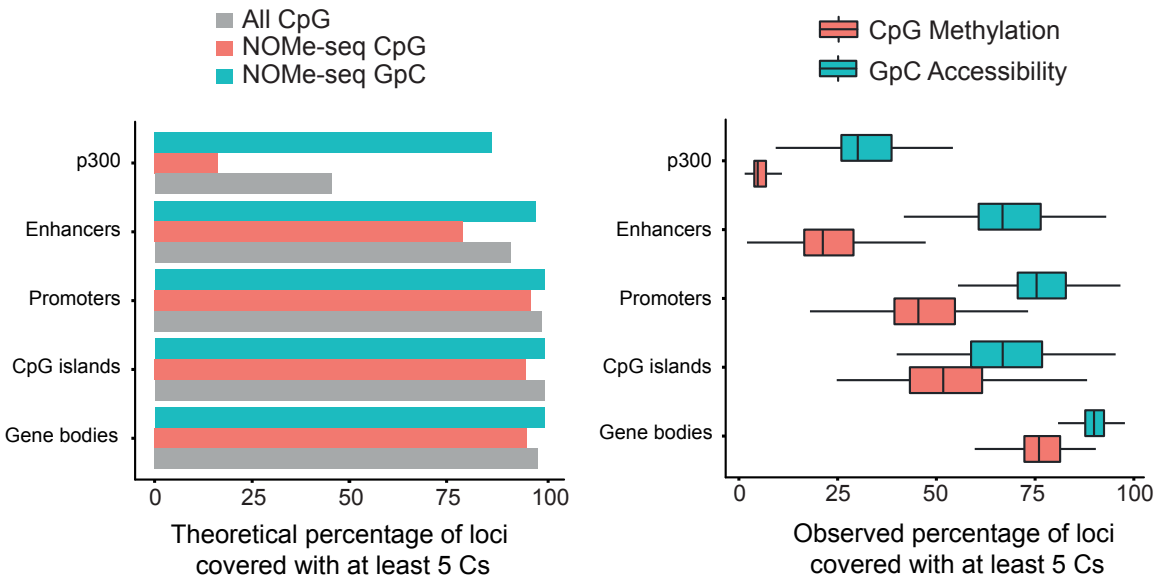



**Figure 1.4: Coverage statistics for CpG DNA methylation and GpC chromatin accessibility**.
(a) Fraction of loci with at least 5 CpG (red) or GpC (blue) dinucleotides (y-axis) per genomic context (x-axis), after exclusion of the conflictive trinucleotides. The grey bar shows the total number of CpGs 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 [6] (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.

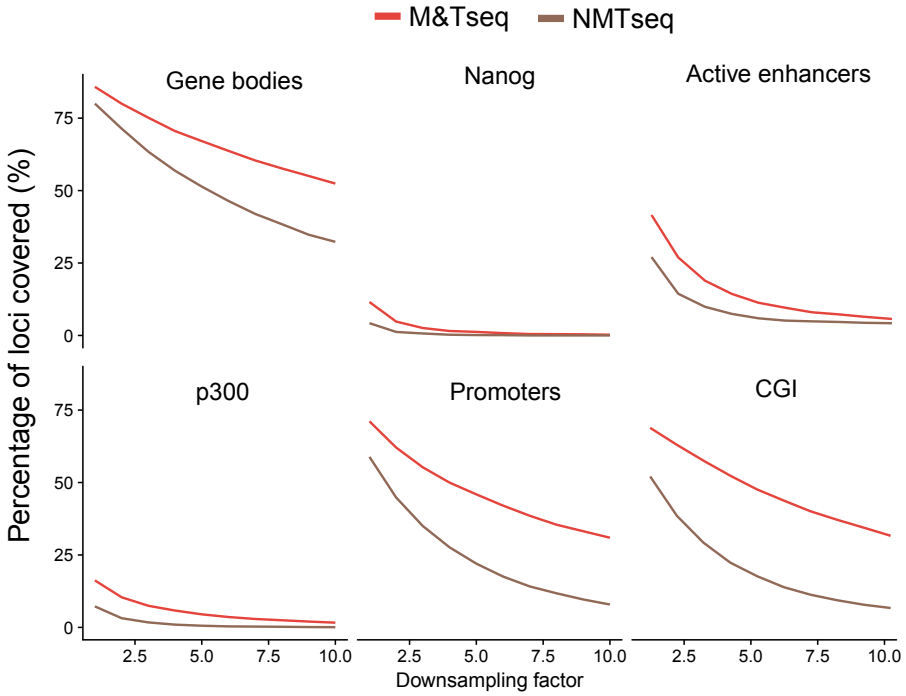



**Figure 1.5: Comparison of the empirical coverage of DNA methylation with scM&T-seq [6].**
The y-axis displays the fraction of loci covered with at least 5 CpG sites. The x-axis displays the downsampling factor. To facilitate the comparison, we selected two cells that were sequenced at equivalent depth.

#### 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[6], scBS-seq[210] and bulk BS-seq[66]. 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 [66]. 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 [6, 210, 66]). Previous studies have sown that female ESCs tend to show lower levels of mean global methylation, which is consistent with a more pluripotent phenotype [258].

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 ($R = 0.74$).

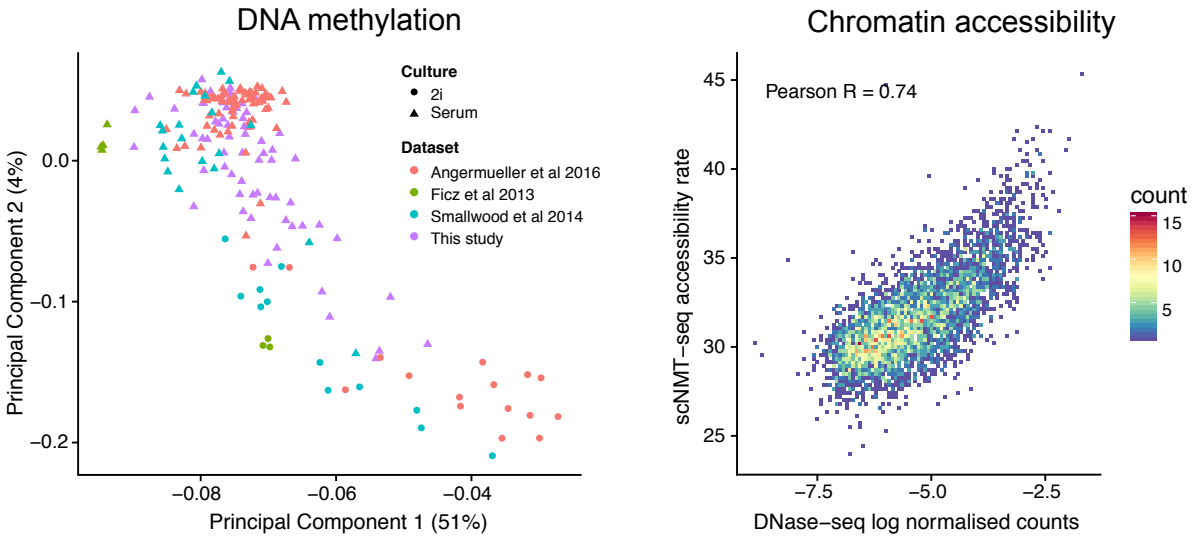



**Figure 1.6: Comparison of unsupervised genome-wide quantifications to published data sets.**
(a) Principal Component Analysis of 1kb running windows. Missing values were imputed using the average methylation rate per locus.
(b) Scatter plot of chromatin accessibility quantified over 10kb running windows 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 [181]. As a control, we observe that cells which did not receive M.CviPI treatment showed globally low GpC methylation levels ( $\approx 2\%$, **??**).

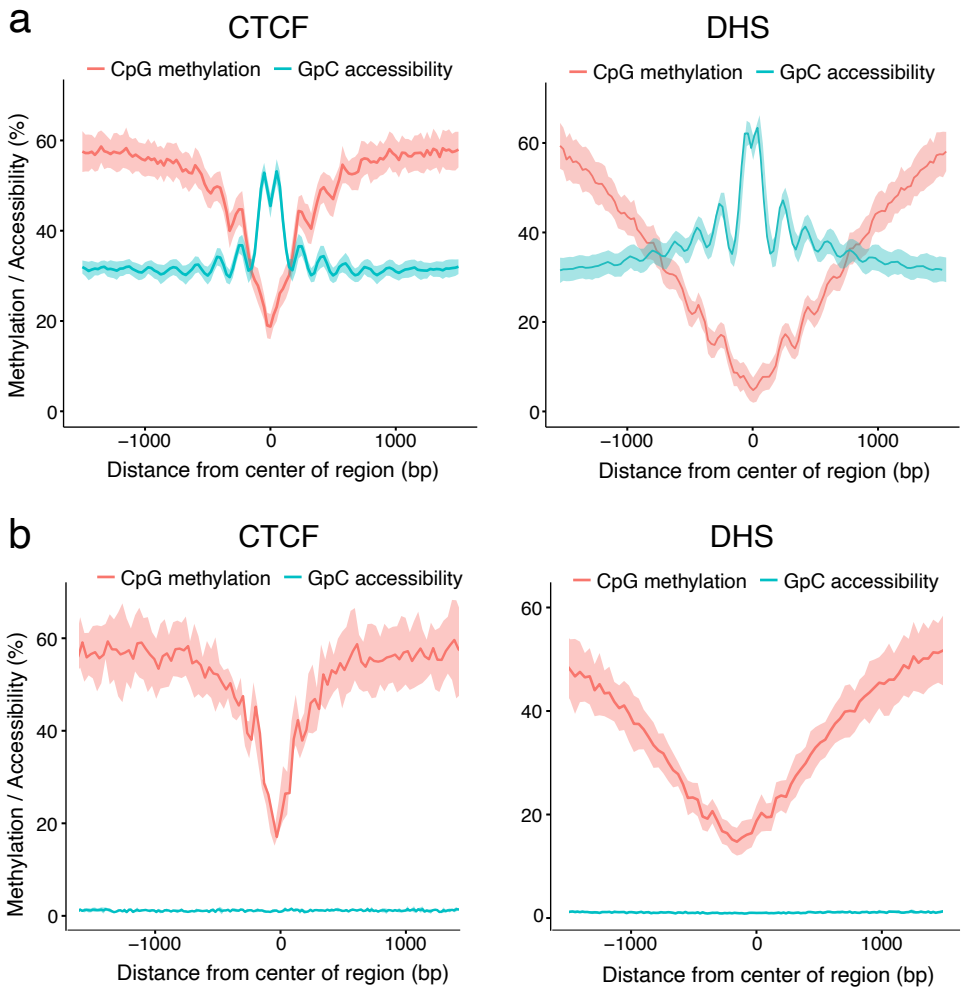



**Figure 1.7: Accessibility and methylation profiles in regulatory genomic contexts.** First, we pseudobulk the data set by pooling information across all cells. Next, we compute running averages of the CpG methylation (red) and the GpC accessibility (blue) in consecutive non-overlapping 50bp windows. 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 [228, 6]. 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.

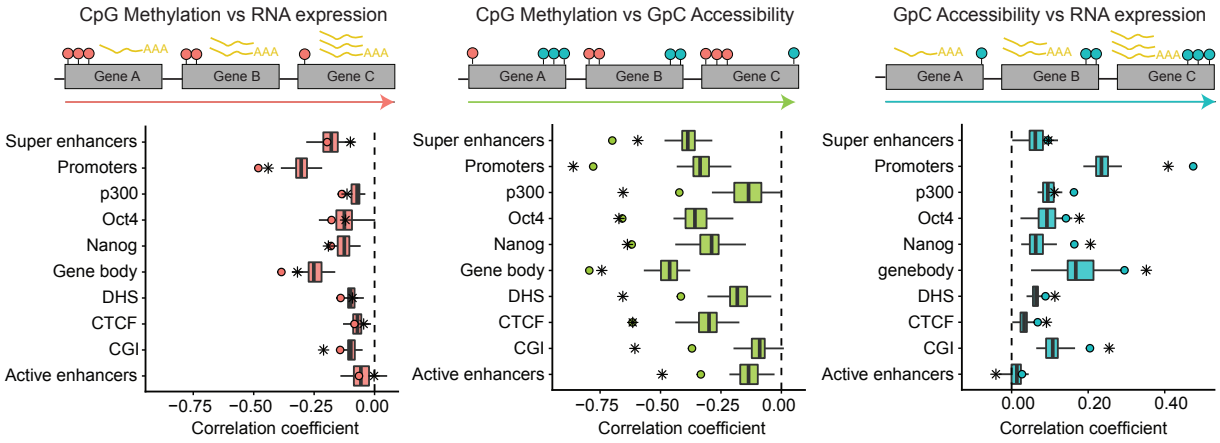



**Figure 1.8: Quantification of linear associations between molecular layers.**
The top diagram illustrates the computation of an association test per cell (across all loci in a given genomic context). The left panel shows DNA methylation versus RNA expression. The middle panel shows DNA methylation 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 [66, 46], 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*).

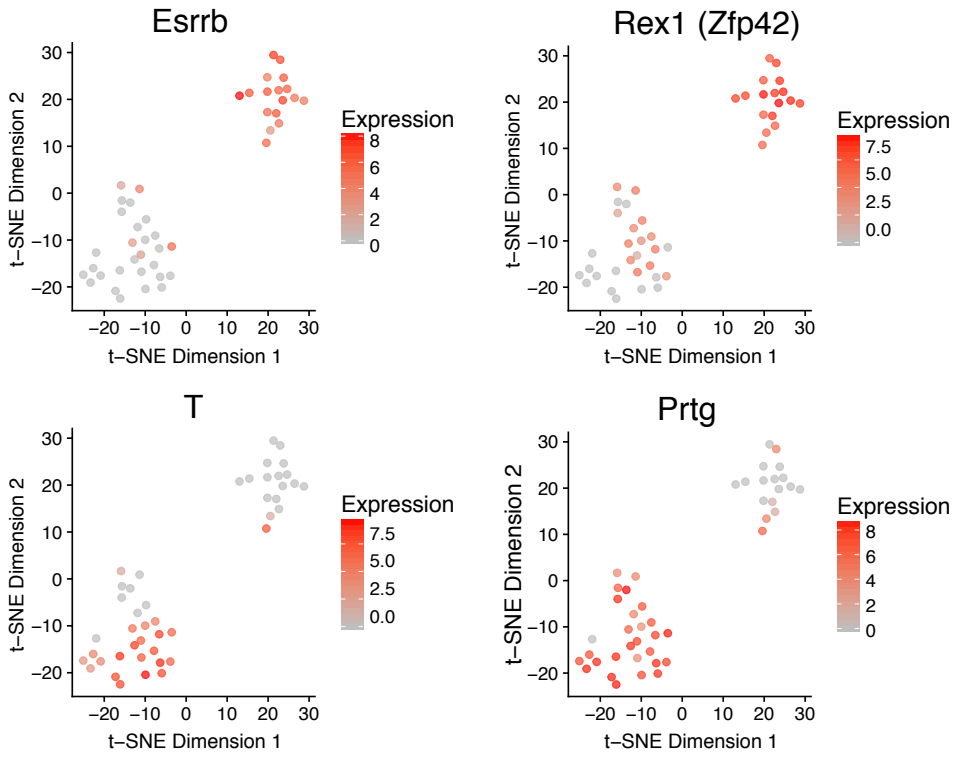



**Figure 1.9: t-SNE Dimensionality reduction on the RNA expression profiles for the embryoid body cells.**
The scatter plots show t-SNE dimensions 1 (x-axis) and 2 (y-axis). Cells are coloured based on expression of pluripotency factors (top) and differentiation markers (bottom).

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 [172]. A previous study [6], 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

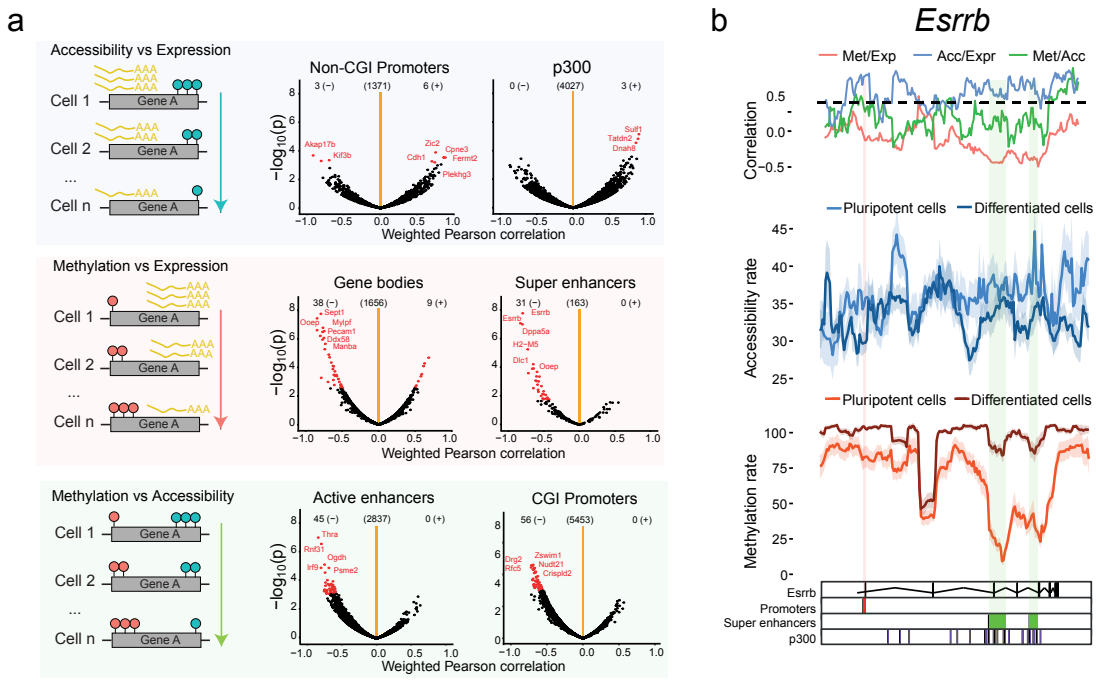



**Figure 1.10**

### 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 [106], 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:

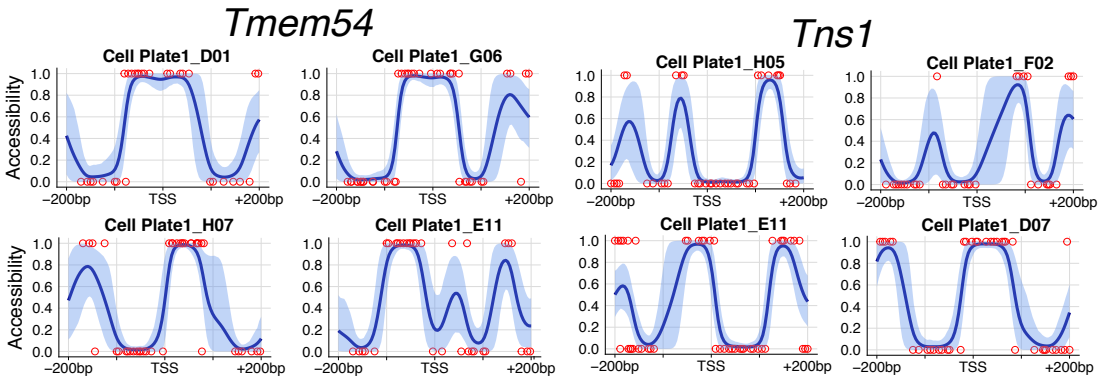



**Figure 1.11: Illustrative examples of single-cell accessibility profiles around the transcription start site**.
Shown are representative profiles for two genes, *Tmem54* and *Tns1*. Each panel corresponds to separate cell. The y-axis displays the binary GpC accessibility values, with 1 being accessibility 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 [106].

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:

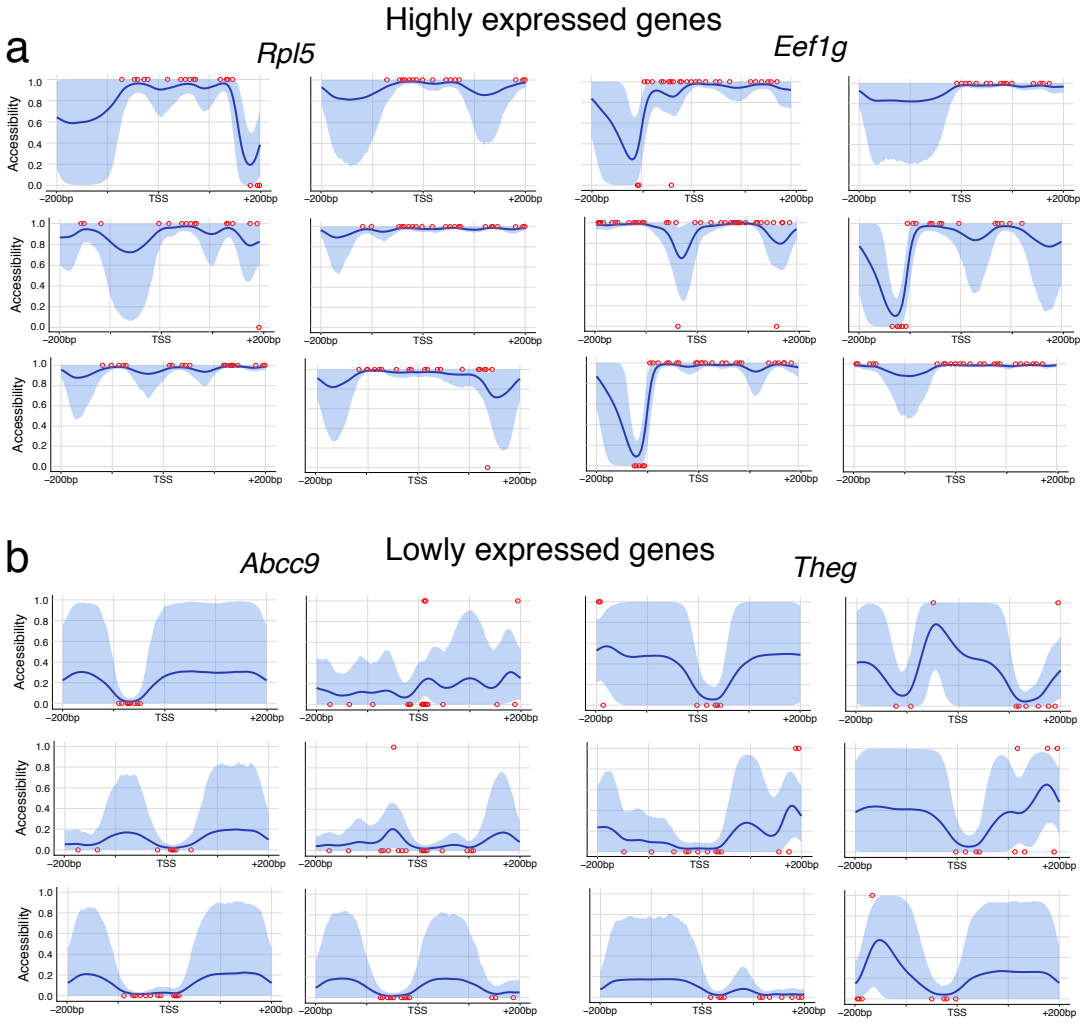



**Figure 1.12: Single-cell accessibility profiles of representative genes with high and low expression levels**
Shown are *Rpl5* and *Eef1g* (high expression levels, top); *Abcc9* and *Theg* (low expression levels, bottom). Each panel corresponds to a separate cell. Axis are the same as in Figure 1.11.

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.

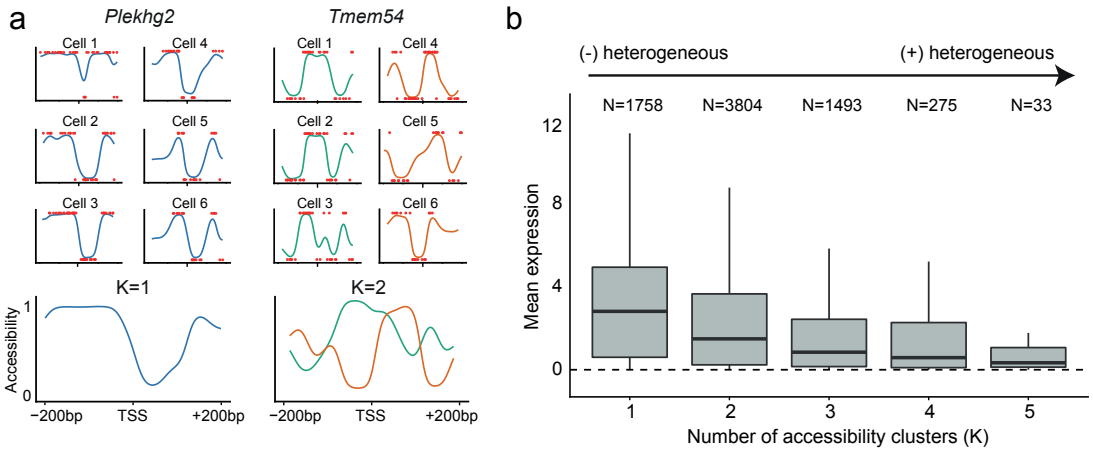



**Figure 1.13**

When relating the number of clusters to the gene expression, we observed that genes with homogeneous accessibility profiles (fewer clusters) were associated with higher average expression levels. Gene Ontology enrichment analysis suggests that this cluster is enriched by genes with houseekeping functions, which are known to display more conserved epigenetic features [209].
In contrast, genes with heterogeneous accessibility (multiple clusters) were associated with lower expression levels and were 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 [238, 16]

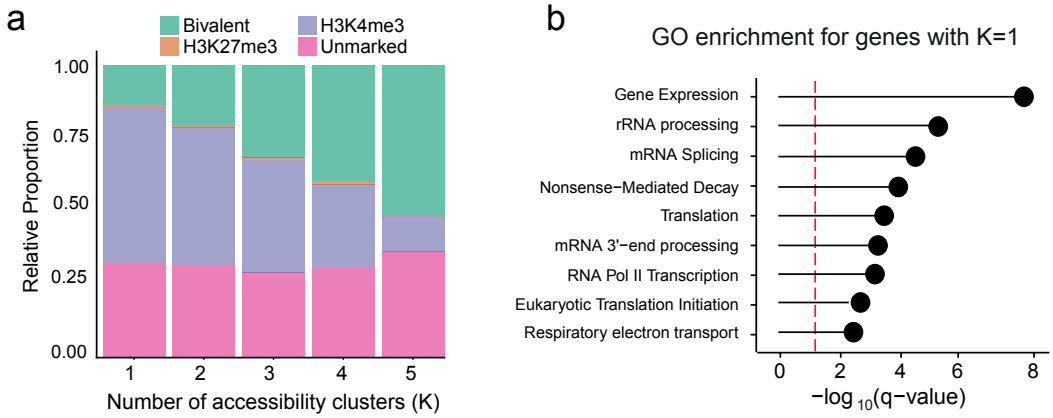



**Figure 1.14**

### 1.2.6 Exploration of epigenome dynamics 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 [233, 80, 202]. 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[80] 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:

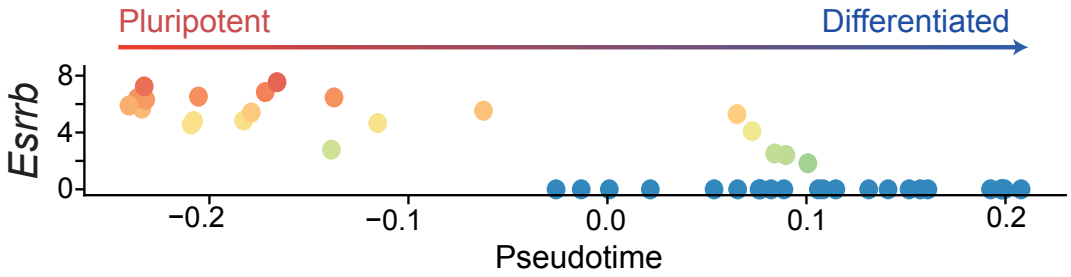



**Figure 1.15: Reconstruction of developmental trajectory in embryoid body cells from the RNA expression data.**
Each dot corresponds to one cell. The y-axis displays expression of *Esrrb*, a canonical pluripotency marker, and the x-axis shows the position of the cells in the first diffusion component.

Next, we investigated whether the strength of association between molecular layers (as calculated in Figure 1.8) are affected along the predicted developmental trajectory. We observe that for DNA methylation and chromatin accessibility, the negative correlation coefficients decreases in practically all genomic contexts Figure 1.16, such that pluripotent cells have a notably weaker methylation-chromatin coupling than differentiated cells.
This triple-omics analysis, which was made possible by the ability to profile three molecular layers, and the continuous nature of single-cell data, indicates that the strength of regulation between two molecular layers can be altered during cell fate decisions.

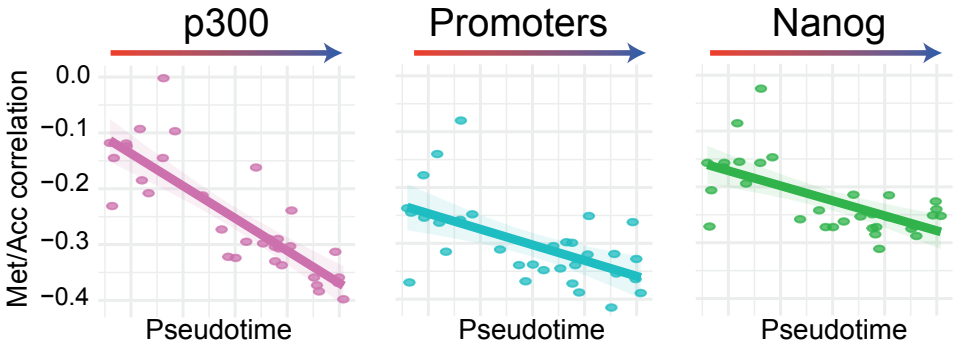



**Figure 1.16**

## 1.3 Open perspectives

Multi-modal sequencing technologies are becoming rapidly available. Yet, compared to scRNA-seq technologies, the field still is at an early stage amd numerous developments are expected to occur in the next years. These are the lines of research that we are considering to improve in scNMT-seq:

- **Scalability**. scRNA-seq protocols are reaching the astonishing numbers of milions of cells per experiment, compared to the limited cell numbers achieved in multi-modal experiments [36, 35, 76]. As in scRNA-seq, the maturation of multi-modal techniques will have a trade-off between sensitivity and scalability [38]. Under the assumption that most of the variance explained by epigenetic layers is driven by cis-regulation [15], 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**, 163].
  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 [223]). 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 [77], an initial round of ATAC protocol to select open chromatin [215] 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 [5]. 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 [5]. A successful single-cell strategy based on deep learning has been proposed (DeepCpG[5]), 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 [107]), 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 [180, 64, 153, 60]. This approach has been successful at dealineating 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[97], allele-specific fractions[54] and RNA velocity information [127].

From the experimental side, NMT-seq can theoretically be combined with novel single-cell assays that quantify transcription factor binding [162] and histone modifications [108].

- **Denoising**. The readouts from bisulfite sequencing are very sensitive [**XX**]. However, in scNMT-seq the CGC positions (27%) suffer from off-target effects of the GpC methylase [110]. 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 and pool information across cells.

- **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 [132] showed that one can obtain phased methyation 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.