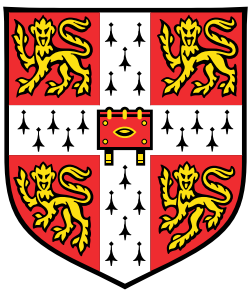
Statistical framework for the integration of single-cell multi-omics data sets



Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2019

Table of contents

Chapter 1

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

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

Next-generation sequencing technologies have revolutionised the study of biological systems by allowing the unbiased profiling of multiple molecular layers, ranging from the genome[63] and the epigenome[64] to the transcriptome[133, 10, 158, 154]. However, bulk sequencing approaches require the pooling of large number of cells to report an average readout, and are hence limited for the study of complex heterogeneous processes, including the immune system, embryonic development or cancer [70, 163, 166].
The progressive development of low-input sequencing techniques resulted in an explosion of single-cell sequencing technologies, mostly for the transcriptome (scRNA-seq). In contrast to bulk protocols, single-cell profiling techniques provide an unprecented opportunity to study the molecular variation associated with cellular heterogeneity, lineage diversification and cell fate commitment [118].

1.1.1 Single-cell RNA sequencing

scRNA-seq protocols differ extensively in terms of scalability, costs and sensitivity [209, 122]. Broadly, there are two groups of methods, plate-based and dropled-based **??**.
In plate-based methods, cells are isolated using micropipettes or flow cytometry into individual wells of a plate, where the library preparation is performed. These class of methods include single-cell tagged reverse transcription sequencing (STRT-seq [95]), Cell Expression by Linear amplification and Sequencing (CEL-seq [77]) and Smart-seq[181, 169]. The main advantage of plate-based methods is the higher quality of libraries and the full length transcript information, in the case of Smart-seq, which enables the quantification of splice variants[92], allele-specific fractions[51] and even RNA velocity information [121]. In contrast, the main drawback lies on the low throughput. Yet, multiplexing techniques, the addition of molecular barcodes to cDNA fragments, allow the parallel processing of multiple experiments, thereby increasing the scale of each experiment [77].

Droplet-based methods are based on the use of droplet microfluidics technology [231]. By capturing cells in individual droplets, each containing all necessary reagents for library preparation, this protocol allows the profiling of thousands or even milions of cells in a single experiment. These class of methods include InDrop [116, 238], Drop-seq[141] and the comercial 10x Genomics Chromium [236]. All three protocols share similar technologies, including the use of unique molecular identifiers (UMI) to correct for biases in PCR amplifications [112]. Differences lie in the barcode design, cDNA amplification step and bead manufacturing [231]. As a trade-off, the increased high throughput of droplet-based approaches comes at the expense of reduced sensitivity[237, 225, 210] (Figure 1.1).

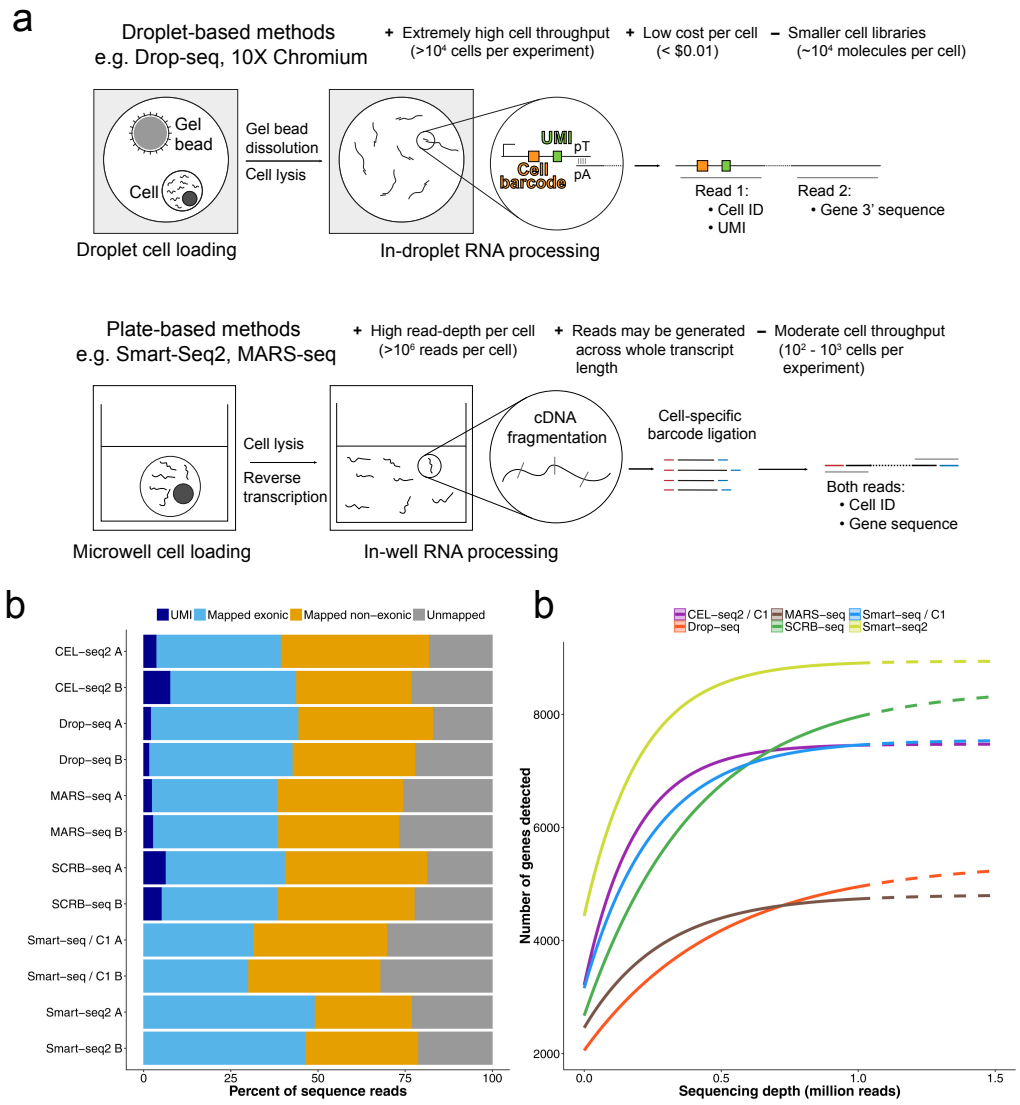


Fig. 1.1 Comparison of quality metrics between different scRNA-seq technologies. (a) Overview of the two types of scRNA-seq methods: plate-based and dropled-based. (b) Fraction of reads mapped to exons (cyan), outside exons (orange) and unmapped (gray). -A and -B denote different replicates. For UMI methods, also shown is the fraction of reads mapping to unique UMIs (dark blue). (c) Median number of genes detected per cell (more than one count) as a function of the (downsampled) sequencing depth. Dashed lines represent extrapolated fits assuming a saturating model. Figure reprinted from [237]

More recently, a third type of scRNA-seq emerged based on a combinatorial cellular indexing strategy [30, 189]. As demonstrated in a study of mouse organogenesis [32], this enables the sequencing

of more than a milion cells in a single experiment for a fraction of the cost of other methods. Notably, the same combinatorial indexing strategy has been successfully applied to other data modalities, providing the unprecented opportunity to study multiple molecular layers in an unbiased and high-throughput setting [222, 179, 156].

1.1.2 Single-cell sequencing of the epigenome

While the large majority of single-cell studies are focused on measuring RNA expression, transcriptomic readouts are just a single dimension of cellular heterogeneity and hence contain limited information to characterise the molecular determinants of phenotypic variation [188]. Consequently, gene expression markers have been identified for a myriad of biological systems, but the accompanying epigenetic changes and the role of other molecular layers in driving cell fate decisions remains poorly understood [70, 107, 16].

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 [100, 165, 124, 199]. Its classical roles include the silencing of repeating elements, inactivation of the X chromosome, gene imprinting, and repression of gene expression [101]. Consistently, the disruption of the DNA methylation machinery is associated with multiple dysfunctions, including cancer [14], autoimmune diseases [135] and neurological disorders [3].
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 [82, 180, 134].

Alongisde developments in scRNA-seq technologies, protocols for the profiling of DNA methylation at the single-cell level also emerged from its bulk counterparts (Figure 1.2), notably bisulfite sequencing (BS-seq) [198, 73, 69, 61]. The underlying principle of BS-seq is the treatment of the DNA with sodium bisulfite before DNA sequencing, which converts unmethylated cytosine residues to uracil (and eventually to thymine, after PCR amplification), leaving 5-methylcytosine residues intact. The resulting C→T transitions are detected by DNA sequencing, thereby yelding DNA methylation information at single-nucleotide resolution [64, 41, 39].

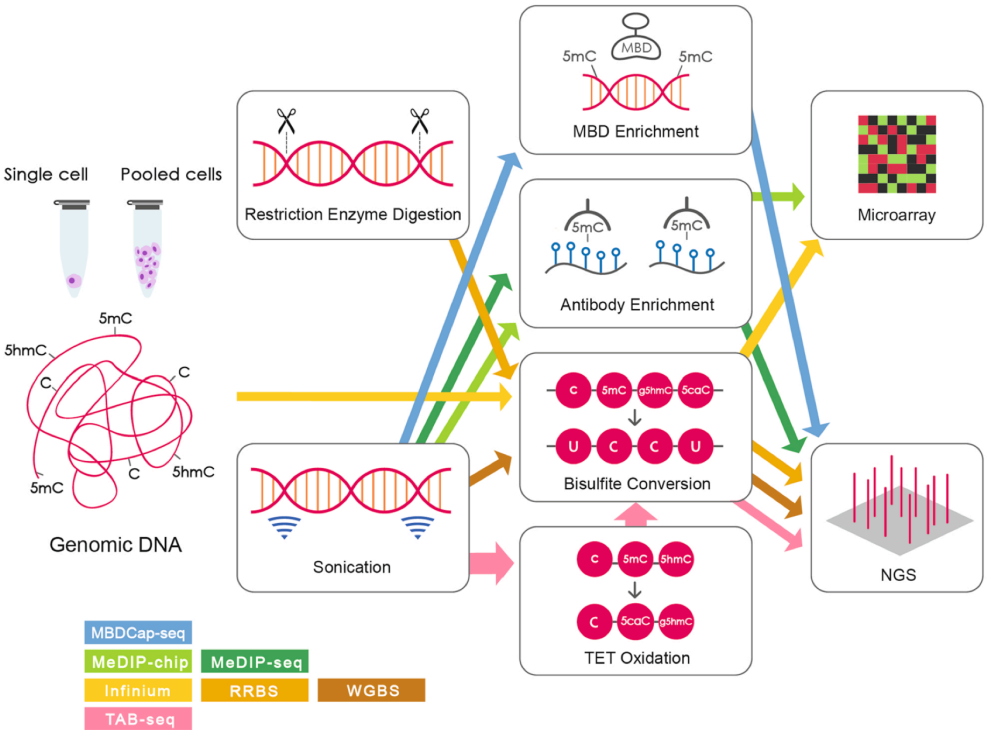


Fig. 1.2 Workflow of DNA methylation profiling protocols. Reprinted from [227]

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, [198] adapted the post-bisulfite adaptor tagging (PBAT) protocol with multiple rounds of 3' random primer amplification (Figure 1.3). When the bisulfite treatment is performed before ligation of adaptors, rather than afterwards, loss of adapter-tagged molecules is minimised, hence enabling the use of scBS-seq with little input material.

In a proof of concept study, [198] applied scBS-seq on ovulated metaphase II oocytes (MIIs) and mouse ESCs. They reported an average coverage of 3.7 million CpG dinucleotides (17.7%) per cell and revealed extensive heterogeneity in the methylome of pluripotent cells.

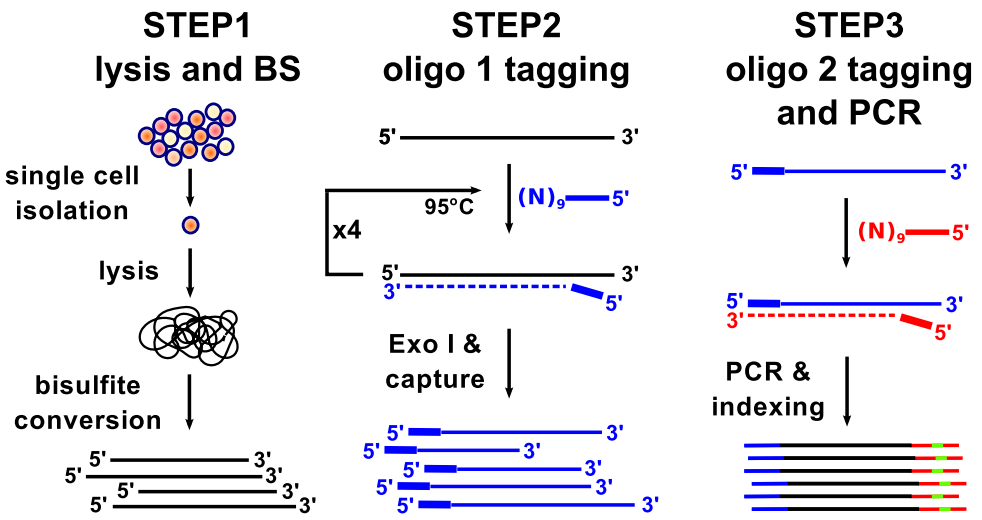


Fig. 1.3 scBS-seq profiling protocol. In the first step, cells are isolated and lysed, followed by bisulfite conversion. In the second step, five round of random priming amplification are performed using the first sequencing adaptor. In the third step, this process is repeated using the second sequencing adaptor. Finally, the resulting fragments are PCR-amplified and sequenced. Figure reprinted from [198].

Alongside scBS, other bulk sequencing methods were also adapted to the single cell resolution, with different trade-offs between coverage and costs. For instance, [72] 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 [147]. 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 [143].

(IMPROVE) More recently, combinatorial indexing technologies such as sci-MET [156] and snmC-seq [138] enabled the profiling of thousands of single cells in parallel, increasing the highthroughput, albeit yelding lower coverage than scBS-seq (maximum of 7.0% observed CpG sites) [156]. The idea behind the combinatorial indexing strategy is to tag the DNA with multiple adaptors such that the probability of two cells harboring the same adatpor is very low. This allows a large range of cells to be processed in parallel, significantly increasing the cell count throughput. Additionally, the proposed combinatorial indexing techniques yield higher mapping rates, on the order of 60% [156] compared to 30% in scBS-seq [198].

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 [117, 218]. The N-terminal tails of the histones emerge from the nucleosome and are a strong hotspot for chemical modifications, including methylation, acetylation, phosphorylation and others [11]. The complex interaction between a

histone modification and the corresponding position, often called the histone code, is an important driver of epigenetic regulation and an active area of research [235].
In addition to the histone modifications, the positioning of the nucleosomes provide another layer of gene expression regulation, mostly by exposing or sheltering transcription factors binding sites [98]. In general, active regulatory regions tend to have low occupancy of nucleosomes, whereas inactive regions show a high density of nucleosomes [207]. Thus, the profiling of DNA accessibility and transcription factor footprints represents an important dimension to understand the regulation of gene expression.
Traditionally, four main experimental approaches have been used to map chromatin accessibility in a genome-wide and high-throughput manner (Figure 1.4): MNase sequencing (MNase-seq) [102], DNase sequencing (DNase-seq) [201], transposase-accessible chromatin followed by sequencing (ATAC-seq) [26] and Nucleosome Occupancy and Methylome-sequencing (NOMe-seq) [106]. A systematic comparison with a controlled experimental design can be found in [161].

- **MNase-seq**: the chromatin is incubated with a micrococcal nuclease (MNase), an enzyme that degrades naked DNA, followed by purification, adapter ligation, PCR amplification and next-generation sequencing. As nucleosomes protect the DNA from digestion, the resulting sequencing fragments reveal nucleosome location, hence providing an inverse measure of chromatin accessibility [102].

- **DNase-seq**: the chromatin is incubated with DNAse, an enzyme that in low concentrations cuts nucleosome-free regions. Hence accessible sites, typically called DNase I hyper-sensitive sites, are released and sequenced [201]. In contrast to MNase-seq, this assay provides a direct measure of chromatin accessibility, and became one of the gold standards to map chromatin accessibility in the human genome by the ENCODE consortium [43, 211]. However, it has been reported that DNase I introduces significant cleaveage biases, thus affecting its reliability to infer transcription factor footprints [81].

- **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. As in DNase-seq, it provides a direct measure of chromatin accessibility. Yet, in the last years it has arguably displaced DNase-seq or MNase-seq as the *de facto* method for profiling chromatin accessibility due to its fast and sensitive protocol [24, 218, 161].

- **NOMe-seq**: it 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 [110]. Hence, after M.CviPI treatment and bisulfite sequencing, 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 marks [106]. 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 (≈ 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 [106, 174, 161]. 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 [161]. 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%).

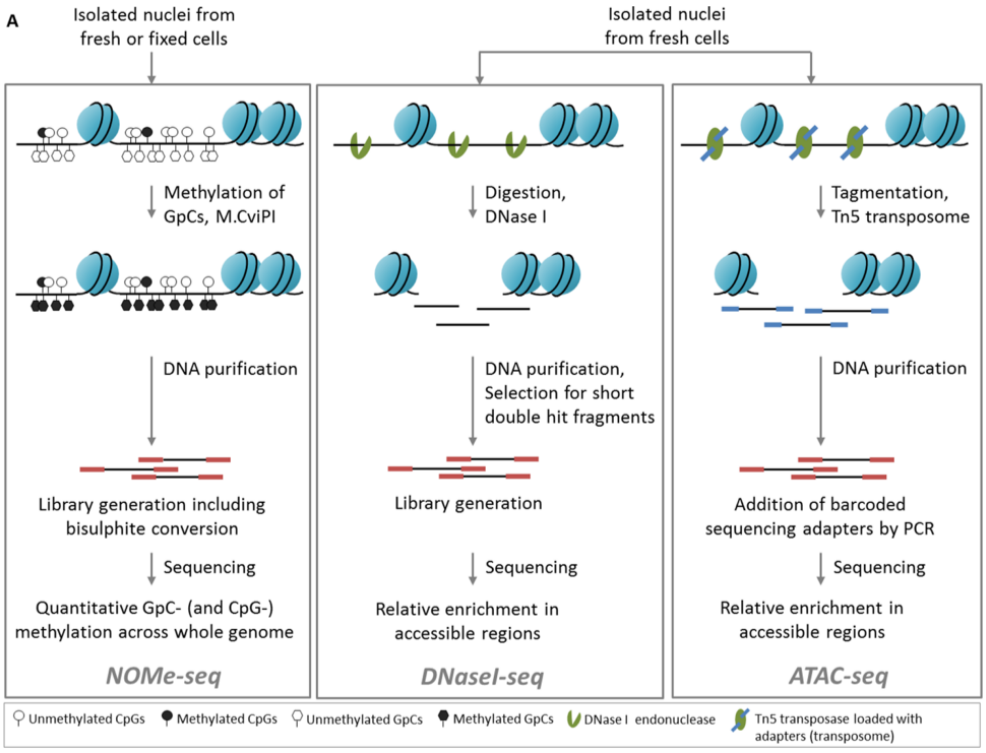


Fig. 1.4 High-level overview of the workflows for the three main chromatin accessibility assays: NOMe-seq, DNase-seq and ATAC-seq. Reprinted from [161].

As with DNA methylation, single-cell profiling methods for chromatin accessibility also emerged from its bulk counterparts, including ATAC-seq[25], NOMe-seq [174] and DNase-seq [99].
Due to its cost-effective strategy, single-cell ATAC-seq has become the most established technique to map open chromatin regions, revealing extensive heterogeneity across different cellular populations [48, 33, 38].
Compared to bulk ATAC-seq, scATAC-seq libraries are notably sparse. In a satured library, [48] reported a range of ≈ 500 to ≈ 70,000 mapped reads per cell, with a median of ≈ 2500. This represents less than 25% of the molecular complexity expected from 500-cell bulk experiments. Yet, despite the low coverage, they showed that cell-type mixtures can be confidently deconvoluted. Later, in a pioneer effort, [47] generated an atlas of chromatin accessibiliy for different mouse tissues, defining the first *in vivo* landscape of the regulatory genome single-cell resolution.
However, the sparse nature of scATAC-seq makes it impractical for the study of heterogeneity in individual regulatory elements. This is addressed mostly by NOMe-seq, which at the expensive of

more expensive sequencing and limited scalability, it provides base-pair resolution readouts, even at the single cell resolution [174].

1.1.3 Multi-modal single-cell sequencing

Cellular phenotypes have been historically characterised using exploratory methods in single molecular layers, most commonly the transcriptome. However, phenotypes result from the combination of multiple sources of biological information, including the genetic background, epigenetic marks, protein levels or lipid composition. 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 processses as well as clinically relevant traits [78, 188]. Motivated by this assumption, multiple omics are being increasingly applied across a wide range of biological domains, including cancer biology [1, 68], regulatory genomics [36], microbiology [111] or host-pathogen interactions [200].
Recent technological advances have enabled the profiling of multiple omics in the same single cell, which has the potential to provide a more comprehensive understanding of biological processess, including mechanistic relationships between the (epi)genome and the transcriptome.
Multi-modal sequencing technologies are becoming rapidly available and its explosion is likely to mirror the trend of scRNA-seq technologies. As reviewed in [208, 35], multi-modal measurements can be obtained using four broad strategies:

- **Application of a non-destructive assay before a destructive one**: the most prominent example is multiparameter fluorescence-activated cell sorting (FACS) followed by destructive high-throughput sequencing [167]. Although simple and efficient, this approach requires prior knowledge of gene expression markers to sort the populations of interest and is limited by the spectral overlap of fluorescence reporters.

- **Physical isolation of different cellular fractions followed by independent high-throughput sequencing**: this technique was pioneered with the simultaneous genome and transcriptome sequencing (G&T-seq) [139]. After cell lysis, the mRNA fraction is separated from the genomic DNA fraction using biotinylated or paramagnetic oligo(dT) beads, followed by conventional uni-modal sequencing of the mRNA and the DNA. This strategy allows the simultaneous profiling of transcriptomic measurements with genome-derived measurements, including DNA sequence, copy number variation, DNA methylation or chromatin accessibility [139, 90, 5, 91]
  A clear advantage of this methodology is its unsupervised nature, which makes it particularly useful for the study of tissues with complex heterogeneity.

- **Conversion of different molecular layers to a common format that can be simultaneously measured using the same readout**: this represents a powerful methodology, as it allows multiple modalities to be profiled in a single workflow. Prominent examples are the simultaneous measure of surface proteins and mRNA expression as in Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq[206]) and RNA expression and protein sequencing assay (REAP-seq[168]). The idea is to incubate cells with antibodies tagged with oligonucleotides that target specific proteins. Subsequently, additional barcodes are introduced

for PCR amplification and a poly(A)-tail selection step is performed, allowing the simultaneous pooling of mRNAs (cDNAs) and antibody tags. Finally, After PCR amplification, both types of molecules can be separated by size and sequenced.
In practice, this principle is significantly more powerful than FACS, as the DNA barcodes can be resolved with very high sensitivity by sequencing and their complexity is limited by the number of nucleotides (4^N where N is the length of the barcode)
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.

Combinations of the three approaches above have also been achieved. One of them, single-cell nucleosome, methylome and transcriptome (scNMT-seq[40]), which is presented in this thesis, combines the principle of physical isolation with the conversion to a common format to achieve a simultaneous measurement of three molecular layers.
(IMPROVE) Although they have been proven successful, single-cell multi-modal approaches still face numerous difficulties, both from the experimental and the computational side, 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. A clear example is scNMT-seq, where an almost %50 decrease of coverage is observed in DNA methylation with respect to scM&T-seq. Similarly, chromatin accessible measurements in sci-CAR[31] showed $\tilde{1}$0-fold less complexity than scATAC-seq.
As cost and limited scalability will likely remain the main barrier for high-resolution multi-modal technologies, we expect novel strategies to be aimed at exploiting the rich amount of information by merging or mapping the multi-modal data sets to large uni-modal atlases [184, 32, 47, 170].
Computational challenges and strategies in multi-modal data will be discussed in Chapters 2 and 4.

1.2 scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation and RNA expression in single cells

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 [118, 70, 163, 166]. Yet, the accompanying epigenetic changes and the role of other molecular layers in driving cell fate decisions 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 [208, 107, 70].

In this chapter we will describe scNMT-seq, an experimental protocol for genome-wide profiling of RNA expression, DNA methylation and chromatin accessibility in single cells. First, we will validate the quality of the triple omics readouts, followed by a comparison with similar available technologies. Subsequently, we will 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 Wolf Reik lab (Babraham Institute, Cambridge, UK). It has been peer-reviewed and published in Clark & Argelaguet et al [40].

The methodology was conceived by Stephen Clark, who performed most of the experiments. Felix Krueger processed and managed sequencing data. I performed the computational analysis, except for the reconsutraction of chromatin accessibility profiles, which was done by Andreas Kapourani. Guido Sanguinetti, Gavin Kelsey, John C. Marioni, Oliver Stegle and Wolf Reik supervised the project.

The article was jointly written by Stephen Clark, John Marioni, Oliver Stegle and Wolf Reik, and me.

1.2.1 Description of the experimental protocol

scNMT-seq builds upon two previous multi-modal single-cell sequencing protocols: single-cell Methylation and Transcriptome sequencing (scM&T-seq) [5] and Nucleosome Occupancy and Methylation sequencing (NOMe-seq) [106, 174]. An overview of the experiment protocol is shown in Figure 1.5.

In the first step (the NOMe-seq step), cells are sorted into individual wells and incubated with a GpC methyltransferase (M.CviPI). As shown *in vitro*, this enzyme labels accessible (or nucleosome depleted) GpC sites via DNA methylation[110, 106]. 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[110, 106].

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[198], whereas the RNA fraction undergoes Smart-seq2 [169].

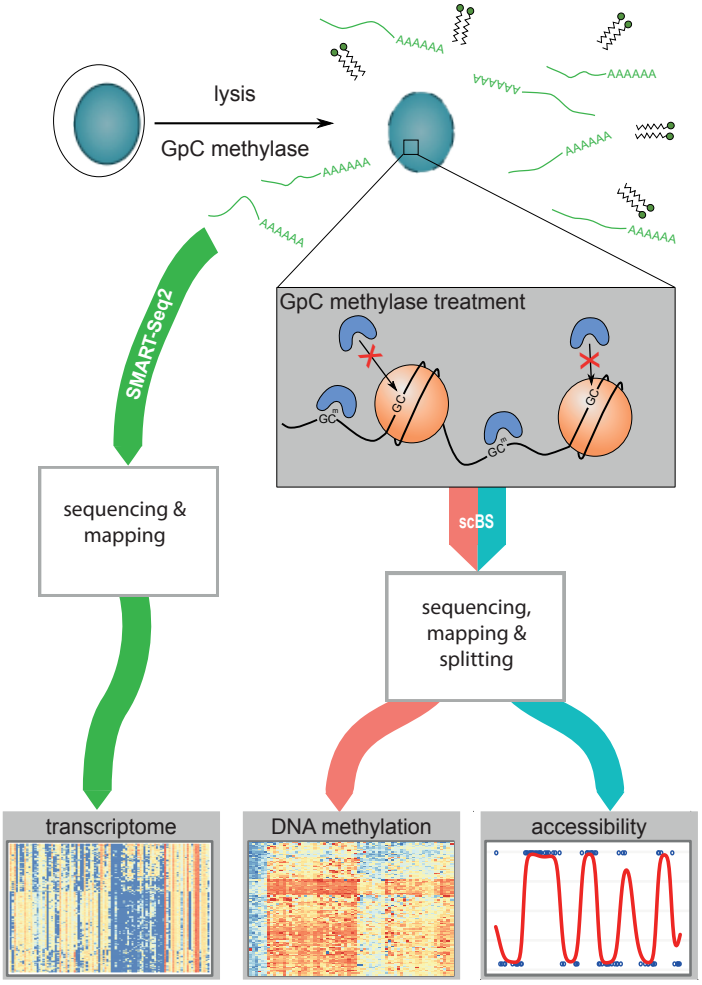


Fig. 1.5 scNMT-seq protocol overview. In the first step, cells are isolated and lysed. Second, cells are incubated with a GpC methyltransferase. Third, the RNA fraction is separated using oligo-dT probes and sequenced using Smart-seq2. The DNA fraction undergoes scBS-seq library preparation and sequencing. 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 [106, 174, 161]. 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 [161].

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 [106]. 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 [120] 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 based on mapping efficiency, bisulfite conversion efficiency and library size.

Finally, methylation calls for each CpG and GpC site are extracted after removal of duplicate alignments. Following the approach of [198], 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.

When quantifying DNA methylation and chromatin accessibility rates over genomic features (i.e. promoters or enhancers), a binomial node is assumed again per cell and feature, but aggregating the counts over all CpG (methylation) and GpC (accessibility) dinucleotides overlapping the genomic feature.

1.2.3 Validation

1.2.3.1 Coverage

We validated scNMT-seq using 70 EL16 mouse embryonic stem cells (ESCs) cultured in serum conditions, together with two controls: negative empty wells and 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 DNA methylation and chromatin accessibility statistics from published data [198, 5, 62].

First, we compared the theoretical maximum coverage with the empirical coverage (Figure 1.6). Despite the reduction in theoretical coverage due to the removal of 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.

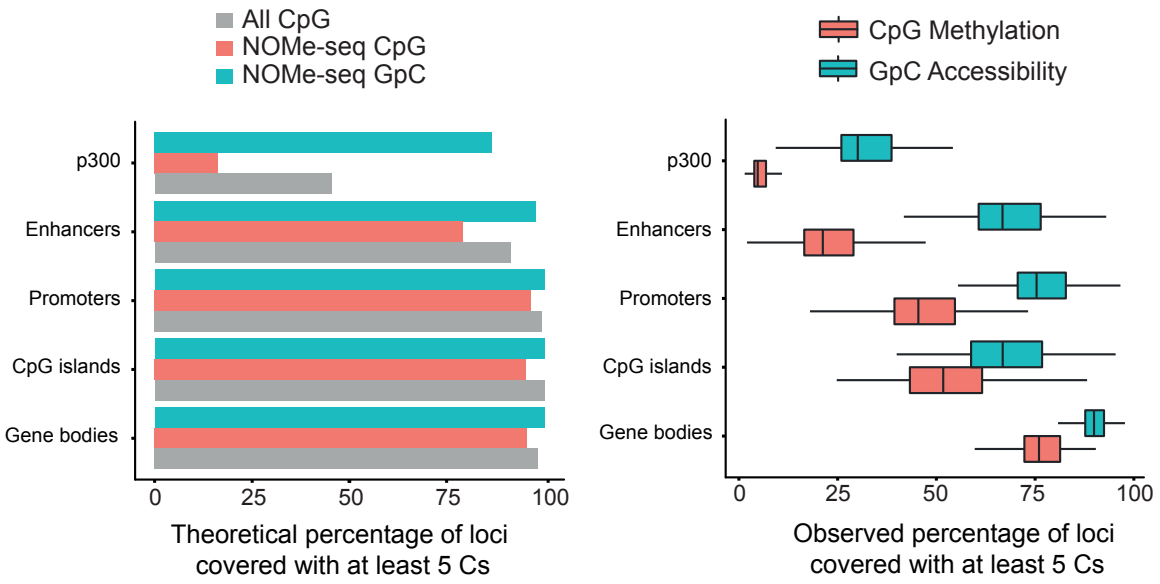


Fig. 1.6 Coverage statistics for CpG DNA methylation and GpC chromatin accessibility. (a) Fraction of loci with at least 5 CpG (red) or GpC (blue) dinucleotides (y-axis) per genomic context (x-axis), after exclusion of the conflictive trinucleotides. The grey bar shows the total number of CpGs without 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 [5] (Figure 1.6), 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 down-sampling the number of reads.

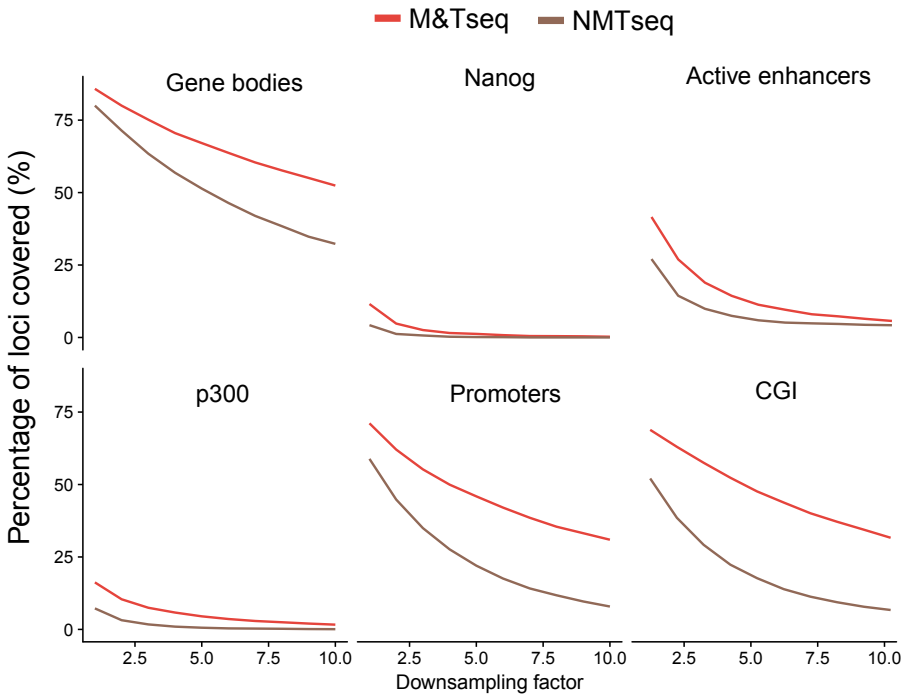


Fig. 1.7 Comparison of the empirical coverage of DNA methylation with scM&T-seq [5]. The y-axis displays the fraction of loci covered with at least 5 CpG sites. The x-axis displays the downsampling factor. To facilitate the comparison, we selected two cells that were sequenced at equivalent depth.

1.2.3.2 Consistency with previous studies

To assess the consistency with previous studies we performed several comparisons.
First, we computationally 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 promoters, DNaseI hypersensitivity sites, enhancer regions and transcription factors binding sites, DNA methylation was decreased while chromatin accessibility was increased, as previously reported [174]. As a control, we observe that cells which did not receive M.CviPI treatment showed globally low GpC methylation levels ($\approx 2\%$, **??**).

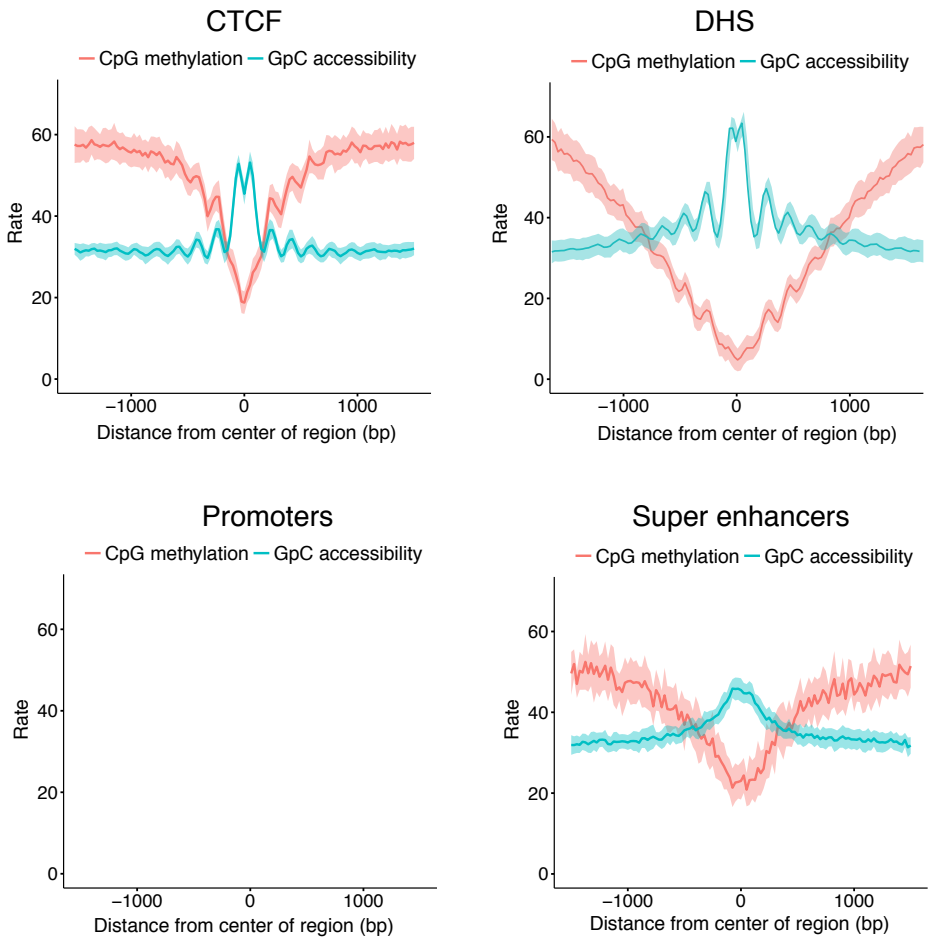


Fig. 1.8 Accessibility and methylation profiles in regulatory genomic contexts. First, we pseudobulk the data set by pooling information across all cells. Next, we compute running averages of the CpG methylation (red) and the GpC accessibility (blue) in consecutive non-overlapping 50bp windows. Solid line displays the mean across all genomic elements within a given annotation and the shading displays the corresponding standard deviation.

Second, instead of interrogating pre-defined genomic annotations, we quantified DNA methylation and chromatin accessibility using a running window throughout the genome. The resulting measurements were compared to data sets from the same cell lines profiled with similar technologies, including scM&T-seq[5], scBS-seq[198] and bulk BS-seq[62].

When comparing the resulting methylomes, we find that most of the variation is not attributed to the technology but to differences in culture condition (serum vs 2i media). Cells grown in 2i media remain in a native pluripotency state with genome-wide DNA hypomethylation [62], whereas cells in serum media transition to a primed pluripotency state poised for differentiation [214].

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 [5, 198, 62]). Previous studies have sown that female ESCs tend to show lower levels of mean global methylation, which is consistent with a more pluripotent phenotype [239].

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 [**XX**], yielding good consistency between datasets (weighted Pearson R = 0.74).

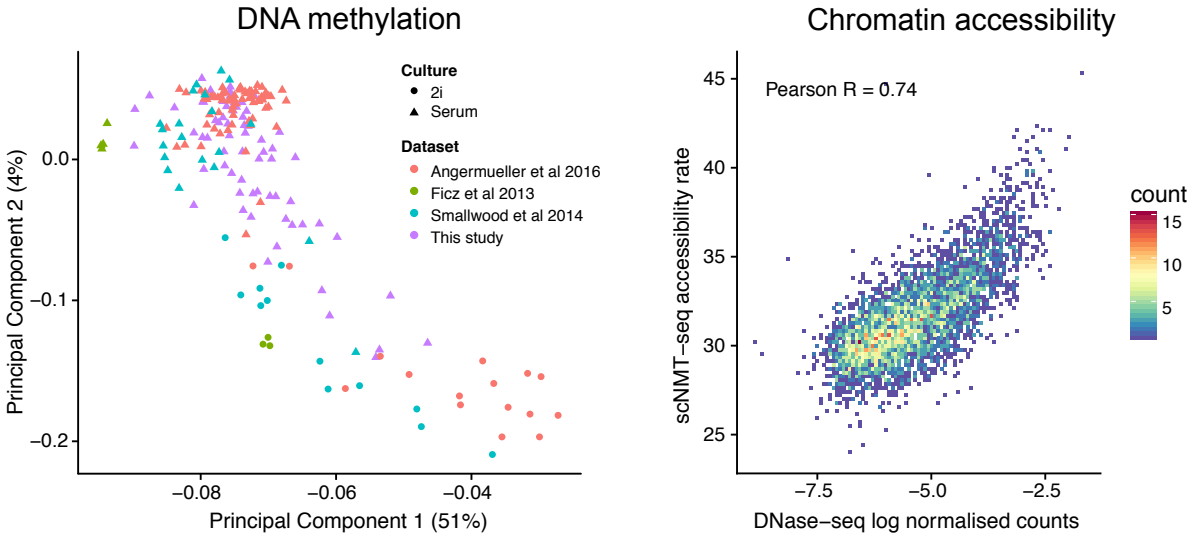


Fig. 1.9 Comparison of unsupervised genome-wide quantifications to published data sets. (a) Principal Component Analysis of 1kb running windows. Missing values were imputed using the average methylation rate per locus. (b) Scatter plot of chromatin accessibility quantified over 10kb running windows of 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.

Finally, 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 [211, 5]. 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.

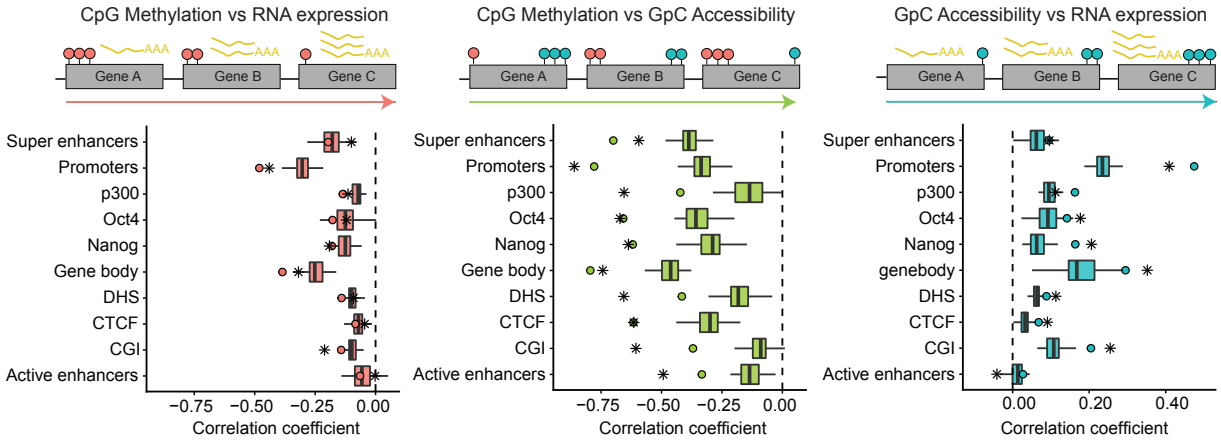


Fig. 1.10 Quantification of linear associations between molecular layer. The top diagram illustrates the computation of an association test per cell (across all loci in a given genomic context). The left panel shows DNA methylation vs RNA expression. The middle panel shows DNA methylation vs chromatin accessibility. The right panel shows RNA expression vs 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 [62, 43], respectively.

Consistently, when stratifiying the loci from Figure 1.8 based on the expression level of the nearest gene, we observe that higher RNA expression is associated with chromatin openness and decreased DNA methylation levels Figure 1.11.

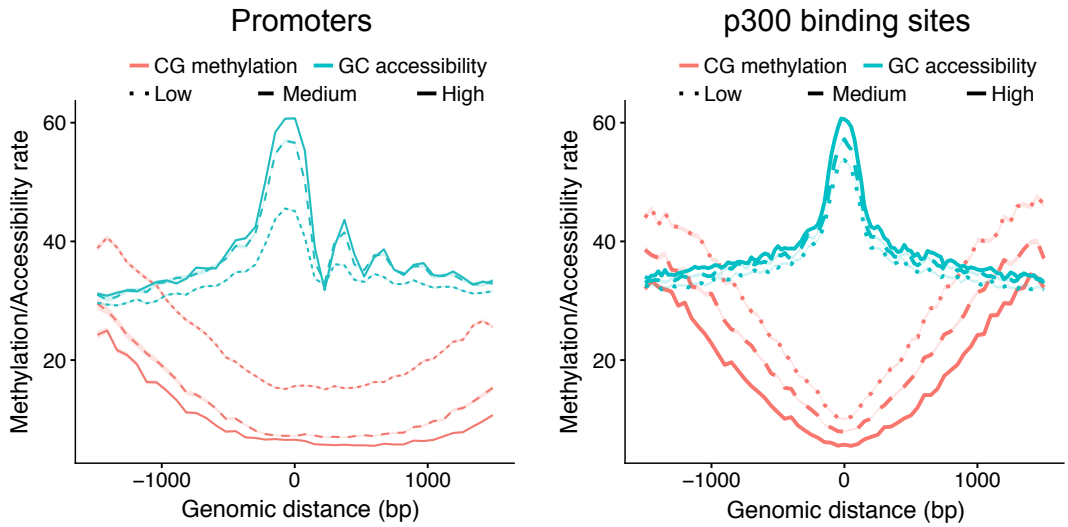


Fig. 1.11 DNA methylation (red) and chromatin accessibility profiles (blue), stratified by RNA expression of the nearest gene. The profiles are quantified as in Figure 1.8. The RNA expression is discretised in three groups: log normalised counts less than 2 (low), between 2 and 6 (medium) and higher than 5 (high)

1.2.4 Identification of genomic elements with coordinated variability across molecular layers

Having validated the quality of scNMT-seq data, we next explored its potential to identify coordinated heterogeneity across different molecular layers.

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

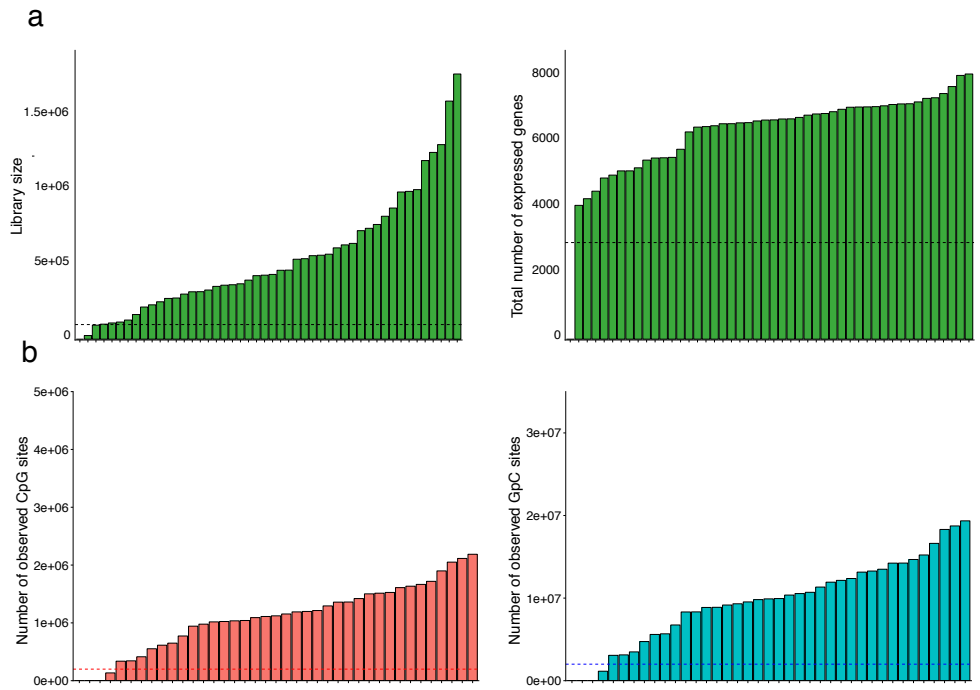


Fig. 1.12 Quality control statistics on the 43 embryoid body cells. (a) number of aligned RNA reads per cell (left) and number of expressed genes (right). (b) Number of CpG sites (DNA methylation) observed. (c) Number of GpC sites (chromatin accessibility) observed. Dashed lines indicate binary thresholds, such that cells below the threshold are excluded for the downstream analysis.

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). This confirms the existence of significant phenotypic heterogeneity that is required for an association analysis.

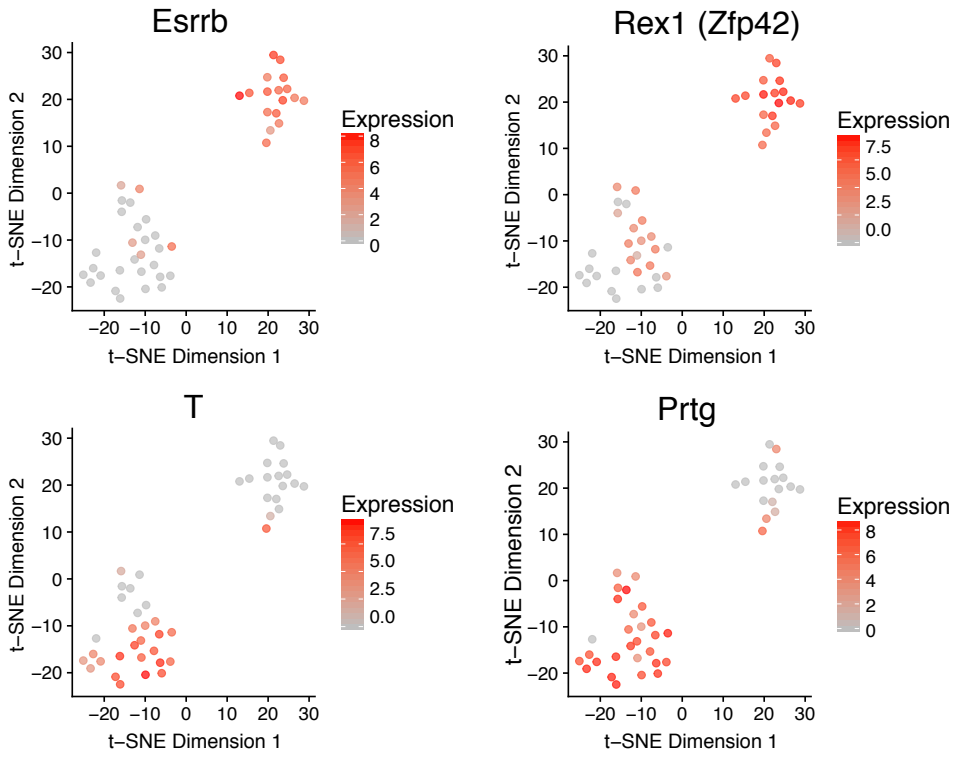


Fig. 1.13 t-SNE Dimensionality reduction on the RNA expression profiles for the embryoid body cells. The scatter plots show t-SNE dimensions 1 (x-axis) and 2 (y-axis). Cells are coloured based on expression of canonical (top) pluripotency factors and (bottom) differentiation markers.

Next, we tested locus-specific linear associations (across cells) between pairwise combinations of molecular layers, using the average CpG rate and GpC rate within a loci as a metric for DNA methylation and chromatin accessibility, respectively.

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, as previously reported [**XX**].

Again, this confirms the expected direction of association between molecular layers, as reported in bulk studies. Yet, the single-cell measurements allow us to inspect the dynamics of single loci (across cells). As an illustrative example, we display the Estrogen Related Receptor Beta (Esrrb) locus, a gene involved in early development and pluripotency [164]. A previous study [5], identified a super enhancer near Esrrb that showed high degree of correlation between DNA methylation and associated 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, hence replicating previous findings. Additionally, we observe a strong negative relationship between DNA methylation and chromatin accesibility, indicating the two epigenetic layers are tightly coupled, consistent with the correlations per cell (across genes), displayed in Figure X.

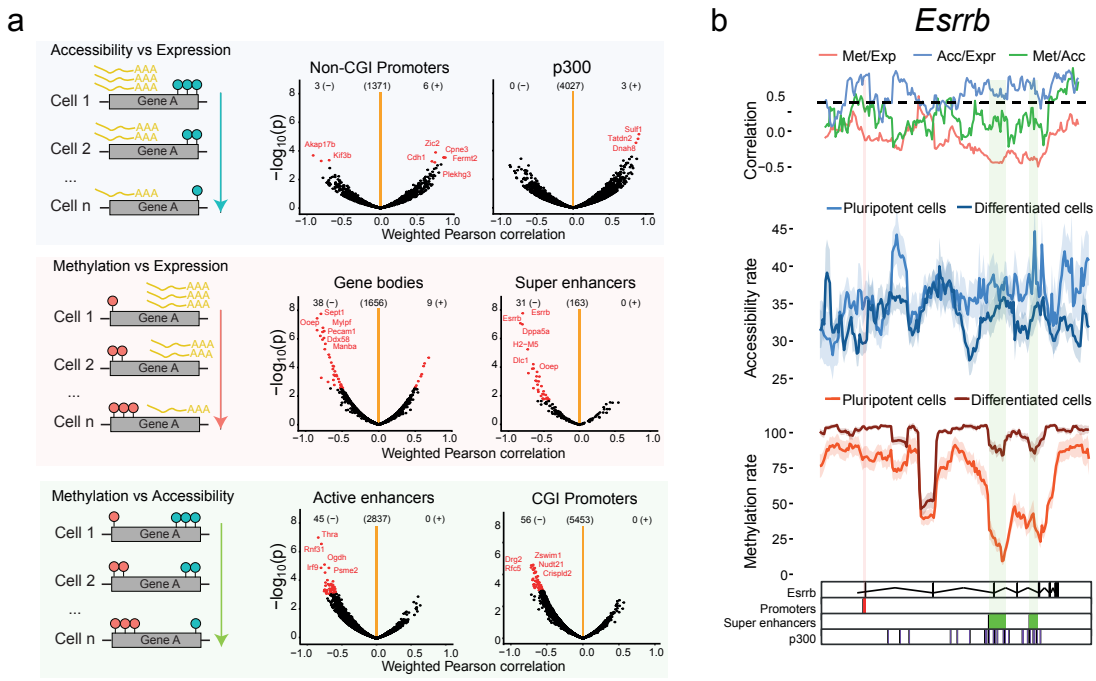


Fig. 1.14

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

A clear advantage of scNMT-seq, compared to other chromatin accessibility technologies, is the high resolution of the its readouts, namely a binary output for each observed GpC dinucleotide. As illustrated in Figure 1, GpC accessibility measurements show complex oscillatory patterns, likely due to presence of nucleosomes, which are not appropriately captured by using an average rate. 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 [103], 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 infered regression patterns are shown in Figure X, showing significant heterogeneity in both the position and the number of nucleosomes:

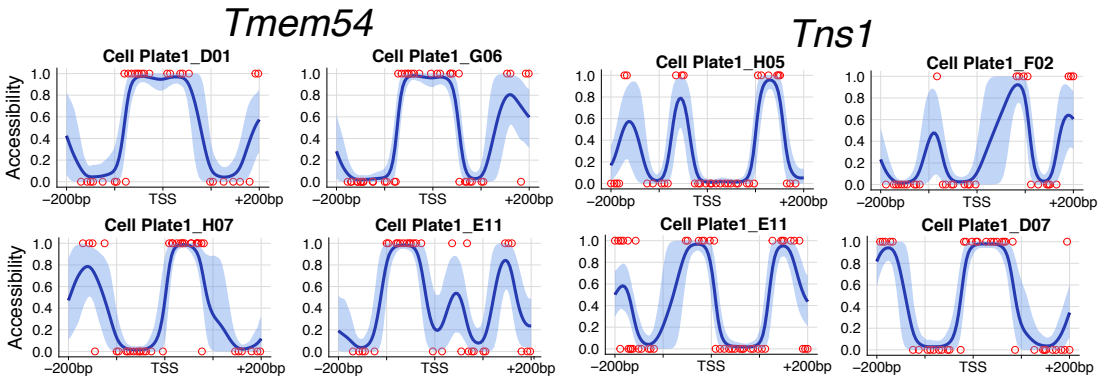


Fig. 1.15 Illustrative examples of single-cell accessibility profiles around the transcription start site. Shown are representative profiles for two genes, *Tmem54* and *Tns1*. Each panel corresponds to separate cell. The y-axis displays the binary GpC accessibility balues, with 1 being accessibility and 0 unaccessible. The x-axis displays the genomic region around the TSS (200bp upstream and downstream). The blue area depicts the infered (non-linear) accessibility profile using the BPRMeth model [103].

As a first validation step, we showed that the accessibility profiles infered around the transcription start site (TSS) lead to a better prediction of RNA expression than using conventional accessibility rates (Figure 1.16).

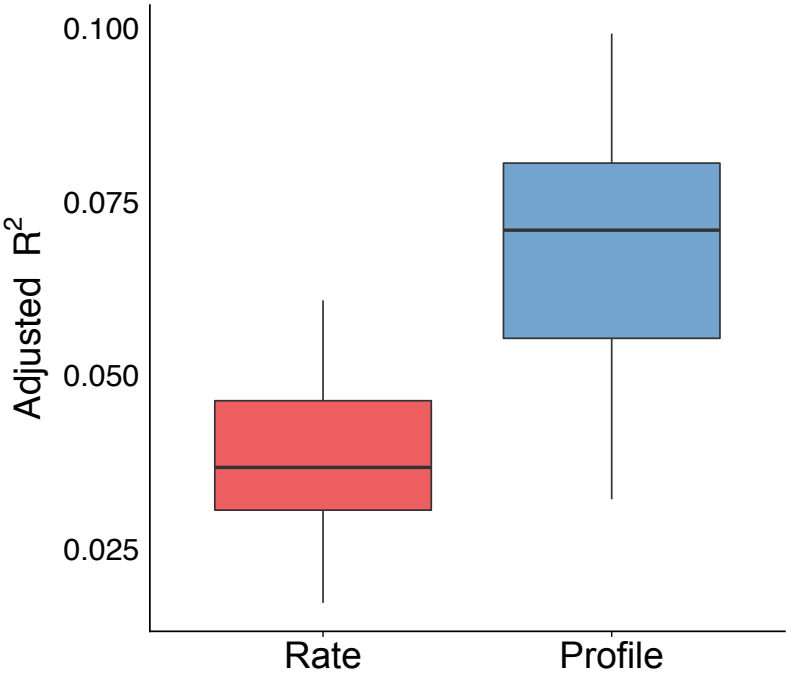


Fig. 1.16 Prediction of RNA expression using conventional accessibility rates (red) and non-linear accessibility profiles (blue). The y-axis displays the adjusted R^2 between observed RNA expression and predicted RNA expression. We fit a linear model per cell (across genes) where the response variable is RNA expression (log normalised counts) and the covariates are either the accessibility rate or the weights of the basis functions in BPRmeth.

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

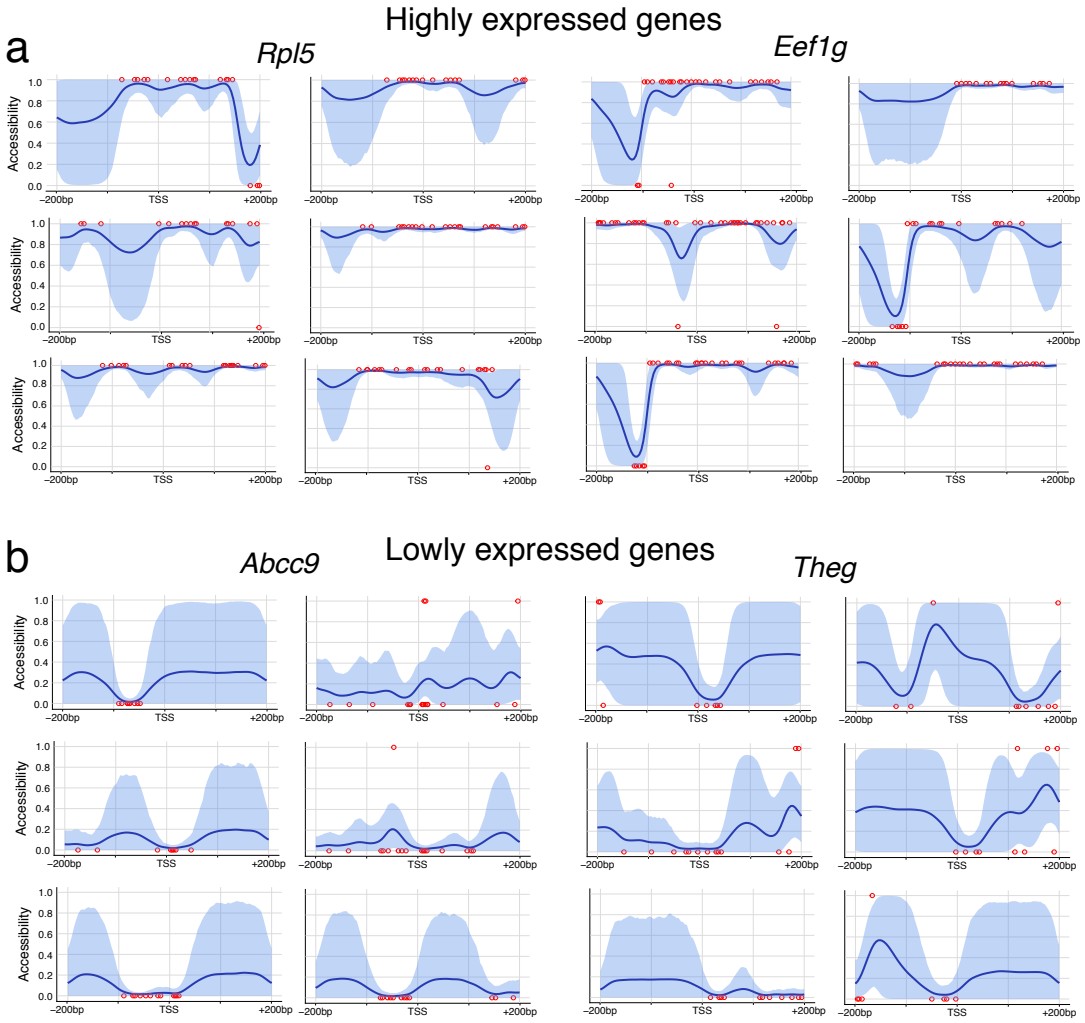


Fig. 1.17 Single-cell accessibility profiles of representative genes with high expression levels (top, *Rpl5* and *Eef1g*) and low expression levels (bottom, *Abcc9* and *Theg*). Each panel corresponds to a separate cell. Axis are the same as in Figure 1.15.

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.

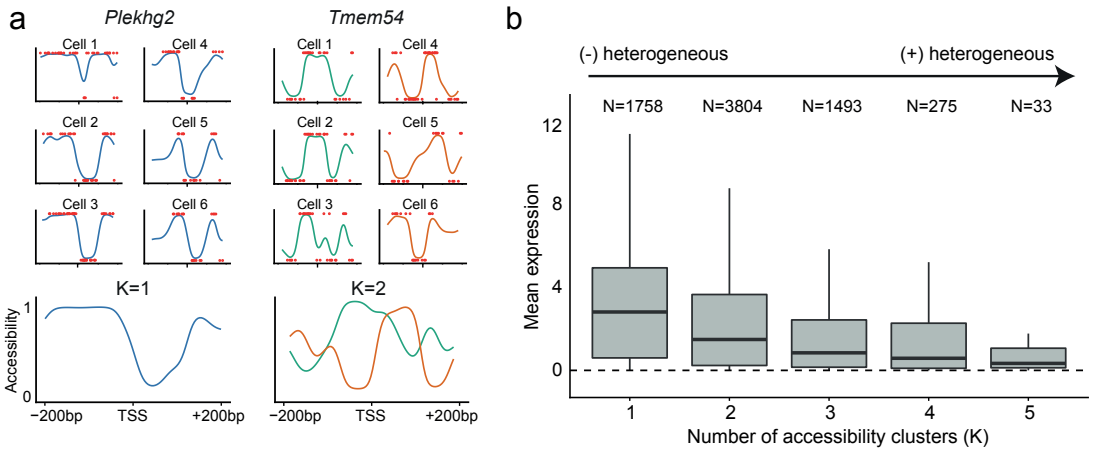


Fig. 1.18

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 [196].
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 [220, 15]

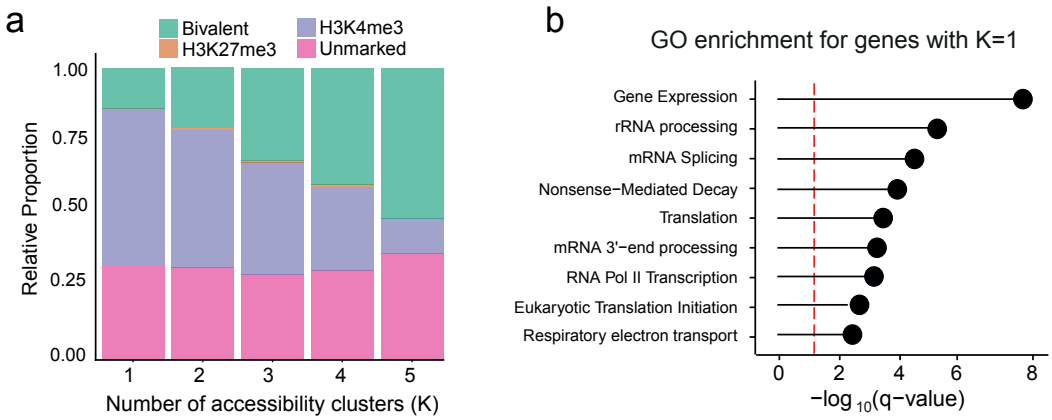


Fig. 1.19

1.2.6 Characterisation 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 [215, 75,

191]. A novel opportunity unveiled by the introduction of single-cell multi-modal technologies is the study of epigenetic dynamics along trajectories infered from the transcriptome.
To explore this idea, we applied a diffusion-based pseudotime method[75] to the EB data set, using the RNA expression of the 500 genes with highest biological overdispersion[137]. The resulting first diffusion component was used to reconstruct a pseudotemporal ordering of cells from pluripotent to differentiated states:

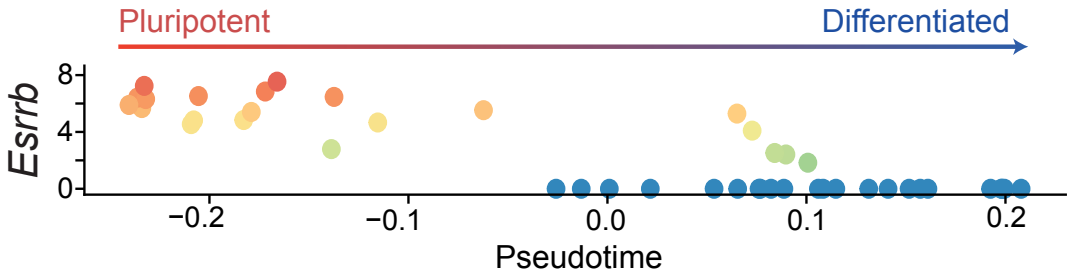


Fig. 1.20 Reconstruction of developmental trajectory in embryoid body cells from the RNA expression data. Each dot corresponds to one cell. The y-axis displays expression of Esrrb, a canonical pluripotency marker, and the x-axis shows the position of the cells in the first diffusion component.

Next, we investigated whether the strength of association between molecular layers (as calculated in Figure 1.10) 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 1.21, 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.

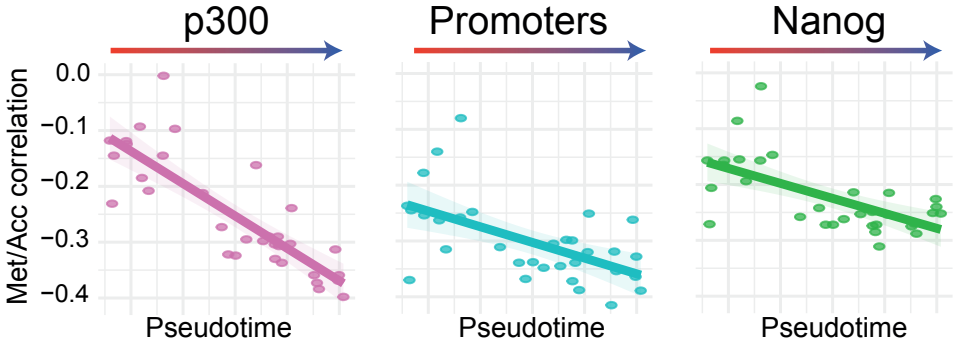


Fig. 1.21

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 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 [32, 33, 71]. As in scRNA-seq, the maturation of multi-modal techniques will have a trade-off between sensitivity and scalability [35]. Under the assumption that epigenetic regulation is more strongly driven by local process compared to global interactions[**XX**], 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 [**droplet**, **sci-met**]. 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.
  Finally, sequencing costs are expected to decrease (even faster than predicted by Moore's law) [209]. Yet, the generation of scNMT-seq libraries remain inexorably more expensive than any scRNA-seq technology. Hence, efforts to decrease the library size by a pre-selection of the genetic material might be indispensable. Examples of such strategies are the digestion by restriction enzymes as in RRBBS [73], an initial round of ATAC protocol to select open chromatin [202] or 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 [4]. 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 [4]. A successful single-cell strategy based on deep learning has been proposed (DeepCpG[4]), 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 [104]), 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 [173, 60, 146, 56]. 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[92], allele-specific fractions[51] and RNA velocity information [121].

From the experimental side, NMT-seq can theoretically be combined with novel single-cell assays that quantify transcription factor binding [155] and histone modifications [105].

- **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 [106]. 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 (???) 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 [125] 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.

Chapter 2

Integrative analysis of single-cell multi-modal data

2.1 Introduction

TO-DO...

2.2 Probabilistic modelling

A scientific model is a simple theoretical representation of a complex natural phenomenon to allow the systematic study of its behaviour. The general idea is that if a model is able to explain some observations, it might be capturing its true underlying laws and can therefore be used to make future predictions.
In particular, statistical models are a powerful abstraction of nature. They consist on a set of observed variables and a set of (hidden) parameters. The procedure of fitting the parameters using a set of observations is called inference or learning.
One of the major challenges of inference when dealing with real data sets is the distinction between signal and noise. An ideal model should learn only the information relevant to gain explanatory power while disregarding the noise. However, this is a non-trivial task in most practical situations. Very complex models will tend to overfit the training data, capturing large amounts of noise and consequently leading to a bad generalisation performance to independent data sets. On the other hand, simplistic models will fit the data poorly, leading to poor explanatory power.
The ideas above can be formalised using the framework of probability and statistics.

2.2.1 Maximum likelihood inference

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

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

This process is called maximum likelihood learning[205, 19, 157]. However, in this setting there is no penalisation for model complexity, making maximum likelihood solutions prone to overfit in cases where the data is relatively sparse. Generalisations that account for model complexity have been proposed and include regularising terms that shrink parameters to small values. However, these are often particular cases of the more general framework of Bayesian statistics [80, 19, 157].

2.2.2 Bayesian inference

In the Bayesian framework, the parameters themselves are treated as random unobserved variables and we aim to obtain probably distributions for $\boldsymbol{\theta}$, rather than a single point estimate. To do so, prior beliefs are introduced into the model by specifying a prior probability distribution $p(\boldsymbol{\theta})$. Then, using Bayes' theorem [13], the prior hypothesis is updated based on the observed data $\mathbf{Y}$ by means of the likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$ function, which yields a posterior distribution over the parameters:

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

where $p(\mathbf{Y})$ is a constant term called the marginal likelihood, or model evidence [19, 157].
The choice of the prior distribution is a key part of Bayesian inference and captures beliefs about the distribution of a variable before the data is taken into account. With asymptotically large sample

sizes, the choice of prior has negligible effects on the posterior estimates, but it becomes critical with sparse data [19, 157, 67].
There are two common considerations when defining the prior distributions. The first relates to the incorporation of subjective information, or predefined assumptions, into the model. For example, one could adapt the prior distribution to match the results from previous experiments (i.e. an informative prior). Alternatively, if no information is available one could set set uninformative priors by following maximum entropy principles [97].
The second strategy is based on convenient mathematical properties to make inference tractable. If the likelihood and the prior distributions do not belong to the same family of probability distributions (they are not conjugate) then inference becomes more problematic [177, 19, 157, 67]. The existence of conjugate priors is one of the major reasons that justify the widespread use of exponential family distributions in Bayesian models [67]. An example is the Automatic Relevance Determination prior discussed in **??**.
Again, the milestone of Bayesian inference is that an entire posterior probability distribution is obtained for each unobserved variable. This has the clear advantage of naturally handling uncertainity in the estimation of parameters. For instance, when making predictions, a fully Bayesian approach attempts to integrate over all the possible values of all unobserved varaibles, effectively propagating uncertainity across multiple layers of the model. Nevertheless, this calculation is sometimes intractable and one has to resort to point estimates [19, 157, 67]. The simplest approximation to the posterior distribution is to use its mode, which leads to the maximum a posteriori (MAP) estimate:

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

This is similar to the maximum likelihood objective function, but with the addition of a term $p(\boldsymbol{\theta})$. When the prior dsitribution is chosen smartly, this term penalises for model complexity. Therefore, in contrast to standard (non-penalised) maximum likelihood inference, Bayesian approaches naturally handle the problem of model complexity and overfitting[19, 157, 67]. At the limit of infinite observations, the influence of the prior to the posterior is negligible and the MAP estimate converges towards the Maximum likelihood estimate, hence providing a rational link between the two inference frameworks.

2.2.3 Deterministic approaches for Bayesian inference

The central task in Bayesian inference is the direct evaluation of the posterior distributions and/or the computation of expectations with respect to the posterior distributions. In sufficiently complex models, closed-form solutions are not available and one has to resort to approximation schemes, which broadly fall into two classes: stochastic or deterministic [67, 21].
Stochastic approaches hinge on the generation of samples from the posterior distribution via a Markov Chain Monte Carlo (MCMC) framework. Such techniques have the appealing property of exact results at the asymptotic limit of infinite computational resources. However, in practice, sampling approaches are computationally demanding and suffer from limited scalability to large data sets [21].
In contrast, deterministic approaches are based on analytical approximations to the posterior

distribution, which often lead to biased results. Yet, given the appropriate settings, these approaches are potentially much faster and scalable to large applications [19, 157, 21].

2.2.3.1 Laplace approximation

The Laplace approximation is probably the simplest of the deterministic tecniques, where the aim is to construct a Gaussian approximation around the mode of the true posterior distribution using a second-order Taylor expansion [19, 157].
Suppose $\mathbf{X}$ contains all unobserved variables. The true posterior distribution can be written as:

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

where $f(\mathbf{X})$ is a function that depends on the unobserved variables and Z is an unknown normalisation constant to ensure that $\int p(\mathbf{X})d\mathbf{X} = 1$.
The second-order Taylor expansion of $\log f(\mathbf{X})$ centered around its (known) mode $\hat{\mathbf{X}}$ is:

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

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

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

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

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

where d is the number of unobserved variables.
Despite its simplicity, the Laplace approximation is a useful strategy that has been successfully applied in practice. Nonetheless, this approximation has notable caveats: first, is limited by its own local definition, ignoring all the density beyond the mode of the posterior. Second, it does not apply to discrete variables. Third, the inversion of the Hessian is very expensive in high-dimensional settings.

2.2.4 Variational inference

Variational inference is a deterministic family of methods that have been receiving widespread attention due to a positive balance between accuracy, speed, and ease of use [21, 230]. The core framework is derived below.

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

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

Note that the KL divergence is not a proper distance metric, as it is not symmetric. In fact, using the reverse KL divergence $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$ defines a different inference framework called expectation propagation [149].
If we allow any possible choice of $q(\mathbf{X})$, then the minimum of this function occurs when $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$. Nevertheless, since the true posterior is intractable to compute, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of distributions $q(\mathbf{X})$ that are tractable to compute and subsequently seek the member of this family for which the KL divergence is minimised.
Doing some calculus it can be shown (see [19, 157]) that the KL divergence $\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y}))$ is the difference between the log of the marginal probability of the observations $\log(\mathbf{Y})$ and a term $\mathcal{L}(\mathbf{X})$ that is typically called the Evidence Lower Bound (ELBO):

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

Hence, minimising the KL divergence is equivalent to maximising $\mathcal{L}(\mathbf{X})$ Figure 2.1:

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X})\Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y})\Big) d\mathbf{X} \\ &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Y})] - \mathbb{E}_q[\log q(\mathbf{X})]\end{aligned} \tag{2.1}$$

The first term is the expectation of the log joint probability distribution with respect to the variational distribution. The second term is the entropy of the variational distribution. Importantly, given a simple parametric form of $q(\mathbf{X})$, each of the terms in Equation (2.1) can be computed in closed form.
In some occasions (see section X), we will use the following form for the ELBO:

$$\mathcal{L}(\mathbf{X}) = \mathbb{E}_q[\log p(\mathbf{Y}|\mathbf{X})] + (\mathbb{E}_q[\log p(\mathbf{X})] - \mathbb{E}_q[\log q(\mathbf{X})]) \tag{2.2}$$

where the first term is the expectation of the log likelihood and the second term is the difference in the expectations of the p and q distributions of each hidden variable.
In conclusion, variational learning involves minimising the KL divergence between $q(\mathbf{X})$ and $p(\mathbf{X}|\mathbf{Y})$ by instead maximising $\mathcal{L}(\mathbf{X})$ with respect to the distribution $q(\mathbf{X})$. The following image summarises the general picture of variational learning:

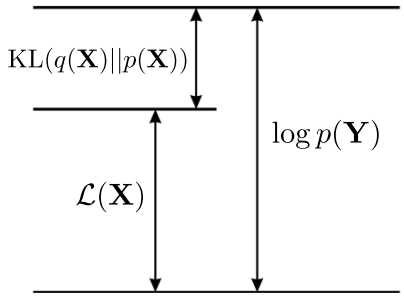


Fig. 2.1 The quantity $\mathcal{L}(\mathbf{X})$ provides a lower bound on the true log marginal likelihood $\log p(\mathbf{Y})$, with the difference being given by the Kullback-Leibler divergence KL$(q||p)$ between the variational distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$

Yet, there are several approaches to define $q(\mathbf{X})$. The two most commonly used are called (unparametric) mean-field and (parametric) fixed-form [230, 21].

2.2.4.1 Mean-field variational inference

The most common type of variational Bayes, known as the mean-field approach, assumes that the variational distribution factorises over M disjoint groups of unobserved variables[193]:

$$q(\mathbf{X}) = \prod_{i=1}^{M} q(\mathbf{x}_i) \tag{2.3}$$

Typically all unobserved variables are assumed to be independent, extensions that consider structured groups of variables have also been considered[12, 194, 84]. Importantly, notice that no parametric assumptions were placed regarding the nature of $q(\mathbf{x}_i)$.
Evidently, in sufficiently complex models where the unobserved variables have dependencies this family of distributions do not contain the true posterior (Figure 2.2). Yet, this is a key assumption to obtain an analytical inference scheme that yields surprisingly accurate results [20, 59, 23].

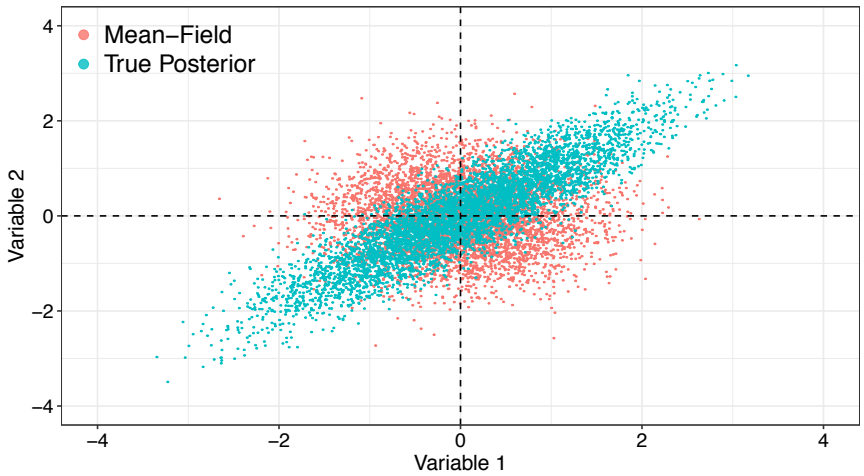


Fig. 2.2 Illustrative example of sampling from a true posterior distribution (blue) versus a fitted mean-field varaitional distribution (red) in a model with two (correlated) unobserved variables. The mean-field approximation wrongly assumes that the unobserved variables are independent.

Using calculus of variations (see [19, 157]), it follows that the optimal distribution $q(\mathbf{X})$ that maximises the lower bound $\mathcal{L}(\mathbf{X})$ is

$$\log \hat{q}_i(\mathbf{x}_i) = \mathbb{E}_{-i}[\log p(\mathbf{Y}, \mathbf{X})] + \text{const} \tag{2.4}$$

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

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

While the form of $\hat{q}(\mathbf{x}_i)$ is not restricted to a specific parametric form, it can be shown that when using conjugate priors, the distributions $\hat{q}_i(\mathbf{x}_i)$ have the same functional form as the priors $\hat{p}(\mathbf{x}_i)$. An example is shown in Appendix X, but a detailed mathematical treatment with derivations of multiple examples can be found in [19, 157, 233].

2.2.4.2 Fixed-form variational inference

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

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \tag{2.5}$$

$$= \mathbb{E}[\log(q(\mathbf{X})) - \log(p(\mathbf{X}, \mathbf{Y}))] \tag{2.6}$$

Numerically optimising this function requires the evaluation of expectations with respect to $q(\mathbf{X})$. In closed form, this is only feasable for a limited group of variational distributions. Alternatively, one can attempt Monte Carlo approximations, but in practice this turns to be slow and leads to high-variance estimates [**XX**].
Typically, one would choose this distribution to factorise over parameters and to be of the same (exponential) family as the prior $p(\mathbf{X})$. In such case there is a closed form coordinate-ascent scheme available, and it turns out that the fixed-form formulation is equivalent to the (non-parametric) mean-field derivation when using conjugate priors.
Unfortunately, for generic models with arbitrary families of distributions, no closed-form variational distributions exist in neither settings [230, 21]. However, while the parametric assumption certainly limits the flexibility of variational distributions, the advantage of this formulation is that it unveils the possibility to use (potentially fast) gradient-based methods for the inference procedure [85, 182, 109].
Furthmore, additional approximations permit generic variational inference methods that can be applied to virtually any model definition[230, 21].

Black box variational inference (TO-DO...)

Reparametrization gradients variational inference (TO-DO...)

2.2.5 Expectation Propagation

Expectation Propagation (EP) is another deterministic strategy with a similar philosophy as the Variational approach. It is also based on minimising the KL divergence between a variational distribution $q(\mathbf{X})$ and the true posterior $p(\mathbf{X}|\mathbf{Y})$, but while variational inference minimises $KL(p||q)$, EP maximises the reverse KL-divergence $KL(q||p)$.
Interestingly, this simple difference leads to an inference scheme with stringkly different properties. This can be understood by inspecting the differences between the two KL divergence formulas:
Variational inference:

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) = -\int_z q(\mathbf{X}) \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} \tag{2.7}$$

Expectation propagation:

$$\mathrm{KL}(p(\mathbf{X}|\mathbf{Y})||q(\mathbf{X})) = -\int_z p(\mathbf{X}|\mathbf{Y}) \log \frac{q(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y})} \tag{2.8}$$

In regions of $\mathbf{X}$ where the true posterior density $p(\mathbf{X}|\mathbf{Y})$ is small, setting a large density for $q(\mathbf{X})$ has a much stronger penalisation in Equation (2.8) than in Equation (2.7), because of the true posterior density being on the denominator. Hence, EP tends to avoid areas where the density $p(\mathbf{X}|\mathbf{Y})$ is very low, even if it does not correspond to areas of very high-density in $p(\mathbf{X}|\mathbf{Y})$. In contrast, in Equation (2.7) there is a strong penalty for having low-density $q(\mathbf{X})$ values.
As discussed in [19], the practical consequences of this duality can be observed when the posterior is multi-modal, as in any sufficiently complex model. In VI, $q(\mathbf{X})$ converges towards areas of high-density in $p(\mathbf{X}|\mathbf{Y})$, namely local optima. In contrast, EP tends to capture as much non-zero density regions from $p(\mathbf{X}|\mathbf{Y})$ as possible, thereby averaging across all optima. In the context of doing predictions, the VI solution is much more desirable than the EP solution, as the average of two good parameter values is not necessarily a good parameter itself.
A detailed mathematical treatment of EP, including derivations for specific examples, can be found in [19, 157, 149]

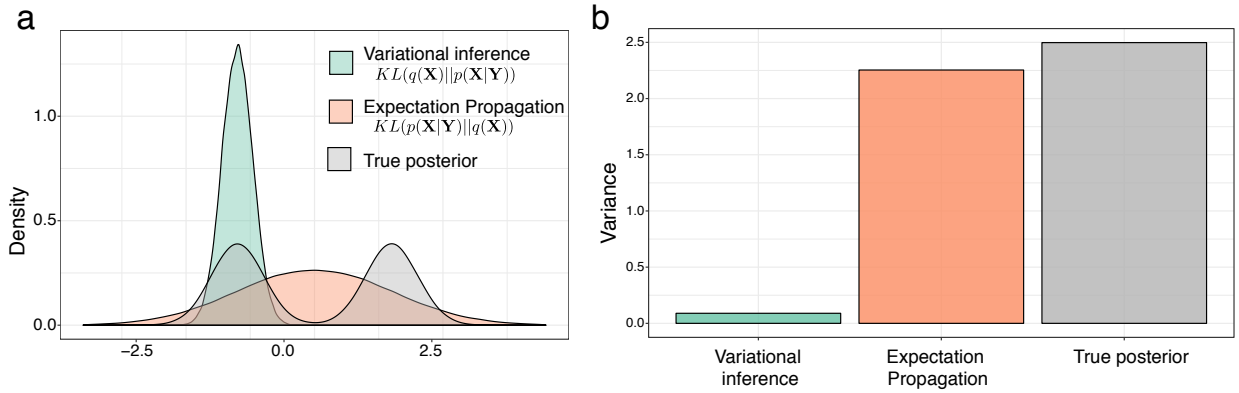


Fig. 2.3 Illustrative comparison of Variational inference and Expectation Propagation. Shown is the (a) Density and (b) Variance of the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$ (grey), the variational distribution (orange) and the expectation propagation distribution (green).

Following the rationale above, it is easy to predict that variational inference tends to be underestimate the variance of the posterior density. Yet, empirical research have shown that this is acceptable, provided that a good model selection is performed [20].

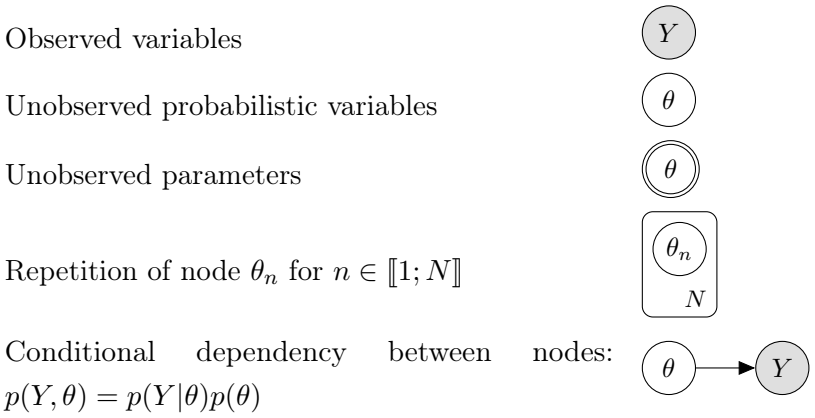
2.2.5.1 Open perspectives

Variational inference is growing in popularity for the analysis of big data sets and it has been applied to a myriad of different problems, including genome-wide association studies [34], population genetics, [178], network analysis [192], computer vision [130] and natural language processing [22]. Yet, despite its success, there is room for improvement. First, the theoretical guarantees are not as well developed as in sampling-based MCMC approaches[21, 230, 159]. For example, although it is surprisingly effective, the mean-field assumption makes strong independence assumptions about the parameters. As we have described, this leads to an underestimation of the true variance in the variational distributions, hence potentially limiting the ability of VI to propagate uncertainity when doing predictions. Additionally, it is not clear in which applications the dependencies between the parameters are important enough than the mean-field approximation could potentially break. Hence, more generally, an open research problem is understanding what are the statistical properties of the variational posterior with respect to the exact posterior [21, 230].
Alternative strategies beyond the mean-field assumption have been considered by allowing some dependencies between the variables, resulting in *structured* mean-field approximations[12, 194, 84]. However, they often lead to very complex (if not intractable) inference frameworks. In this thesis we make use of a structured mean-field assumption for the spike-and-slab prior (see Section 2.6.1.5), as initially proposed in [213].
Another area of extensive research is how to extend the applicability of VI to non-conjugate models. As discussed in section Section 2.2.3, the ELBO of non-conjugate models contains intractable integrals and closed-form variational updates cannot be derived. Amenable inference hence requires the use of either stochastic Monte Carlo approximations [**XX**] or deterministic Taylor/Laplace approximations that introduce additional lower bounds [230, 195, 55]. In this thesis we follow this rationale to derive an inference framework for a model with general likelihoods (see Section 2.6.7.4).

2.3 Graphical notations for probabilistic models

Probabilistic models can be represented in a diagrammatic format (i.e. a graph or a network) that offers a compact visual representation of complicated systems of probability distributions [19].
In a graphical model the relationship between the nodes becomes more explicit, namely their conditional independence properties which allow the joint distribution over all variables to be factorised into a series of simpler products involving subsets of variables [19].
The basic unit of a network is the node, which represents the different types of variables, including observed variables, unobserved probabilistic variables and unobserved parameters. The nodes are connected by unidirectional edges (arrows) which capture the conditional independence relationship between the variables.

For this thesis we adapted the graphical notations from [53].



2.4 Latent variable models for genomics

With the exponential growth in the use of high-throughput genomics, biological data sets are increasingly high dimensional, both in terms of samples and features. A key principle of biological data sets is that variation between the features results from differences in underlying, often unobserved, processes. Such processes, whether driven by biological or technical effects, are manifested by coordinated changes in multiple features. This key assumption sets off an entire statistical framework of exploiting the redundancy encoded in the data set to learn the (latent) sources of variation in an unsupervised fashion. This is the aim of dimensionality reduction techniques, or latent variable models (LVMs) [119, 203, 126, 175, 123, 204, 148].

2.4.1 General mathematical formulation

Given a dataset $\mathbf{Y}$ of N samples and D features, LVMs attempt to exploit the dependencies between the features by reducing the dimensionality of the data to a potentially small set of K latent variables, also called factors. The mapping between the low-dimensional space and the high-dimensional space is performed via a function $f(\mathbf{X}|\mathbf{\Theta})$ that depends on some parameters $\mathbf{\Theta}$.

The choice of $f(\mathbf{X}|\mathbf{\Theta})$ is essentially the field of dimensionality reduction. A trade-off exists between complexity and interpretation: while non-linear functions such as deep neural networks provide

more explanatory power, this leads to a considerable challenges in interpretation [232]. Hence, for most applications where interpretability is important, $f(\mathbf{X}|\mathbf{\Theta})$ is assumed to be linear [**XX**]:

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T \tag{2.9}$$

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

The inference procedure consists in learning the values of all unobserved variables, including factors and weights. As we shall demonstrate, different inference schemes and assumptions on the prior distributions lead to significantly different model outputs [183].

2.4.2 Principal component Analysis

Principal Component Analysis (PCA) is the most popular technique for dimensionality reduction [88, 186].
Starting from Equation (2.9), two formulations of PCA exist [19]. In the maximum variance formulation, the aim is to infer an orthogonal projection of the data onto a low-dimensional space such that variance explained by the projected data is maximised:

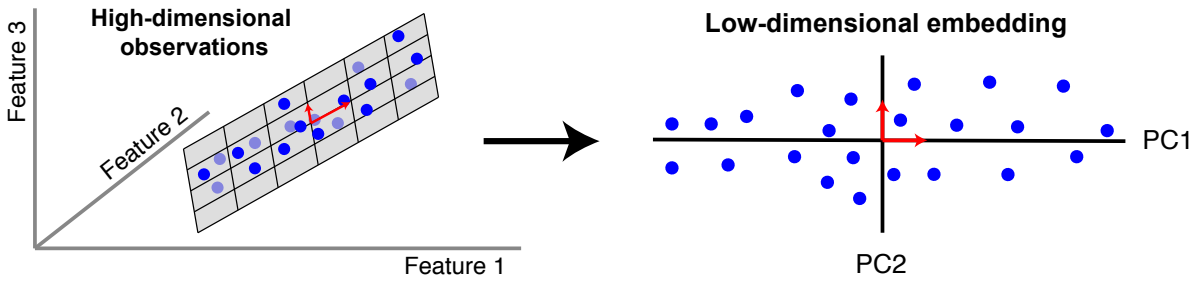


Fig. 2.4

For a single principal component, the optimisation problem is:

$$\underset{\|\mathbf{w}\|=1}{\arg\max} = \mathbf{w}_1^T\mathbf{Y}^T\mathbf{Y}\mathbf{w} \tag{2.10}$$

where $\mathbf{Y}^T\mathbf{Y} = \mathbf{S} \in \mathbb{R}^{D\times D}$ is the data covariance matrix and $\mathbf{w}_1^T$ is the vector of loadings.
The k-th principal component can be found by subtracting from $\mathbf{Y}$ the reconstructed data by the previous $k-1$ principal components. If we define $\mathbf{z}_k = \mathbf{w}_k^T\mathbf{Y}$ to be the k-th principal component:

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

Re-applying Equation (2.10) defines the new optimisation problem.

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

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

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

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

$$\mathbf{S}\mathbf{w}_k = \lambda_k \mathbf{w}_k \tag{2.11}$$

Hence, the loading vectors $\mathbf{w}_k$ are the eigenvectors of $\mathbf{S}$, which can be computed via singular value decomposition [19].

The reason why the maximum variance solution and the minimum reconstruction error solution are the same can be understood by applying Pythagoras theorem to the right triangle defined by the projection of a sample $\mathbf{y}_n$ to a loading vector $\mathbf{w}$ (Figure 2.5). Assuming again centered data, the variance of $\mathbf{y}_n$ is $\|\mathbf{y}_n\| = \mathbf{y}_n^T \mathbf{y}_n$. This variance decomposes as the sum of the variance in the latent space $\|\mathbf{z}_n\| = \mathbf{z}_n^T \mathbf{z}_n$ and the residual variance after reconstruction $\|\mathbf{y}_n - \mathbf{z}_n \mathbf{w}^T\|$:

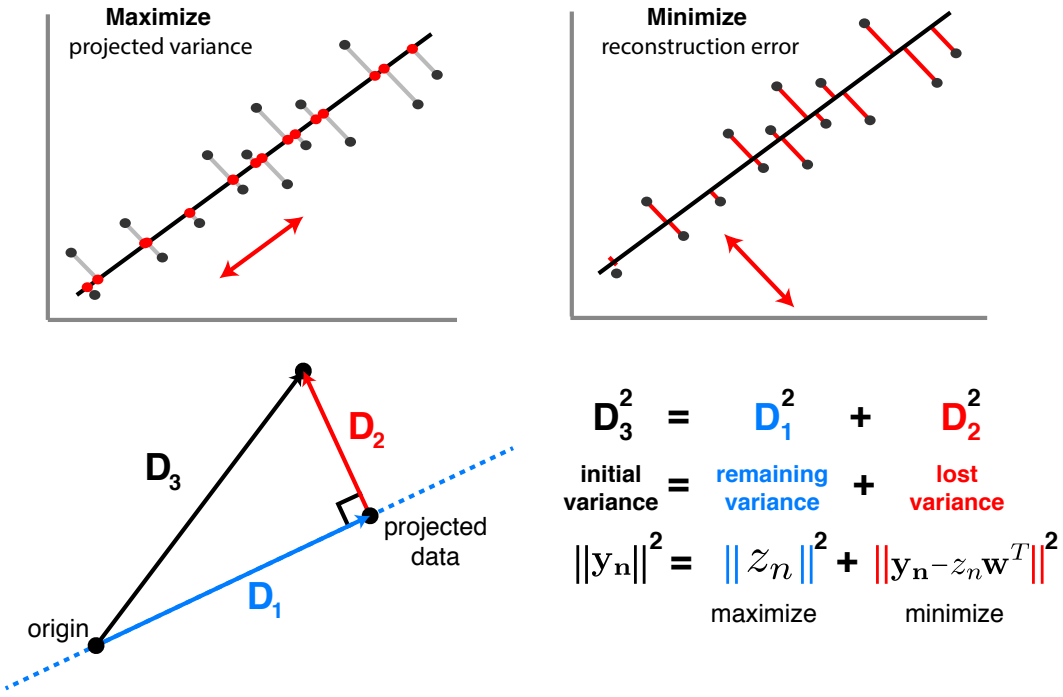


Fig. 2.5 In the maximum variance formulation we aim at maximising the variance of the projected data (blue line), whereas in the minimum error formulation we are aimed at minimising the residual variance (red line). Given a fixed total variance (black line), both strategies are equivalent

The main strength of PCA relies on its simplicity and closed form solution. Additionally, the linear mapping has the advantage of yielding interpretable loadings, so that inspection of $\mathbf{w}_k$ reveals which features are jointly affected by the k-th principal component.
However, PCA suffers from serious drawbacks when applying it to real data sets [129]. First, biological measurements are inherently noisy, and there is no explicit account of noise in PCA. In practice, high variance components are often asociated with signal whereas low-variance components are assumed to be noise, but an ideal model should explicitly disentangle the uncoordinated variability that is attributed to noise from the coordinated variability that is characterised as signal. Second, in its original formulation, no missing data is allowed [94]. Third, there is no rationality on how to evaluate the fit and perform model selection. Finally, it does not offer a principled way of modelling prior information about the data.

2.4.3 Probabilistic Principal Component Analysis and Factor Analysis

A probabilistic version of PCA was initially proposed in [212]. It can be formulated by converting some (or all) fixed parameters into random variables and adding an explicit noise term to Equation (2.9):

$$\mathbf{Y} = \mathbf{W}\mathbf{Z} + \boldsymbol{\epsilon} \tag{2.12}$$

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

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

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

$$p(\mathbf{Y}|\mathbf{W},\mathbf{Z},\sigma) = \prod_{n=1}^{N}\prod_{d=1}^{D}\mathcal{N}\left(y_{n,d}\,|\,\mathbf{w}_{:,k}^{T}\mathbf{z}_{n,:},\sigma^2\mathbf{I}\right) \tag{2.13}$$

The corresponding graphical model is:

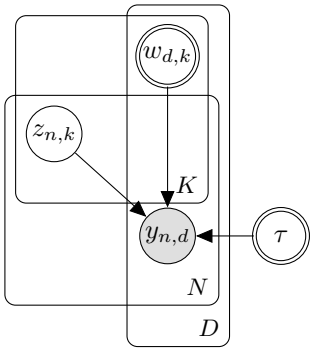


Fig. 2.6 Graphical model for probabilistic PCA. The latent variables are modelled as random variables, whereas the loadings and the noise are modelled as deterministic parameters.

Importantly, the choice of the distribution for ϵ implies that the noise of each feature is independent but restricted to have the same variance σ. In practice this is a limiting assumption, as different features are expected to show different degrees of noise, albeit this constraint can be relaxed and forms the basis of Factor Analysis [190, 19].

The inference procedures involves learning the parameters $\mathbf{W}$, and σ^2 and a posterior probability distribution for $\mathbf{Z}$. As the model depends on latent variables, inference can be performed using the iterative Expectation-Maximisation (EM) algorithm [190, 19]. In the expectation step, the posterior distribution for $\mathbf{Z}$ is computed in closed form (due to conjugacy between the likelihood and the prior), given current estimates for the parameters $\mathbf{W}$, and σ^2. In the maximisation step, the parameters are calculated by maximising the expectation of the joint log likelihood under the posterior distribution of $\mathbf{Z}$ found in the E step [212].

Interestingly, the EM solution of probabilistic PCA lies in the same subspace than the traditional PCA solution [212], but the use of a probabilistic framework brings several benefits. First, model selection can be performed by comparing likelihoods across different settings of parameters. Second, missing data can naturally be accounted for by ignoring the missing observations from the likelihood. Finally, the probabilistic formulation sets the core framework for a Bayesian treatment of PCA, enabling a broad range of principled extensions tailored different types of data sets.

2.4.4 Bayesian Principal Component Analysis and Bayesian Factor Analysis

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

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

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

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

where a_0 and b_0 are fixed hyperparameters. As in Equation (2.13), this results in a Normal likelihood:

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

The corresponding graphical model is:

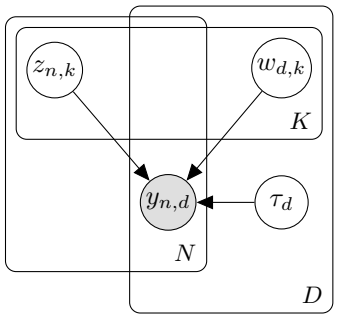


Fig. 2.7 Graphical model for Bayesian Factor Analysis. All unobserved variables are modelled as random variables.

2.4.4.1 Hierarchical priors: Automatic relevance determination

A key advantage of the full Bayesian treatment is that it explicitly captures uncertainity on the estimation of all unobserved variables, as opposed to the probabilistic PCA model [18, 17]. Yet, more importantly, the use of (hierarchical) prior distributions allow different modelling assumptions to be encoded, providing a flexible and principled approach to extend PCA to a myriad of modelling scenarios, including multi-view generalisations [113, 221, 115, 29, 108, 234].

As an example, a major challenge in PCA is how to determine the dimensionality of the latent space (i.e. the number of principle components). As we will show, the use of hierarchical prior distributions allows the model to introduce sparsity assumptions on the loadings in such a way that the model automatically learns the number of factors.
In the context of Factor Analysis, one the first sparsity priors to be proposed was the Automatic Relevance determination (ARD) prior [160, 140, 18, 17].

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

The aim of this prior is two-fold. First, the zero-mean normal distribution specifies that, *a priori*, no information is available and all features are *inactive*. When exposed to some data, the posterior distribution for $\mathbf{W}$ will be estimated by weighting the contribution from the likelihood, potentially allowing features to escape from the zero-centered prior (Figure 2.8).
Second, performing inference on the variable $\boldsymbol{\alpha} = \{\alpha_1,\cdots,\alpha_k\}$ enables the model to discard inactive factors. To understand this, let us assume that only $K = 5$ true factors exist, but the model is initialised with $K = 20$ factors. In such case, inactive factors can be prunned out by driving the corresponding α_k to infinity. In turn, this causes the posterior $p(\mathbf{w}_{:,k}|\mathbf{Y})$ to be sharply peaked at zero, resulting in the inactivation of all its weights Figure 2.9.

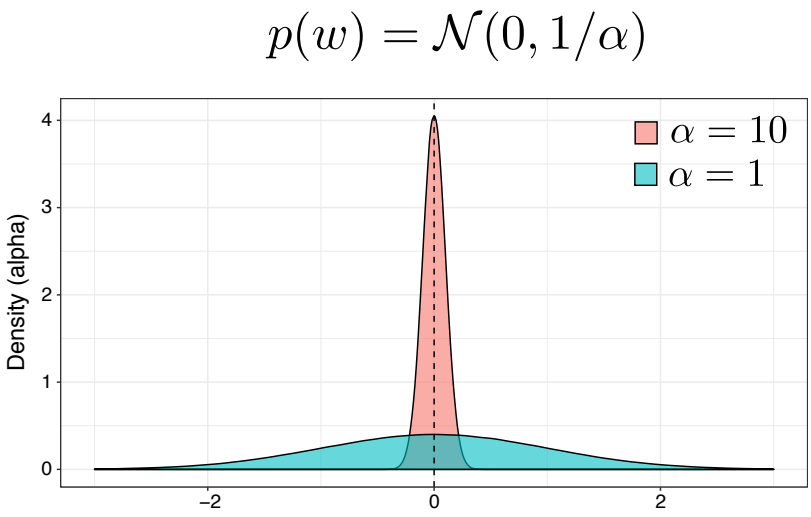


Fig. 2.8 Visualisation of the sparsity-inducing Automatic Relevance Determination prior

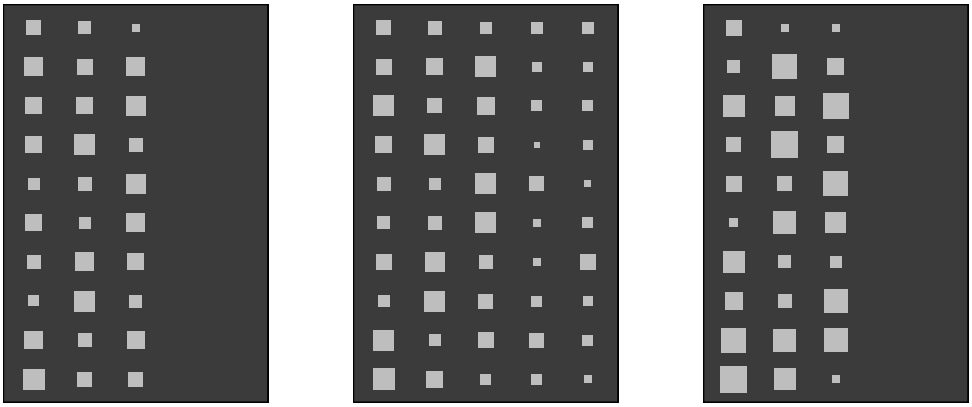


Fig. 2.9 Hinton plots display the values of the loading matrix, similar to a heatmap, where bigger squares depict larger loadings. Shown are the Hinton plots for (a) the true weights, (b) the infered weights by a Factor Analysis model with no ARD prior (middle), and (c) the infered weights by a Factor Analysis model with ARD prior per factor.
This figure was generated using simulated data with $N = 100$ samples, $D = 10$ features and $K = 3$ factors.

2.4.4.2 Hierarchical priors: Spike-and-slab prior

Sparse extensions of the Bayesian factor analysis model have been proposed as a regularisation mechanism but also to model inherent assumptions regarding the sparse nature of biological data [203, 66].

The variability observed in biological data is driven both by technical factors and biological factors. Thechnical factors (i.e. batch effects) tend to be relatively strong and alter the expression of a large proportion of genes, whereas the biological factors are potentially weak effects driven by changes in small gene regulatory networks [66]. Hence, a practical factor analysis model should be able to learn factors with different degrees of sparsity.

The ARD prior proposed in Section 2.4.4.1 allows entire factors to be dropped out from the model,

but it provides a weak degree of regularisation when it comes to inactivating individual loadings within the active factors.

A sparse generalisation of the Factor Analysis model proposed above can be achieved by combining the ARD prior with a spike-and-slab prior [150, 213]:

$$p(w_{d,k} \mid \alpha_k, \theta_k) = (1-\theta_k)\mathbb{1}_0(w_{d,k}) + \theta_k \mathcal{N}\left(w_{d,k} \mid 0, \alpha_k^{-1}\right) \tag{2.14}$$

$$p(\theta_k) = \text{Beta}\left(\theta_k \mid a_0^\theta, b_0^\theta\right) \tag{2.15}$$

$$p(\alpha_k) = \mathcal{G}\left(\alpha_k \mid a_0^\alpha, b_0^\alpha\right) \tag{2.16}$$

The corresponding graphical model is:

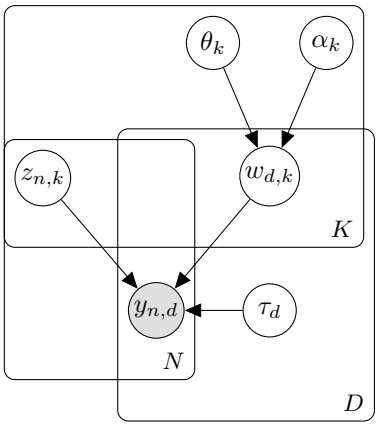


Fig. 2.10 Graphical model for Bayesian sparse Factor Analysis. A double sparsity-inducing prior is used on the loadings: an ARD prior to prune inactive factors and a spike-and-slab prior to inactive individual features within the active factors.

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

2.5 Multi-view factor analysis models

Probabilistic PCA and Factor Analysis perform dimensionality reduction from a single input matrix. However, in some occasions data is collected from multiple data sources that exibit heterogeneous statistical properties, resulting in a structured data set where features are naturally partitioned into groups [226, 128, 228]. A clear biological example is multi-omics data, where, for the same set of samples, multiple molecular layers are profiled, including genetic background, epigenetic information, protein levels or lipid composition [93, 79]. Each of the data modalities can be analysed separately using conventional (single-view) methods, but the challenge is the use of a single model that integrates all these layers of information using a flexible and principled approach. This is

refered to as the multi-view learning problem [226, 128].
A tempting approach to circumvent the multi-view learning problem is to simply concatenate all different data sets before applying conventional (single-view) latent variable models [188]. However, this is prone to fail for several reasons. First, heterogeneous data modalities cannot always be modelled using the same likelihood function. For example, the normal likelihood is commonly used for analysing continuous data, yet the statistical properties of binary readouts and count-based traits are not appropriately modelled by this distribution [171]. Second, even if all views are modelled with the same likelihood, differences in the scale and the magnitude of the variance can lead to some views being overrepresented in the latent space, hence requiring strong normalisation steps. Finally, for a structured input data, composed of different groups of features, one can also expect a structured latent representation, with factors capturing signal in different subsets of views [221, 115]. Not taking this behaviour into account can lead to challenges in the interpretability of the latent space.

A comprehensive review of multi-view machine learning methods can be found in [226] and a more genomics-based perspective can be found in [188]. For the purpose of this thesis, we will describe multi-view methods based on latent variable models.

2.5.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between two datasets [89, 76].
Given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_2}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ with maximal cross-correlation. For the first pair of canonical variables, the optimisation problem is:

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

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

Given the similarity with PCA, both methods share statistical properties, including the linear mapping between the low-dimensional space and the high-dimensional space, and the closed-form solution using singular value decomposition [89, 76].
Because of its simplicity and efficient computation, CCA has widespread use as a dimensionality reduction technique [76]. Yet, as expected, CCA suffers from the same pitfalls as PCA: difficulties in selecting the number of components, lack of sparsity in the solutions and absence of probabilistic formulation. In addition, CCA have been shown to overfit for datasets where $D >> N$ [145, 74]. Hence, probabilistic versions with sparsity assumptions that reduce overfitting and improve interpretability followed.

2.5.2 Probabilistic Canonical Correlation Analysis

Following the derivation of probabilistic PCA [212], a similar effort enabled a probabilistic formulation of CCA as a generative model [9].
In this model, the two matrix of observations $\mathbf{Y}^1$ and $\mathbf{Y}^2$ are decomposed in terms of two loading matrices $\mathbf{W}^1$ and $\mathbf{W}^2$ but a joint latent matrix $\mathbf{Z}$:

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

With the following prior probability distributions:

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

As in [212], the loadings and the variance of the noise are assumed to be non-probabilistic parameters, whereas the factors are probabilistic unobserved variables. This yields the following likelihood functions:

$$p(\mathbf{Y}^1|\mathbf{W}^1, \mathbf{Z}, \tau_1) = \prod_{n=1}^{N}\prod_{d=1}^{D_1} \mathcal{N}\left(y^1_{n,d} \,|\, (\mathbf{w}^1_{:,k})^T\mathbf{z}_n, \tau_1^{-1}\right) \tag{2.17}$$
$$p(\mathbf{Y}^2|\mathbf{W}^2, \mathbf{Z}, \tau_2) = \prod_{n=1}^{N}\prod_{d=1}^{D_2} \mathcal{N}\left(y^2_{n,d} \,|\, (\mathbf{w}^2_{:,k})^T\mathbf{z}_n, \tau_2^{-1}\right)$$

The corresponding graphical model is:

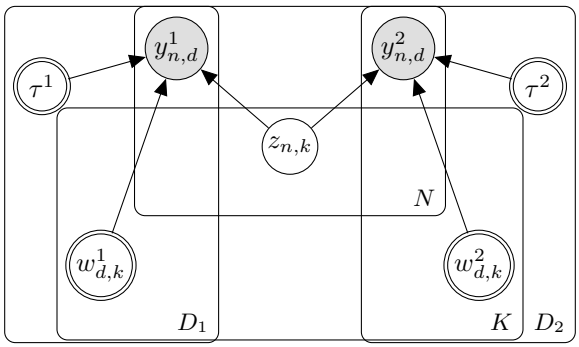


Fig. 2.11 Graphical model for probabilistic Canonical Correlation Analysis

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

Analogously to probabilistic PCA, the expected value of the posterior distribution $p(\mathbf{Z}|\mathbf{Y}^1, \mathbf{Y}^2)$ span the same subspace as standard CCA [9]. Nonetheless, this formulation enables a broad range of principled extensions into larger probabilistic models.

2.5.3 Bayesian Canonical Correlation Analysis

A fully Bayesian treatment of CCA followed based on exactly the same principle presented in Section 2.4.4 by introducing prior distributions to all unobserved variables [224, 114]:

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

Resulting in the same likelihood model as in Equation (2.17). Yet, notice that an ARD is introduced per factor, allowing an automatic inference of the dimensionality in the latent subspace. Also, there is some flexibility in the definition of noise. An independent noise term can be defined per view or per per feature. One could also model correlated noise by generalising the distribution to a multivariate gaussian with full-rank covariance. [224, 114].

The corresponding graphical model is:

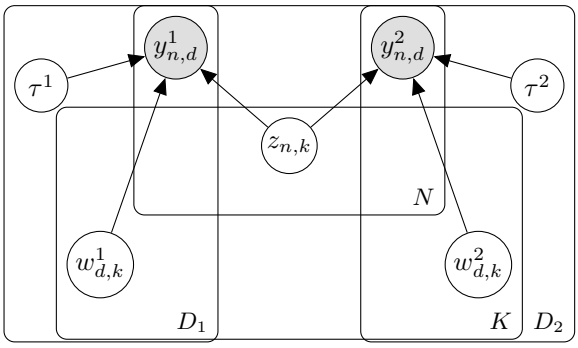


Fig. 2.12 Graphical model for Bayesian Canonical Correlation Analysis

As expected, in practice this yields a more sparse solution than traditional CCA (Figure 2.13):

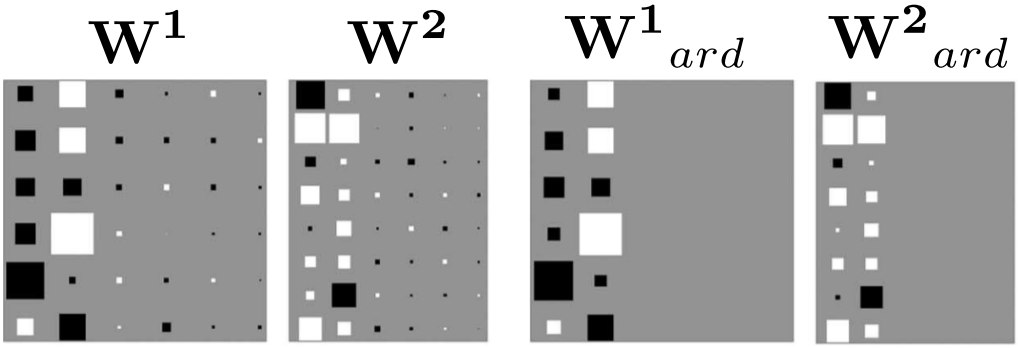


Fig. 2.13 Comparison of the hintons diagram of $\mathbf{W}^1$ and $\mathbf{W}^2$ for the maximum likelihood CCA model (two left plots) and the variational bayes CCA model (two right plots). Reprinted from [224].

2.5.4 Group Factor Analysis

Group Factor Analysis (GFA) is the natural generalisation of Bayesian Canonical Correlation Analysis to an arbitrary number of views. The original idea was originally presented in [221] and a series of generalisations followed, tailored with specific assumptions for different applications [115, 127, 29, 108, 234, 185].
In this section we will outline the core principle of GFA and describe some of the applications that followed.

Given a data set of M views $\mathbf{Y}_1, \cdots, \mathbf{Y}_M$, the task of GFA is to find K factors that capture the inter-variability between views as well as the intra-variability within views. In other words, some factors should explain variance across multiple views whereas some factors should explain variance within a single view.

The starting point from the model is Bayesian CCA Section 2.5.3. Generalising it to M views:

$$\begin{aligned}\mathbf{Y}^1 &= \mathbf{W}^2\mathbf{Z} + \epsilon^1 \\ \mathbf{Y}^2 &= \mathbf{W}^2\mathbf{Z} + \epsilon^2 \\ &\cdots \\ \mathbf{Y}^M &= \mathbf{W}^3\mathbf{Z} + \epsilon^3\end{aligned}$$

The key in learning the GFA solution lies on the sparsity structure of the loadings. If using the same ARD prior per factor as in Bayesian CCA, the factors would be exclusively restricted to have the same activity across all views. In GFA, however, the aim is to find components that capture intra-specific variation (active in a single view) and components that capture inter-specific dependencies (active in multiple views). This can be achieved by generalising the ARD sparsity as

follows:

$$p(\mathbf{W}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_{:,k}^{m} \,|\, 0, \frac{1}{\alpha_k^m}\right) \tag{2.18}$$

$$p(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{2.19}$$

Effectively, setting an ARD prior per factor and group defines a matrix $\boldsymbol{\alpha} \in \mathbb{R}^{M \times K}$ with four types of factors: (1) Inactive factors that do not explain variance in any view, such that all values $\boldsymbol{\alpha}_k$ are large. (2) Fully shared factors that explain variance across all views, such that all values $\boldsymbol{\alpha}_k$ are small. (3) Unique factors that explain variance in a single view, such that all values $\boldsymbol{\alpha}_k$ are very large, except one. (4) Partially shared factors that explain variance in a subsets views, such that some values in $\boldsymbol{\alpha}_k$ are large whereas others are small.

The following figure illustrates a representative GFA solution for three views:

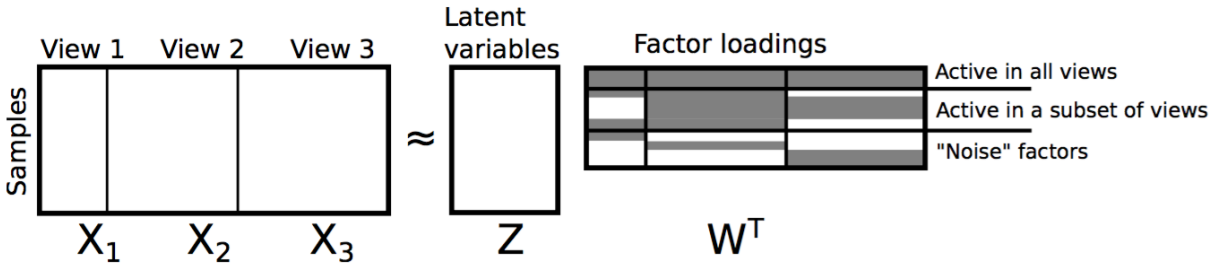


Fig. 2.14 (RE-DO FIGURE) Illustration of the GFA solution for $M = 3$. The data sets are concatenated into a single matrix $\mathbf{Y}$, which is factorised into a product of the latent variable matrix $\mathbf{Z}$ and a concatenated loading matrix $\mathbf{W}$. The shade in $\mathbf{W}$ depicts the factor and view-wise sparsity, where gray shading indicates high activity (small α_m^k value) and no shading indicates low activity (large α_m^k value). Note that some factors are active across all views (top), some factors are active in subsets of views (middle) and some factors are active in a single view (bottom).

Finally, notice that if $M = 1$ the model reduces to Bayesian PCA (Section 2.4.4). However, if $M = 2$ the model does not reduce to Bayesian CCA, because in the GFA setting some factors are also allowed to capture intra-specific variability.

2.5.4.1 Similar approaches

2.5.4.2 Extensions

Multiple extensions from the initial GFA framework have been proposed:

- Klami et al, 2015 [115]: in the GFA framework outlined above, $\boldsymbol{\alpha} \in \mathbb{R}^{M,K}$ is drawn from a flat gamma prior where views and factors are assumed independent. To encourage the model to find correlations between views, they specify a low-rank decomposition for $\boldsymbol{\alpha}$:

$$\log \boldsymbol{\alpha} = \mathbf{U}\mathbf{V}^T + \boldsymbol{\mu}_u^T + \boldsymbol{\mu}_v^t$$

with corresponding priors for $\mathbf{U}$ and $\mathbf{V}$. This representation directly captures correlations between the view activation profiles and is particularly useful when dealing with large number of views [**Klami2014**].

- Zhao et al, 2016 [234]: in the GFA framework outlined above, $\mathbf{W}^M$ have no element-wise sparsity. In this study, they implemented a more sophisticated structured sparsity prior called the three parameter beta prior [7]. The advantage of this prior distribution is three-fold: (1) it regularises the matrix, globally, removing factors that do not explain variation in any view. (2) Shrinks columns $\mathbf{w}^m_{:,k}$ to obtain the desired factor and group-wise sparsity. (3) Introduces feature-wise sparsity for each element $w^m_{d,k}$

2.6 Multi-Omics Factor Analysis: a framework for unsupervised integration of multi-omics data sets

The work described in this chapter results from a collaboration with the Multi-omics and statistical computing group lead by Wolfgang Huber at the EMBL (Heidelberg, Germany). It has been peer-reviewed and published in [6].

The method was conceived by Florian Buettner, Oliver Stegle and me. I performed most of the mathematical derivations and implementation, but with significant contributions from Damien Arnol and Britta Velten. The single-cell application was led by me whereas the CLL data application was led by Britta Velten, but with joint contributions in either cases. Florian Buettner, Wolfgang Huber and Oliver Stegle supervised the project.
The article was jointly written by Britta Velten, Florian Buettner, Wolfgang Huber, Oliver Stegle and me.

2.6.1 Model description

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

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m \tag{2.20}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times K}$ is a matrix that contains the factor values and $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ are M matrices that contain the loadings that relate the high-dimensional space to the low-dimensional latent representation. Finally, $\boldsymbol{\epsilon}^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic:

$$p(\epsilon_d^m) = \mathcal{N}\left(\epsilon_d^m \,|\, 0, 1/\tau_d^m\right) \tag{2.21}$$

Non-gaussian noise models can also be defined (see Section 2.6.6). Unless otherwise stated, we will always assume Gaussian noise.
Altogether, this results in the following likelihood:

$$p(\mathbf{Y}|\mathbf{W}, \mathbf{Z}, \mathbf{T}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{n=1}^{N} \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_n^T \mathbf{w}_d^m, 1/\tau_d^m\right) \tag{2.22}$$

2.6.1.1 Interpretation of the factors

Each factor ordinates cells along a one-dimensional axis centered at zero. Samples with different signs indicate opposite phenotypes, with higher absolute value indicating a stronger effect. For example, if the k-th factor captures the variability associated with cell cycle, we could expect cells in

the Mitosis state to be at one end of the factor (irrespective of the sign, only the relative positioning being of importance). In contrast, cells in G1 phase are expected to be at the other end of the factor. Cells with intermediate phenotype, or with no clear phenotype (i.e. no cell cycle genes profiled), are expected to be located around zero, as specified by the prior distribution.

2.6.1.2 Interpretation of the loadings

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

2.6.1.3 Interpretation of the noise

The use of a probabilistic framework allows the model to explicitly disentangle the signal (i.e. the explained variance) from the noise (i.e. unexplained variance). Large values of τ_d^m indicate high certainity on the observations for the feature d in view m, as predicted by the latent variables. In contrast, small values of τ_d^m are indicative of low predictive power by the latent variables.

2.6.1.4 Missing values

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

2.6.1.5 Prior distributions for the factors and the loadings

The key determinant of the model is the regularization used on the prior distributions of the factors and the weights.
For the factors, we define an isotropic Gaussian prior:

$$p(z_{nk}) = \mathcal{N}(z_{nk} \,|\, 0, 1) \tag{2.23}$$

which effectively assumes a continuous latent space and independent samples and factors.
For the weights we encode two levels of sparsity, a (1) view- and factor-wise sparsity and (2) an individual feature-wise sparsity. The aim of the factor- and view-wise sparsity is to disentangle the activity of factors to the different views, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not explain any variation in view m.
In addition, we place a second layer of sparsity which encourages inactive weights on each individual

feature. Mathematically, we express this as a combination of an Automatic Relevance Determination (ARD) prior [140] for the view- and factor-wise sparsity and a spike-and-slab prior [150] for the feature-wise sparsity: However, this formulation of the spike-and-slab prior contains a Dirac delta function, which makes the inference procedure troublesome. To solve this we introduce a re-parametrization of the weights w as a product of a Gaussian random variable $\hat{w}$ and a Bernoulli random variable s, [213] resulting in the following prior: In this formulation α_k^m controls the activity of factor k in view m and θ_k^m controls the corresponding fraction of active loadings (i.e. the sparsity levels).

Finally, we define conjugate priors for θ and α:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{2.24}$$

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{2.25}$$

with hyper-parameters $a_0^\theta, b_0^\theta = 1$ and $a_0^\alpha, b_0^\alpha = 1e^{-3}$ to get uninformative priors.
Posterior values of θ_k^m close to 0 implies that most of the weights of factor k in view m are shrinked to 0 (sparse factor). In contrast, a value of θ_k^m close to 1 implies that most of the weights are non-zero (non-sparse factor). A small value of α_k^m implies that factor k is active in view m. In contrast, a large value of α_k^m implies that factor k is inactive in view m.

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

$$\begin{aligned} p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = & \prod_{m=1}^{M}\prod_{n=1}^{N}\prod_{d=1}^{D_m}\mathcal{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right) \\ & \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K}\mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right)\text{Ber}(s_{d,k}^m|\theta_k^m) \\ & \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \\ & \prod_{m=1}^{M}\prod_{k=1}^{K}\text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \\ & \prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \\ & \prod_{m=1}^{M}\prod_{d=1}^{D_m}\mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right). \end{aligned} \tag{2.26}$$

and the corresponding graphical model is shown in Figure 2.16. This completes the definition of the MOFA model.

2.6.2 Downstream analysis

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

- **Variance decomposition**: calculate the variance explained (R^2) by each factor in each view.

- **Ordination of the samples in the latent space**: scatterplots or beeswarm plots of factors, colored or shaped by sample covariates can reveal the main drivers of sample heterogeneity.
- **Inspection of loadings**: the weights (or loadings) can be interpreted as an activity score for each gene on each factor. Hence, inspecting the top loadings reveals the genes (or other genomic features) that underlie each factor.
- **Imputation**: MOFA generates a condensed and denoised low-dimensional representation of the data. As discussed in Section X, the data can be reconstructed from the latent space by a simple matrix multiplication: $\hat{\mathbf{Y}} = \mathbf{Z}\mathbf{W}^T$.
- **Feature set enrichment analysis**: when a factor is difficult to characterise based only on the inspection of the top loadings, one can compute a statistical test for enrichment of biological pathways using predefined gene-set annotations.

The downstream functionalities implemented in MOFA are highlighted in Figure 2.15.

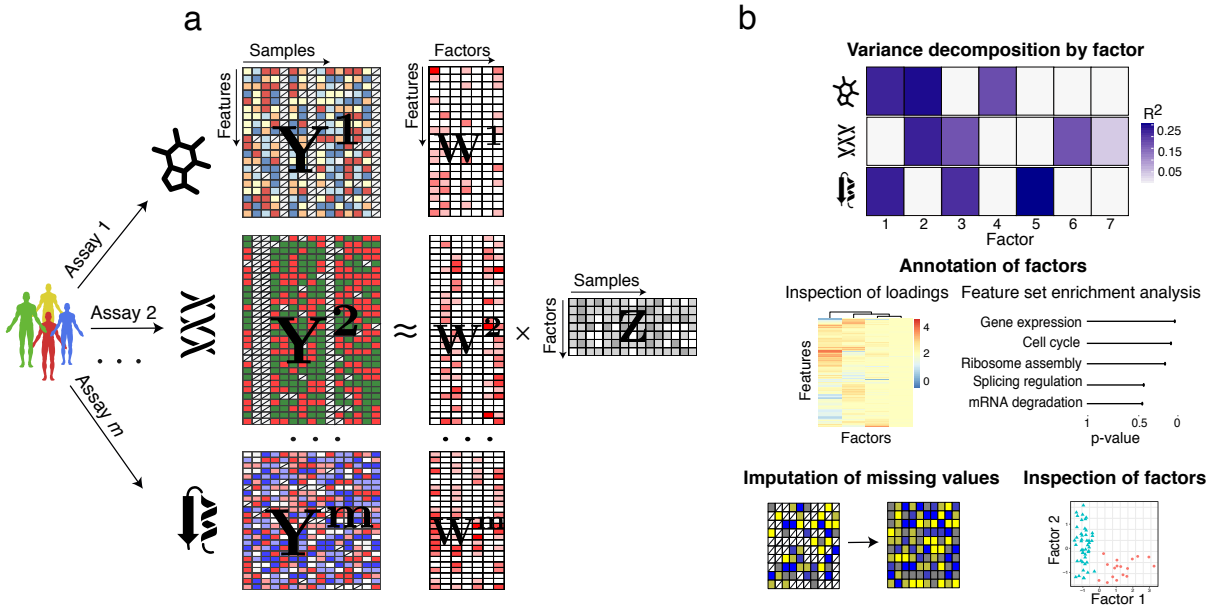


Fig. 2.15 MOFA overview. The model takes M data matrices as input $(\mathbf{Y}^1, \cdots, \mathbf{Y}^M)$, one or more from each data modality, with co-occurrent samples but features that are not necessarily related and can differ in numbers. MOFA decomposes these matrices into a matrix of factors $(\mathbf{Z})$ and M weight matrices, one for each data modality $(\mathbf{W}^1, \cdots, \mathbf{W}^M)$. White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denotes missing values. The fitted MOFA model can be queried for different downstream analyses, including a variance decomposition to assess the proportion of variance explained by each factor in each data modality.

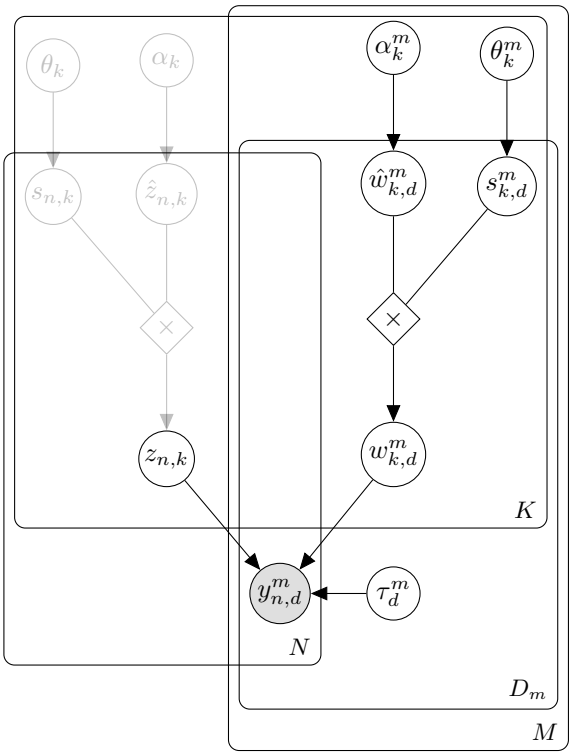


Fig. 2.16 Graphical model for MOFA. The white circles represent hidden variables that are infered by the model, whereas the grey circles represent the observed variables. There are a total of four plates, each one representing a dimension of the model: M for the number of views, N for the number of samples, K for the number of factors and D_m for the number of features in the m-th view. The use of transparency in the top left nodes is intentional and becomes clear in Chapter 4 where we implement a spike-and-slab prior on the factors.

2.6.2.1 Inference

To make the model scalable to large data sets we adopt a Variational inference framework with a structured mean field approximation. A detailed overview is given in section XX, and details on the variational updates for the MOFA model are given in Appendix XX.
To enable efficient inference for non-Gaussian likelihoods we employ local bounds [96, 195]. This is described in detail in Section 2.6.6

2.6.3 Monitoring convergence

An attractive property of Variational inference is that the objective function, the Evidence Lower Bound (ELBO), increases monotonically at every iteration. This provides a simple way of monitoring convergence Figure 2.17. This is indeed one of the reasons why we opted for this inference framework over Expectation Propagation or sampling-based approaches.

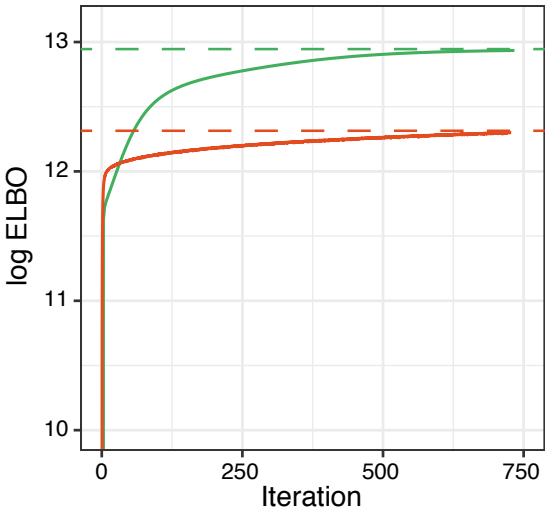


Fig. 2.17 Training curve for two diffent initialisations of MOFA. The y-axis displays the log of the ELBO, with higher values indicating a better fit. The x-axis displays the iteration number. The horizontal dash lines mark the value of the ELBO upon convergence.

2.6.4 Model selection and consistency across random initilizations

The variational optimisation problem in MOFA is not convex and the posterior distributions will vary depending on the initialisation of the model. Thus, it becomes mandatory to perform model selection and assess the consistency of the factors across different trials.

The strategy we adopted in this work is to train several MOFA models (e.g. 10 trials) under different parameter initialisations, and after training we select the model with the highest ELBO for downstream analysis Figure 2.18. In addition, we evaluate the robustness of the factors by plotting the Pearson correlations between factors across all trials. Figure 2.18.

A similar strategy has also been proposed in [87, 86].

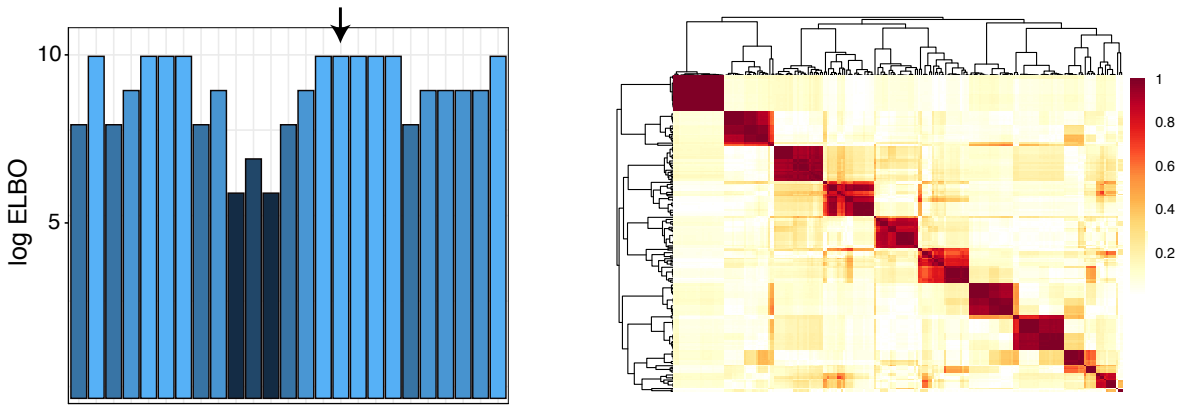


Fig. 2.18 Model selection and robustness analysis in MOFA. The left plot the log ELBO (y-axis) for 25 model instances (x-axis). The arrow indicates the model with the highest ELBO that would be selected for downstream analysis. The right plot displays the absolute value of the Pearson correlation coefficient between pairwise combinations of all factors across the 25 model instances. A block-diagonal matrix indicates that factors are robustly estimated regardless of the initialisation.

2.6.5 Learning the number of factors

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

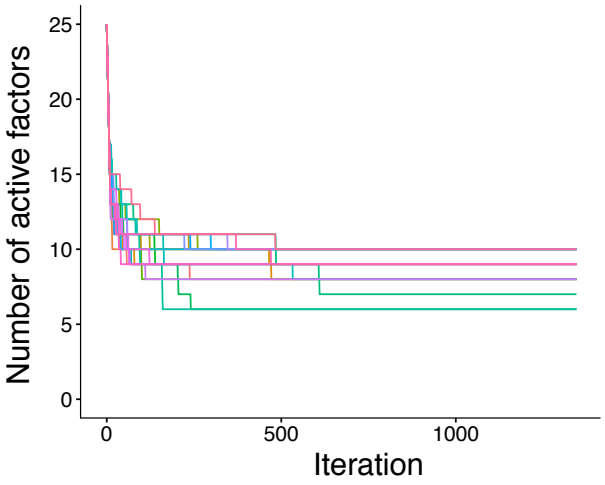


Fig. 2.19 Training curve for the number of active factors across 25 different model instances. The y-axis displays the number of active factors. The x-axis displays the iteration number. Different lines denote different model instances.

2.6.6 Modelling and inference with non-Gaussian data

To implement efficient variational inference in conjunction with a non-Gaussian likelihood we adapt prior work from [195] using local variational bounds. The key idea is to dynamically approximate non-Gaussian data by Gaussian pseudo-data based on a second-order Taylor expansion. To make the approximation justifiable we need to introduce variational parameters that are adjusted alongside the updates to improve the fit.
Denoting the parameters in the MOFA model as $\mathbf{X} = (\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\theta})$, recall that the variational framework approximates the posterior $p(\mathbf{X}|\mathbf{Y})$ with a distribution $q(\mathbf{X})$, which is indirectly optimised by optimising a lower bound of the log model evidence. The resulting optimization problem can be re-written as

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

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

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

with $f_{nd}(c_{nd}) = -\log p(y_{nd}|c_{nd})$. We dropped the view index m to keep notation uncluttered.
Extending [195] to our heteroscedastic noise model, we require $f_{nd}(c_{nd})$ to be twice differentiable and bounded by κ_d, such that $f''_{nd}(c_{nd}) \leq \kappa_d \, \forall n, d$. This holds true in many important models as for example the Bernoulli and Poisson case. Under this assumption a lower bound on the log likelihood can be constructed using Taylor expansion,

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

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

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

The algorithm propsed in [195] then alternates between updates of $\boldsymbol{\zeta}$ and q($\boldsymbol{\Theta}$). The update for $\boldsymbol{\zeta}$ is given by

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

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

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

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

Bernoulli likelihood for binary data

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

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

where $\sigma(a) = (1 + e^{-a})^{-1}$ is the logistic link function and $\mathbf{Z}$ and $\mathbf{W}$ are the latent factors and weights in our model, respectively.
In order to make the variational inference efficient and explicit as in the Gaussian case, we aim to approximate the Bernoulli data by a Gaussian pseudo-data as proposed in [195] and described above which allows to recycle all the updates from the model with Gaussian views. While [195] assumes a homoscedastic approximation with a spherical Gaussian, we adopt an approach following [96], which allows for heteroscedaticity and provides a tighter bound on the Bernoulli likelihood.

Denoting $c_{nd} = (\mathbf{Z}\mathbf{W}^T)_{nd}$ the Jaakkola upper bound [96] on the negative log-likelihood is given by

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

with λ given by $\lambda(\zeta) = \frac{1}{4\zeta}\tanh\left(\frac{\zeta}{2}\right)$.
This can easily be derived from a first-order Taylor expansion on the function $f(x) = -\log(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) = \frac{x}{2} - \log(\sigma(x))$ in x^2 and by the convexity of f in x^2 this bound is global as discussed in [96]. In order to make use of this tighter bound but still be able to re-use the variational updates from the Gaussian case we re-formulate the bound as a Gaussian likelihood on pseudo-data $\hat{\mathbf{Y}}$.
As above we can plug this bound on the negative log-likelihood into the variational optimization problem to obtain

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

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

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

as described in [96], [19].
For the variational distribution $q(\mathbf{Z}, \mathbf{W})$ we observe that the Jaakkola bound can be re-written as

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

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

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

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

Poisson likelihood for count data

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

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

where $\lambda(c) > 0$ is the rate function and has to be convex and log-concave in order to ensure that the likelihood is log-concave.
As done in [195], here we choose the following rate function: $\lambda(c) = \log(1 + e^c)$.
Then an upper bound of the second derivative of the log-likelihood is given by

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

The pseudodata updates are given by

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

2.6.7 Model validation with simulated data

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

2.6.7.1 Recovery of the latent space

First, we tested the ability of MOFA to recover simulated factors, varying the number of views, the number of features, the number of factors and the fraction of missing values.
For every simulation scenario we initialised a model with a high number of factors ($K = 100$), and inactive factors were automatically dropped during model training using a threshold of 1% variance explained. In addition, to test the robustness under different initialisations, ten models were trained for every simulation scenario.
We observe that in most settings the model accurately recovers the correct number of factors (Figure 2.20). Exceptions occur when the dimensionality of the latent space is too large (more than 50 factors) or when an excessive amount of missing values (more than 80%) is present in the data.

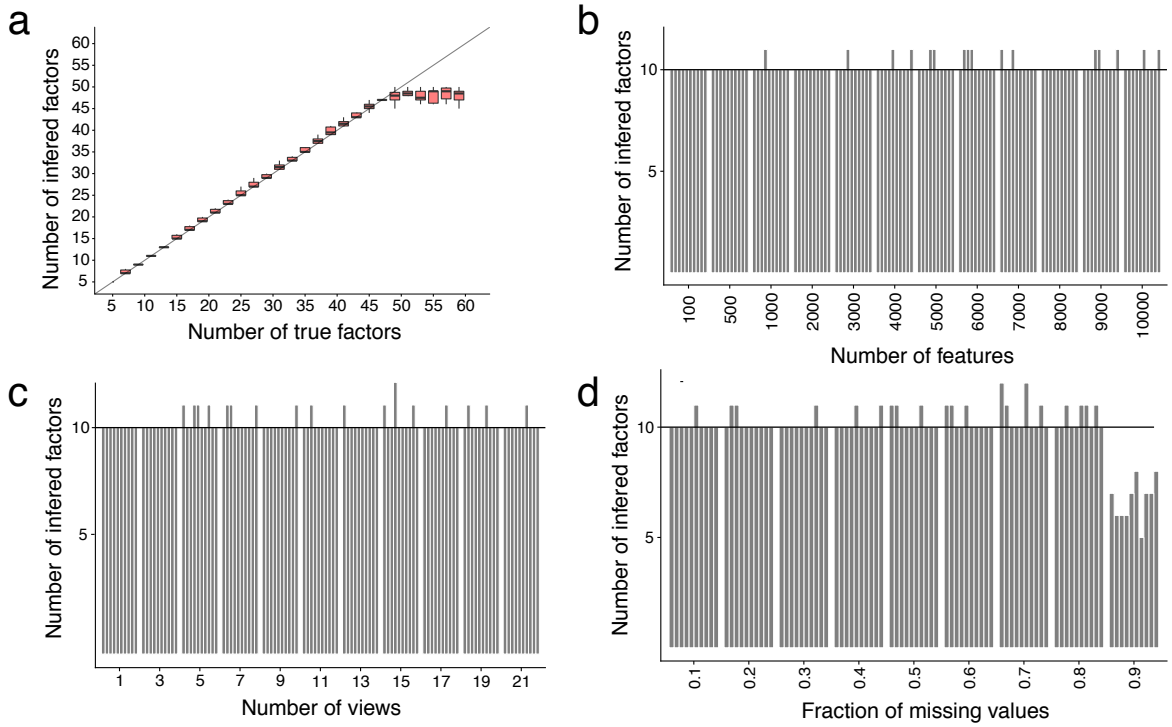


Fig. 2.20 Assessing the ability of MOFA to recover simulated latent spaces. In all plots the y-axis displays the number of infered factors. (a) x-axis displays the number of true factors, and boxplots summarise the distribution across 10 model instances. For (c-d) the true number of factors was set to $K = 10$ and each bar corresponds to a different model instance. (b) x-axis displays the number of features, (c) x-axis displays the number of views, (d) x-axis displays fraction of missing values.

2.6.7.2 Group-wise sparsity on the loadings

One of the most important statistical assumptions underlying MOFA (and other Group Factor Analysis methods) is the sparsity prior aimed at disentangling the activity of factors across views. To evaluate this feature we simulated data from the generative model were the factors were clearly set to be active or inactive in specific views. We compared the performance with two other methods: the iCluster+ model [151] and a GFA implementation [127].

the GFA implementation shares the same factor and view-wise sparsity as MOFA, and is therefore expected to show similar performance. On the other hand, iCluster is a model that is aimed at clustering and it only contains a sparsity constraint (in a penalised maximum likelihood setting) per factor, shared across all views.

On simulated data, MOFA and GFA correctly infer the true activity pattern of the factors whereas iCluster infers incorrect sharedness of factors across views, especially with increasing dimensionality of the latent space.

Fig. 2.21 Evaluating the ability of MOFA, iCluster and GFA to recover sparse factor activity patterns across views. The leftmost plot displays the true activity pattern, with factors being (strongly) active in different subsets of views. The remaining three plots show, for each model, the fraction of variance explained (R^2) by each factor in each view.

2.6.7.3 Feature-wise sparsity on the loadings

A key aspect of MOFA is the use of a spike-and-slab prior distribution to enforce feature-wise sparsity on the loadings, which yields a more interpretable solution (see **??**).

To assess the effect of the spike-and-slab prior we fit a group of models with a spike-and-slab prior and another group of models only with Automatic Relevance Determination prior. We further compared both solutions to a conventional Principal Component Analysis fit on the concatenated data set.

As expected, we observe that the spike-and-slab prior induces more zero-inflated weights, although the ARD prior provided a moderate degree of regularisation. The PCA solution was notably more dense than both bayesian models (Figure 2.22).

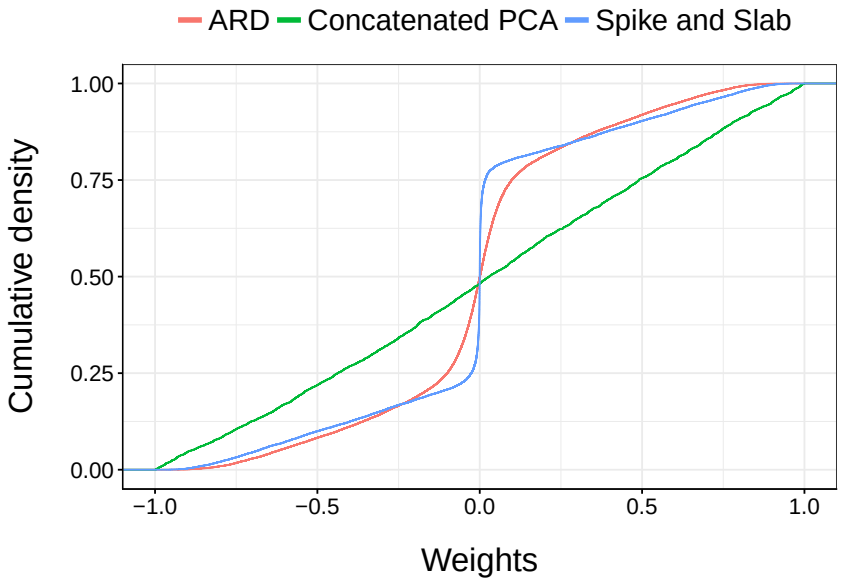


Fig. 2.22 Assessing sparsity on the loadings in MOFA. The plot shows the empirical cumulative density function of the loadings for an arbitrary factor in a single view. The loadings were simulated with a sparsity level of $\theta_k^m = 0.5$ (50% of active features.)

2.6.7.4 Non-gaussian likelihoods

A key improvement of MOFA with respect to previous methods is the use of non-Gaussian likelihoods to integrate multiple data modalities. As described in section XX, we implemented a Bernoulli likelihood to model binary data and a Poisson likelihood to model count data.

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

We observe that although both likelihoods are able to recover the true number of factors, the models with the non-Gaussian likelihoods clearly result in a better fit to the data (**????**).

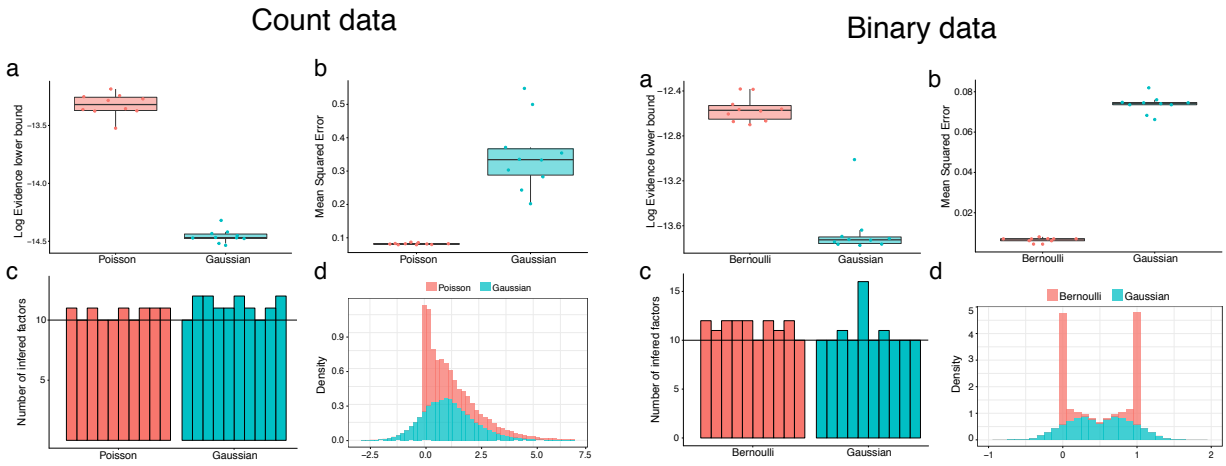


Fig. 2.23 Validation of the non-gaussian likelihood models implemented in MOFA on simulated data. The four plots on the left assess the Poisson and the Gaussian likelihoods applied to count data. The four plots on the right assess the Bernoulli and the Gaussian likelihoods applied to binary data. (a) The y-axis displays the ELBO for each model instance (x-axis). (b) The y-axis displays the mean reconstruction error for each model instance (x-axis). (c) The y-axis displays the number of estimated factrors for each model instance (x-axis). The horizontal dashed line marks the true number of factors $K = 10$. (d) Distribution of reconstructed data.

2.6.7.5 Scalability

Finally, we evaluated the scalability of the model when varying each of its dimensions independently (**??**), and we compared the speed with a Gibbs sampling implementation of GFA [127] and iCluster+ [151].

Overall, we observe that MOFA scales linear with respect to all dimensions and is significantly faster than any of the three evaluated techniques.

As a real application showcase, the training on the CLL data Figure 2.25 required 25 minutes using MOFA, 34 hours with GFA and 5-6 days with iCluster.

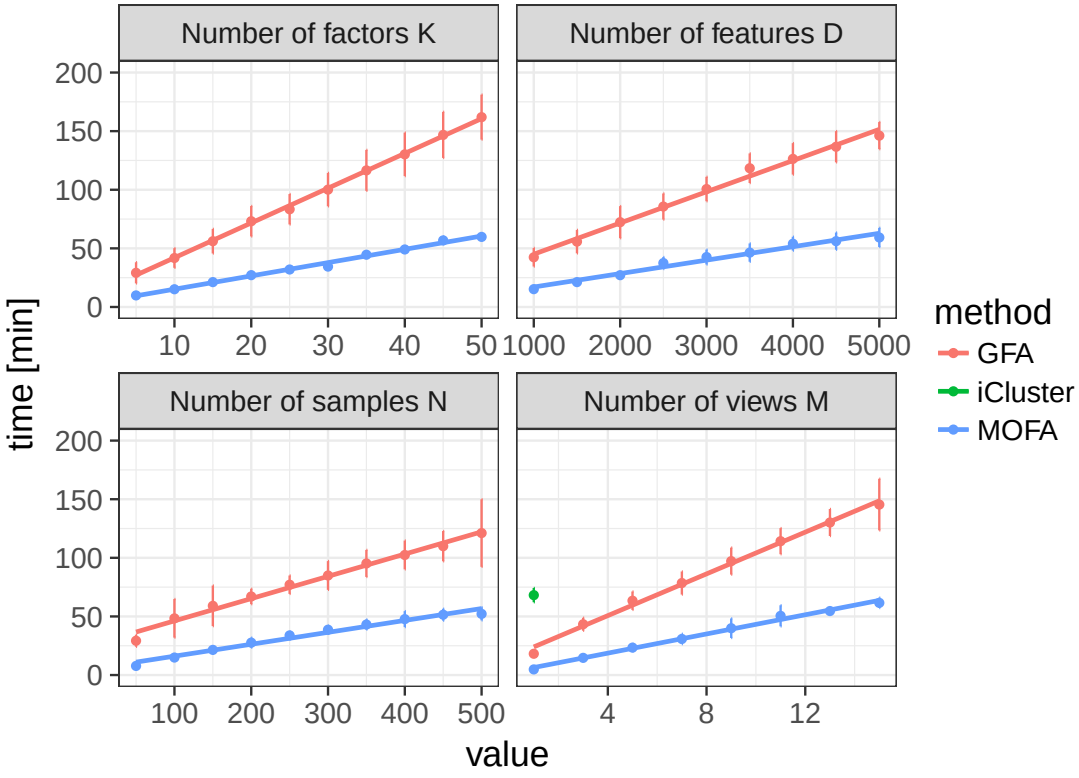


Fig. 2.24 Evaluation of speed and scalability in MOFA. The y-axis displays the time required for convergence. The x-axis displays the value of the dimension that was tested, either number of factors (K), number of features (D), number of samples (N) and number of views (M). Baseline parameters were $M = 3, K = 10, D = 1000, N = 100$. Each line represents a different model, GFA (red), MOFA (blue) and iCluster (green). Default convergence criteria where used for all methods. Each dot displays the average time across 10 trials with error bars denoting the standard deviation. iCluster is only shown for one value as all other settings required more than 200min for convergence.

2.6.8 Application to chronic lymphocytic leukaemia

Personalised medicine is an attractive field for the use of multi-omics, as dissecting heterogeneity across patients is a major challenge in complex diseases, and requires integration of information from multiple biological layers [37, 44, 2].

In most cases, predicting patient survival and response to a treatment is still not reliable due to a lack of predictive biomarkers and our incomplete understanding of the mechanisms underlying response heterogeneity. Identification of the main drivers of inter-patient variation and their molecular basis is an important step towards personalized treatment decisions.

To demonstrate the potential of the method, we applied MOFA to a study of 200 patient samples of chronic lymphocytic leukaemia (CLL) profiled for somatic mutations, RNA expression, DNA methylation and ex-vivo drug responses[52] Figure 2.25.

This data set was selected for five main reasons.

- The large number of cases to benchmark the speed of MOFA against other common integrative methods.

- The rich literature in this type of cancer which provides a good resource for the interpretation of the factors.

- The complex missing data structure of the study, with nearly 40% samples having incomplete assays Figure 2.25, in addition to the missing values present within some assays. hence, this study is ideal to benchmark MOFA's capabilities to deal with missing entries. As described in Section X, the inference framework we implemented allows the model to cope with this setting by merely ignoring missing entries and, when possible, pooling information from other molecular layers in order to infer the factor values.

- The different data modalities: after data processing, three assays had continuous observations whereas for the somatic mutations the observations were binary. As described in section XX, MOFA can combine different likelihood models to integrate multiple data types.

- The existence of clinical covariates: after model fitting, the factors can be associated with additional covariates. This provides an excellent test to assess whether the MOFA factors can capture the clinical phenotypes better than other dimensionality reduction techniques.

2.6.8.1 Model overview

In this data set, MOFA recovered 10 factors explaining a minimum of 3% of variance. Among these, the first two factors (sorted by variance explained) were active across most views, indicating a strong effect across multiple molecular layers. Other factors such as Factor 3 or Factor 5 explained variation in two data modalities, whereas Factor 4 was only in the RNA expression data.
Overall, the then MOFA factors inferred explained 41% of variance in the drug response data, 38% in the mRNA expression, 24% in the DNA methylation and 24% in somatic mutations.
Inspection of the top weights in the somatic mutation view revealed that Factor 1 was strongly associated with the mutation status of the immunoglobulin heavy-chain variable (IGHV) region, while Factor 2 was aligned with a trisomy of chromosome 12 (Figure X). In a completely unsupervised fashion, MOFA identified the two major axes of molecular disease heterogeneity being indeed the two most important clinical markers in CLL [58, 229].
A scatterplot based on these factors shows a clear separation of patients by their IGHV status on the first factor and presence or absence of trisomy 12 on the second factor. Remarkably, the IGHV status of a fraction of patients was missing (grey dots). Yet, as the two factors were shared across multiple views, MOFA was able to pool information from the other molecular layers to map those samples to the latent space.

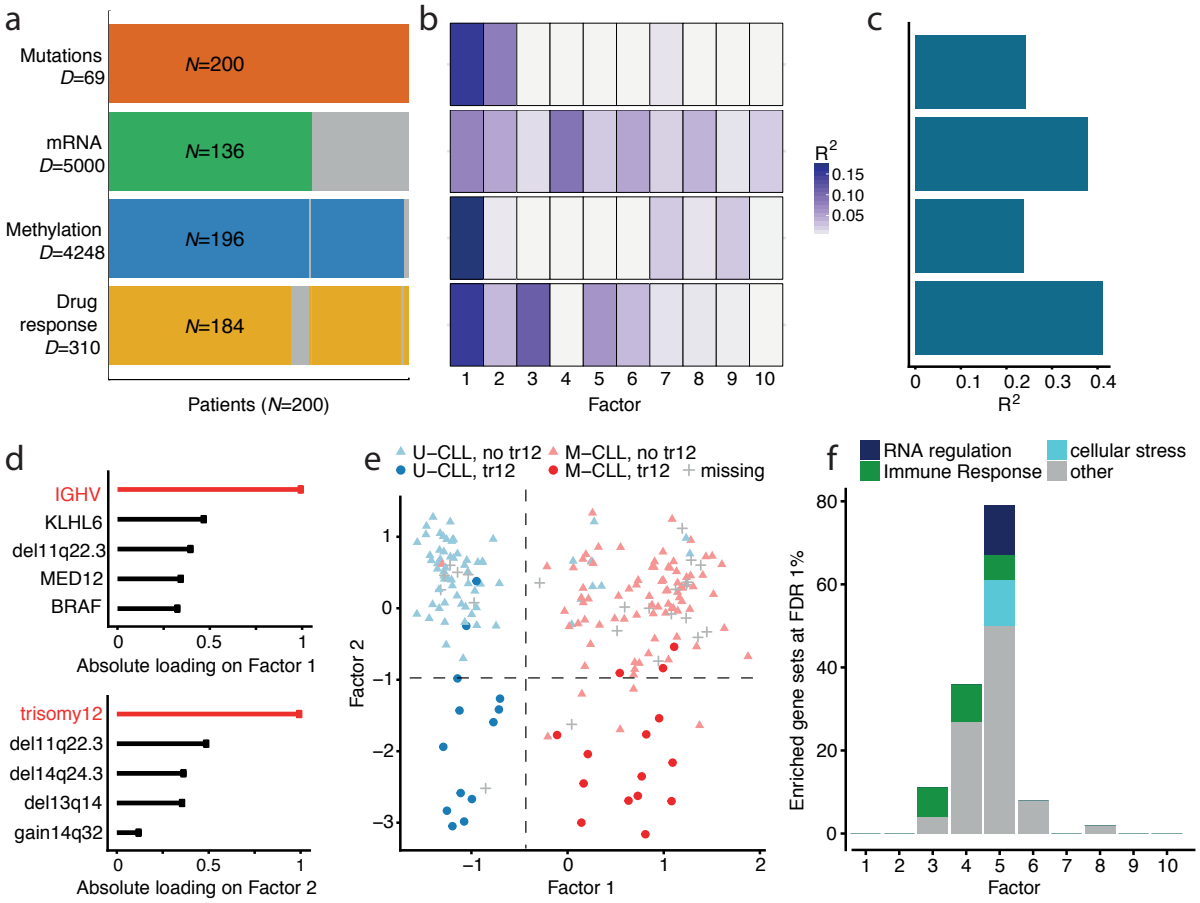


Fig. 2.25 XX

2.6.8.2 Characterisation of Factor 1

IGHV status is probably the most important prognostic marker in CLL and has routinely been used to distinguish between two distinct subtypes of the disease. Molecularly, it is a surrogate of the level of activation of the B-cell receptor, which is in turn related to the differentiation state of the tumoral cells. Multiple studies have associated mutated IGHV with a better resposne to chemoimmunotherapy, whereas unmutated IGHV patients have a worse prognosis [58, 28, 46, 50].

In clinical practice, the IGHV status has been generally considered binary. However, our results suggest a more complex structure with at least three groups or a potential underlying continuum, as also suggested in [162, 176].
Interestingly, there is some discrepancy between the IGHV status predicted by MOFA and the IGHV status reported in the clinical data. Out the 200 patients, MOFA classifies 176 in accordance with the clinical label, it classifies 12 patients that lacked the clinical marker and it re-classifies 12 patients to the opposite group.
To validate the MOFA-based classification, we inspected the molecular profiles. sample-to-sample correlation matrices for the individual layers suggest that for 3 of the cases where the inferred factor disagrees with the clinical label, the molecular data supports the predicted label. The other 9 cases showed intermediate molecular sgnatures now well captured by the binary classification.

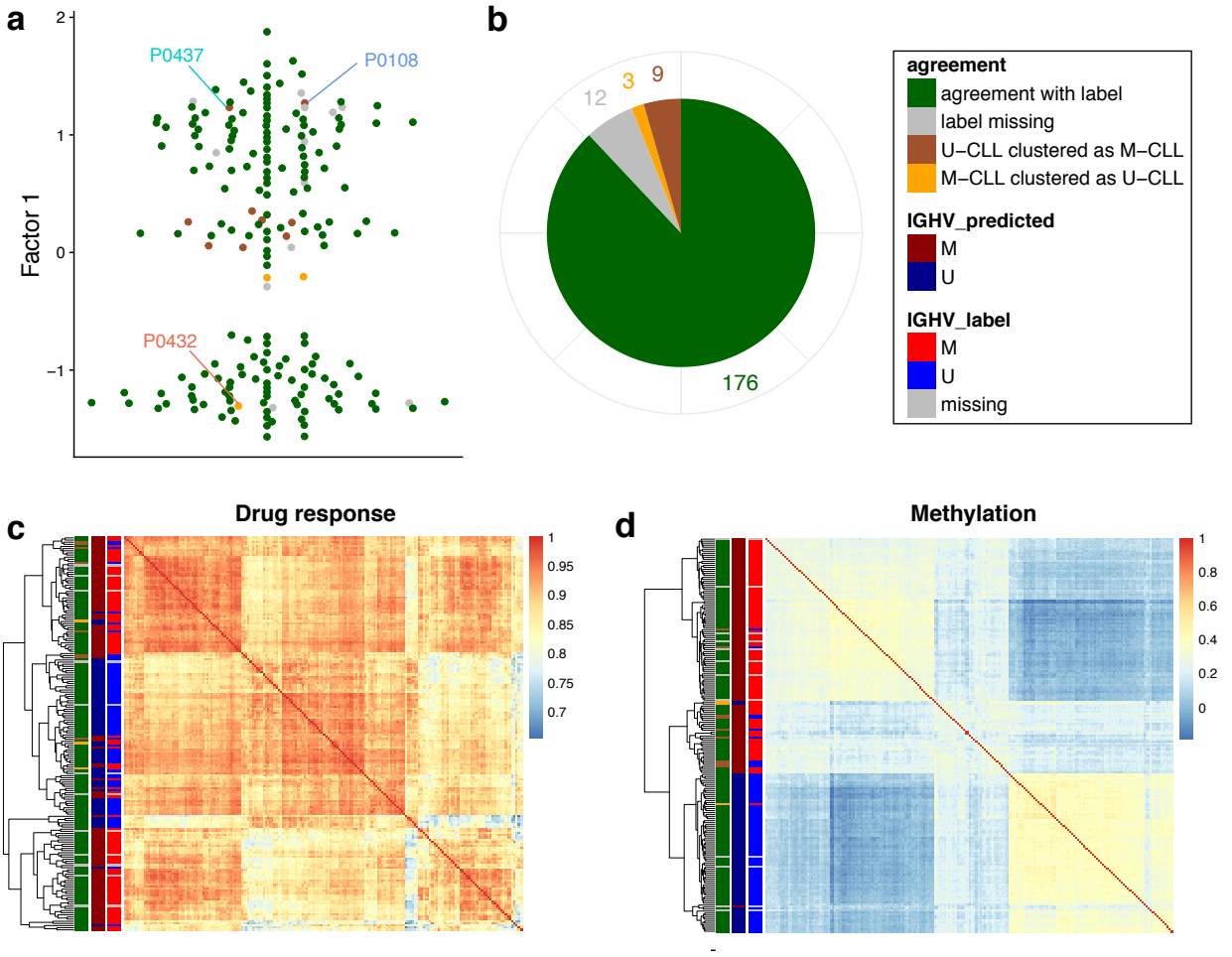


Fig. 2.26 XX

Finally, we characterised in more detail the molecular changes associated with IGHV status, as predicted by MOFA.

On the RNA expression, inspection of the top weights pinpoint genes that have been previously associated to IGHV status, some of which have been proposed as clinical markers[219, 142, 216, 153, 172]. Heatmaps of the RNA expression levels for these genes reveals clear differences between samples when ordered according to the corresponding Factor 1 values.

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

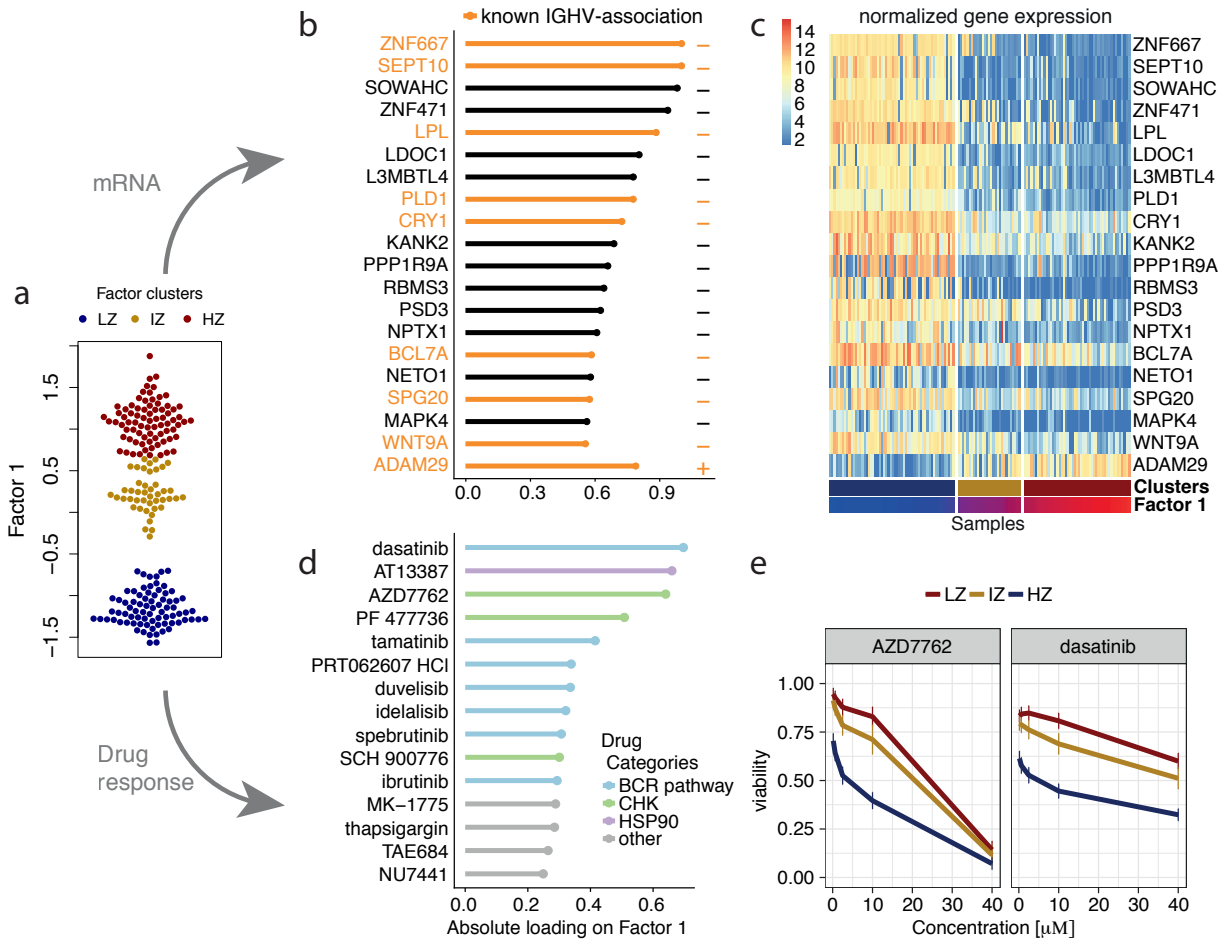


Fig. 2.27 XX

2.6.8.3 Characterisation of other Factors

2.6.8.4 Prediction of clinical outcomes

We conjectured that the integration of the multiple molecular layers could allow a better prediction of clinical response compared to using single omics or naive data integration.
To evaluate the utility of the MOFA factors as predictors of clinical outcomes we fit Cox regression models [45] using the patients' time to next treatment (TTT) as a response variable. Two types of analysis were performed: a univariate analysis where each Factor was independently associated with TTT, and a multivariate analysis where the combination of all Factors were used to predict TTT.

(CHECK) In the univariate Cox models, we observe that Factor 1 (IGHV status), Factor 7 (associated with chemo-immunotherapy treatment prior to sample collection) and Factor 8 (Wnt signalling) were significant predictors of TTT. Accordingly, when splitting patients into binary groups based on the corresponding factor values, we observe clear differences in the survival curves.

In the multivariate Cox model, MOFA (Harrell's C-Index C=0.78) outperformed all other input settings, including single-omic data (C=0.68-0.72), individual genetic markers (C=0.66) as well the concatenated data matrix (C=0.74).

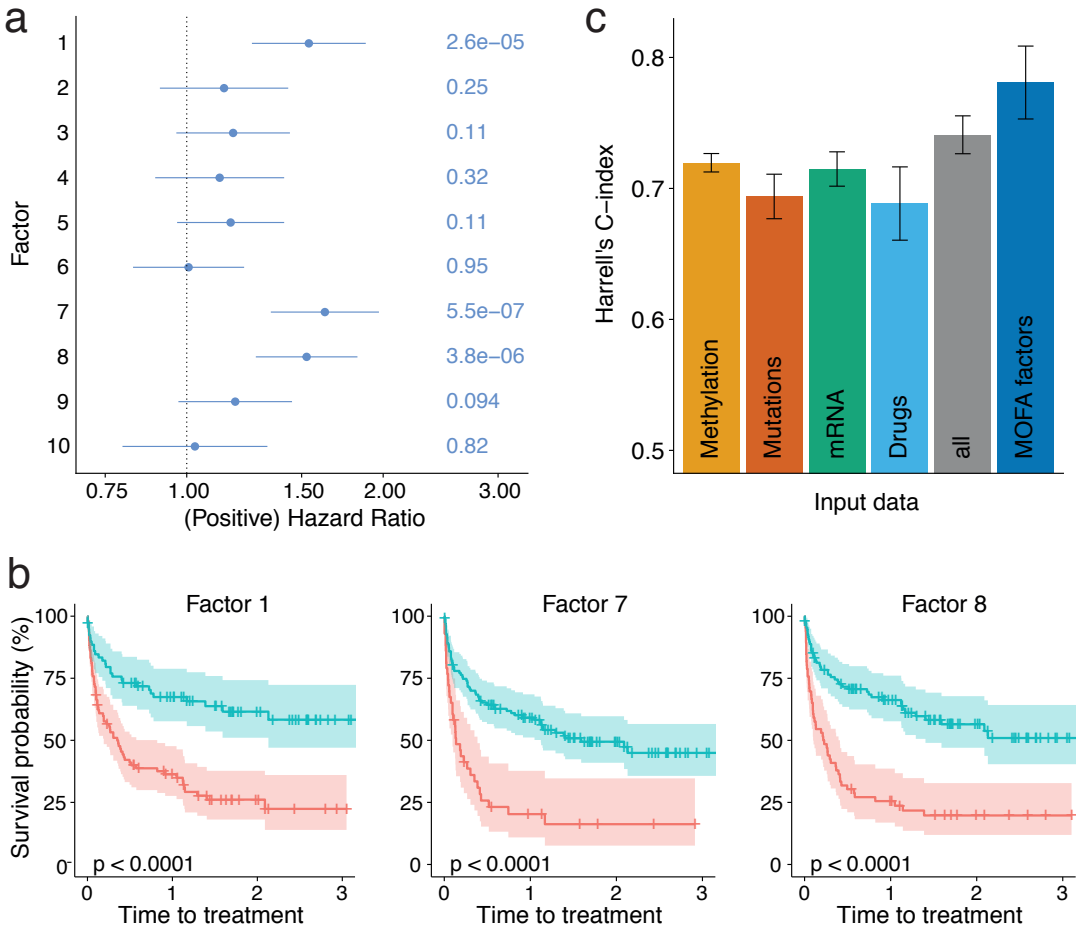


Fig. 2.28 XX

2.6.8.5 Imputation of missing values

A promising application of MOFA is the imputation of missing values as well as entire missing assays, which could massively reduce experimental costs.
The principle of imputation in MOFA follows the same logic as simulating from the generative model: if the factors and weights are available one can reconstruct the data by a simple matrix multiplication:

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

where $\mathbb{E}[\mathbf{Z}]$ and $\mathbb{E}[\mathbf{W}]$ denote the expected values of the variational distributions for the factors and the loadings, respectively. Instead of point estimates, more advanced and fully Bayesian posterior predictive distribution could be obtained by propagating the uncertainity [67]. Yet, given that the variance of the variational distributions tend to be heavily underestimated (see section XX), we did not attempt this approach.

To assess the imputation performance, we trained MOFA models on patients with complete measurements after masking parts of the drug response data. In a first experiment, we masked values at random, and in a second experiment we masked the entire drug response data. We compared the results to some established imputation strategies, including imputation by feature-wise mean,

SoftImpute [144], a k-nearest neighbour method [217].

For both imputation tasks, MOFA consistently yielded more accurate predictions, albeit the differences became less pronounced in the imputation of full assays, a more challenging task.

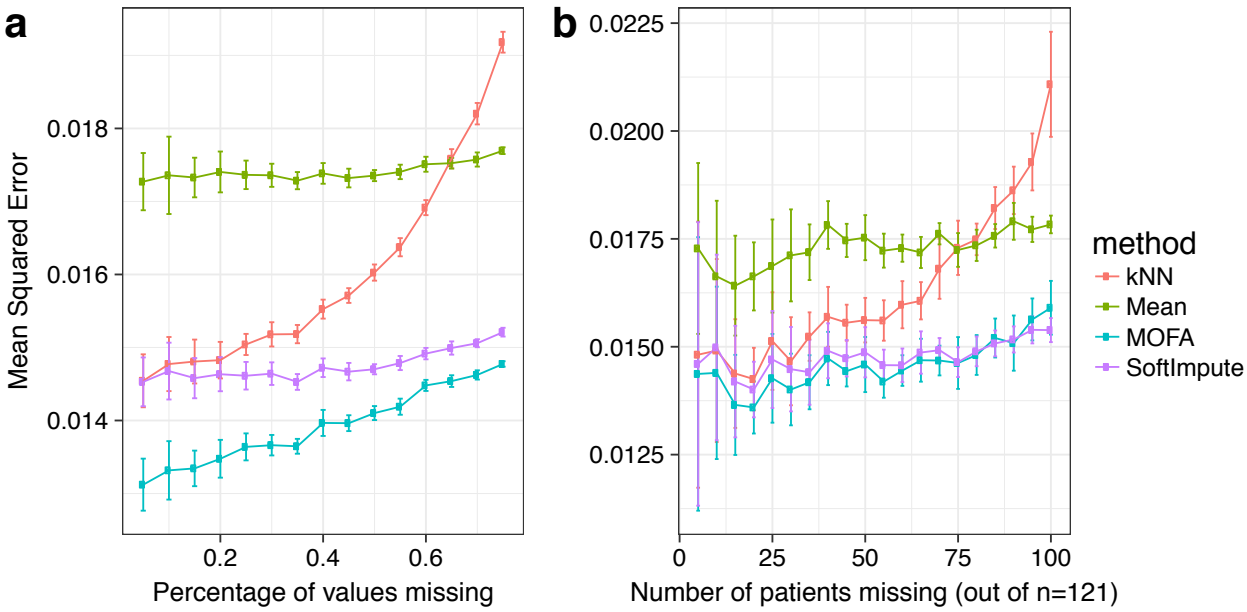


Fig. 2.29 Evaluation of imputation performance in the drug response assay of the CLL data. The y-axis shows averages of the mean-squared error across 15 trials for increasing fractions of missing data (x-axis). Two experiments were considered: (a) values missing at random and (b) entire assays missing at random. Error bars represent standard deviations.

2.6.9 Application to single-cell multi-omics

The emergence of single-cell multi-modal techniques has created open opportunities for the development of novel computational techniques that integrate data sets across multiple modalities [208, 42, 35]. Here, we investigated the potential of MOFA to unravel the heterogeneity in one of the earliest single-cell multi-omics experiments [5].

The data set consists on 87 embryonic stem cells (ESCs) where RNA expression and DNA methylation were simultaneously measured using single-cell Methylation and Transcriptome sequencing (scM&T-seq). Two populations of ESCs were profiled: the first one contains 16 cells grown in 2i media, which induces a native pluripotency state with genome-wide DNA hypomethylation [62]. the second population contains 71 cells grown in serum media, which triggers a primed pluripotency state poised for differentiation [214].

The RNA expression data was processed using standard pipelines to obtain log normalised counts, followed by a selection of the top 5,000 most overdispersed genes [137].

The DNA methylation data was processed as described in Chapter 1. Briefly, for each CpG site, we calculated a binary methylation rate from the ratio of methylated read counts to total read counts. Next, CpG sites were classified by overlapping with genomic contexts, namely promoters, CpG islands and enhancers (defined by the presence of distal H3K27ac marks). Finally, for each annotation we selected the top 5,000 most variable CpG sites with a minimum coverage of 10%

across cells.
Each of the resulting matrices was defined as a separate view for MOFA. The methylation data was modelled with a Bernoulli likelihood and the RNA expression data was modelled with a Gaussian likelihood.

In this data set, MOFA learnt 3 factors (minimum explained variance of 1%). Factor 1 captured the transition from naive to primed pluripotent states, which MOFA links to widespread coordinated changes between DNA methylation and RNA expression (Figures 2.30 and 2.31. Inspection of the gene loadings for Factor 1 pinpoints important pluripotency markers including Rex1/Zpf42 or Essrb [152]. As previously described both in vitro [5]) and in vivo [8], the dynamics of DNA methylation are driven by a genome-wide increase in DNA methylation levels.
Factor 2 captured a second dimension of heterogeneity driven by the transition from a primed pluripotency state to a differentiated state, with RNA loadings enriched with canonical differentiation markers including keratins and annexins [65].

Jointly, the combination of Factors 1 and 2 reconstruct the coordinated changes between the transcriptome and the epigenome along the differentiation trajectory from naive pluripotent cells to differentiated cells. When applying popular integrative clustering algorithms [223, 197, 151], the trajectory is not recovered Figure 2.32, illustrating the importance of learning continuous latent spaces before applying clustering methods.

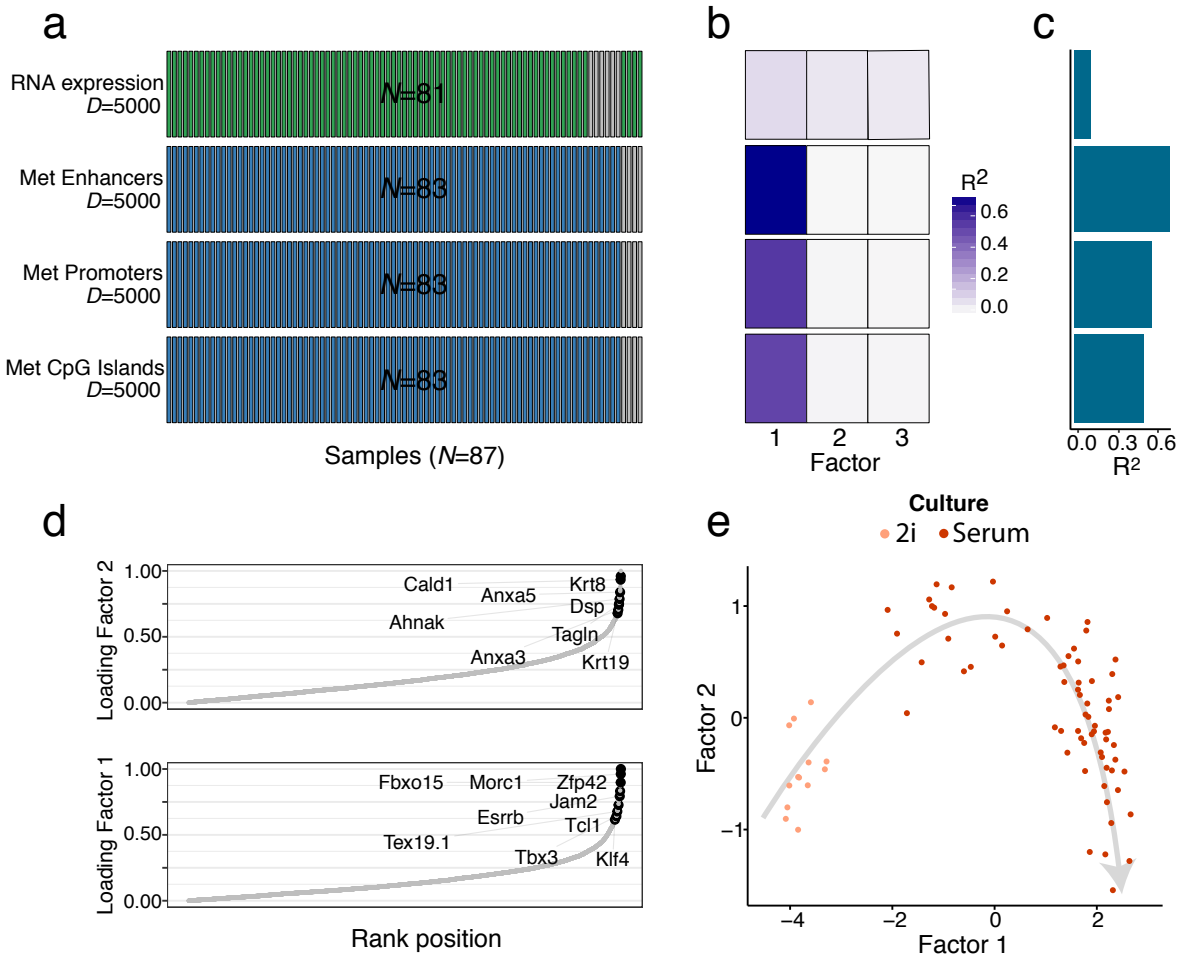

a
RNA expression
D=5000
N=81
Met Enhancers
D=5000
N=83
Met Promoters
D=5000
N=83
Met CpG Islands
D=5000
N=83
Samples (N=87)
b
R2
0.6
0.4
0.2
0.0
1
2
3
Factor
c
0.0
0.3
0.6
R2
d
Loading Factor 2
1.00
0.50
0.00
Cald1
Anxa5
Krt8
Dsp
Ahnak
Tagln
Anxa3
Krt19
Loading Factor 1
1.00
0.50
0.00
Fbxo15
Morc1
Zfp42
Jam2
Esrrb
Tcl1
Tex19.1
Tbx3
Klf4
Rank position
e
Culture
2i
Serum
Factor 2
1
0
−1
−4
−2
0
2
Factor 1

Fig. 2.30 XX

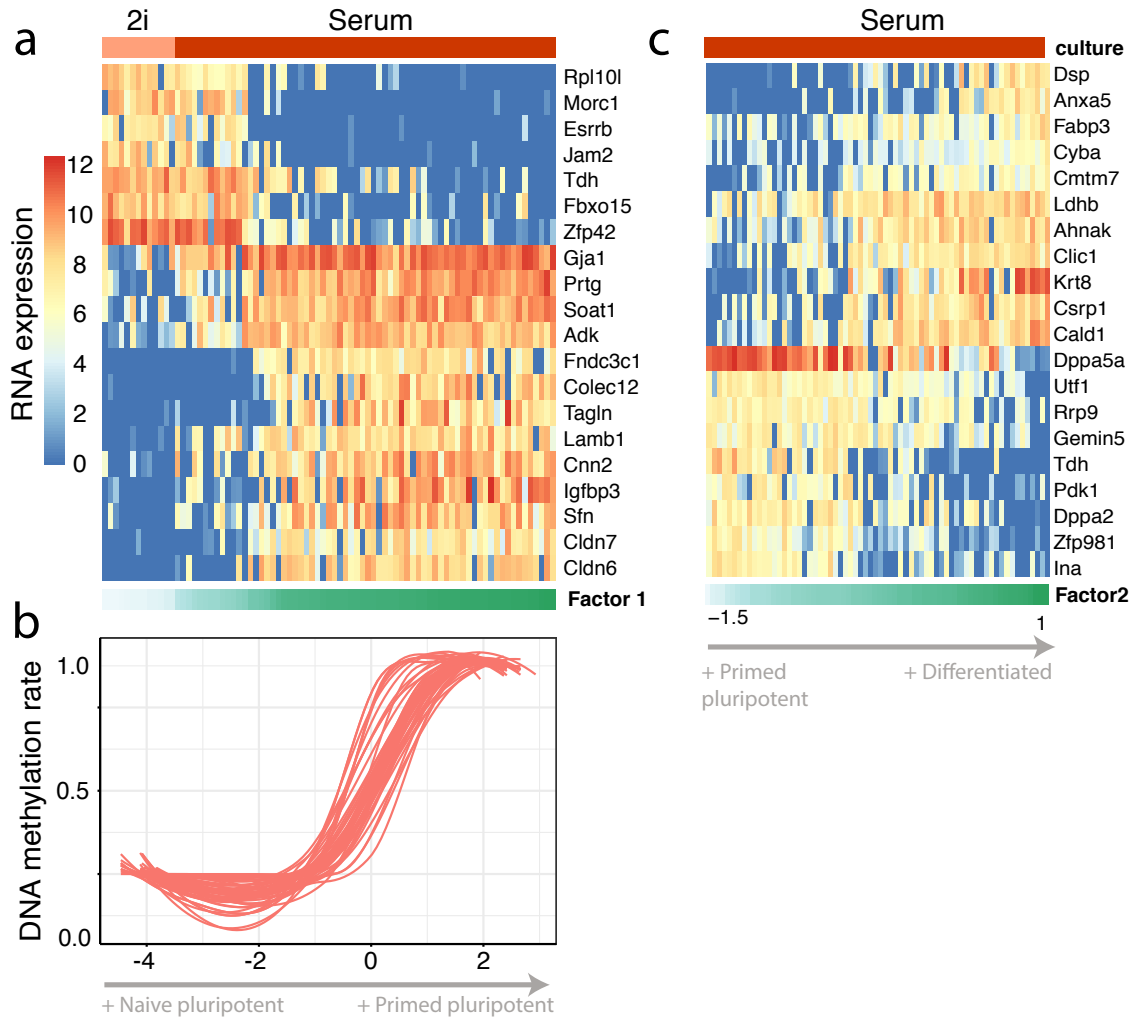


Fig. 2.31 XX

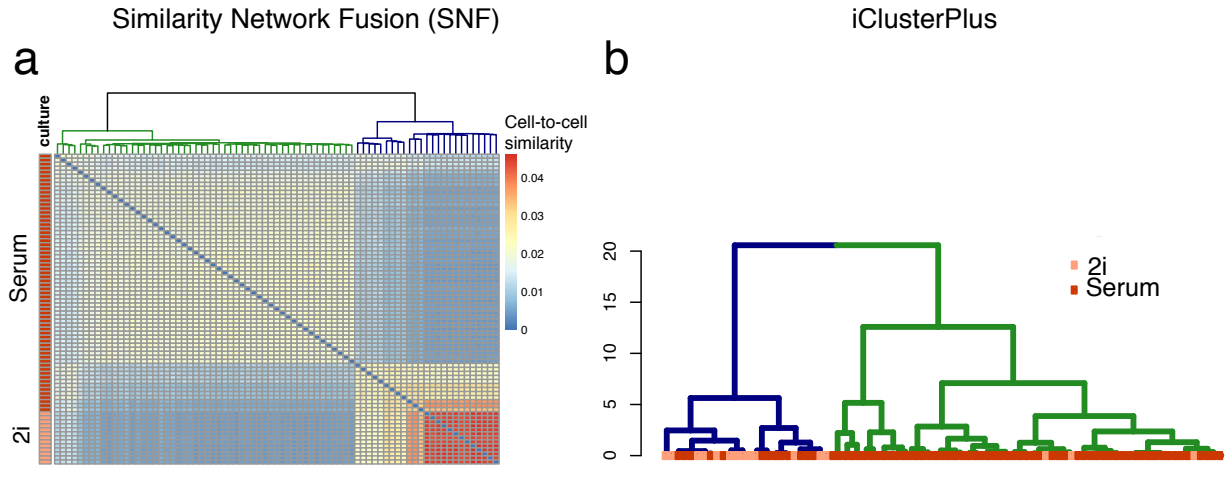


Fig. 2.32 XX

2.6.10 Open perspectives

MOFA addresses important challenges for the integrative analysis of multi-omics applications. Yet, the model is not free of limitations and there are open possibilities for future research, some of which we followed in Chapter 4.

- **Linearity**: this is an assumption that is critical for obtaining interpretable feature loadings. Yet, as in any machine learning technique, there is often a trade-off between explanatory power and interpretability[**XX**]. Non-linear appproaches are particularly appealing in biology, where the drivers of variation often result from the complex (non-linear) interaction between potentially simple (and linear) processes. As such, non-linear approaches, including deep neural networks or variational autoencoders have shown promising results when it comes to dimensionality reduction [131, 54, 136], batch correction[136], denoising [57] or imputation [132]. Notably, few non-linear multi-view factor analysis models exist [49], making it an interesting line of research.

- **Scalability**: the size of biological datasets is rapidly increasing, particularly in the field of single cell sequencing, with some studies reporting more than a milion cells[209, 32].
  When comparing to previous methods that make use of maximum likelihood or sampling-based Bayesian methods, the variatonal framework implemented in MOFA yields a notable improvement in scalability. Yet, in its vanilla form, variational inference becomes prohibitively slow with very large datasets [83, 21, 84], hence motivating the development of even more efficient inference schemes that potentially scale to milions of samples. This line of research is followed in Chapter 4, with the development of a stochastic version of the variational inference algorithm.

- **Sample independence assumption and generalisations to multi-group structures**: the sparsity assumptions in MOFA are based on the principle that features are structured into well-defined views. As such, the activity of the infered factors is also expected to be structured, so that different factors explain variability in different subsets of views (Figure 2.15). Following the same logic, many studies contain structured samples, as either multiple experiments or conditions. The integration of multiple sample groups requires breaking the assumption of independent samples and introducing a prior that captures the structured sparsity at the sample level. This line of research is followed in Chapter 4, with the introduction of a symmetric multi-group and multi-view sparsity prior.

- **Tailored likelihoods for single-cell analysis**: MOFA enables the modular extension to arbitrary non-gaussian likelihoods, provided that they can be locally bounded and integrated into the variational framework (see Section 2.6.6). New likelihood models such as zero-inflated negative binomial distributions [187] could make MOFA more suited to the analysis of single-cell data.

- **Bayesian treatment of predictions**: in the current implementation, after inference we extract point estimates for each variable, namely expectations. While convienient for plotting, this ignores the uncertainity associated with the estimates, one of the main strength of

Bayesian methods. Future extensions could attempt a more comprehensive Bayesian treatment that propagates uncertainity in the downstream analyses, mainly when it comes to making predictions and imputation [67].

- **Incorporation of prior information**: an unsupevised approach is appealing for discovering the principal axes of variation, but sometimes this can yield challenges in the interpretation of factors. Future extensions could exploit the rich information encoded in pathway databases, similar to the approach proposed in [27].

Chapter 3

Single cell multi-omics profiling reveals a hierarchical epigenetic landscape during mammalian germ layer specification

3.1 Introduction

TO-DO...

Chapter 4

A scalable statistical framework for the integrative analysis of single-cell -omics across experiments and data modalities

4.1 Introduction

TO-DO...

Bibliography

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

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

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

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

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

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

[7] A. Armagan, D. B. Dunson, and M. Clyde. "Generalized Beta Mixtures of Gaussians". In: *arXiv e-prints*, arXiv:1107.4976 (July 2011), arXiv:1107.4976. arXiv: 1107.4976 [stat.ME].

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

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

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

[11] A. J. Bannister and T. Kouzarides. "Regulation of chromatin by histone modifications". In: *Cell Research* 21 (Feb. 2011).

[12] D. Barber and W. Wiegerinck. "Tractable Variational Structures for Approximating Graphical Models". In: *NIPS*. 1998.

[13] T. Bayes. "LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S". In: *Philosophical Transactions of the Royal Society of London* 53 (1763), pp. 370–418. DOI: 10.1098/rstl.1763.0053.

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

[15] B. E. Bernstein et al. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells". In: *Cell* 125.2 (2006), pp. 315–326.

[16] P. Bheda and R. Schneider. "Epigenetics reloaded: the single-cell revolution". In: *Trends in Cell Biology* 24.11 (2014), pp. 712–723. ISSN: 0962-8924. DOI: https://doi.org/10.1016/j.tcb.2014.08.010.

[17] C. Bishop. "Variational Principal Components". In: *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*. Vol. 1. IEE, Jan. 1999, pp. 509–514.

[18] C. M. Bishop. "Bayesian PCA". In: *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*. Cambridge, MA, USA: MIT Press, 1999, pp. 382–388. ISBN: 0-262-11245-0.

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

[20] D. M. Blei and M. I. Jordan. "Variational inference for Dirichlet process mixtures". In: *Bayesian Anal.* 1.1 (Mar. 2006), pp. 121–143. DOI: 10.1214/06-BA104.

[21] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *arXiv e-prints*, arXiv:1601.00670 (Jan. 2016), arXiv:1601.00670. arXiv: 1601.00670 [stat.CO].

[22] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[23] M. Braun and J. McAuliffe. "Variational inference for large-scale models of discrete choice". In: *arXiv e-prints*, arXiv:0712.2526 (Dec. 2007), arXiv:0712.2526. arXiv: 0712.2526 [stat.ME].

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

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

[26] J. D. Buenrostro et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In: *Nature Methods* 10 (Oct. 2013).

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

[28] P. Bulian et al. "Mutational status of IGHV is the most reliable prognostic marker in trisomy 12 chronic lymphocytic leukemia". In: *Haematologica* 102.11 (2017), e443–e446. ISSN: 0390-6078. DOI: 10.3324/haematol.2017.170340. eprint: http://www.haematologica.org/content/102/11/e443.full.pdf.

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

[30] J. Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. In: *Science* 357.6352 (2017), pp. 661–667. ISSN: 0036-8075. DOI: 10.1126/science.aam8940. eprint: http://science.sciencemag.org/content/357/6352/661.full.pdf.

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

[32] J. Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. In: *Nature* 566.7745 (2019), pp. 496–502. DOI: 10.1038/s41586-019-0969-x.

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

[34] P. Carbonetto and M. Stephens. “Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies”. In: *Bayesian Anal.* 7.1 (Mar. 2012), pp. 73–108. DOI: 10.1214/12-BA703.

[35] L. Chappell, A. J. Russell, and T. Voet. “Single-Cell (Multi)omics Technologies”. In: *Annual Review of Genomics and Human Genetics* 19.1 (2018). PMID: 29727584, pp. 15–41. DOI: 10.1146/annurev-genom-091416-035324. eprint: https://doi.org/10.1146/annurev-genom-091416-035324.

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

[37] R. Chen and M. Snyder. “Promise of personalized omics to precision medicine”. In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 5.1 (2013), pp. 73–82. DOI: 10.1002/wsbm.1198. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wsbm.1198.

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

[39] S. J. Clark et al. “Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq)”. In: *Nature Protocols* 12 (Feb. 2017).

[40] S. J. Clark et al. “scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells”. In: *Nature Communications* 9.1 (2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-03149-4.

[41] S. J. Clark et al. “Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity”. In: *Genome Biology* 17.1 (2016), p. 72.

[42] M. Colome-Tatche and F. Theis. “Statistical single cell multi-omics integration”. In: *Current Opinion in Systems Biology* 7 (2018). Future of systems biology Genomics and epigenomics, pp. 54–59. ISSN: 2452-3100. DOI: https://doi.org/10.1016/j.coisb.2018.01.003.

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

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

[45] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246.

[46] J. Crombie and M. S. Davids. "IGHV mutational status testing in chronic lymphocytic leukemia". In: *American Journal of Hematology* 92.12 (2017), pp. 1393–1397. DOI: 10.1002/ajh.24808. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.24808.

[47] D. A. Cusanovich et al. "A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility". In: *Cell* 174.5 (2018), 1309–1324.e18. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2018.06.052.

[48] D. A. Cusanovich et al. "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing". In: *Science* 348.6237 (2015), pp. 910–914. ISSN: 0036-8075. DOI: 10.1126/science.aab1601. eprint: http://science.sciencemag.org/content/348/6237/910.full.pdf.

[49] A. Damianou, N. D. Lawrence, and C. H. Ek. "Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis". In: *arXiv preprint arXiv:1604.04939* (2016).

[50] R. N. Damle et al. "Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia." In: *Blood* 94 6 (1999), pp. 1840–7.

[51] Q. Deng et al. "Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells". In: *Science* 343.6167 (2014), pp. 193–196. ISSN: 0036-8075. DOI: 10.1126/science.1245316. eprint: http://science.sciencemag.org/content/343/6167/193.full.pdf.

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

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

[54] J. Ding, A. Condon, and S. P. Shah. "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models". In: *Nature Communications* 9.1 (2018), p. 2002.

[55] M. Emtiyaz Khan and W. Lin. "Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models". In: *arXiv e-prints*, arXiv:1703.04265 (Mar. 2017), arXiv:1703.04265. arXiv: 1703.04265 [cs.LG].

[56] M. Enge et al. "Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns". In: *Cell* 171.2 (2017), 321–330.e14.

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

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

[59] C. Faes, J. T. Ormerod, and M. P. Wand. "Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data". In: *Journal of the American Statistical Association* 106.495 (2011), pp. 959–971. ISSN: 01621459.

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

[61] M. Farlik et al. "Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics". In: *Cell Reports* 10.8 (2015), pp. 1386–1397. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2015.02.001.

[62] G. Ficz et al. "FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency". In: *Cell Stem Cell* 13.3 (2013), pp. 351–359.

[63] R. Fleischmann et al. "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd". In: *Science* 269.5223 (1995), pp. 496–512. ISSN: 0036-8075. DOI: 10.1126/science.7542800. eprint: http://science.sciencemag.org/content/269/5223/496.full.pdf.

[64] M. Frommer et al. "A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands." In: *Proceedings of the National Academy of Sciences* 89.5 (1992), pp. 1827–1831. ISSN: 0027-8424. DOI: 10.1073/pnas.89.5.1827. eprint: https://www.pnas.org/content/89/5/1827.full.pdf.

[65] E. Fuchs. "Keratins as biochemical markers of epithelial differentiation". In: *Trends in Genetics* 4.10 (1988), pp. 277–281. ISSN: 0168-9525. DOI: https://doi.org/10.1016/0168-9525(88)90169-2.

[66] C. Gao, C. D. Brown, and B. E. Engelhardt. "A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects". In: *arXiv e-prints*, arXiv:1310.4792 (2013), arXiv:1310.4792. arXiv: 1310.4792 [stat.AP].

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

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

[69] S. Gravina et al. "Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome". In: *Genome Biology* 17.1 (2016), p. 150.

[70] J. A. Griffiths, A. Scialdone, and J. C. Marioni. "Using single-cell genomics to understand developmental processes and cell fate decisions". In: *Molecular Systems Biology* 14.4 (2018). DOI: 10.15252/msb.20178046. eprint: http://msb.embopress.org/content/14/4/e8046.full.pdf.

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

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

[73] H. Guo et al. "Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing". In: *Genome Research* 23.12 (Dec. 2013), pp. 2126–2135.

[74] Y. Guo et al. "Sufficient Canonical Correlation Analysis". In: *Trans. Img. Proc.* 25.6 (June 2016), pp. 2610–2619. ISSN: 1057-7149. DOI: 10.1109/TIP.2016.2551374.

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

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

[77] T. Hashimshony et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (2012), pp. 666–673. ISSN: 2211-1247. DOI: https://doi.org/10.1016/j.celrep.2012.08.003.

[78] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83. DOI: 10.1186/s13059-017-1215-1.

[79] Y. Hasin, M. Seldin, and A. Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017), p. 83.

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

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

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

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

[84] M. D. Hoffman and D. M. Blei. "Structured Stochastic Variational Inference". In: *arXiv e-prints*, arXiv:1404.4114 (Apr. 2014), arXiv:1404.4114. arXiv: 1404.4114.

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

[86] V. Hore. "Latent Variable Models for Analysing Multidimensional Gene Expression Data". PhD thesis. University of Oxford, 2015.

[87] V. Hore et al. "Tensor decomposition for multiple-tissue gene expression experiments". In: *Nature Genetics* 48.9 (2016), pp. 1094–1100.

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

[89] H. Hotelling. "Relations between two sets of variates". In: *Biometrika* 28.3-4 (Dec. 1936), pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. eprint: http://oup.prod.sis.lan/biomet/article-pdf/28/3-4/321/586830/28-3-4-321.pdf.

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

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

[92] S. Huang, K. Chaudhary, and L. X. Garmire. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods". In: *Frontiers in Genetics* 8 (2017), p. 84. ISSN: 1664-8021. DOI: 10.3389/fgene.2017.00084.

[93] S. Huang, K. Chaudhary, and L. X. Garmire. "More Is Better: Recent Progress in Multi-Omics Data Integration Methods". In: *Front. Genet.* 8 (June 2017), p. 1005.

[94] A. Ilin and T. Raiko. "Practical Approaches to Principal Component Analysis in the Presence of Missing Values". In: *J. Mach. Learn. Res.* 11 (Aug. 2010), pp. 1957–2000. ISSN: 1532-4435.

[95] S. Islam et al. "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq". en. In: *Genome Res.* 21.7 (July 2011), pp. 1160–1167.

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

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

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

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

[100] Z. Jin and Y. Liu. "DNA methylation in human diseases". In: *Genes & Diseases* 5.1 (2018), pp. 1–8. ISSN: 2352-3042. DOI: https://doi.org/10.1016/j.gendis.2018.01.002.

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

[102] N. Kaplan et al. "The DNA-encoded nucleosome organization of a eukaryotic genome". In: *Nature* 458 (Dec. 2008).

[103] C.-A. Kapourani and G. Sanguinetti. "BPRMeth: a flexible Bioconductor package for modelling methylation profiles". In: *Bioinformatics* 34.14 (Mar. 2018), pp. 2485–2486. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bty129. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/34/14/2485/25138157/bty129.pdf.

[104] C.-A. Kapourani and G. Sanguinetti. "Melissa: Bayesian clustering and imputation of single cell methylomes". In: *bioRxiv* (2018). DOI: 10.1101/312025. eprint: https://www.biorxiv.org/content/early/2018/05/02/312025.full.pdf.

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

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

[107] G. Kelsey, O. Stegle, and W. Reik. "Single-cell epigenomics: Recording the past and predicting the future". In: *Science* 358.6359 (2017), pp. 69–75. ISSN: 0036-8075. DOI: 10.1126/science.aan6826. eprint: http://science.sciencemag.org/content/358/6359/69.full.pdf.

[108] S. A. Khan et al. "Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis". In: *Bioinformatics* 30.17 (2014), pp. i497–i504.

[109] E. Kiciman, D. Maltz, and J. C. Platt. "Fast Variational Inference for Large-scale Internet Diagnosis". In: *Advances in Neural Information Processing Systems 20*. Ed. by J. C. Platt et al. Curran Associates, Inc., 2008, pp. 1169–1176.

[110] J. A. Kilgore et al. “Single-molecule and population probing of chromatin structure using DNA methyltransferases”. In: *Methods* 41.3 (2007). Methods Related to the Structure and Function of Eukaryotic Chromatin, pp. 320–332. ISSN: 1046-2023. DOI: https://doi.org/10.1016/j.ymeth.2006.08.008.

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

[112] T. Kivioja et al. “Counting absolute numbers of molecules using unique molecular identifiers”. In: *Nature Methods* 9 (Nov. 2011).

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

[114] A. Klami, S. Virtanen, and S. Kaski. “Bayesian Canonical Correlation Analysis”. In: *J. Mach. Learn. Res.* 14.1 (Apr. 2013), pp. 965–1003. ISSN: 1532-4435.

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

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

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

[118] A. A. Kolodziejczyk et al. “The Technology and Biology of Single-Cell RNA Sequencing”. In: *Molecular Cell* 58.4 (2015), pp. 610–620. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2015.04.005.

[119] S. Komili and P. A. Silver. “Coupling and coordination in gene expression processes: a systems biology view”. In: *Nat. Rev. Genet.* 9 (Jan. 2008), p. 38.

[120] F. Krueger and S. R. Andrews. “Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications”. In: *Bioinformatics* 27.11 (Apr. 2011), pp. 1571–1572. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btr167. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/27/11/1571/13845683/btr167.pdf.

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

[122] A. Lafzi et al. “Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies”. In: *Nature Protocols* 13.12 (2018), pp. 2742–2757.

[123] N. D. Lawrence et al. “Efficient inference for sparse latent variable models of transcriptional regulation”. In: *Bioinformatics* 33.23 (Aug. 2017), pp. 3776–3783. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx508. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/33/23/3776/25168082/btx508.pdf.

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

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

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

[127] E. Leppäaho and S. Kaski. "GFA: exploratory analysis of multiple data sources with group factor analysis". In: *Journal of Machine Learning Research* 18 (2017), pp. 1–5.

[128] Y. Li, F.-X. Wu, and A. Ngom. "A review on machine learning principles for multi-view biological data integration". In: *Briefings in Bioinformatics* 19.2 (Dec. 2016), pp. 325–340. ISSN: 1477-4054. DOI: 10.1093/bib/bbw113. eprint: http://oup.prod.sis.lan/bib/article-pdf/19/2/325/25524236/bbw113.pdf.

[129] Z. Li, S. E. Safo, and Q. Long. "Incorporating biological information in sparse principal component analysis with application to genomic data". In: *BMC Bioinformatics* 18.1 (2017), p. 332.

[130] A. C. Likas and N. P. Galatsanos. "A variational approach for Bayesian blind image deconvolution". In: *IEEE Transactions on Signal Processing* 52.8 (Aug. 2004), pp. 2222–2233. ISSN: 1053-587X. DOI: 10.1109/TSP.2004.831119.

[131] C. Lin et al. "Using neural networks for reducing the dimensions of single-cell RNA-Seq data". In: *Nucleic Acids Research* 45.17 (July 2017), e156–e156. ISSN: 0305-1048. DOI: 10.1093/nar/gkx681. eprint: http://oup.prod.sis.lan/nar/article-pdf/45/17/e156/25366829/gkx681.pdf.

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

[133] R. Lister et al. "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". In: *Cell* 133.3 (2008), pp. 523–536. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2008.03.029.

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

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

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

[137] A. Lun, D. McCarthy, and J. Marioni. "A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2; referees: 3 approved, 2 approved with reservations]". In: *F1000Research* 5.2122 (2016). DOI: 10.12688/f1000research.9501.2.

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

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

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

[141] E. Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5 (2015), pp. 1202–1214.

[142] K. Maloum et al. "IGHV gene mutational status and LPL/ADAM29 gene expression as clinical outcome predictors in CLL patients in remission following treatment with oral fludarabine plus cyclophosphamide". en. In: *Ann. Hematol.* 88.12 (Dec. 2009), pp. 1215–1221.

[143] D. E. Martin-Herranz et al. "cuRRBS: simple and robust evaluation of enzyme combinations for reduced representation approaches". In: *Nucleic Acids Research* 45.20 (Sept. 2017), pp. 11559–11569. ISSN: 0305-1048. DOI: 10.1093/nar/gkx814. eprint: http://oup.prod.sis.lan/nar/article-pdf/45/20/11559/21743171/gkx814.pdf.

[144] R. Mazumder, T. Hastie, and R. Tibshirani. "Spectral Regularization Algorithms for Learning Large Incomplete Matrices". In: *J. Mach. Learn. Res.* 11.Aug (2010), pp. 2287–2322.

[145] S. D. McCabe, D.-Y. Lin, and M. I. Love. "MOVIE: Multi-Omics VIsualization of Estimated contributions". In: *bioRxiv* (2018). DOI: 10.1101/379115. eprint: https://www.biorxiv.org/content/early/2018/07/29/379115.full.pdf.

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

[147] A. Meissner et al. "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis". In: *Nucleic Acids Research* 33.18 (Jan. 2005), pp. 5868–5877. ISSN: 0305-1048. DOI: 10.1093/nar/gki901. eprint: http://oup.prod.sis.lan/nar/article-pdf/33/18/5868/7126600/gki901.pdf.

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

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

[150] T. J. Mitchell and J. J. Beauchamp. "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[151] Q. Mo et al. "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proceedings of the National Academy of Sciences* 110.11 (2013), pp. 4245–4250. DOI: 10.1073/pnas.1208949110.

[152] H. Mohammed et al. "Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation". In: *Cell Reports* 20.5 (), pp. 1215–1228.

[153] F. Morabito et al. "Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment". en. In: *Leuk. Res.* 39.8 (Aug. 2015), pp. 840–845.

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

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

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

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

[158] U. Nagalakshmi et al. "The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing". In: *Science* 320.5881 (2008), pp. 1344–1349. ISSN: 0036-8075. DOI: 10.1126/science.1158441. eprint: http://science.sciencemag.org/content/320/5881/1344.full.pdf.

[159] S. Nakajima and S. Watanabe. "Variational Bayes Solution of Linear Neural Networks and Its Generalization Performance". In: *Neural Computation* 19.4 (2007), pp. 1112–1153.

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

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

[162] C. C. Oakes et al. "DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia". en. In: *Nat. Genet.* 48.3 (Mar. 2016), pp. 253–264.

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

[164] B. Papp and K. Plath. "Pluripotency re-centered around Esrrb". In: *The EMBO Journal* 31.22 (2012), pp. 4255–4257. ISSN: 0261-4189. DOI: 10.1038/emboj.2012.285. eprint: http://emboj.embopress.org/content/31/22/4255.full.pdf.

[165] A. Parle-Mcdermott and A. Harrison. "DNA Methylation: A Timeline of Methods and Applications". In: *Frontiers in Genetics* 2 (2011), p. 74. ISSN: 1664-8021. DOI: 10.3389/fgene.2011.00074.

[166] A. P. Patel et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". In: *Science* 344.6190 (2014), pp. 1396–1401. ISSN: 0036-8075. DOI: 10.1126/science.1254257. eprint: http://science.sciencemag.org/content/344/6190/1396.full.pdf.

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

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

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

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

[171] M. Pilling. "Handbook of Applied Modelling: Non-Gaussian and Correlated Data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4 (2018), pp. 1264–1265. DOI: 10.1111/rssa.12402. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12402.

[172] H. Plesingerova et al. "COBLL1, LPL and ZAP70 expression defines prognostic subgroups of chronic lymphocytic leukemia patients with high accuracy and correlates with IGHV mutational status". en. In: *Leuk. Lymphoma* 58.1 (Jan. 2017), pp. 70–79.

[173] O. Poirion et al. "Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage". In: *Nature Communications* 9.1 (2018), p. 4892.

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

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

[176] A. C. Queirós et al. "A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact". en. In: *Leukemia* 29.3 (Mar. 2015), pp. 598–605.

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

[178] A. Raj, M. Stephens, and J. K. Pritchard. "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". In: *Genetics* 197.2 (2014), pp. 573–589. ISSN: 0016-6731. DOI: 10.1534/genetics.114.164350. eprint: http://www.genetics.org/content/197/2/573.full.pdf.

[179] V. Ramani et al. "Massively multiplex single-cell Hi-C". In: *Nature Methods* 14 (Jan. 2017).

[180] B. H. Ramsahoye et al. "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a". In: *Proceedings of the National Academy of Sciences* 97.10 (2000), pp. 5237–5242. ISSN: 0027-8424. DOI: 10.1073/pnas.97.10.5237. eprint: https://www.pnas.org/content/97/10/5237.full.pdf.

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

[182] R. Ranganath, S. Gerrish, and D. M. Blei. "Black Box Variational Inference". In: *arXiv e-prints*, arXiv:1401.0118 (Dec. 2013), arXiv:1401.0118. arXiv: 1401.0118 [stat.ML].

[183] M. Rattray et al. "Inference algorithms and learning theory for Bayesian sparse factor analysis". In: *Journal of Physics: Conference Series* 197 (Dec. 2009), p. 012002. DOI: 10.1088/1742-6596/197/1/012002.

[184] A. Regev et al. "Science Forum: The Human Cell Atlas". In: *eLife* 6 (Dec. 2017), e27041. ISSN: 2050-084X. DOI: 10.7554/eLife.27041.

[185] S. Remes, T. Mononen, and S. Kaski. "Classification of weak multi-view signals by sharing factors in a mixture of Bayesian group factor analyzers". In: *arXiv preprint arXiv:1512.05610* (2015).

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

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

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

[189] A. B. Rosenberg et al. “Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding”. In: *Science* 360.6385 (Apr. 2018), p. 176.

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

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

[192] G. Sanguinetti, N. D. Lawrence, and M. Rattray. “Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities”. In: *Bioinformatics* 22.22 (Sept. 2006), pp. 2775–2781. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btl473. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/22/22/2775/16851948/btl473.pdf.

[193] L. K. Saul, T. Jaakkola, and M. I. Jordan. “Mean Field Theory for Sigmoid Belief Networks”. In: *arXiv e-prints*, cs/9603102 (Feb. 1996), cs/9603102. arXiv: cs/9603102.

[194] L. K. Saul and M. I. Jordan. “Exploiting Tractable Substructures in Intractable Networks”. In: *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press, 1996, pp. 486–492.

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

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

[197] R. Shen, A. B. Olshen, and M. Ladanyi. “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”. In: *Bioinformatics* 25.22 (2009), pp. 2906–2912. DOI: 10.1093/bioinformatics/btp543.

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

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

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

[201] L. Song and G. E. Crawford. “DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells”. In: *Cold Spring Harbor Protocols* 2010.2 (Feb. 2010), pdb.prot5384.

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

[203] O. Stegle et al. “Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses”. en. In: *Nat. Protoc.* 7.3 (Feb. 2012), pp. 500–507.

[204] G. L. Stein-O’Brien et al. “Enter the Matrix: Factorization Uncovers Knowledge from Omics”. In: *Trends in Genetics* 34.10 (2018), pp. 790–805.

[205] S. M. Stigler. "The Epic Story of Maximum Likelihood". In: *arXiv e-prints*, arXiv:0804.2996 (Apr. 2008), arXiv:0804.2996. arXiv: 0804.2996 [stat.ME].

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

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

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

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

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

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

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

[213] M. K. Titsias and M. Lázaro-Gredilla. "Spike and slab variational inference for multi-task and multiple kernel learning". In: *Advances in neural information processing systems*. 2011, pp. 2339–2347.

[214] M. Tosolini and A. Jouneau. "Acquiring Ground State Pluripotency: Switching Mouse Embryonic Stem Cells from Serum/LIF Medium to 2i/LIF Medium". In: *Embryonic Stem Cell Protocols*. Springer New York, 2016, pp. 41–48. ISBN: 978-1-4939-2954-2. DOI: 10.1007/7651_2015_207.

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

[216] A. Trojani et al. "Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia". en. In: *Cancer Biomark.* 11.1 (2011), pp. 15–28.

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

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

[219] Y. Vasconcelos et al. "Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes". en. In: *Leukemia* 19.11 (Nov. 2005), pp. 2002–2005.

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

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

[222] S. A. Vitak et al. “Sequencing thousands of single-cell genomes with combinatorial indexing”. In: *Nature Methods* 14 (Jan. 2017).

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

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

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

[226] C. Xu, D. Tao, and C. Xu. “A Survey on Multi-view Learning”. In: *arXiv e-prints*, arXiv:1304.5634 (Apr. 2013), arXiv:1304.5634. arXiv: 1304.5634 [cs.LG].

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

[228] I. S. L. Zeng and T. Lumley. “Review of Statistical Learning Methods in Integrated Omics Studies (An Integrated Information Science)”. In: *Bioinformatics and Biology Insights* 12 (2018).

[229] T. Zenz et al. “From pathogenesis to treatment of chronic lymphocytic leukaemia”. en. In: *Nat. Rev. Cancer* 10.1 (Jan. 2010), pp. 37–50.

[230] C. Zhang et al. “Advances in Variational Inference”. In: *arXiv e-prints*, arXiv:1711.05597 (Nov. 2017), arXiv:1711.05597. arXiv: 1711.05597.

[231] X. Zhang et al. “Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems”. In: *Molecular Cell* 73.1 (2019), 130–142.e5. ISSN: 1097-2765. DOI: https://doi.org/10.1016/j.molcel.2018.10.020.

[232] Z. Zhang et al. “Opening the black box of neural networks: methods for interpreting neural network models in clinical applications.” In: *Annals of translational medicine* 6 11 (2018), p. 216.

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

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

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

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

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

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

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