Inaugural dissertation for

obtaining the doctoral degree of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht-Karls-University Heidelberg

Presented by

M.Sc. Danila Bredikhin

born in Moscow, Russia

Oral examination: 20.03.2023

Structured data abstractions and interpretable latent representations for single-cell multimodal genomics

Referees: Prof. Dr. Henrik Kaessmann

Dr. Judith Zaugg

Summary

Single-cell multimodal genomics involves simultaneous measurement of multiple types of molecular data, such as gene expression, epigenetic marks and protein abundance, in individual cells. This allows for a comprehensive and nuanced understanding of the molecular basis of cellular identity and function. The large volume of data generated by single-cell multimodal genomics experiments requires specialised methods and tools for handling, storing, and analysing it.

This work provides contributions on multiple levels. First, it introduces a single-cell multimodal data standard — MuData — designed to facilitate the handling, storage and exchange of multimodal data. MuData provides interfaces that enable transparent access to multimodal annotations as well as data from individual modalities. This data structure has formed the foundation for the multimodal integration framework, which enables complex and composable workflows that can be naturally integrated with existing omics-specific analysis approaches.

Joint analysis of multimodal data can be performed using integration methods. In order to enable integration of single-cell data, an improved multi-omics factor analysis model (MOFA+) has been designed and implemented building on the canonical dimensionality reduction approach for multi-omics integration. Inferring later factors that explain variation across multiple modalities of the data, MOFA+ enables the modelling of latent factors with cell group-specific patterns of activity. MOFA+ model has been implemented as part of the respective multi-omics integration framework, and its utility has been extended by software solutions that facilitate interactive model exploration and interpretation.

The newly improved model for multi-omics integration of single cells has been applied to the study of gene expression signatures upon targeted gene activation. In a dataset featuring targeted activation of candidate regulators of zygotic genome activation (ZGA) — a crucial transcriptional event in early embryonic development, — modelling expression of both coding and non-coding loci with MOFA+ allowed to rank genes by their potency to activate

a ZGA-like transcriptional response. With identification of *Patz1*, *Dppa2* and *Smarca5* as potent inducers of ZGA-like transcription in mouse embryonic stem cells, these findings have contributed to the understanding of molecular mechanisms behind ZGA and laid the foundation for future research of ZGA in vivo.

In summary, this work's contributions include the development of data handling and integration methods as well as new biological insights that arose from applying these methods to studying gene expression regulation in early development. This highlights how single-cell multimodal genomics can aid to generate valuable insights into complex biological systems.

Zusammenfassung

Die multimodale Einzelzellgenomik umfasst die gleichzeitige Messung mehrerer Arten molekularer Daten, wie z. B. Genexpression, epigenetische Markierungen und Proteinhäufigkeit, in einzelnen Zellen. Dies ermöglicht ein umfassendes und nuanciertes Verständnis der molekularen Grundlagen von Zellidentität und -funktion. Die großen Datenmengen, die bei multimodalen Einzelzellgenomikexperimenten anfallen, erfordern spezielle Methoden und Werkzeuge für ihre Verarbeitung, Speicherung und Analyse.

Diese Arbeit liefert Beiträge auf mehreren Ebenen. Erstens wird ein Standard für multimodale Einzelzelldaten - MuData - eingeführt, der die Handhabung, die Speicherung und den Austausch von multimodalen Daten erleichtern soll. MuData bietet Schnittstellen, die einen transparenten Zugriff auf multimodale Annotationen sowie auf Daten aus einzelnen Modalitäten ermöglichen. Diese Datenstruktur bildet die Grundlage für das multimodale Integrationsframework, das komplexe und zusammengestellte Arbeitsabläufe ermöglicht, die einfach in bestehende omics-spezifische Analyseansätze integriert werden können.

Die gemeinsame Analyse multimodaler Daten kann mithilfe von Integrationsmethoden durchgeführt werden. Um die Integration von Einzelzelldaten zu ermöglichen, wurde ein verbessertes Multi-omics-Faktoranalysemodell (MOFA+) entwickelt und implementiert, das auf dem Ansatz der kanonischen Dimensionalitätsreduktion für die Multi-omics-Integration aufbaut. MOFA+ ermöglicht die Modellierung latenter Faktoren mit zellgruppenspezifischen Aktivitätsmustern, indem es spätere Faktoren ableitet, welche die Variation über mehrere Modalitäten der Daten erklären. Das MOFA+-Modell wurde als Teil des entsprechenden Multi-omics-Integrationsframework implementiert, und sein Nutzwert wurde durch Softwarelösungen erweitert, welche interaktive Modellexploration und -interpretation erleichtern.

Das neu verbesserte Modell für die Multi-omics-Integration einzelner Zellen wurde auf die Untersuchung von Genexpressionssignaturen bei gezielter Genaktivierung angewendet. In einem Datensatz mit gezielter Aktivierung von potenziellen Regulatoren für die Aktivierung des zygotischen Genoms (ZGA) - ein entscheidendes Transkriptionsereignis in der frühen

Embryonalentwicklung - ermöglichte die Modellierung der Expression sowohl kodierender als auch nicht kodierender Loci mit MOFA+ die Einstufung von Genen nach ihrer Fähigkeit, eine ZGA-ähnliche Transkriptionsreaktion zu aktivieren. Mit der Identifizierung von *Patz1*, *Dppa2* und *Smarca5* als starke Auslöser der ZGA-ähnlichen Transkription in embryonalen Stammzellen der Maus haben diese Ergebnisse zum Verständnis der molekularen Mechanismen hinter der ZGA beigetragen und die Grundlage für die künftige Erforschung der ZGA in vivo gelegt.

Zusammenfassend lässt sich sagen, dass der Beitrag dieser Arbeit in der Entwicklung von Methoden zur Datenverarbeitung und -integration sowie in neuen biologischen Erkenntnissen besteht, die sich aus der Anwendung dieser Methoden zur Untersuchung der Genexpressionsregulation in der frühen Entwicklung ergeben haben. Dies zeigt, wie die multimodale Einzelzellgenomik dazu beitragen kann, wertvolle Einblicke in komplexe biologische Systeme zu gewinnen.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

From measuring RNA abundance in tissues to quantifying information across multiple layers of gene regulation in individual cells, a rapid progress of experimental and analytical techniques has deepened our understanding of cell and systems biology. This chapter provides an overview of the key experimental and computational advances behind single-cell omics data acquisition, storage, analysis and interpretation. The figures in this chapter were created by myself unless stated otherwise.

1.1 A brief history of single-cell genomics

By no means comprehensive, this section aims to provide a concise introduction into single-cell RNA sequencing and single-cell multimodal omics.

1.1.1 Single-cell transcriptomics

1.1.1.1 RNA sequencing

Under the modern model of transcription — DNA-dependent RNA synthesis — gene expression is stochastic in nature, i.e. gene expression does fluctuate in constant environmental conditions (Kærn et al., 2005; Marinov et al., 2014). The key steps of transcription are orchestrated by several biochemical reactions including transitions between repressed and active promoter states, mRNA synthesis, splicing and degradation (Kærn et al., 2005).

Single-stranded copies of genes are produced in the process of transcription at the scale of thousands to hundreds thousands molecules per cell (Marinov et al., 2014). Quantifying the relative abundance of different transcripts can serve as a description of cell's identity as well as its future activity and has become a major phenotyping method to describe biological systems with molecular resolution (Mortazavi et al., 2008; Zhong Wang et al., 2009). For instance, gene expression regulation and its relation to phenotype has been extensively studied in development, with perturbations such as RNA interference (RNAi) (Spitz and Furlong, 2006) or lately with single-cell RNA sequencing (Griffiths et al., 2018).

Differences in RNA abundance have been shown to be mediators of genetic variants located in expression quantitative trait loci (eQTLs) — genomic regions carrying variation that influences gene expression (Albert and Kruglyak, 2015). These differences can then propagate to the variation at the protein level. In addition to genetic variation by which protein product is altered, gene expression regulation changes can also lead to phenotypic change and phenotypic evolution (Harrison et al., 2012). Beyond changes in RNA abundance, phenotypic traits depend on RNA processing (splicing, polyadenylation) as well as its stability, translation and structure (Manning and Cooper, 2017).

RNA sequencing (RNA-seq) is the technology that enables quantitative transcriptome-wide gene expression profiling, which became an indispensable way to study biological systems (Stark et al., 2019; Zhong Wang et al., 2009). During RNA-seq, a pool of RNAs is converted to complementary DNA (cDNA) molecules, which are then sequenced in a high-throughput manner (Zhong Wang et al., 2009). As a methodological approach, RNA-seq

has combined a few important advancements that arguably predetermined its success over the predating techniques. Unlike microarray methods, which are based on the hybridisation of oligonucleotide probes targeting specific genes, adaptor-based approach makes RNA-seq unbiased; the latter also has a higher dynamic range. In contrast to Sanger sequencing of cDNA libraries, high-throughput DNA sequencing technologies have provided a way to quantify individual transcripts at scale. Overall, as an approach which is more generic and scalable — both vertically (transcriptome-wide) and horizontally (more samples), — combined with numerous methodological and technical improvements (Stark et al., 2019), RNA-seq has defined the landscape of genome-wide scale molecular biology in the beginning of the XXI century.

1.1.1.2 Single-cell RNA-sequencing

While RNA-seq has historically been used to profile mixtures of cells («in bulk»), seminal work by F. Tang et al., 2009 has demonstrated that it can also be applied to a single mouse blastomere. Since then, single-cell RNA sequencing (scRNA-seq) has witnessed an exponential growth in terms of the number of studies and publications (Svensson, da Veiga Beltrame, et al., 2020) as well as in terms of the scale of the experimental design (Svensson, Vento-Tormo, et al., 2018).

The key to scRNA-seq technological advances combines ability to deal with small quantities of RNA with increased throughput. While the general idea of transcriptome profiling on the level of single-cell holds for different experimental techniques, there's value in discussing similarities and major differences across commonly used approaches and platforms. With these approaches trying to achieve untargeted amplification of the whole transcriptome in individual cells, they start with converting RNA to cDNA, which is then amplified via in vitro transcription (IVT) or polymerase chain reaction (PCR) (Svensson, Vento-Tormo, et al., 2018). Notably, for the latter, an oligo(dT) primer targeting poly(A)-tail is frequently used to generate the first cDNA strand (Kolodziejczyk et al., 2015).

Cells are typically isolated in individual wells on a plate or droplets on a microfluidic chip (Klein et al., 2015; Kolodziejczyk et al., 2015). Each well or droplet contains necessary components such as sequencing adapters, cell barcodes, primers. In order to pull and analyse material from multiple cells and samples, molecular barcodes are added to cDNA. Strategies like combinatorial indexing allow for multiplex barcoding of thousands of cells per experiment (Cusanovich et al., 2015).

One of the challenges of early single-cell assays has been an amplification bias that arises during the PCR, which affects RNA quantification (Ziegenhain et al., 2017). A key solution to address this has been the introduction of unique molecular identifiers (UMIs) (Islam et al., 2014; Kivioja et al., 2012). Their addition makes it possible to distinguish PCR duplicates from biological variation and computationally remove the former by collapsing reads that map to the same place in the transcriptome and share their UMI. Computational correction for PCR duplicates using UMIs has proven to improve the power and false discovery rate of differential gene expression analyses (Parekh et al., 2016).

Arising from low amount of starting material, another challenge of single-cell assays is low capture efficiency (Islam et al., 2014) resulting in high data sparsity. Addressing it requires experimental advances leading to improved capture efficiency (Hagemann-Jensen, Ziegenhain, P. Chen, et al., 2020) as well as specific computational approaches (see Section 1.2.2.1).

Overall, single-cell transcriptome sequencing has been shown to provide accurate quantitative measurements of RNA abundance in individual cells and to recapitulate bulk transcriptome complexity when large amounts of cells are sequenced (A. R. Wu et al., 2014). Taken together, the discussed advances of transcriptome-wide profiling have opened the door to applying scRNA-seq to various biological systems, and in the past decade we witnessed a rapid growth of the scale of scRNA-seq experiments (Svensson, Vento-Tormo, et al., 2018). Using the database with information about single-cell studies (Svensson, da Veiga Beltrame, et al., 2020), I visualised the reported number of cells in them in the figure ~1.1. Highlighted are some key technologies such as CEL-Seq (Hashimshony et al., 2012), Smart-seq2 (Picelli, Björklund, et al., 2013), inDrop (Klein et al., 2015), Drop-seq (Macosko et al., 2015), Perturb-Seq (Dixit et al., 2016), Microwell-Seq (Han, R. Wang, et al., 2018), Slide-seq (Rodriques et al., 2019), Smart-seq3xpress (Hagemann-Jensen, Ziegenhain, and Sandberg, 2022) as well as largest studies such as the ones on mouse brain (Saunders et al., 2018; Zeisel et al., 2018), mouse organogenesis (J. Cao, Spielmann, et al., 2019) and human development (J. Cao, O'Day, et al., 2020).

Two widely established (Svensson, da Veiga Beltrame, et al., 2020) technologies for single-cell transcriptomics are Chromium by 10x Genomics (Zheng et al., 2017) and Smart-seq2 (Picelli, Faridani, et al., 2014). With each technology providing its own set of advantages (Kashima et al., 2020; X. Wang et al., 2021), the preference of one over another is dictated by the research question and the respective aims of the study.

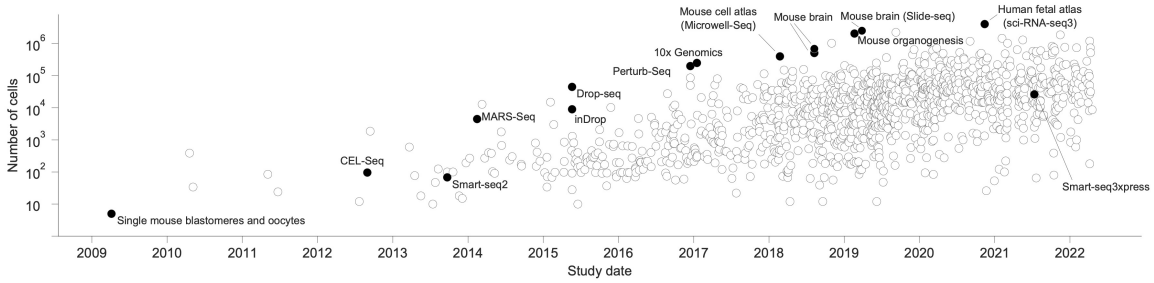

Fig. 1.1 Increasingly large-scale transcriptome profiling enabled by scRNA-seq technological advances
Reported number of cells in each study is plotted against the publication date.

Smart-seq2 (Picelli, Faridani, et al., 2014), an updated version of the Smart-seq protocol (Ramsköld et al., 2012), provides full-length RNA sequencing. This is achieved by taking advantage of the template-switching activity of Moloney murine leukemia virus (MMLV) reverse transcriptase (Y. Y. Zhu et al., 2001), as I illustrate with Figure 1.2. As it is a plate-based method, its applications are limited to hundreds of cells (dictated by being based on 96-well plates). The full-length nature of Smart-seq2 data enables applications like detecting allele-specific expression at stages of mouse preimplantation development (Q. Deng et al., 2014) or identifying splicing events (Arzalluz-Luque and Conesa, 2018; Huang and Sanguinetti, 2021). The most recent improvements to the Smart-seq2 protocol such as Smart-seq3 (Hagemann-Jensen, Ziegenhain, P. Chen, et al., 2020), Smart-seq3xpress (Hagemann-Jensen, Ziegenhain, and Sandberg, 2022) and FLASH-seq (Hahaut et al., 2022) also incorporate UMIs as part of the workflow and provide greater scalability.

Microdroplet-based Chromium system by 10x Genomics (Zheng et al., 2017) offers scalability to many cells (thousands) at the expense of the amount of information per cell: only $3'$- or $5'$-ends of mRNA are targeted with a few dozen thousand reads per cell (Figure 1.2). With its low cost and high throughput, this technology enabled complex experiments at scale, as discussed further in section 1.1.2.

Higher scale and lower cost for single-cell RNA methods has also been achieved by targeted sequencing, which includes enrichment of transcripts of interest (Mercer et al., 2014; Pokhilko et al., 2021). Such targeted approaches have recently enabled selective amplification of genes of interest in screens coupling single-cell transcriptomics with genetic perturbations (Schraivogel et al., 2020).

Illumina short-read sequencing is typically used to effectively convert the library to the digital information that can be analysed (Picelli, Faridani, et al., 2014; Zheng et al., 2017).

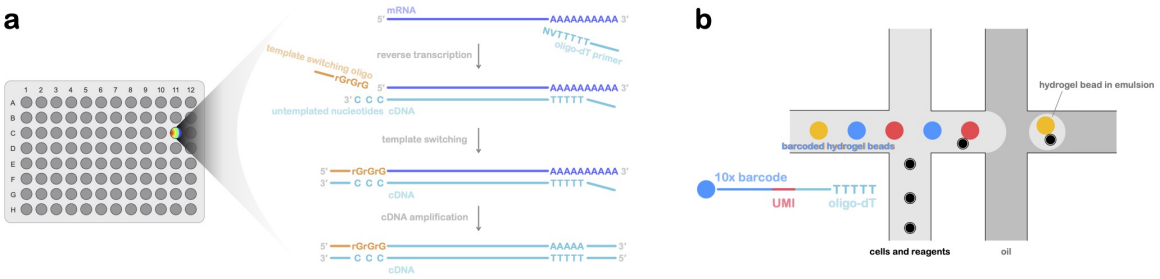

Fig. 1.2 Schematic illustration of plate-based and microdroplet-based approaches
a: Plate-based full-length scRNA-seq. Reactions happen in wells. Depicted are reverse transcription, first strand synthesis with template switching and cDNA amplification. **b:** Microdroplet-based 3′ scRNA-seq. Barcoded beads are combined with cells in droplets.

The libraries generated by either RNA sequencing method can also be used in conjunction with long-read sequencing technologies such as Nanopore (Lebrigand et al., 2020) or PacBio (Gupta et al., 2018) as well as for sequencing with the most recently developed techniques such as mostly-natural sequencing by synthesis by Ultima Genomics (Simmons et al., 2022). Moreover, direct RNA sequencing is also possible on an array of nanopores. This holds potential to enable the collection of more information (RNA modifications) as well as to remove the PCR bias from the sequencing data (Garalde et al., 2018).

1.1.2 Applications of single-cell transcriptomics

Single-cell transcriptomics has become widely adopted for basic and translational biomedical research. Across these use cases, single-cell technologies have increased resolution and enabled the characterisation of molecular changes at new dimensions ranging from qualitative changes (cell type composition) to quantitative variation (gene expression) to dynamic insights into biological processes (cell lineages).

1.1.2.1 Immunology

Immunology has seen great advances driven by scRNA-seq in studying the cellular heterogeneity of the immune system as well as its development and disease (H. Chen et al., 2019; Domínguez Conde et al., 2022; Giladi and Amit, 2018; Papalexi and Satija, 2018). Static picture of leukocyte composition and their transcriptional landscape constructed with scRNA-seq have allowed to propose updated cell type classifications, such as the one for monocytes and dendritic cells (Villani et al., 2017). Undoubtedly, the fact that peripheral

blood is easier to obtain than many other (human) tissue samples without requiring complex tissue dissociation protocols has contributed to the emergence of single-cell studies of peripheral blood mononuclear cells (PBMCs). Furthermore, constructing an atlas of tissue-resident immune cells across multiple donors and organs offered unique opportunities to dissect cell types and cell states in myeloid, B cell and T cell compartments (Domínguez Conde et al., 2022). As part of that atlas, the study of the heterogeneity of lymphocyte populations has been complemented with VDJ sequencing of their T- and B-cell receptors (TCRs and BCRs). As information about the TCR repertoire being crucial for health and disease (Attaf et al., 2015), there has been focus on targeted single-cell TCR sequencing (T. D. Wu et al., 2020). For a system that spans across the whole organism, the emergence of cross-organ studies (Han, R. Wang, et al., 2018) as well as cross-tissue immune cell atlases (Domínguez Conde et al., 2022) is important.

Single-cell transcriptomics has also allowed to profile cell type composition and expression changes in autoimmune diseases such as multiple sclerosis, which targets the central nervous system (Schafflick et al., 2020), or systemic lupus erythematosus (Nehar-Belaid et al., 2020). Notably, single-cell resolution allows to query how genetics affects expression in cell type-specific manner (expression quantitative trait loci, eQTLs) (Perez et al., 2022), and for complex diseases like that, immune cells across tissues such as blood and cerebrospinal fluid have to be studies (Schafflick et al., 2020).

Another direction of immunological single-cell studies has been concerned with the development of the immune system (H. Chen et al., 2019). The goal of such studies devoted to cell differentiation is to reconstruct hematopoietic lineage and to identify cell fate decisions during hematopoietic stem cell differentiation (Paul et al., 2015). Combining single-cell transcriptomics with other assays allowed to reconstruct the whole immune system in development, from yolk sac to primary hematopoietic sites to peripheral organs (Suo et al., 2022).

1.1.2.2 Developmental biology

With global changes to the transcriptional landscape in early development, early embryogenesis in model organisms such as mice has been studied through the lens of single-cell transcriptomics (Pijuan-Sala et al., 2019). Dissection of major cell types in developing mouse embryos with particularly large amounts of cells profiled (2 million cells) was made

possible by newly advanced techniques such as combinatorial indexing (sci-RNA-seq3) (J. Cao, Spielmann, et al., 2019).

Beyond model organisms, whole-animal single-cell transcriptomics has made it possible to dissect complex animals such as a flatworm *Schmidteamediterranea* (Plass et al., 2018) or a hydrozoan medusa *Clytia hemisphaerica* (Chari, Weissbourd, et al., 2021) with single-cell resolution.

1.1.2.3 Neurobiology

Single-cell genomics aided advances in understanding identities of brain cells with the transcriptomes of the murine neocortical areas (Tasic et al., 2018) and developing mouse brain and spinal cord (Rosenberg et al., 2018) being just a few examples. Combined with epigenomics measurements as well as physiological, anatomical and morphological properties (Gouwens et al., 2020; Scala et al., 2021; Z. Zhang et al., 2021), gene expression measurements with single-cell resolution now allow for comprehensive evaluation of neural circuits (Armand et al., 2021). More, these techniques have contributed to our understanding of the evolutionary relationships between neuronal cell types and brain regions such as mammalian cerebellum (Sepp et al., 2021) and enabled comparative transcriptomics on cerebral organoids, e.g. comparing the ones from chimpanzee, macaque and human (Kanton et al., 2019).

1.1.2.4 Atlases

Last but not least, cross-tissue and cross organ atlases allow to create a map of transcriptional activity of most of the cell types in complex organisms such as humans and mice. This scale has been achieved thanks to global initiatives and consortia such as Human Cell Atlas (Rozenblatt-Rosen et al., 2017), Tabula Sapiens (Consortium* et al., 2022) and Human Developmental Cell Atlas (Haniffa et al., 2021). Analogous scale has been achieved on mice with 20 mouse organs profiled as part of the Tabula Muris Consortium (Schaum et al., 2018), later extended to mice of different age (Almanzar et al., 2020). Such efforts inevitably come with technological challenges, both experimental and analytical, driving new advancements such as Microwell-seq for high-throughput and low-cost scRNA-seq, which has been used to describe mouse (Han, R. Wang, et al., 2018) and human (Han, Zhou, et al., 2020) cell landscapes. There also have been more focused efforts to profile complex systems such as the immune system, developing (Suo et al., 2022) or adolescent (Domínguez Conde et al., 2022),

or the nervous system, also developing (La Manno, Siletti, et al., 2021) or adolescent (Ortiz et al., 2020; Zeisel et al., 2018), in order to produce atlases with high cell type resolution.

1.1.3 Single-cell multimodal omics

With high-throughput techniques providing insights to all three cornerstones of the central dogma of molecular biology, DNA, RNA and protein (Crick, 1970), at the level of individual cells, combining these layers of information (Figure 1.3) brings new opportunities for gaining a holistic view of a cell (Efremova and Teichmann, 2020).

Described in the previous section, scRNA-seq is destructive by the nature of the techniques. While in bulk studies there is an option of making measurements in another modality on the part of the same sample, e.g. adjacent tissue, single-cell resolution requires involvement of non-destructive techniques and/or ingenious separation of cell compartments, e.g. nucleus from the cytoplasm, in order to be able to perform multiple destructive measurements on the same cell. An alternative is to first apply a non-destructive technique such as cytometry before sequencing (Efremova and Teichmann, 2020).

There are a few key challenges in obtaining mutlimodal measurements (Efremova and Teichmann, 2020). Depending on the type of the omics measurements, the scalability and sensitivity of assays might be limited, and combining destructive assays requires further developments and innovations for signal amplification from more molecular layers. Furthermore, the data coverage in multimodal omics assays is typically sparse, and the resulting data provides further analysis challenges with its high sparsity and noise levels.

Importantly, single-cell multimodal omics disentangles stochastic processes from cell-to-cell heterogeneity (C. Zhu et al., 2020). This multimodal techniques allows to define relations between molecular layers from the data as opposed to inferring them from unimodal measurements (C. Zhu et al., 2020), and coupling between data modalities is regarded to be one of the key challenges in multimodal integration (Argelaguet, Cuomo, et al., 2021; Lähnemann et al., 2020). Simultaneously capturing different molecular layers in an omics-wide manner is a serious technical challenge, and the progress at resolving it will allow to better understand generative processes behind different molecular layers as well as their interactions.

1.1.3.1 Single-cell transcriptomics combined with epigenomics

ScRNA-seq has been increasingly frequently combined with other techniques such as genome sequencing (Macaulay, Haerty, et al., 2015) and, more prominently, epigenomics including bisulphite sequencing to assess DNA methylation (Smallwood et al., 2014) or assay for transposase-accessible chromatin followed by sequencing (ATAC-seq) to quantify accessibility of chromatin (Buenrostro et al., 2015).

The idea behind ATAC-seq and similar protocols is to preserve chromatin structure and DNA-binding proteins and to then expose chromatin to a transposase enzyme Tn5, which fragments DNA and inserts sequencing adapters — oligonucleotides later used for amplification and sequencing (Grandi et al., 2022). ATAC-seq allows to generate data similar to DNase-seq or MNase-seq with reduced cost, resources and time consumption and with a lower input material requirements (less cells) (Grandi et al., 2022). Genomic loci enriched for transposition events are then views as more accessible and can correspond to promoters, enhancers and other functional elements.

As transcription is driven by cis- and trans-regulatory elements, being able to access the state of the chromatin state in addition to the gene expression profile in the same cell is pivotal for modelling and understanding gene expression (J. Cao, Cusanovich, et al., 2018; S. Chen et al., 2019; S. Ma et al., 2020). For instance, it makes it possible to link cis-regulatory sites in the genome to their target genes on the basis of the chromatin-expression covariance across large numbers of cells and cell populations and conditions (J. Cao, Cusanovich, et al., 2018).

An important observation is there's a good congruence between cell types defined based on their chromatin accessibility and the ones defined based on their gene expression (S. Ma et al., 2020). Such data can then be used to infer gene regulatory modules, e.g. by quantifying the interactions between transcription factors (TFs) and their targets (Fiers et al., 2018). Digging deeper into these gene regulatory networks, the activity of transcription factors and how they regulate expression of the downstream genes can be then studied (Schep et al., 2017). Importance of TFs makes sense in the light of evolution: core regulatory complexes of transcription factors seem to define cell type identity (Arendt et al., 2016).

DNA modifications such as cytosine methylation in DNA (5mC) also play a crucial role in gene expression regulation and transcriptional programs maintenance and can be profiled with single-nucleotide resolution assays such as scBS-seq (Smallwood et al., 2014). DNA methylation is commonly measured using the bisulfite conversion — only unmethylated

cytosine residues are deaminated into uracils, which can be then identified in sequencing reads. Bisulfite conversion has also been effectively combined with other assays to produce multiple levels of information from one experiment such as chromatin conformation in Methyl-HiC (G. Li et al., 2019) or gene expression in scM&T-seq (Angermueller, S. J. Clark, et al., 2016). Building on the latter, labelling accessible chromatin with a methyltransferase before bisulfite conversion and sequencing enables trimodal measurements for each cell in scNMT-seq (S. J. Clark et al., 2018).

1.1.3.2 Single-cell transcriptomics combined with protein measurements

Combining measurements of RNA with measuring protein abundance allows to better resolve cell populations and, more broadly, cell phenotypes with subtle differences at the level of transcripts. As such, transcriptomics measurements were combined with measuring abundance of surface protein markers (epitopes) in CITE-seq (Stoeckius et al., 2017) and REAP-seq (Peterson et al., 2017), intracellular (phospho-)proteins in RAID (Gerlach et al., 2019) or intracellular protein activity in INs-seq (Katzenelenbogen et al., 2020).

The rapid progress in making it possible to profile more modalities per cell has been lately demonstrated by the emergence of trimodal assays such as TEA-seq (Swanson et al., 2021) and DOGMA-seq (Mimitou et al., 2020), that combine gene expression with chromatin accessibility and protein epitope measurements, NEAT-seq (A. F. Chen et al., 2022), in which intranuclear proteins are measured.

1.1.3.3 Novel opportunities offered by single-cell multimodal omics

Naturally, high-throughput (unimodal) single-cell methods applied to variety of biological contexts as described above have been further extended to include multiple modalities. These multimodal methods can provide a higher cell state resolution and enable exploring relationships between modalities and mechanisms behind them. For instance, links between the transcriptome, DNA methylation and DNA accessibility have been studied in differentiating mouse embryonic stem cells, a multi-view perspective enabled by scNMT-seq (S. J. Clark et al., 2018), which was later applied to mouse gastrulation (Argelaguet, S. J. Clark, et al., 2019). Similarly, scM&T-seq has been used to study age-related changes in heterogeneity of murine stem cells, measured at the level of DNA methylation and gene expression (Hernando-Herraez et al., 2019).

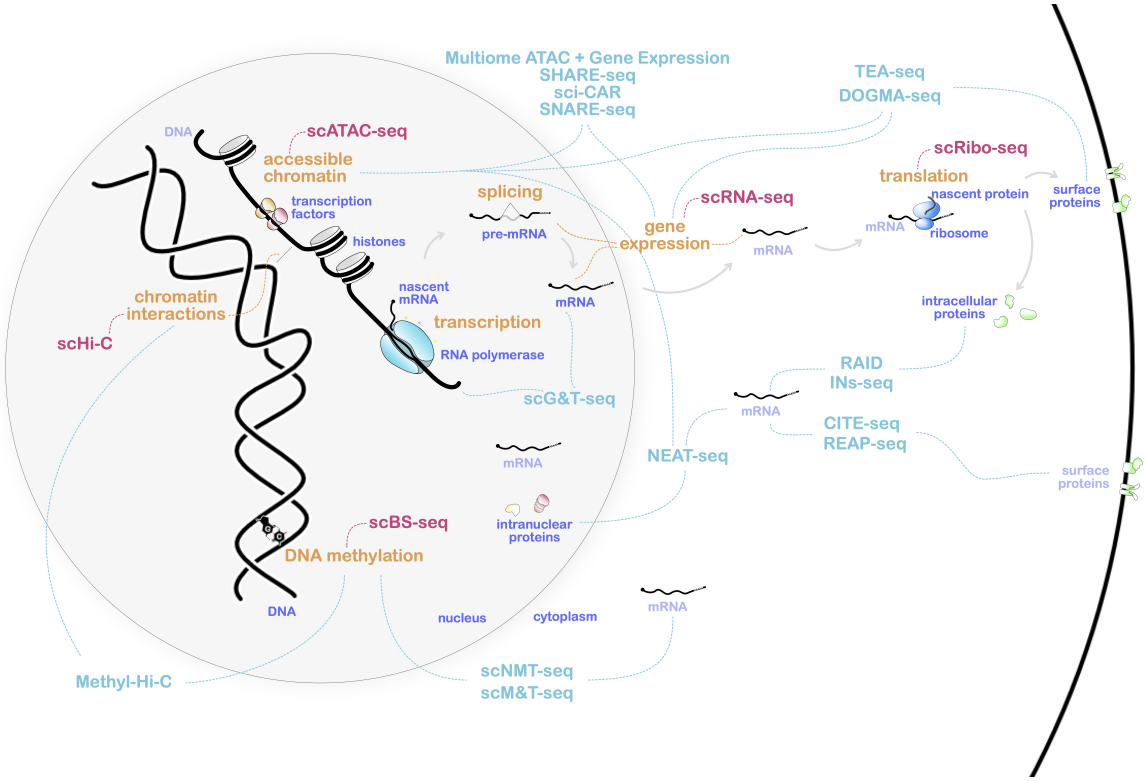

Fig. 1.3 Single-cell assays allow to capture information across multiple molecular layers Molecular structures are labelled in blue, molecular processes in orange, unimodal single-cell assays in red, multimodal single-cell assays in turquoise. Assays depicted here are not exhaustive and have been chosen for illustrative purposes.

More detailed molecular understanding of chromatin state changes and genomic regions crucial to corticogenesis has been derived using with chromatin accessibility measurements combined with RNA profiling of developing human neocortex (16 to 24 post-conception weeks) (Trevino et al., 2021). Joint ATAC and RNA profiles from mouse skin have shed light on how changes in chromatin accessibility drive lineage commitment (S. Ma et al., 2020). High-throughput nature of such screens (that study profiled almost 35 thousand cells from more than 20 cell types) allows to define regulatory chromatin regions that are important for *cis*-regulatory interactions.

Ability to simultaneously record complementary types of information for each cells is bearing the potential to provide unprecedented insights into the function of the immune system across different ages and physiological conditions (Giladi and Amit, 2018). This has been notable with the relevant progress in generating reference maps of human blood and bone marrow (Triana et al., 2021) (effectively a proteo-genomic hematopoiesis atlas) and with applications to immuno-oncology (Leader et al., 2021; A. Ma et al., 2022).

1.2 Single-cell omics data analysis

1.2.1 From sequencing reads to count matrices

Computational processing of the data is an indispensable part of contemporary biology. For scRNA-seq and sequencing-based assays in general, it starts with raw reads in FASTQ format. These reads are then mapped to the reference genome or transcriptome, and the reads mapped to specified genomic locations (e.g. genes) are counted to generate a cell-by-feature matrix (Stark et al., 2019). Protocols incorporating UMIs allow to account for duplicates that arose during amplification by counting molecules rather than individual reads. In addition to methods that are designed to align non-contiguous sequences to the reference genome such as STAR (Dobin et al., 2013) and STARsolo (Kaminow et al., 2021) or transcriptome such as BWA (Heng Li and Durbin, 2009), there are computationally efficient (Brüning et al., 2022; Vieth et al., 2019) tools that perform alignment and quantification in one step by directly relating reads to the transcriptome: kallisto bustools (Melsted et al., 2021), alevin-fry (He et al., 2022). The latter enable workflows with substantially reduced computational time (Everaert et al., 2017; He et al., 2022; Melsted et al., 2021).

1.2.2 Processing and modelling considerations

1.2.2.1 Modelling RNA counts

Transcript or gene abundance estimates in the form of count matrices are the canonical input for the vast majority of downstream analysis tools. Prior to modelling, these count matrices are typically normalised to account for factors such as sequencing depth per cell, with a variance-stabilising transformation applied afterwards (Luecken and Theis, 2019; Vallejos et al., 2017) although there are also workflows that operate on original counts (Lopez et al., 2018). The best way to conduct normalisation steps so that it's also computationally feasible is a point of discussions (Ahlmann-Eltze and Huber, 2021; Hafemeister and Satija, 2019; L. Lun et al., 2016; Lause et al., 2021; Townes, Hicks, et al., 2019). In addition to that, some tools like Salmon also allow to estimate abundance uncertainty (Patro et al., 2017). For full-length protocols, it might be relevant to normalise for gene length and GC content.

With the high sparsity of scRNA-seq count matrices, particularly pronounced for droplet-based methods, models that account for this sparsity with zero inflation have been proposed (Risso et al., 2018). However it has recently been argued that counts in droplet-based

scRNA-seq data are consistent with Gamma-Poisson (negative binomial) distribution while residual zero inflation can be explained by biological differences in cell types and cell states (Svensson, 2020). Also, the extent of zero inflation is severely affected by the choice of protocols, with the ones having no UMIs retaining substantial zero inflation (Jiang et al., 2022).

1.2.2.2 Note on the quantification resolution

Expression of genes rather than transcripts has been the predominant focus of single-cell transcriptomics. This is in part due to not having access to full-length transcripts to then reconstruct isoforms (Hagemann-Jensen, Ziegenhain, and Sandberg, 2022) in many of the popular platforms. In this scenario, a read often maps to multiple isoforms leading to high isoform quantification uncertainty. However, when there is sufficient evidence for one or another isoform, such transcripts can be quantified separately as parts of different transcript groups with low uncertainty (Sarkar et al., 2020). Such data-driven approach to flexibly define the resolution of quantification might provide a notable improvement over rigid gene-level analysis while allowing transcript-level information where possible. Moreover, transcript-level data has also been shown to increase sensitivity and accuracy of gene-level differential analysis (Yi et al., 2018).

1.2.2.3 Note on measuring accessibility and counting proteins

Count values in different modalities arise from different latent generative processes and feature different noise sources and technical biases. This motivates the application of tailored solutions for generating and modelling feature counts in each modality.

For scATAC-seq data, it seems appropriate to model fragment counts (Martens et al., 2022) with Poisson distribution, which also follows best practices for bulk ATAC-seq (F. Yan et al., 2020) and ChIP-seq data (Y. Zhang et al., 2008).

Protein measurements readout using DNA-barcoded antibodies in approaches like CITE-seq (Stoeckius et al., 2017) have specific properties such as added background noise and relatively small fraction of unique features measured (dozens and hundreds of proteins). To tackle those challenges, there have been approaches proposed such as denoising and scaling by background (dsb) (Mulè et al., 2022). This addresses ambient antibody capture as a major source of noise in protein abundance measurements, and empty droplets — based on their RNA content — provide an estimate of the ambient background.

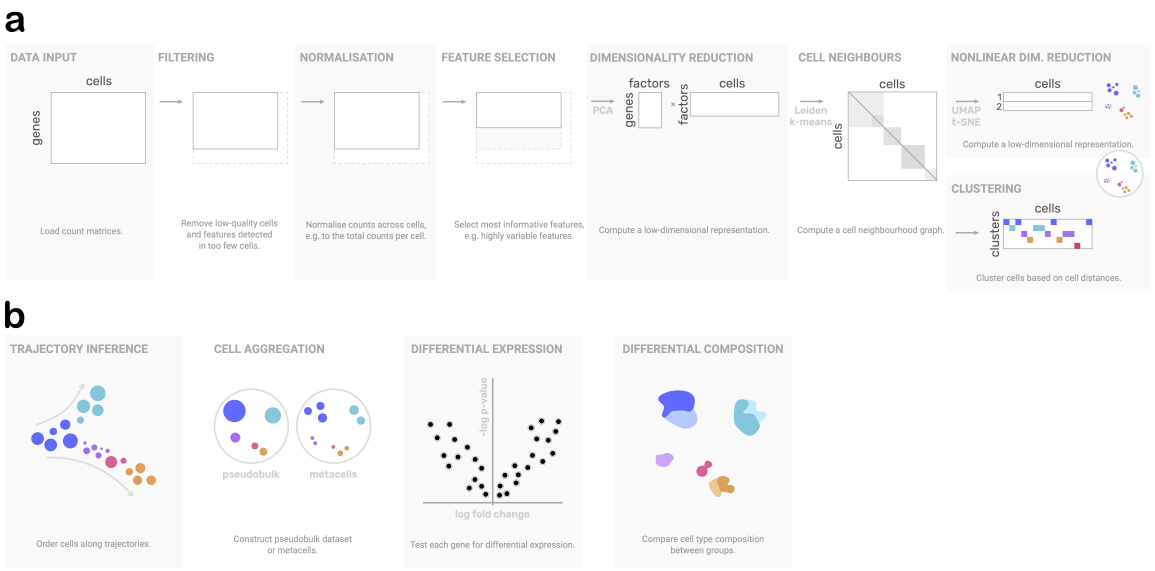

Fig. 1.4 Data processing workflow
a: Consecutive steps of a canonical scRNA-seq processing workflow. Data matrix or its derivations at each step is shown as white rectangles of specified dimensions. **b:** Analyses of scRNA-seq data tailored to specific tasks such as trajectory inference, cell aggregation, differential expression or differential cell type composition analysis.

1.2.3 Downstream processing of count matrices

The best practices for single-cell transcriptomics (Luecken and Theis, 2019) include quality control at cell (barcode) and feature (gene) level, with possible control of signal from ambient RNA and of doublet rate, normalisation, removing unwanted variation, feature selection, dimensionality reduction and guidance for downstream analyses such as clustering and differential expression.

1.2.3.1 Quality control

A distribution over the number of counts across cells is a typical way of representing sequencing depth of a dataset, with outliers being filtered out by a data-dependent threshold (Luecken and Theis, 2019). Apart from low count per cell, cells can also feature high fraction of counts from mitochondrial genes, which can indicate that cytoplasmic RNA has leaked through a broken membrane and put such cells on a list of barcodes to remove. Cell outliers with high counts can signify doublets.

Accounting for doublets and multiplets is particularly important in droplet-based high-throughput experiments as the multiplet rate depends on the amount of input cells (Bloom,

2018). Multiplet detection and removal has been addressed in specific methods (Xi and J. J. Li, 2021). One of the ways that these methods tackle the doublet identification is computationally constructing artificial doublets from the data and then trying to identify similar droplets in the dataset (McGinnis et al., 2019).

1.2.3.2 Count normalisation

An mRNA molecule is counted after it has been successfully captured, reverse transcribed and sequenced, and due to sampling effects, the resulting sequencing depth varies from cell to cell (Luecken and Theis, 2019). It is frequently corrected for during the normalisation step. The recent comparison of normalisation strategies argues that a simple and commonly strategy to log-transform library size-normalised values with a pseudocount combined with principal component analysis perform well in many situations:

$$y_i = log(\frac{x_i}{\sum_j x_j} + 1),$$

where x_i is a raw count of gene i in a cell and $\sum_j x_j$ is total sum of counts for this cell.

1.2.3.3 Covariates

Covariates can either be known or unknown (hidden). If they are unknown, they need to be estimated from the data first. Technical or unwanted biological covariates might be removed with a simple linear regression or with more complex models (Buettner, Natarajan, et al., 2015; Buettner, Pratanwanich, et al., 2017). Specific covariates such as the signal arising from ambient RNA can be removed with specially developed methods (Young and Behjati, 2020).

One typical covariate is a batch effect, and consequently, linear models (Johnson et al., 2007) and linear mixed models (Tung et al., 2017) to separate the batch effect have been developed as well as approaches building on the detection of mutual nearest cell neighbours (Haghverdi, Lun, et al., 2018).

1.2.3.4 Feature selection

With dozens of thousands of features (about 25000 genes in human or mouse studies), the unfiltered feature space is inevitably tainted by the «curse of dimensionality» (Bellman, 2015). In order to reduce the noise in the data (e.g. by filtering the most sparse genes out) and

to make downstream computations computationally feasible, informative genes are selected. The choice of such genes is dictated by the complexity of the dataset but a typical strategy suggests selecting highly variable genes for the downstream analysis (Luecken and Theis, 2019).

1.2.3.5 Dimensionality reduction

Dimensionality reduction is a first-line processing step in a vast number of workflows. An important notion behind applying dimensionality reduction methods to omics datasets is the «low dimensional» nature of gene expression data, i.e. there is a low-dimensional space where the biological signals can be accurately represented by a small number of basis vectors (Heimberg et al., 2016). Indeed, if most genes can be grouped into modules with correlated expression levels across cell populations, we can reduce the data representation from the *gene* level to the *module* level (Trapnell, 2015). It has been shown that transcriptional states in single-cell data can be reconstructed with a small fraction (less than a thousand) of transcripts per cell (Heimberg et al., 2016). This tolerance to noise is arguably due to the gene expression covariance structure.

Principal component analysis (PCA) is a cornerstone method for such analyses, with well-defined theory behind it, its accessibility via efficient implementations, linearity and hence interpretability (Alter et al., 2000; Ringnér, 2008). PCA is also at the basis of many other unsupervised methods such as kernel PCA, which extends PCA with kernel methods (Ham et al., 2004; Mika et al., 1998).

Principal components effectively define a new coordinate system maximising variation along them, with the first principal component being the direction of the largest variation in samples. While the full singular value decomposition (SVD), which is typically used to calculate principal components, of the original matrix is lossless, in practice, only a few components of interest, signifying the largest variation, are computed and used to represent the dataset (Luecken and Theis, 2019). Each principal component is a linear combination of original features, which makes it possible to gauge the importance of each feature (e.g. gene) for each axis of the embedding and identify the most information-rich features.

While PCA provides a convenient linear reduced space to work with the data, which is still interpretable, it can be challenging to represent main variation in the data with principal components when plotting the data in two dimensions. More generally, high-dimensional data visualisation is a challenging problems across different domains.

For data representation it is desired that some of the properties in high-dimensional space are preserved in the embedding. One of such properties can be local similarity that methods such as t-Distributed Stochastic Neighbour Embedding (t-SNE) attempt to preserve (Van der Maaten L and Hinton G, 2008). More recently introduced methods provide alternatives that argue to preserve more global structure with superior run time performance, e.g. Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) or viSNE (Amir et al., 2013). Their further developments also try to preserve local density of data points building on top of the state-of-the-art methods such as t-SNE and UMAP (Narayan et al., 2021). The idea of taking density into account echoes diffusion maps (Farrell et al., 2018; Haghverdi, Buettner, et al., 2015; Haghverdi, Maren Büttner, et al., 2016) that were developed to visualise differentiation trajectories and establish pseudotemporal ordering of single cells.

While methods like t-SNE and UMAP provide immense value for representing high-dimensional data, they can also be misleading and sensitive to hyperparameter choice (Wattenberg et al., 2016). It has also been argued that non-linear embeddings of the data to two or three dimensions should be used with care when drawing conclusions based on them (Bergen, R. A. Soldatov, et al., 2021; Chari, Banerjee, et al., 2021), which is of particular importance for the omics analyses, where biological insights are attempted to be generated from the data.

Another family of methods to visualise high-dimensional data has arisen thanks to the progress of deep neural networks and their combination with Bayesian models in models such as Variational Autoencoders (VAEs) (Kingma and Welling, 2014; Lopez et al., 2018). Parallels to PCA have also been drawn as VAEs have been reported to use differences in variance in order to generate latent data representation and to collapse to PCA in the linear case (Rolinek et al., 2019).

1.2.3.6 Definition of cell neighbourhoods and clustering

Another pivotal step is constructing a distance matrix. Commonly, Euclidean distances are calculated in the principal component (PC) space (Luecken and Theis, 2019). Alternatives like cosine similarity are also used (Haghverdi, Lun, et al., 2018). The resulting cell-cell distance matrix can be utilised for a variety of downstream analyses including UMAP (McInnes et al., 2018) and clustering.

Clustering is used for discovering the natural groupings of cells based on their features: in the case of scRNA-seq, relying on their transcriptional profiles. In most of the relevant

applications this clustering is unsupervised, i.e. without any labels providing the group truth (Kiselev et al., 2019). Typically, a cell distance matrix is used for defining cell similarity, which a clustering algorithms is applied on (Kiselev et al., 2019): the widely used ones include *k*-means, hierarchical clustering and more scalable graph-based community detection such as Louvain (Blondel et al., 2008) and Leiden (Traag et al., 2019), named after a Belgian and a Dutch city. For the latter family of clustering algorithms, densely connected components in the *k*-nearest-neighbours graph are identified (Lancichinetti and Fortunato, 2009). Irrespective of the clustering method, it is usually applied on principal components rather than on individual features, and methods that integrate PCA and hierarchical clustering more tightly have also been developer (žurauskienė and Yau, 2016).

1.2.3.7 Differential expression

Differential expression analysis in the context of single-cell transcriptomics is typically concerned with identifying genes that vary their expression level between states or conditions. When these condition represent phenotypic change, we would be then concerned with identifying the ones responsible for that change between these conditions (Trapnell, 2015). For the method performance, it has been argued that accounting for biological replicates is important, with a failure to do so leading to false discoveries (Squair et al., 2021). The best performance was achieved by aggregating cells within a biological replicate («pseudobulk») (Crowell et al., 2020; Soneson and Robinson, 2018; Squair et al., 2021) and with the methods originally developed for bulk sequencing (edgeR (Robinson et al., 2010) - DESeq2 (Love et al., 2014), both based on Gamma-Poisson mixture, or negative binomial distribution). More tailored methods have been proposed with such as MAST based on generalised linear model (Finak et al., 2015), a Bayesian approach to single-cell differential expression (Kharchenko et al., 2014), where the observed abundance of each gene in a cell was modelled as a mixture of Poisson («drop-out») and negative binomial («amplification»), or a trajectory-based differential expression analysis (Van den Berge et al., 2020).

1.2.3.8 Temporal modelling

For the data where the dynamic processes such as differentiation are of interest, the focus is on trajectory inference, or pseudo-time analysis, methods that allow to order cells along a trajectory based on their similarity (Ding et al., 2022; Saelens et al., 2019; Trapnell et al., 2014). Such methods have also been extended to be incorporate auxiliary information for

cell fate mapping such as RNA velocity (Ding et al., 2022; Lange et al., 2022) or metabolic labelling (X. Qiu et al., 2022).

1.2.3.9 Cell composition

Single-cell data have naturally encouraged the development of methods for compositional analysis. An applied motivation for such methods is modelling cell type proportions and comparing them between conditions such as individuals with and without the disease. Methods like that account for uncertainty in cell type proportions estimations and enable differential abundance testing and are frequently formulated as generalised linear models (M. Büttner et al., 2021; Y. Cao et al., 2019) or graph-based models (Dann et al., 2022).

1.2.3.10 Cell aggregation

In single-cell datasets, multiple observations are obtained from the same biological entity, which leads to a fundamental challenge of designing appropriate models to account for biological replicates (Squair et al., 2021). Aggregation of the data in disjoint groups frequently referred to as *pseudobulk* allows to address at least some of these challenges (Lun and Marioni, 2017; Squair et al., 2021). A more fine-grained resolution of the aggregated data is offered by metacells — groups of scRNA-seq profiles that represent distinct cell states in the data such that original cells in a group represent repeated sparse sampling, and various algorithms to construct metacells have been proposed (Baran et al., 2019; Ben-Kiki et al., 2022; Bilous et al., 2022; Persad et al., 2022).

1.2.4 Integration of multimodal single-cell data

The ever increasing variety of experimental multimodal omics techniques motivates the exploration and development of corresponding analytical approaches. First of all, multimodal omics data analysis is based on understanding and choosing appropriate models for individual modalities, and the lessons learned for single-cell transcriptomics remain relevant. Then, multimodal-specific challenges come into play.

First and foremost, there are statistical challenges for multimodal data integration and interpretation (Argelaguet, Cuomo, et al., 2021). With the increase in the number of modalities, the number of features increases hence the need for appropriate regularisation strategies as well as for the methods that can scale to the ever growing feature space. Underlying

different molecular layers, there are different generative processes, and combining different likelihoods in one model is not a trivial task. Sparsity structure, the nature of missing values and the sources of noise can also be different for various data-generating methods.

Many integration methods for multimodal omis are following the original idea of matrix factorisation and extend it to multiple modalities, or views (Argelaguet, Cuomo, et al., 2021; Xu et al., 2013). Methods based on joint factorisation include, but not limited to, multiple co-inertia analysis (Meng et al., 2014), joint non-negative matrix factorisation (Z. Yang and Michailidis, 2016; S. Zhang et al., 2012), joint and individual variation explained (Lock et al., 2013), multi-omics factor analysis (MOFA) (Argelaguet, Velten, et al., 2018). Other family of methods extend nearest neighbour search to work across multiple modalities (Hao et al., 2021) or integrate neighbour graphs of individual modalities (Bo Wang et al., 2014). Other approaches like LIGER (C. Gao et al., 2021; Welch et al., 2019) and canonical correlation analysis (CCA) (Stuart, Butler, et al., 2019) first require modalities to be defined in the same feature space, e.g. epigenomics signal is to be aggregated for genes to be integrated with gene expression. Multi-view clustering methods has been proposed in the form of a Dirichlet process mixture model (Bachireddy et al., 2021; Burdziak et al., 2019) or multiplex community detection (Mucha et al., 2010; Traag et al., 2019).

Corresponding models based on variational autoencoders have also been proposed for multimodal integration such as the ones for CITE-seq data (Gayoso, Steier, et al., 2021) as well as for chromatin and transcription (Ashuach et al., 2021).

1.2.5 Analysis ecosystems and workflows

As illustrated by a rich variety of computational approaches to model, process and interpret single-cell data, there's a notably complexity in analysis workflows for such data. This motivates abstractions that would allow to handle original data and derived information such as embeddings and annotations with minimal overhead. At the same time, with a wide range of experimental techniques, methods and their applications, the choice of programming language typically dictates the choice of analytical tools and, currently, the choice of the data storage model (Luecken and Theis, 2019).

1.2.5.1 Data storage within R ecosystem

A widely popular ecosystem of packages for the R programming language with a focus on biology is Bioconductor. Its motivating principles include streamlined user experience

Fig. 1.5 Data standards for single-cell genomics
a: Data storage standards adopted for scRNA-seq datasets. SingleCellExperiment and Seurat are R packages, AnnData is a Python package. For illustrative purposes, only abstractions for matrices derived from a single modality are shown. N and D denote dimensions in terms of the number of cells and genes, respectively. **b:** Data storage standards adopted for multimodal storage. MultiAssayExperiment extends SingleCellExperiment in a modular fashion while Seurat, in its most recent versions, provides a single object for unimodal or multimodal data storage. N, N1, N2 refer to the number of cells; D, D1, D2 refer to the number of features.

and support of an open scientific community (Amezquita et al., 2020). such abstraction is an object structure (SingleCellExperiment) that contains primary and transformed counts matrices (assays) as well as provides slots to store feature and cell annotations and cell embeddings (Huber et al., 2015). SingleCellExperiment object sticks to the convention of storing features in rows and allows to incorporate matrices and metadata such as experimental covariates or derived cell labels. Following the modularity principle of Bioconductor, analysis and methods packages such as scran (L. Lun et al., 2016), scater (McCarthy et al., 2017) and DropletUtils (Lun, Riesenfeld, et al., 2019) can operate on the Bioconductor data standards.

Another commonly used toolkit for single-cell analysis is Seurat (Satija et al., 2015). Together with defining the in-memory Seurat object structure to store data, it provides a set of tools for loading, processing, integrating and visualising single-cell transcriptomics data. Lately this format and framework have been extended to chromatin assays (Stuart, Srivastava, et al., 2021).

1.2.5.2 Data storage within Python ecosystem

Scientific Python ecosystem has been constructed with NumPy — a library for multidimensional array programming — as its foundation (Harris et al., 2020). Briefly, it provides a multidimensional array data structure as well as functions to perform vectorised calculations on arrays. With its central position in the ecosystem, NumPy effectively specifies a well defined API that other specialised array implementations can rely on. Lately, JAX library for high-performance numerical computing and machine learning has been introduced featuring, among other things, an API that largely follows NumPy (Bradbury et al., 2018).

Built on top of NumPy, there are software packages to deal with in-memory tabular data (McKinney, 2011). On top of that, another layer of abstraction has been build in order to present annotated datasets (AnnData) (Virshup et al., 2021) or N-dimensional labeled arrays (xarray) (Hoyer and Hamman, 2017).

1.2.5.3 Data processing workflows

On top of the data standards such as MultiAssayExperiment, Seurat object and AnnData, analysis workflows are built (Zappia et al., 2018). These workflows collate individual tools and provide a convenient interface making these tools interoperable through a data standard. Part of the Bioconductor ecosystem, scran (L. Lun et al., 2016) and scater (McCarthy et al., 2017) operate on SingleCellExperiment objects. Seurat framework is not separated from the

data standard it ships with and defines operations on a stateful Seurat object that cover many types of analysis for single-cell transcriptomics (Butler et al., 2018; Hao et al., 2021; Satija et al., 2015; Stuart, Butler, et al., 2019) as well as, lately, chromatin accessibility (Stuart, Srivastava, et al., 2021). In Python, Scanpy (single-cell analysis in Python) (Wolf et al., 2018) provides improved scalability, better separation of individual processing steps, which increases modularity, and easier access to interfaces with machine learning toolboxes such as PyTorch (Paszke et al., 2019).

1.2.5.4 Multimodal omics data storage

With multimodal omics becoming a more widely used data type at single-cell scale, new computational and analytical solutions are required to handle such data (Argelaguet, Cuomo, et al., 2021). Such experiments pose considerable challenges for data management, processing, integration, visualisation and exchange (Conesa and Beck, 2019; Wilkinson et al., 2016).

Existing open-source solutions for comprehensive multimodal omics data storage have been implemented exclusively for the R language ecosystem in libraries such as MultiAssayExperiment (Ramos et al., 2017) and Seurat (Hao et al., 2021). While providing the necessary functionality and supporting arbitrary numbers of modalities, these libraries have their own limitations showing that their design might not be generic enough for modern single-cell multimodal omics applications. As such, Seurat requires a common set of cells to be profiled for all the modalities and lacks separation of multimodal annotations from the ones based on individual modalities. MultiAssayExperiment lacks multimodal annotations whatsoever. In practice, the analysis results obtained by these libraries are saved in a binary file that is interfaced only with the R language.

MultiAssayExperiment provides a generic solution to manage multi-omics data in the R programming language as part of the Bioconductor ecosystem (Ramos et al., 2017). In particular, it introduced a data structure to represent and manipulate multi-omics datasets, with three main components of it being (1) sample-level table, (2) list of experiments (assays), (3) sample map to relate samples to individual observations in assays. This makes it possible, for instance, to store multiple observations per sample in an assay, which has been relevant for replicates in patient-level measurements.

1.2.5.5 Multimodal data processing workflows

The development and abundance of multimodal techniques have posed grand challenges to the analytical tools and computational approaches to handle and interpret such data.

While individual multimodal integration techniques were discussed above, the only popular framework that natively handles multimodal omics data is Seurat (Hao et al., 2021). In addition to its data format that allows to store multiple arbitrary-sized feature sets profiled across a single set of cells, it implements an integration method to calculate weighted nearest neighbours (WNN) across modalities.

It is worth pointing out that individual modalities of multimodal datasets can be used to guide decisions and analyses. For instance cell type composition can be explored in an unbiased manner with scRNA-seq and cell types can be selected to perform further experiments and profile alternative modalities. Another example is clustering cells based on their gene expression profiles to then guide peak calling in the scATAC-seq modality in individual cell populations (Granja et al., 2021). For such strategies to be seamlessly integrated into user workflows, multimodal frameworks have to cooperate with analysis toolboxes for individual modalities.

1.3 Discussion and new perspectives

Single-cell multimodal omics is a powerful tool for studying the molecular properties of biological systems. Complex multimodal datasets enable the analysis of multiple molecular layers within individual cells as well as the characterisation of their relationships. These approaches typically do not capture, however, cell interactions or their spatial context, which can have significant impacts on cell state and function. Another challenge is to study biological processed at the single-cell level in time. This can be particularly important for understanding how different molecular layers interact and influence one another as correlated or consequential patterns may occur across different layers with different temporal resolution or delay. Addressing these challenges is necessary in order to fully leverage the potential of single-cell multimodal omics. Some of the recent developments in these directions are discussed below.

1.3.1 Transcriptomics and multimodal omics across time and space

1.3.1.1 Transcriptional dynamics

Beyond a static picture of RNA content in a cell, its dynamics can be profiled as well (Stark et al., 2019). For instance, in scNT-seq (Q. Qiu et al., 2020) and scSLAM-seq (Erhard et al., 2019), nascent transcription is measured using metabolic labelling of newly transcribed RNA molecules with 4-thiouridine (4sU) and consequent conversion of 4sU to cytidine analogues. Availability of such data has also required and triggered the development of new analytical methods (X. Qiu et al., 2022). More than transcription, translation of RNA molecules in individual cells has also been profiled with single-cell Ribo-seq, which achieves close to single-codon resolution (VanInsberghe et al., 2021). Furthermore, longitudinal profiling of living cells has been recently made possible with non-disruptive spatiotemporal imaging with immunofluorescence (J. Ko et al., 2022).

1.3.1.2 Spatial omics

In multicellular organisms, the position of their building blocks — cells — in space in relation to other cells can determine its state and function (Bodenmiller, 2016). The importance of studying cells in their spatial environment has been highlighted in various contexts, for instance for linking connectivity networks of the nervous system to the transcriptomics (Lein et al., 2017), spatial bone marrow niche organisation (Baccin et al., 2020) or due

the recognised role of tumour microenvironment for cancer development and progression (Hanahan and Coussens, 2012). And while many studies are performed on isolated cells with their tissue context being destroyed, there has been significant progress in developing and performing high-throughput studies while preserving spatial information (Burgess, 2019).

Spatial location for individual genes (thousands) can be measured simultaneously with their expression using imaging techniques such as *in situ* hybridisation (Eng et al., 2019; Moffitt et al., 2016; Shah et al., 2016). Transcriptome-wide spatial quantification has become possible with technologies that transfer mRNAs to an array with oligonucleotides for spatial barcoding (Efremova and Teichmann, 2020; Ståhl et al., 2016; Sanja Vickovic et al., 2019). The latter currently comes at the cost of resolution, albeit special computational approaches have been developed to resolve cell types in spatial transcriptomics data (Kleshchevnikov et al., 2022).

Imaging techniques have also been extended with other modalities such as with microscopy techniques as has recently been done in the STcEM, which combined scanning electron microscopy with fluorescent in situ hybridisation (Androvic et al., 2022), or with neuronal activity to track functional maturation of neurons (Wan et al., 2019).

Following spatial transcriptomics, spatially-resolved epigenomics assays are being developed such as spatially resolved ATAC-seq (Y. Deng et al., 2021; Thornton et al., 2021). Such assays have been applied to provide a new, spatial dimension to the sequencing of human and mouse cortex tissues (Thornton et al., 2021) and of mouse embryos (Y. Deng et al., 2021).

With advances in spatial information acquisition, it will be increasingly common to multimodal omics sequencing techniques to be spatial as well (L. Tang, 2021). Now, platforms like DBiT-seq (Y. Liu et al., 2020) and SM-Omics (S. Vickovic et al., 2022) provide high-throughput spatially resolved transcriptomics combined with antibody-based protein measurements with 10–100 μm spot size.

1.3.2 RNA velocity

As common scRNA-seq protocols generate reads from both spliced and unspliced mRNA, it has also been generally possible to estimate a time derivative of gene expression for each gene, which has been termed «RNA velocity» (La Manno, R. Soldatov, et al., 2018). The balance of unspliced and spliced mRNA abundances indicates the future state of mature mRNA abundance making it possible to effectively make a prediction about the future state of the cell. RNA velocity improvements have been proposed with a likelihood-based dynamical

model (Bergen, Lange, et al., 2020) and unified latent time across the transcriptome (M. Gao et al., 2022). RNA velocity can be further combined with cell similarity to model cell-cell transition probabilities (Lange et al., 2022).

Moreover, the kinetic model of RNA velocity has been expanded to proteins (Gorin, Svensson, et al., 2020) and multi-omics, the latter incorporating chromatin accessibility in the velocity estimates (C. Li et al., 2021). The current approaches to RNA velocity, however, require rigour when applying the method and can be sensitive to parameter choices and have behaviour dependent on the underlying biology of the system (Bergen, R. A. Soldatov, et al., 2021; Gorin, Fang, et al., 2022).

1.3.3 Modular workflows and emerging languages

The complexity of multimodal data calls for appropriate data structures and integration methods. In order to accommodate the diversity of molecular assays and their combinations, data standards should be generic and extensible. Modularity and composability of analytical workflows, designed around these data standards, is their important feature, as demonstrated by the workflows in Python and R described above. This can allow for more diversification of programming languages, frameworks and tools while keeping workflows practical and accessible.

While standards haven't been established for the newer programming languages, it is worth noting a few of them seem to attract a growing interest from the community. That includes Julia, a language that is focused on performant numerical computations (Perkel, 2019), and Rust, a more low-level language, which increasingly frequently underlies high-performant tools (Perkel, 2020).

Chapter 2

Multimodal data abstractions and operations

With multimodal datasets becoming more widely used and larger in size, there has been a pronounced need for tailored computational solutions to store, manage and exchange such data. Some existing solutions have been described in the section 1.2.5, which also highlighted a glaring need for standardising multimodal data handling in Python. To address this, I designed and implemented multimodal data (MuData) objects alongside a suite of analysis methods that build on it. In this chapter, I describe the design of this multimodal data abstraction that is naturally integrated into the existing single-cell genomics workflows. To handle MuData objects and perform data integration, I created MUON — a multimodal omics analysis Python framework. Last but not least, I provide an outlook on a scalable multi-dataset multimodal data storage.

The contents of this chapter concerning MuData and MUON have been published in Bredikhin et al., 2022. The concept and the design of MuData and MUON as well as their primary reference implementations are my own contribution.

2.1 Multimodal Data (MuData)

Here, I introduce MuData as a standard to operate multimodal datasets that features a seamless interface with the existing *de facto* standards for single-cell transcriptomics. MuData is a data abstraction around a collection of two-dimensional datasets: it offers an interface to access data properties and to apply transformations while not exposing implementation details to its user. Notably, it is a generic data structure around observations (e.g. samples, individuals or cells) and features (e.g. genes, proteins or genomic locations). For the purposes of this chapter, its description below will be tailored to single-cell multimodal datasets.

2.1.1 Data standard for multimodal omics

2.1.1.1 Design considerations

Experiments generating data with multiple modalities observed for the same cells pose novel analysis and data management challenges. In order to address them in a principled manner, data abstractions are required. The result of a multimodal experiment is a collection of datasets for the respective modalities but also cross-modality information that is relevant beyond individual modalities. Moreover, analysis and derived annotations can be either modality-specific or connected to several modalities or to the whole dataset.

Other considerations include efficient unimodal access and providing interfaces similar to unimodal omics. Unimodal access implies that there should be a transparent way of working with just one modality of the few ones in the multimodal experiment. It also means that existing software solutions should be able to access individual modalities within a multimodal dataset using established interfaces. This way the tools developed for certain data types such as scRNA-seq are ensured to be readily available when such data types are combined with other data types in a multimodal experiment. Moreover, this also means individual modalities are stored on disk in a way they would have been stored for unimodal experiments.

The outputs of the processing steps for individual modalities are confined within them thus providing separation of concerns and reflecting the input structure. With integrative analysis of multimodal data, multiple input modalities give rise to outputs that are to be stored at the multimodal level. As analysis outputs for multimodal integration do frequently resemble the analysis results for individual modalities (e.g. embeddings or neighbourhood graphs), it is also desirable they are to be stored in the same way providing similar interfaces to access similar outputs.

A multimodal omics dataset will be defined below as one with cells profiled across several sets of features (e.g. transcript abundance, gene expression, protein abundance, genomic locus accessibility, methylation state of individual nucleotides). These features come from distinctive groups so that each feature d comes from a single modality m_d, i.e.

$$\forall d : \exists m, m \in \{1, ..., M\} : 1_m(d) = \begin{cases} 1, \text{ m} = \text{m}_d, \\ 0, \text{ m} \neq m_d, \end{cases}$$

where $1_m(d)$ is an indicator of a feature d belonging to the modality m from all the modalities in the dataset $\{1, ..., M\}$.

A typical property of a multimodal experiment is to have most or all cells sampled across all modalities. More formally, if N is a total number of cells in the experiment, each cell i profiled at least for one modality of M, we would expect

$$\frac{\sum_{i,m} 1^i_m}{M} \approx N,$$

where 1^i_m is an indicator of a cell i having been profiled for modality m. Another way to formulate this would be to consider a sorted set of cells in a dataset so that first k cells are profiled for all the modalities and cells from $k+1$ to N are profiled for $1, ..., M' < M$ modalities. We then expect $N - k \ll N$.

With the generic design of MuData, this is, however, not enforced or required, which becomes important in practice, especially when managing derived data where filtering has been performed.

2.1.1.2 Hierarchical design of MuData

MuData provides a level of abstraction on top of the ones offered by the unimodal data standards. This hierarchical model allows to generalise existing data formats for single omics, mainly transcriptomics, which revolve around a two-dimensional matrix with counts, while preserving compatibility with existing toolchains and standards. In particular, the structured format of MuData leverages and builds upon the AnnData standard for annotated data (Virshup et al., 2021) adopted by the single-cell community.

Nested structure enables storage of both modality-specific and dataset-specific information in a uniform manner. In detail, cell- and feature-specific annotations, including multidimensional and pairwise annotations, are available at the level of individual modalities

and the whole dataset (Figure 2.1). Thus MuData can store both original and derived data and analysis outputs, for instance cell type labels or embeddings of cells (Figure 2.1b).

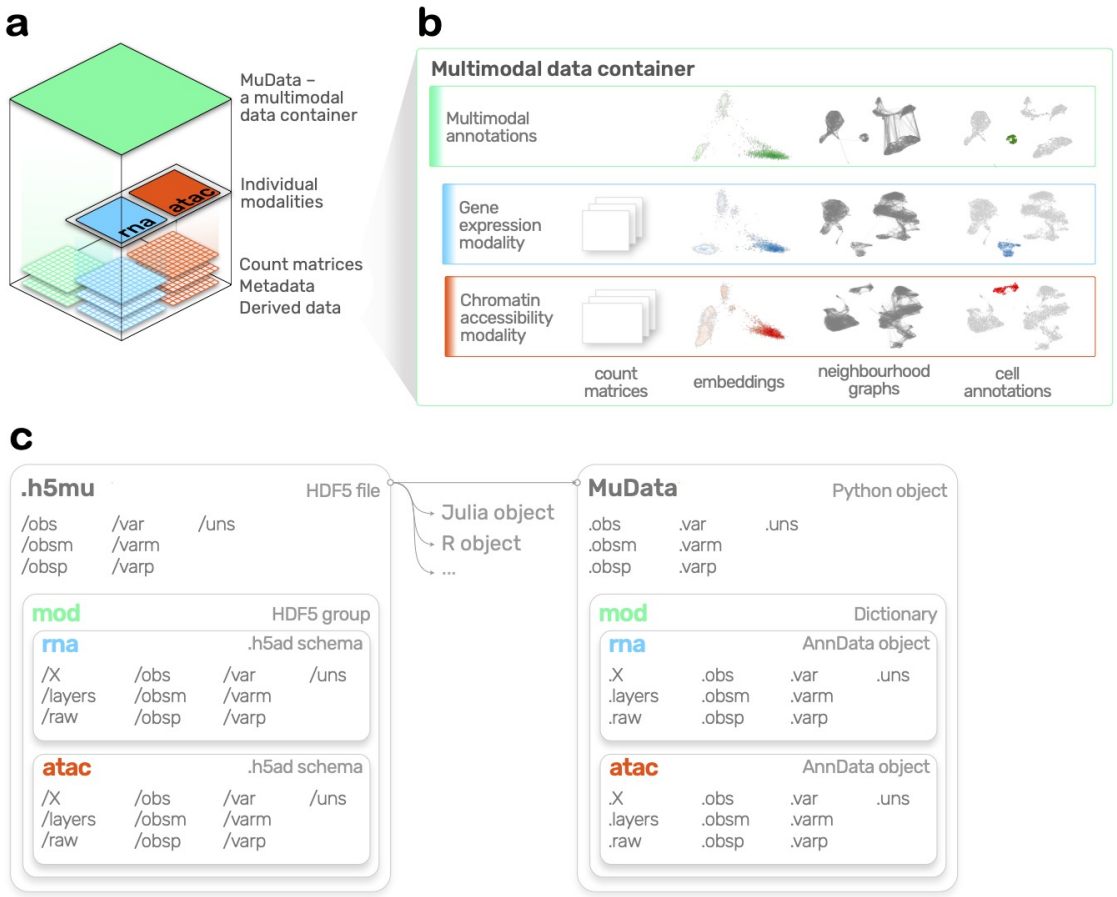

Fig. 2.1 MuData object structure
a: Schematic representation of a MuData object with RNA (blue) and ATAC (red) modalities as an example. Matrices with original and transformed counts and associated and derived metadata are represented as arrays and tables. Green denotes multimodal information. **b:** Example contents in a MuData object. Count matrices, embeddings, neighbourhood graphs and cell annotations for individual modalities (blue and red) as well as for the multimodal level are shown. **c:** Schematic representation of MuData object on disk in the HDF5 format (left) and in memory in Python (right). Plates denote different levels of hierarchy.

MuData offers multiple serialisation options. In-memory MuData objects are typically serialised to HDF5 (The HDF Group, 2000–2010) files on disk, hereinafter referred to as H5MU files. HDF5 (Hierarchical Data Format 5) was designed for organised storage of large volumes of data and has been widely used across many applications and domains

(Blanas et al., 2014; De Carlo et al., 2014; Dougherty et al., 2009; M. M. Hoffman et al., 2010; S. Lee et al., 2022). This is also the format that AnnData takes advantage of by default and is thus a natural default serialisation option for MuData due to their hierarchical relationship. The schema of the on-disk MuData storage is reminiscent of the AnnData's one (Figure 1.5) and can serve as a generalisation to multimodal data (Figure 2.1c). By serialising individual modalities using the AnnData serialisation mechanism, such MuData design thus permits direct access to individual modalities and makes them usable with existing (unimodal) analysis toolchains.

MuData functionality includes, but is not limited to: object instantiation from a map of modality names and AnnData objects; access to individual modalities as AnnData objects; subsetting cells and/or features. Subsetting cells works across modalities, i.e. cell selection is applied across all the modalities as well as to the multimodal annotations (Figure 2.2). Feature relations across modalities can be stored in the MuData object as a sparse multimodal graph (Figure 2.2b).

The interface of a MuData object offers access to its properties through a set of predefined attributes discussed below together with their respective functionality.

2.1.1.3 Annotation of cells and features

The data container makes sure that annotations of cells and features are unambiguous by validating the dimensions of the respective annotation tables. Following terminology of AnnData, annotations for cells and features are accessible via `.obs` (observations) and `.var` (variables) attributes, respectively (Figure 2.2a).

Multimodal annotations for cells and features are accessible via `.obsm` and `.varm` attributes, respectively. MOFA factors or a UMAP space are examples of multimodal embeddings (Figure 1.4), and MOFA factors loadings are essentially a multimodal feature annotation.

2.1.1.4 Cell and feature relations

Cell neighbours and analogous structures of cell relations can be represented as graphs. In practice, these graphs are typically sparse and are stored in a sparse format as for each cell, only distances or relations to its k neighbours are recorded. Multimodal cell neighbours computed, for instance, based multimodal latent factors (Argelaguet, Arnol, et al., 2020) or

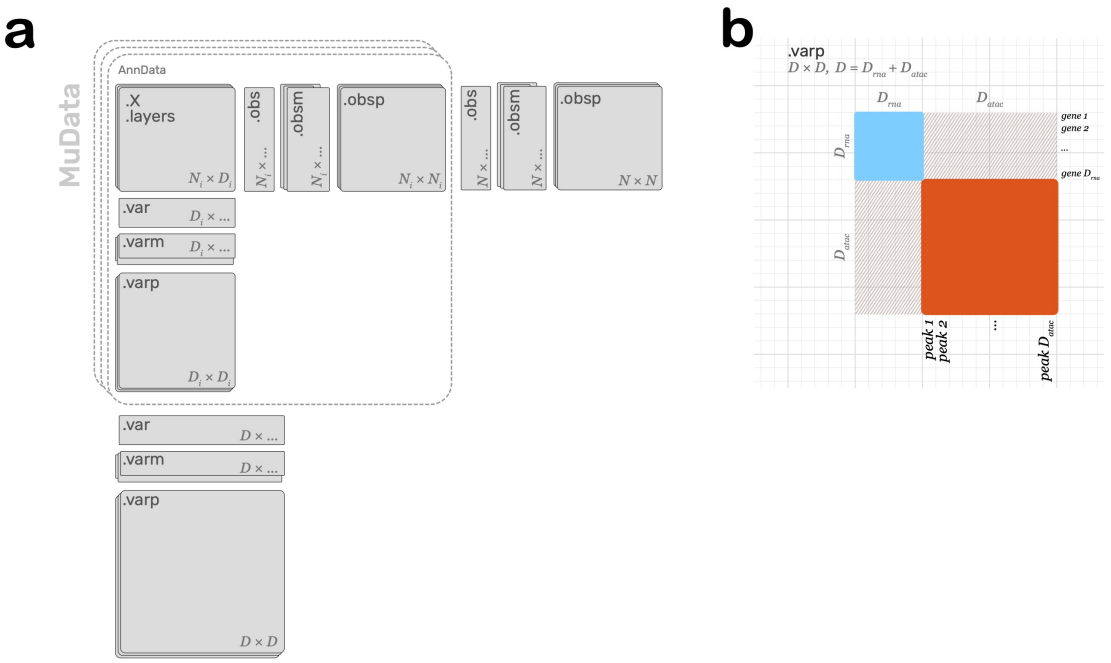

Fig. 2.2 MuData schema visualisation
a: Representation of the schema for MuData. Matrices and tables are shown as rectangles with their dimensions and access attributes denoted. Corresponding dimensions are shown to be aligned. Individual modalities are AnnData objects. In addition to that, there's data at the multimodal level, both for cells and features. **b:** Illustration of intra- (solid colours) and inter-modality (stripes) feature relations using genes (RNA, in blue) and peaks (ATAC, in red) as an example. Global feature dimensions are the sum of the feature dimensions of individual modalities.

with multimodal graph methods (Hao et al., 2021; Bo Wang et al., 2014) are accessible via `.obsp` attributes following the way unimodal relations are stored in AnnData (Figure 2.2a).

Feature relations in `.varp` can be viewed in a similar way. With the multimodal `.varp` attribute, such relations are stored for features across modalities. This is particularly valuable for cross-modality relations such as feature correlation structure, and MuData format allows to store information such as peak-gene binary correspondence, peak-gene distance, correlation between accessibility and expression, etc. (Figure 2.2b).

2.1.1.5 Auxiliary information

Auxiliary unstructured data can be stored in the `.uns` slot following AnnData's design (Virshup et al., 2021). This feature is practically useful for storing which parameters were used to compute derived annotations, software or genome assembly and annotation versions as well as additional statistics of performed computations such as variance explained by each principal component or, for the multimodal `.uns` attribute, each MOFA factor.

2.1.1.6 Synchronising MuData container

By design, MuData stores pointers to AnnData objects that contain data from individual modalities. It also contains information derived from individual modalities such as cell and feature names in order to store higher-level (multimodal) information. This means the container needs to be synchronised with the individual modalities.

To keep the information in the container up-to-date, the identities of cells or features have to be collated across modalities, which equates to the *union* operation on sets. Analogously, the *intersection* of cell sets across modalities can be performed in order to derive a dataset with all the cells profiled across all modalities. In order to handle the use case of multimodal datasets, the same identifier (e.g. molecular barcode) in different modalities is considered to point to a single cell in the dataset. Features are, however, considered to always be disjoint. Having the global cells and feature sets collated identifies correspondence between annotations for the experiment and for individual modalities (Figure 2.4a).

A natural way to collate items when dealing with tabular data are *join* operations. In some typical scenarios, however, this operation can be performed faster than a *join* operation. Below, a few scenarios are outlined that I have identified in the common analysis workflows and optimised the procedure to collate data for. First one concerns count matrices that are generated for multimodal omics experiments in a way that a set of cells with unique barcodes

is used across all the modalities. That means when the equality of the sets of cell barcodes and their order for all modalities is verified, there's a one-to-one correspondence between global cell identities and cell identities of individual modalities. Second, when user-defined feature names can be used as respective identities, i.e. names in each modality are unique as well as names across modalities, the items can be collated faster as no new identity has to be generated that would ensure a stable *join* operation.

2.1.2 MuData implementation and serialisation

2.1.2.1 MuData implementation

I wrote the reference implementation of MuData[1] in the Python programming language (Van Rossum and Drake Jr, 1995). Following AnnData (Virshup et al., 2021), MuData relies on the numerical Python stack with NumPy for arrays (Harris et al., 2020), Pandas for tabular data (data frames) (Wes McKinney, 2010). Importantly, MuData does not bring more dependencies to the dependency tree of AnnData, making it a lean library to build infrastructure or analysis tools on.

Cross-language and cross-platform support of the MuData standard is achieved through serialisation and corresponding libraries. To demonstrate the practical efficiency of this approach, in addition to the Python software, I implemented the MuDataSeurat library[2] for the R programming language (R Core Team, 2013) for read and write operations for Seurat objects (Satija et al., 2015) in memory and H5MU or H5AD files on disk. Together with Ilia Kats, we have also implemented the MuData R library[3] for the Bioconductor ecosystem (Huber et al., 2015) and the Muon.jl[4] package for Julia (Bezanson et al., 2017).

2.1.2.2 MuData serialisation

Serialisation is concerned about transmitting object as a series of bytes, e.g. writing it to a file or directory on a local hard drive or a remote machine. Importantly, individual modalities are serialised natively with AnnData, and multimodal annotations largely follow the serialisation strategies of annotations in individual modalities (Figure 2.1c). I implemented MuData serialisation to HDF5 files (The HDF Group, 2000–2010) and also outlined and implemented

[1] https://pypi.org/project/mudata/
[2] https://github.com/pmbio/MuDataSeurat
[3] https://www.bioconductor.org/packages/MuData/
[4] https://github.com/scverse/Muon.jl

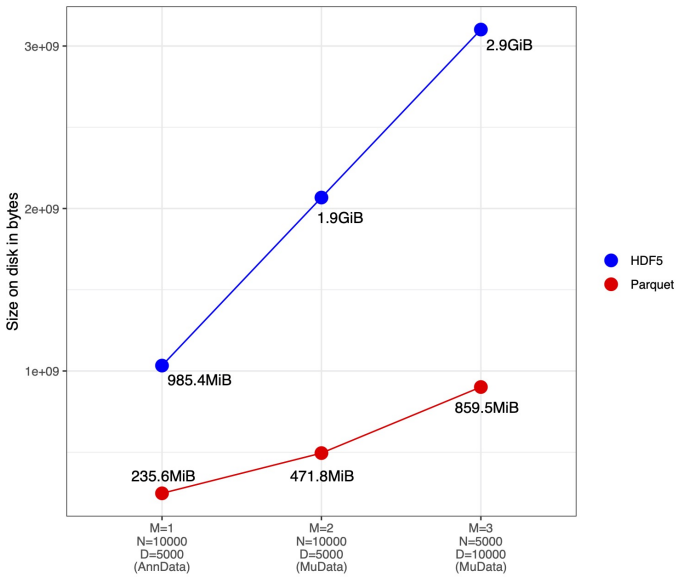

Fig. 2.3 Different serialisation formats for MuData and their size on disk

MuData serialisation using Parquet files allows to achieve notable file size reduction. Here, I simulated 3 datasets of variable size, where M is a number of modalities, N is the number of cells (rows) and D is the number of features (columns). For each dataset, three matrices were generated and stored: a sparse matrix from the Poisson($\lambda = 1$) distribution (1), which was then library size-normalised and log-transformed (2) and scaled to zero mean and unit variance (3).

a new serialisation strategy for AnnData, MuData and similar file formats using Parquet files (*Apache Parquet* 2022).

With HDF5 programming interfaces available across many languages, H5MU files are accessible in different programming environments – as apposed to e.g. R-specific RDS files (Huber et al., 2015; Satija et al., 2015) – solidifying the standard and allowing data exchange and consistent format definition across platforms. HDF5 made it possible to implement disk backing in order to access data in MuData files without loading count matrices from individual modalities. The latter are stored on disk inside an H5MU file following AnnData's native H5AD schema, and that makes it possible to directly read individual modalities.

Apache Parquet is an open-source data file format that offers efficiency storage for column-oriented data (*Apache Parquet* 2022). Taking advantage of the format and its optimisations, I implemented serialisation strategy for MuData using such files to store matrices and tables of data while keeping them in directories with structure matching in-memory object structure (Figure 2.1c). This proved to be beneficial with regard to storage optimisation achieving 70% – 75% reduction in file size with the naïve serialisation implementation (Figure 2.3). Another key advantage of this serialisation is accessibility of original data and its annotation as individual files to the user and to the external tools. For instance, Parquet files can be efficiently queried with DuckDB (Raasveldt and Mühleisen, 2019).

2.1.2.3 Comparison of MuData with alternative data standards

MuData builds on AnnData's concepts and code and incorporates and extends its ideas in a modular fashion akin to the modular infrastructure of Bioconductor (Huber et al., 2015). In fact, MuData strives to be the most general solution for multimodal omics data storage to date (Table 2.1).

2.1.2.4 Versatile use of MuData objects

MuData finds its primary use in multimodal omics data handling. Its flexible design allowed me however to extend its potential use cases and take advantange of MuData objects in diverse applications. I outline a few of those usage scenarios below.

Linear epigenomics features such as chromatin accessibility are frequently split per chromosome for efficient storage and access (Granja et al., 2021; Stovner and Sætrom, 2020). As this satisfies the original idea of multimodality (Figure 2.4a), such datasets can be split e.g. per chromosome, with each chromosome represented by an AnnData object, and used to create a MuData object.

MuData's design and implementation allow to specify an axis that is shared among the contained AnnData objects. This enables its configuration for datasets with the shared features space, e.g. data from multiple scRNA-seq experiments (Figure 2.4b).

Subsets of the data naturally arise in processing workflows. For instance, for many downstream applications, only selected features (e.g. highly variable genes) are kept. Other analyses, however, demand original data for the profiled features. This and similar use cases can also be addressed with MuData by denoting all the axes as shared. As such, this is also applicable to subsets of observations (cells) allowing to store e.g. subsampled data with a subset of cells in the same object (Figure 2.4c).

Table 2.1 MuData comparison to other data standards

	MuData	**AnnData**	**Seurat**	**Multi-Assay-Experiment**
Main programming language	Python	R	R	R
Object can contain data on disk (out of memory)	Yes	Yes	No[5]	Yes[6]
Default serialisation format	H5MU	H5AD	RDS	RDS
Default serialisation accessible in another programming language	Yes	Yes	No[7]	No[8]
Support for multiple modalities	Yes	No	Yes	Yes
Support for data missing in some modalities	Yes	NA	No	Yes
Support for multimodal embeddings	Yes	NA	No	No
Support for inter-modality feature relations	Yes	NA	No	No

[5]With SeuratDisk library, in-memory Seurat objects can be constructed from parts of the data stored in HDF5 files.

[6]Only possible with HDF5Array library for matrices stored in external HDF5 files.

[7]With SeuratDisk library, in-memory Seurat objects can be exported to HDF5 files.

[8]Only matrices stored in external HDF5 files, exported with HDF5Array library, can be accessed.

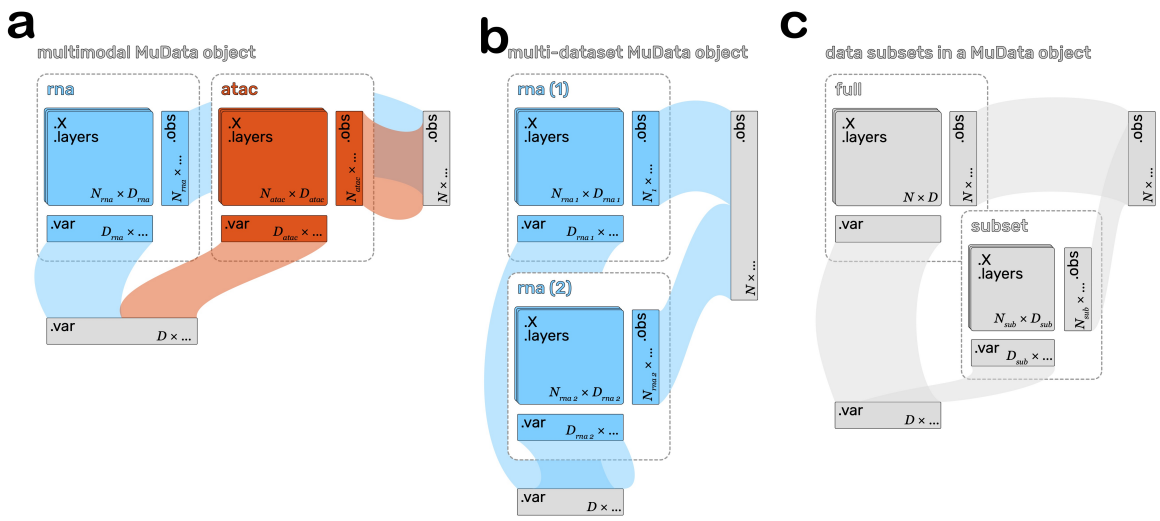

Fig. 2.4 Advanced use of MuData containers
a: Schematic representation of collating global sets of cells (observations) and features (variables) for a multimodal dataset with RNA (blue) and ATAC (red) modalities. Global set of features is a combination of feature sets of individual modalities. Global set of cells is a union of cells present in individual modalities (some modalities can have some cells missing). **b:** Schematic representation of collating global sets of cells and features for a collection of two scRNA-seq datasets, both in blue showing the same modality for both datasets. Global set of features is a union of genes (some datasets can have some genes missing). Global set of cells is a combination of cells of individual datasets. **c:** Schematic representation of collating global sets of cells and features for subsets of the same dataset. Both cells and features axes are shared in this case. Global set of features and global set of cells correspond to features and cells of the full dataset.

2.2 Multimodal omics analysis framework (MUON)

MuData enables interoperability through data standardisation. When analysing the data, most of the time one interacts with layers of abstraction on top of the data format — tools and frameworks. Here, I present a framework for multimodal data integration (MUON) that has been designed around MuData and enables multimodal data processing and integration, extending current best practices for scRNA-seq analysis to multimodal omics.

2.2.1 MUON: a framework for multimodal omics data

The MUON framework has been designed around MuData containers to manage, process, and visualise multimodal omics data. For the analysis of individual modalities stored in the container, existing workflows can be reused. For instance, in a multimodal dataset with gene expression and chromatin accessibility information from the same cell, gene expression can be processed and analysed with scanpy (Wolf et al., 2018). This way canonical processing steps, which include quality assessment, filtering cells, count normalisation and feature selection, are reused in MUON workflows (Figure 2.5a).

One of the key features of the MUON workflows is their modularity, which allows to define and combine alternative data processing strategies. With the major focus of MUON on multimodal integration, this means that the necessary components for this integration (e.g. embeddings of individual modalities) can be computed in independent ways as well as an integration strategy of choice can be applied.

For instance, matrix decomposition methods such as principal component analysis (PCA) or factor analysis (see Section 3.1) can be applied individually on gene expression and chromatin accessibility count matrices. Alternatively, joint matrix factorisation methods such as multi-omics factor analysis (MOFA) (Argelaguet, Arnol, et al., 2020; Argelaguet, Velten, et al., 2018) allow to perform decomposition on count matrices from multiple modalities simultaneously. Both approaches allow to obtain low-dimensional data embeddings. These embeddings can be then used for efficiently constructing a cell neighbour graph, either for individual modalities or for multiple modalities. The latter can be achieved via various strategies such as estimating cell neighbour graph based on MOFA factors, which capture information across modalities, calculating multimodal neighbours based on individual embeddings (Hao et al., 2021) or fusing multiple neighbour graphs derived for individual modalities (Bo Wang et al., 2014). These neighbourhood representations can be used directly

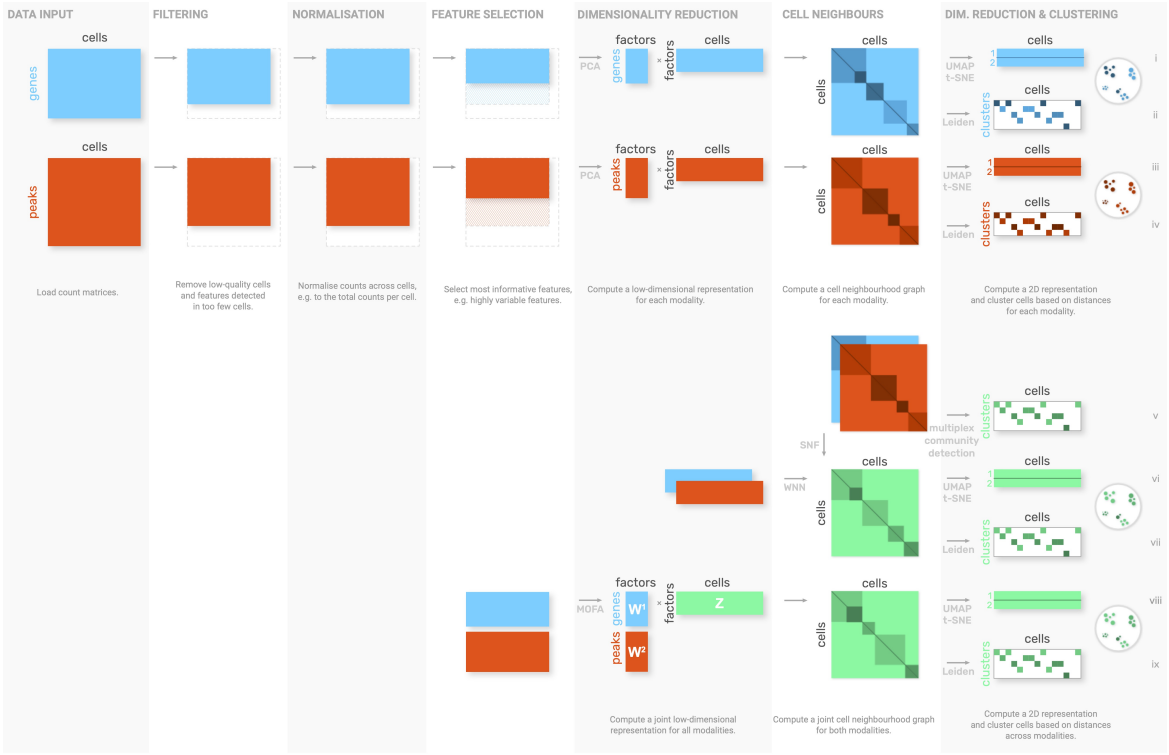

Fig. 2.5 Examples of joint scRNA-seq and scATAC-seq workflows that can be implemented end-to-end in MUON

Schematic representation of analytical workflows for multimodal data using joint RNA and ATAC profiling as an example. Here, features are represented by genes (in blue) and peaks (in red) in the two respective modalities. Matrices with counts or derived annotations are represented with rectangles, all stored in the MuData container. From left to right, processing steps for individual omics are followed by multimodal integration where information from multiple modalities is combined (in green). For processing individual omics, existing methods and frameworks can be reused. For multimodal integration, MUON enables tailored integration by making it possible to combine alternative analysis and integration steps. The outputs depicted here, from top to bottom, are UMAP space and cell cluster labels computed based on RNA (*I* and II, respectively) or ATAC (III and IV, respectively), cell cluster labels based on modality-specific neighbour graphs (V), UMAP space and cell cluster labels based on WNN (VI, VII) or MOFA (VIII, IX) output.

for downstream methods to construct non-linear embeddings for visualisation (McInnes et al., 2018) or to perform clustering of cells.

I have implemented the respective modular architecture of MUON as well as the interfaces to the aforementioned methods. Importantly, these interfaces extend and generalise the ones applicable for individual modalities in tools such as SCANPY (Wolf et al., 2018). For instance, the multiplex community detection (Mucha et al., 2010) interface has been exposed at `muon.tl.leiden` mimicking `scanpy.tl.leiden` for community detection algorithms for individual neighbour graphs (Traag et al., 2019).

2.2.2 Example applications of MUON

To illustrate the functionality of MUON, I have used it to analyse the latest single-cell multimodal datasets. In particular, in conjunction with SCANPY and AnnData and MuData as data structures, MUON interface was used to handle all the operations from loading counts matrices to performing quality control, data filtering, feature selection, cross-modality integration, visualisation and to storing the processed data and analysis results.

2.2.2.1 Single-cell RNA + ATAC sequencing

Simultaneous scRNA-seq and scATAC-seq profiling as performed using the Chromium Single Cell Multiome ATAC + Gene Expression protocol by 10x Genomics (*Single Cell Multiome ATAC + Gene Expression, 10x Genomics* 2022) allows to gather both gene expression and chromatin accessibility information from individual cells cell. Integrative analysis of such data can be performed with some of the existing toolchains such as the R-based Seurat framework (Hao et al., 2021), which implements a nearest neighbours-based integration strategy. Below I will highlight how analytical multimodal workflows for joint scRNA-seq and scATAC-seq profiling, including the ones implemented elsewhere, can be encoded in MUON end-to-end.

First, I processed and applied alternative integration strategies to the dataset of about ten thousand human peripheral blood mononuclear cells (PBMCs) (*PBMC Granulocyte Sorted 10k, 10x Genomics* 2022). For instance, MOFA yields a factor space, which we can use to visualise the dataset (Figure 2.6a), interpret it on the level of individual features (see Section 3.2.4) or use for downstream analyses (Figure 2.6b). The first MOFA factors, which explain the largest proportion of variance, capture biological variation along the myeloid-lymphoid and cytotoxicity axes (Figure 2.6a).

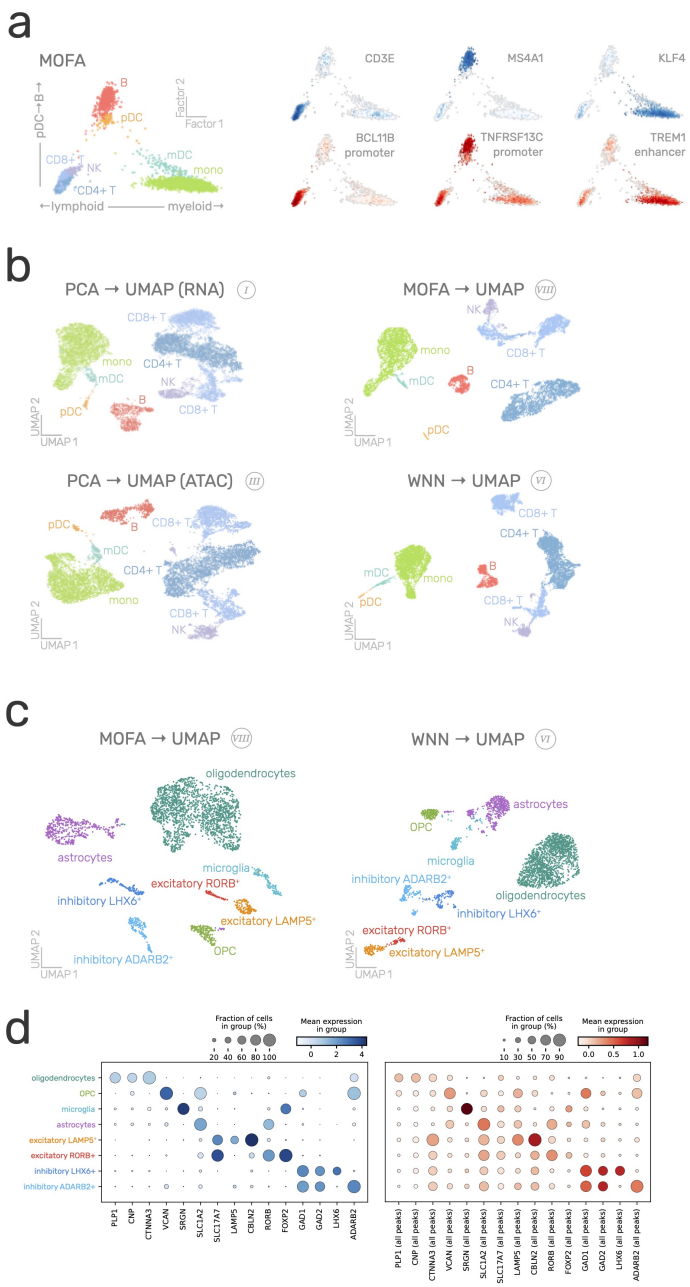

Fig. 2.6 Integration of RNA and ATAC modalities with MUON
a: MOFA factors for ten thousand human PBMCs. Cells are colours by coarse-grained cell type on the left and, on the right, by gene expression (in blue) or peak accessibility (in red). Illustrated genes and peaks have been chosen to reflect cell type-specific variability in the factor space. **b:** UMAP space for the same dataset constructed from principal components for individual modalities (left) or from MOFA factors (top right) or from cell WNN graph (bottom right). Same coarse-grained cell type annotation was used to colour the cells. c: UMAP space for the three thousand human brain cells constructed from MOFA factors (left) or from cell WNN graph (right). Cells are colours by coarse-grained cell type (top) and gene expression and peak accessibility in blue and red, respectively (bottom). d: Gene expression (left, in blue) for the cell type markers in the same dataset. Accessibility (right, in red) is shown for all the peaks that correspond to the respective marker genes.

Alternative strategies illustrated in Figure 2.5 can be successfully applied here, and UMAP space can be computed to visualise the information from individual modalities or from both modalities, for instance, either through MOFA factors or the weighted nearest neighbours graph (Figure 2.6b).

As these workflows do not make assumptions about the biological nature of the data, they are portable across different collections of cell types, tissues and species. In order to showcase the utility of MUON in a different setting, I also applied analogous MUON workflows to the exploration of another dataset — three thousand human brain cells (*Flash-Frozen Human Healthy Brain Tissue (3k), 10x Genomics* 2022) (Figure 2.6c). Leveraging visualisation capabilities of SCANPY and MUON, I generated a map of gene markers for different cell types in this dataset (Figure 2.6d). For this data representation, MUON can aggregate accessibility across defined genomic loci such as genes.

2.2.2.2 CITE-seq, human blood cells

I then designed, implemented and applied a MUON workflow to analyse CITE-seq data, in which features in two modalities are represented by gene and protein counts (Stoeckius et al., 2017).

To process protein counts, special strategies to denoise and scale them have been developed such as *dsb* originally implemented in R (Mulè et al., 2022). Together with Ilia Kats, I reimplemented this method in Python providing MUON with a native Python solution to process CITE-seq data. These protein counts can then be used for defining cell types using pairwise plots (Figure 2.7a) as visualisation resembling gating in flow cytometry (Mattanovich and Borth, 2006).

Integrative CITE-seq data analysis can be performed with some of the existing frameworks such as Seurat (Hao et al., 2021) as well as with the packages exclusively focused on CITE-seq data such as CiteFuse (H. J. Kim et al., 2020), both featuring custom protein count normalisation approaches and implemented in R. MUON enables CITE-seq data handling in Python and allows to implement analogous workflows, mimicking the processing and analysis steps discussed in the previous section. To demonstrate it, I used MUON to load, preprocess, annotate and visualise the CITE-seq PBMC dataset (Figure 2.7b).

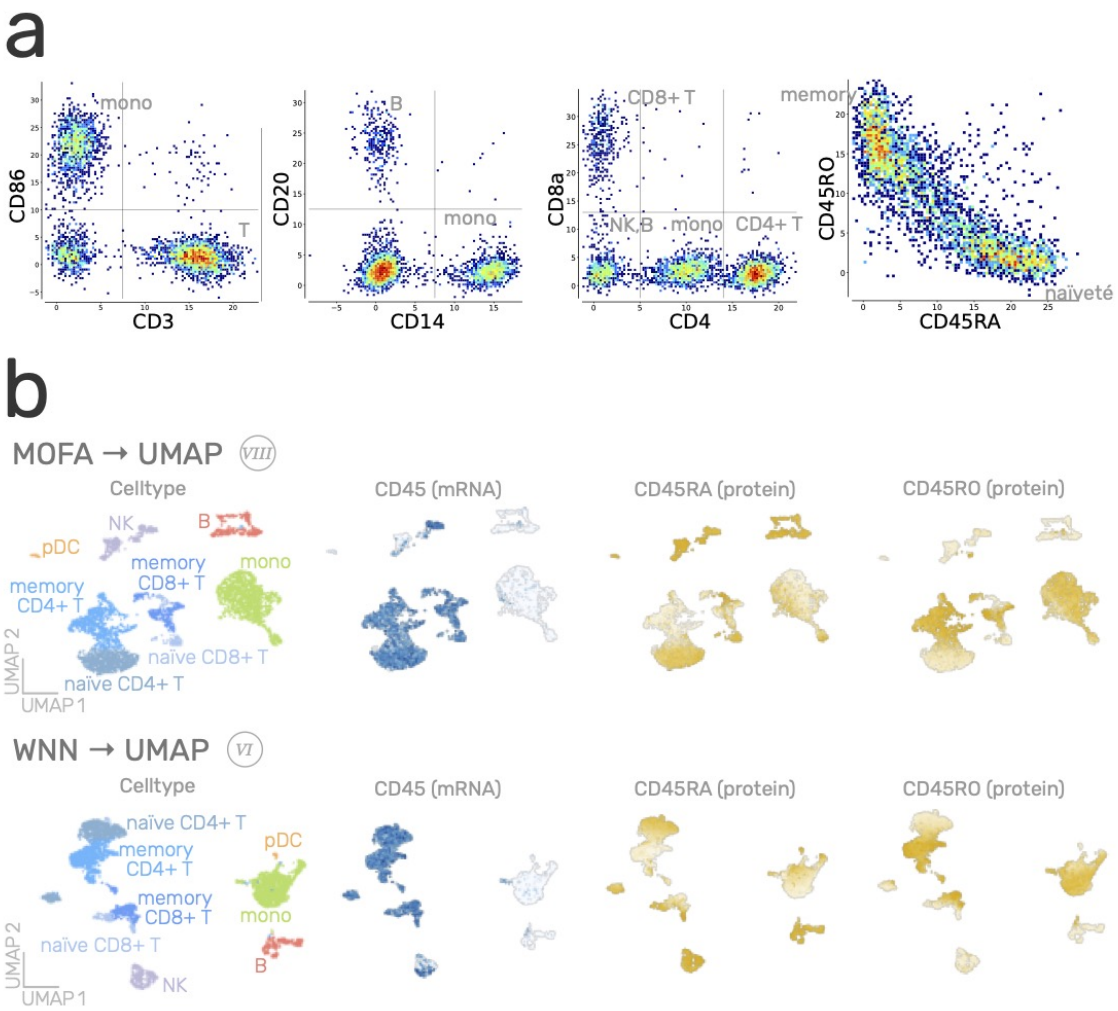

Fig. 2.7 Integration of CITE-seq data with MUON
a: Protein abundance values for some of the features in the CITE-seq PBMC dataset. Colours correspond to relative local density of cells, red denotes higher density, blue denotes lower density. Values after the dsb normalisation are plotted. **b:** UMAP latent space for the same dataset from MOFA factors (top) or WNN graph (bottom). Cells are coloured by coarse-grained cell type (left) and features values (gene expression in blue, protein abundance in yellow).

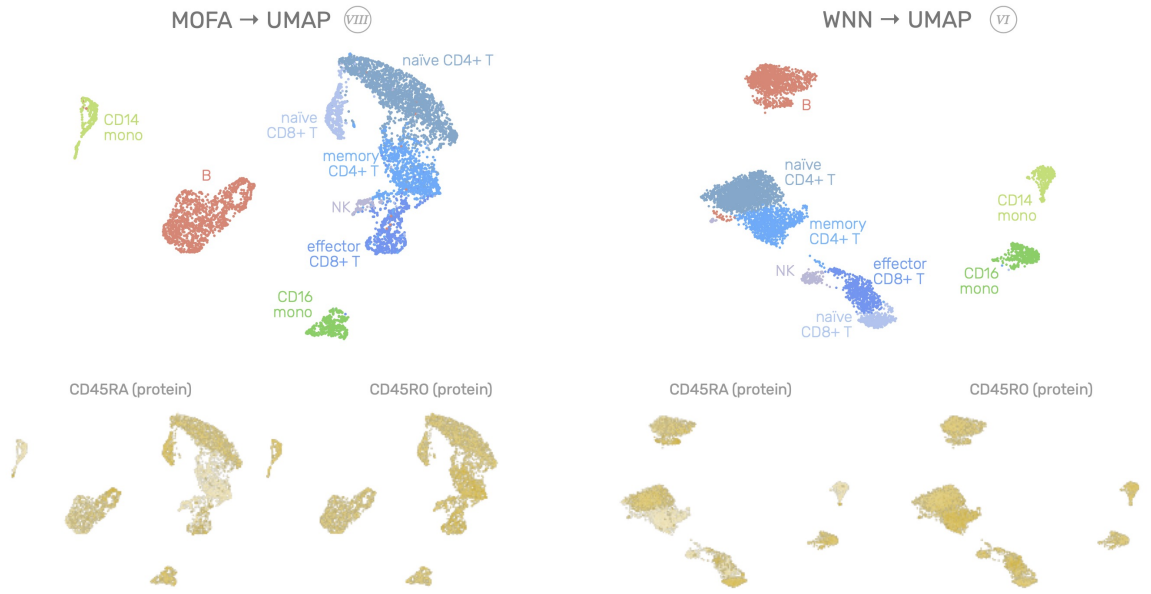

Fig. 2.8 Integration of a trimodal dataset with MUON
UMAP latent space for the TEA-seq profiling of PBMCs. Left and right parts of the figure show the latent space constructed on MOFA factors or WNN graph, respectively. Colour corresponds to coarse-grained cell types (top) or protein counts for CD45RA and CD45RO proteins.

2.2.2.3 Trimodal assays

As MUON as well as MuData is designed to handle an arbitrary number of modalities, trimodal assays can also be analysed with this workflow. For instance, transcriptome (T) with epitopes (E) and chromatin accessibility (A) were rofiled with TEA-seq for peripheral mononuclear blood cells (PBMCs) (Swanson et al., 2021).

This example essentially combines the Multiome and CITE-seq processing, and the analysis results allow to distinguish basic cell populations based on a few protein markers as well as on molecular features across modalities (Figure 2.8).

2.2.3 Methods

2.2.3.1 MUON implementation

I implemented MUON[9] in the Python programming language (Van Rossum and Drake Jr, 1995). MUON builds on the software stack for numerical and scientific computing,

[9]https://pypi.org/project/muon/

in particular NumPy (Harris et al., 2020), SciPy (P. Virtanen et al., 2020), Scikit-learn (Pedregosa et al., 2011), Pandas (McKinney et al., 2011), h5py (Collette, 2013), AnnData (Virshup et al., 2021), SCANPY (Wolf et al., 2018). MUON interfaces MOFA+ (Argelaguet, Arnol, et al., 2020) for the corresponding multimodal integration and comes with a Python reimplementation of WNN following (Hao et al., 2021) and its generalisation to arbitrary number of modalities from (Swanson et al., 2021). For data visualisation, MUON relies on matplotlib (Hunter, 2007) and seaborn (Waskom, 2021).

I exposed the functionality of MUON via interfaces that resemble the ones in SCANPY (Wolf et al., 2018). Briefly, functions are organised in modules by their purpose, e.g. `muon.pp` for preprocessing, `muon.tl` for computational methods, `muon.pl` for plotting. Operations and methods that are specific to individual modalities are provided in corresponding modules such as `muon.atac` and `muon.prot` for chromatin accessibility and protein abundance, respective.

2.2.3.2 Processing RNA + ATAC datasets

Chromium Single Cell Multiome ATAC + Gene Expression data was provided by 10x Genomics (*Single Cell Multiome ATAC + Gene Expression, 10x Genomics* 2022). Data for PBMCs from a healthy donor with granulocytes removed through cell sorting had been processed with the ARC 1.0.0 pipeline (*PBMC Granulocyte Sorted 10k, 10x Genomics* 2022). The dataset was loaded with MUON. SCANPY and MUON were used to normalise and log-transform counts for both gene expression and chromatin accessibility, to identify highly variable features, and then to centre them to zero mean and unit variance. Principal components for each modality were computed with PCA as implemented in Scikit-learn and SCANPY. MOFA factors and WNN graph were computed with MUON.

I used the same workflow to process data from the frozen human healthy brain tissue (*Flash-Frozen Human Healthy Brain Tissue (3k), 10x Genomics* 2022) and I relied on (Hodge et al., 2019) to identify coarse-grained cell types (Figure 2.6d).

2.2.3.3 Processing CITE-seq data

CITE-seq data for PBMCs from a healthy donor were provided by 10x Genomics (*PBMC Protein 5k v3, 10x Genomics* 2022). Protein counts were denoised and scaled using the dsb procedure (Mulè et al., 2022) as implemented in MUON. Gene expression counts were

library size-normalised and log-transformed prior to scaling and identification of highly variable genes. MOFA factors and WNN graph were computed with MUON.

2.2.3.4 Processing TEA-seq data

I used MUON to load, process and visualise a sample from the trimodal dataset (Swanson et al., 2021). Individual modalities were processed as discussed in previous sections essentially making this processing workflow a combination of RNA + ATAC and CITE-seq processing workflows described above for other datasets.

2.3 Discussion

Analytical tasks discussed in Section 1.2 dictate the requirements for efficient data representations and storage. I formulated the following considerations that are to be met by such data standards:

- defined data structure permitting automation;
- efficient storage with respect to file sizes providing reasonable read and write speed;
- shareable files accessible from different programming environments;
- data access without loading it all into memory (scalability concern);
- possible to read in a few years' time (stability and legacy concern);
- adopted by data providers and users in the field;
- expandable to new modalities such as spatial and temporal domain, and to cross-modality interactions.

In the previous sections I demonstrated how these challenges are addressed by MuData (Section 2.1), which implements efficient and extensible approach to multimodal data handling. MuData builds on the existing ecosystem for representing individual modalities, and this continuity has also been realised in MUON (Section 2.2) — a multimodal omics analysis framework built around MuData.

While many current analysis workflows are focusing on either horizontal (e.g. different datasets) or vertical (different modalities) integration, the complexity and the scale of multimodal single-cell datasets dictate the necessity of designing new instruments for generalised multi-dataset multimodal data storage. I will provide my considerations for the key points behind potential solutions to that below.

2.3.1 Considerations for multimodal multi-dataset collections

This section is aimed to provide a collection of ideas and solutions to some of the challenges in representing and handling collections of multiple multimodal single-cell genomics datasets.

Existing data abstractions for single-cell and multimodal genomics, including MuData, tackle the issue of handling individual datasets comprised of numerical data such as gene abundance estimates and of annotations, from clinical information to experimental summary and quality control to derived data such as cell labels. Multi-dataset analysis scenarios currently lack an abstraction layer however they have been arising with increasing frequency,

for instance when comparing individual experiments with large annotated datasets (*atlases*) such as the one by Consortium* et al., 2022, performing benchmarks for new methods or generating new insight using an aggregation of multiple available datasets. Here, I outline some considerations for new data abstractions that focus on such scenarios as well as propose the design of the respective solutions.

2.3.1.1 Generalising aligned axes

AnnData and MuData feature a concept of aligned axes: core components of those objects have explicit constraints on their dimensions, e.g. the matrix with reduced dimensions should have the number of rows equal to the number of cells in the dataset.

AnnData, and consequently MuData, are inherently two-dimensional: designed around a count matrix, there are only dimensions of observations (cells) and of variables (features, e.g. genes). Complex analysis workflows motivate going beyond two dimensions. For instance, when analysing chromatin accessibility, one deals with multiple dimensions such as cells, peaks, genes, transcription factors and their motifs. Then, during the analysis, there's a matrix of peaks by transcription factors generated as well as peaks by motifs and motifs by transcription factors.

The solution is to allow the user to configure named axes, which will then constrain their dimensions upon defining them, and allow an arbitrary number of those. Data retrieval can then be implemented using these names:

```
data["cells","peaks"]
data["peaks","tf_motifs"]
```

In practice, it is important to preserve data at each step of the data transformations workflow, for instance to store original gene counts prior to normalisation as they are typically used as input for count-based models and analyses such as differential expression with generalised linear models (Robinson et al., 2010) or autoencoders (Lopez et al., 2018). This can be addressed by extending data retrieval to the layered storage for the matching dimensions, e.g. original, normalised and scaled counts:

```
data["cells","genes","scaled"]
```

This can be further extrapolated to nested structured storage, for instance for organising spliced and unspliced count matrices (see Section 1.3.2), each of which can also have original, normalised and scaled variants:

```
data["cells","genes","unspliced","normalised"]
```

This provides a practical improvement over existing solutions that only provide access to count matrices via predefined or user-defined labels. Such solution also contributes to the ongoing research of improving clarity and reliability in tensor manipulation expanding it to annotated datasets (Rogozhnikov, 2022).

2.3.1.2 Metadata groups

Metadata can be derived at different stages of data analysis workflows. For instance, there can be patient information available in the beginning of the workflow, statistical properties calculated and added at the quality control stage and annotated biological entities as one of the analysis results. Semantic separation of these metadata groups can be beneficial both for structuring analysis workflows and for data sharing purposes. Current solutions, however, do not make such distinctions.

As a consequence of the generalisation of aligned axes, metadata groups can be represented by respective tables with corresponding dimensions and axes. Some of such tables such as with quality control information can be defined across datasets. Importantly for multi-dataset workflows, this allows storing and querying data statistics across datasets and modalities, for example for multiple scRNA-seq datasets (Figure 2.4b):

```
data["dataset1","qc_table"]
data["dataset2","qc_table"]
```

2.3.1.3 Multi-dataset multimodal storage

Generalised aligned axes enable multi-dataset infrastructure so that multiple datasets can be stored and then queried across multiple feature spaces. Figure 2.9 provides an example of such storage.

```
data["reference","genes"]
data["reference","peaks"]
data["query","peaks"]
```

Querying across datasets and modalities can be useful for data integration methods as allows to get latent representations of all the datasets and modalities in the uniform fashion

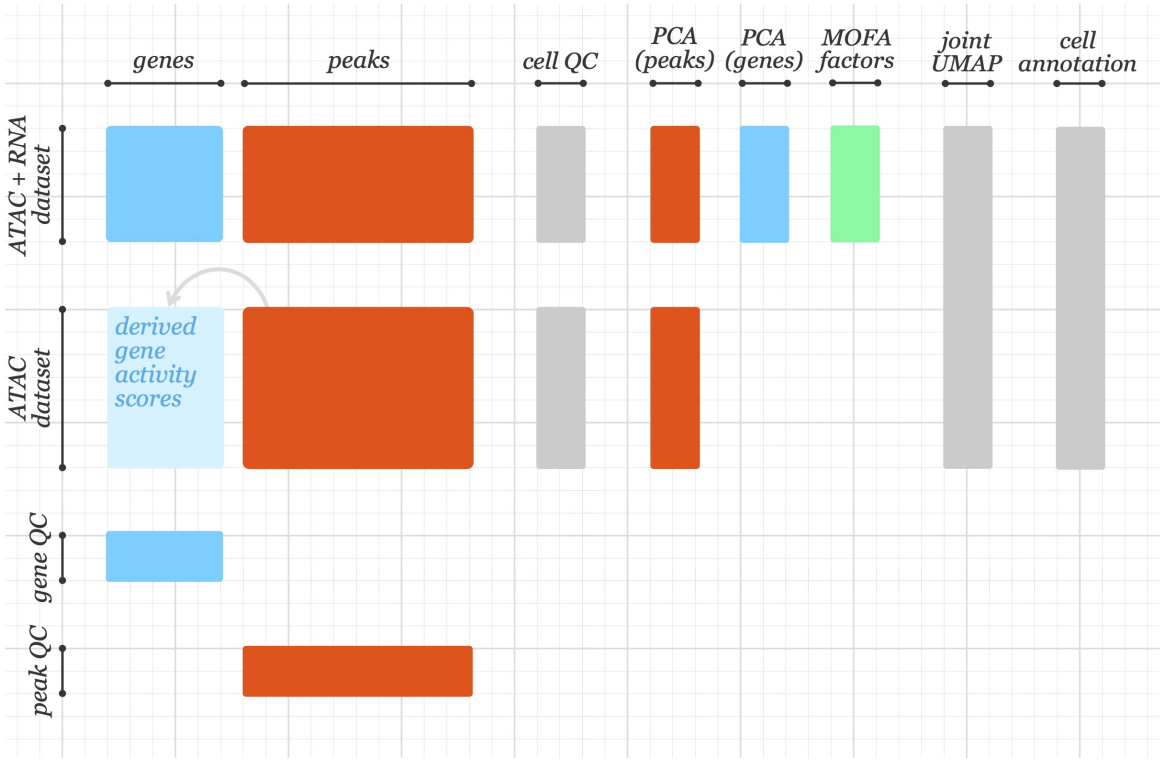

Fig. 2.9 Multi-dataset multimodal storage structure
Structured storage for two-dimensional data including numerical objects (counts) and tables (annotations) is demonstrated on two datasets, one of multimodal ATAC + RNA profiling and another of ATAC profiling. RNA- and ATAC-specific information is colour-coded in blue and red, respectively. Multimodal annotation for the first dataset is shown in green. Matrices and tables are depicted with rectangles and are aligned according to their dimensions.

and then store the integrated representation in the same object. Such joint representation would then have a union of the original axes as one its dimensions permitting it to be queried with original axes as well to return dataset-specific parts of the joint matrix.

```
data["reference","umap"]
data["query","umap"]
data[("reference","query"),"umap"]  # integrated
```

2.3.1.4 Data immutability

Count matrices and their downstream derivatives are frequently treated as immutable in analysis workflows, e.g. individual values in the data are not modified and the derived multimodal annotations are produced by algorithms for the whole dataset.

Hash functions can be used to derive unique identifiers for the matrices so that different matrices will be mapped to different values (Bakhtiari et al., 1995). Derived outputs such as embeddings can then refer to the hash values uniquely identifying the data that they originated from as well as carry a uniquely identifying hash value themselves.

Treating matrices and tables as objects identified by their hash values allows to define computationally efficient workflows as well as to then devise object merging operations so that the resulting data container would include all the objects with unique hash values. To implement this, an additional data structures can be employed such as conflict-free replicated data types (CRDTs) (Shapiro et al., 2011). The utility of such functionality is especially notable for incrementally updating data on disk rather than loading it all into memory to then write the whole dataset to disk. More so, CRDTs can be employed in multi-dataset and multimodal settings where different people can work on different parts – datasets or modalities — of the data.

2.3.1.5 Delayed operations

Dataset collections and even individual single-cell datasets have been growing larger in size so that their size is larger than random access memory available on personal computing devices. Thus efficient data abstractions should make sure individual tables are only read into memory when required — and can be easily discarded after the computation. Moreover, analytical solutions for querying databases and files can further reduce the memory requirements for computational workflows (Raasveldt and Mühleisen, 2019).

2.3.1.6 Idempotent operations

In analysis workflows, operations are frequently chained, e.g. a UMAP representation (see Section 1.2.3.5) requires computing cell neighbourhoods (Section 1.2.3.6), which are frequently computed on principal components (Section 3.1.1.1). In practice these operations represent a graph rather than a linear sequence as analytical tasks can be solved with multiple alternative operations. Thus it is natural to build the analysis graph with idempotent operations, i.e. operations $f()$ such that

$$f(f(x)) = f(x).$$

Current analytical interfaces such as in SCANPY (Wolf et al., 2018) are not based on idempotent operations thus limiting alternative multi-dataset computations. This can be addressed for instance by specifying the expected outputs of the computation:

```
f(data, in_layer="counts", out_layer="lognorm")
```

More broadly, such operations can be bound to data objects. This idea can be illustrated with principal component analysis applied to a scaled count matrix:

```
data.bind(
    f=pca,
    in_axes=("cells","genes","scaled"),
    out_axes=("cells","components"),
)
```

Such binding operations can then be applied across multiple datasets stored in the object thus offering a transparent interface for horizontal scalability.

2.3.1.7 Interoperability with relational databases

Storing individual tables with known dimensions means that analytics can be generalised to any n-dimensional objects, most importantly tables that can be stored in relational databases. Relational databases typically store data as collections of rows (or columns) and feature various optimisations. This is particularly relevant for data that is not dataset-specific, e.g. gene annotations such as names, intervals and sequences. For example, the computational model to calculate GC content for each gene in the annotation that also contains gene sequence in this case could be represented as follows:

```
SELECT
    GCcontent(sequence)
FROM
    data["GRCH38_genes","genes_metadata"]
```

Such approach to handling tabular data can also increase the accessibility of annotations and cross-dataset statistical properties that are readily available on local or remote servers.

Chapter 3

Factor analysis for single-cell multi-omics

Multi-omics factor analysis (MOFA) is a Bayesian framework that has been previously proposed for multimodal data integration, which can be viewed as a generalisation of PCA to multi-omics data. Based on group factor analysis, MOFA uses automatic relevance determinations (ARD) and spike-and-slab priors to address data sparsity and facilitate factor interpretability.

Single-cell datasets feature increasingly large numbers of cells but also increasingly sophisticated structure of the data with cells coming from different experimental batches, samples or donors, etc. To address this, I proposed and implemented ARD and spike-and-slab priors for factors complementing the respective priors on weights and making the model symmetrical. This enables integration of multiple modalities (views) across multiple groups of cells where factor activity can vary from group to group. To make such models more accessible across a range of platforms, I also improved existing and implemented new functionality for the MOFA R package, implemented an interactive web-based platform for model interrogation as well as conceived, designed and implemented a Python package for downstream analysis and visualisation of MOFA models. The results in this chapter represent my own work unless stated otherwise, and most of them have been published in Argelaguet, Arnol, et al., 2020, which describes our joint work together with Ricard Argelaguet, Damien Arnol and others.

3.1 Introduction to latent variable models

3.1.1 Basic latent variable models

Many datasets can be described with a manifold of lower dimensionality than the original data, and that also applies to gene expression data (Heimberg et al., 2016, **?**]). This motivates latent variable models (LVMs) with continuous latent variables. Many of those models assume Gaussian distribution for both observed and latent variables and use a linear dependence of observed variables on the latent ones (Bishop, 2006). Such models are frequently formulated as matrix factorisation problems, with principal component analysis (PCA) being the cornerstone technique. PCA and other linear as well as nonlinear latent variable models are used across different domains for dimensionality reduction and lossy data compression («latent space») as well as for data visualisation.

Notably, as demonstrated in the section 1.2.3, a lot of downstream analysis tasks use linear manifolds as their input, which highlights their utter importance for high-dimensional genomics data analysis. Moreover, while latent variable models such as PCA or factor analysis fall under the category of unsupervised methods, they have been extended to incorporate knowledge via structured sparse priors in models such as f-scLVM (Buettner, Pratanwanich, et al., 2017) or MuVI (Qoku and Buettner, 2022).

3.1.1.1 Principal component analysis

PCA is a linear model that transforms the data into a set of orthogonal components that explain the largest amount of variance. The data matrix Y is factorised into a product of loadings W and components Z: $Y \sim WZ^T$. The former makes it possible to interpret the transformed axes: features (e.g. genes) with high loadings strongly influence respective components. In practice, full decomposition is not necessary and too computationally intensive to compute, and only some of the components are calculated, e.g. relying on truncated SVD (Halko et al., 2010). Numerous extensions of PCA have been developed that enable better modelling of more complex data. As such, Bayesian formulation as shown below in the section 3.1.2 allows to incorporate sparsity. It is also important to note that efficient and open-source implementations are available and are widely used such as `sklearn.decomposition.PCA` in Python (Pedregosa et al., 2011) and `stats::prcomp` in R (R Core Team, 2013).

3.1.1.2 Nonnegative matrix factorisation

Nonnegative matrix factorisation (NMF) has been proposed as a method for multivariate data decomposition under the constraint of non-negativity of components (D. Lee and H. S. Seung, 2000; D. Lee and S. Seung, 1999). As this limits the combinations of components to be additive (no subtractions), it can provide improved interpretability (D. Lee and H. S. Seung, 2000). As a testament to this, widely used characterisation of mutational processes signatures in human cancers was established with NMF (Alexandrov et al., 2013). NMF has also been extended by being combined with sparsity conditions (Eggert and Korner, 2004) and Bayesian inference (Ali Taylan Cemgil, 2009) and served as a foundation for other methods such as nonnegative spatial factorisation (Townes and Engelhardt, 2021).

NMF can be used for dimensionality reduction of Gaussian- or Poisson-distributed data (Févotte and A. Taylan Cemgil, 2009). Sparse Poisson-distributed data $y \sim \text{Poisson}(\lambda)$ can be handled with models such as hierarchical Poisson matrix factorisation (HPF) (Gopalan et al., 2014). HPF has also been adopted for dimensionality reduction of scRNA-seq data avoiding the requirement of prior normalisation (Levitin et al., 2019).

3.1.1.3 Independent component analysis

When the latent distribution is not Gaussian and factorises so that

$$p(\mathbf{z}) = \prod_{k=1}^{K} p(z_k),$$

this gives rise to the family of models known as independent component analysis (Bishop, 2006; Comon and Jutten, 2010; Hyvärinen and Oja, 2000). Such models were applied in various context including scRNA-seq (Biton et al., 2014; Francesconi et al., 2019; Liebermeister, 2002; Macaulay, Svensson, et al., 2016; Sastry et al., 2021; Sompairac et al., 2019; Trapnell et al., 2014; W. Wang et al., 2021), and there's been progress in defining generalisations of ICA including independent vector analysis (Hyvarinen and Morioka, 2016; T. Kim et al., 2006; Winther and Petersen, 2007).

3.1.2 Probabilistic formulation of factor analysis

3.1.2.1 Principal component analysis

Briefly, PCA is an orthogonal projection of data onto a lower dimensional space so that the variance of the projection is maximised (Figure 3.1). While various alternative formulations of PCA can be proposed (Bishop, 2006), probabilistic formulation allows to define it in a Bayesian fashion that can be transparently extended later. In this formulation, the prior distribution over the continuous latent variable $\mathbf{z}$ is given by a standard Gaussian distribution:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I}).$$

Observed variable $\mathbf{x}$ conditioned on the values of $\mathbf{x}$ is also Gaussian:

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I}).$$

The observed counts are thus defined by a linear transformation of the latent variable $\mathbf{z}$ with an additive Gaussian noise:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \mu + \varepsilon.$$

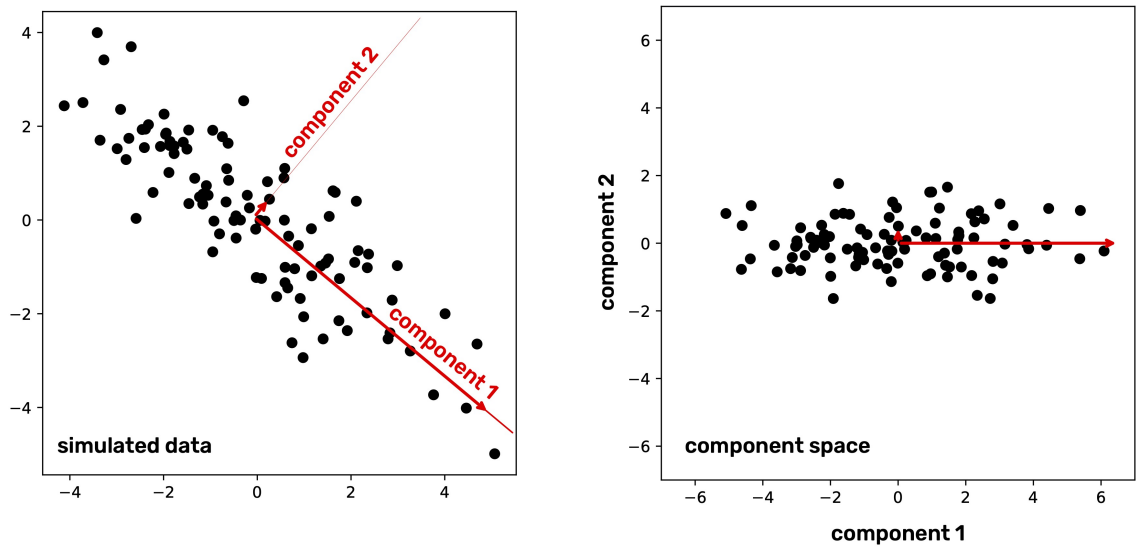

Fig. 3.1 Demonstration of PCA of simulated data
PCA was applied on simulated data (left) and identified the first component as the axis of the largest variance in the data (right).

PCA can be also interpreted as a Gaussian process that maps embedding to the data space giving rise to Gaussian Process LVM (GPLVM) formulation (Lawrence, 2003). Such models have been adopted and adapted for omics data for instance with scLVM (Buettner, Natarajan, et al., 2015) and tGPLVM (Verma and Engelhardt, 2020).

3.1.2.2 Sparsity

So-called automatic relevance determination (ARD) can be used for pruning out extra components (S. Virtanen et al., 2012). For that, an additional prior can be defined:

$$p(w_{dk} \mid \alpha_k) = \mathcal{N}(w_{dk} \mid 0, \alpha_k^{-1}),$$

i.e.

$$p(\mathbf{W} \mid \alpha) = \prod_{i=1}^{K} (\frac{\alpha_i}{2\pi})^{K/2} \exp(-\frac{1}{2}\alpha_i \mathbf{w}_i^T \mathbf{w}_i),$$
$$p(a_k) = \mathcal{G}(a_k \mid a_0^{\alpha}, b_0^{\alpha})$$

Biological processes would be expected to stem from the activity of a fraction of all the genes, and this notion can be incorporated in the model as a sparsity assumption. In order to accommodate sparsity at the level of individual features, suitable prior formulations for sparse data can be used. One way to accommodate sparsity is to add a Bernoulli-Gaussian prior commonly referred to as spike-and-slab model.

$$p(\hat{w}_{dk}, s_{dk} \mid \alpha_k, \theta_k) = \mathcal{N}(\hat{w}_{dk}^m \mid 0, \alpha_k^{-1}) Ber(s_{kd} \mid \theta_k), i.e.$$

$$p(\hat{w}_{dk}, s_{dk} \mid \alpha_k, \theta_k) = (1 - \theta_k)\mathbf{1}_0(\hat{w}_{dk}) + \theta_k \mathcal{N}(\hat{w}_{dk} \mid 0, \alpha_k^{-1})$$
$$p(\theta_k) = \mathcal{B}(\theta_k \mid \alpha_0^{\theta}, b_0^{\theta})$$

In this model, parameter θ_k for each factor k will reflect the level of sparsity per factor, with the value close to zero corresponding to a sparser factor (most weights are close to 0).

3.1.2.3 Factor analysis

Stemming from probabilistic PCA, factor analysis defines conditional distribution of the observed variable $\mathbf{x}$ given the latent variable $\mathbf{z}$ as a Gaussian with a diagonal covariance Ψ:

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \Psi).$$

3.1.2.4 Group factor analysis

When variable groups (*views* of the same data) are to be considered, the factors should provide weights $\mathbf{W}_m$ across the views m. This has been formulated in group factor analysis that also incorporates view-wise ARD prior (Klami et al., 2014; S. Virtanen et al., 2012):

$$p(w^m_{d,k} \mid \alpha_{m,k}) = \mathcal{N}(w^m_{d,k} \mid 0, \alpha^{-1}_{m,k}).$$

In the case of omics data, data views are different omics modalities. Multi-omics factor analysis (MOFA) model advances group factor analysis with both group-wise and feature-wise sparsity and comes with an accessible interface for model interpretation in R (Argelaguet, Velten, et al., 2018).

3.1.3 Variational inference

Inference in group factor analysis models such as MOFA is performed using mean-field variational Bayes (David M. Blei et al., 2017; Saul et al., 1996). In variational inference (VI), the intractable posterior distribution $p(\mathbf{z} \mid \mathbf{y})$ is approximated by a variational distribution $q(\mathbf{z})$ from a predefined family of distributions. The mean-field assumption states that this distribution factorises as

$$q(\mathbf{z}) = \prod_{j=1}^{J} q_j(z_j).$$

The difference between the true and the approximate posterior distributions is calculated as Kullback–Leibler (KL) divergence:

$$\mathrm{KL}(q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x})) = -\int_z q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}.$$

The goal is thus to find $q(\mathbf{Z})$ that minimises the KL divergence, which is equivalent to maximising the Evidence Lower Bound (ELBO) $\mathscr{L}(\mathbf{z})$:

$$\begin{aligned}\mathrm{KL}(q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x})) &= \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z} \mid \mathbf{x})] \\ &= \log p(\mathbf{x}) - \mathscr{L}(\mathbf{z}), \\ \mathscr{L}(\mathbf{z}) &= \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] - \mathbb{E}[\log q(\mathbf{z})].\end{aligned}$$

Maximising ELBO thus means maximising a lower bound on the true marginal log-likelihood $\log p(\mathbf{x})$ with the difference between the true posterior $\log p(\mathbf{z} \mid \mathbf{x})$ and the variational distribution $\log q(\mathbf{z})$ being given by the KL divergence.

3.2 MOFA+

MOFA+ is a group factor analysis model suited for integrating multi-omics data via a common sample space with features coming from distinct modalities. It infers a low-dimensional representation of the data in terms of a small number of latent factors that capture the source of variability. MOFA+ benefits from ARD to disentangle variation shared across modalities from modality-specific variability as well as from sparsity assumptions on weights, which facilitate factor interpretation. Joint modelling of multiple groups of cells profiled across multiple modalities is enabled in MOFA+ by ARD priors on factors.

3.2.1 MOFA+ overview

MOFA+ builds on the group factor analysis framework MOFA (Argelaguet, Velten, et al., 2018) with view-wise and feature-wise sparsity. In order to apply group factor analysis models to multimodal single-cell data, I devised and implemented structured sparsity for groups of cells (ARD) and individual cells (spike-and-slab) thus defining the MOFA+ model. Consequently, I implemented the storage of MOFA+ models with new sparsity options based on MOFA model serialisation.

The input of MOFA+ is a collection of matrices with cells and features grouped in cell groups and modalities, or views, respectively. In practice, cell groups can correspond to different experiments, experimental conditions or sequencing batches. A trained MOFA+ model contains K latent factors that explain the major axes of variation in the dataset with associated feature weight matrices (Figure 3.2). The latter can be inspected and visualised in order to provide interpretation to the inferred factors (Figure 3.2b).

Moreover, MOFA+ also inherits features from the first version of the method including inference with non-Gaussian likelihoods such as Poisson for count data or Bernoulli for binary data (Seeger and Bouchard, 2012).

To facilitate the use and the adoption of the method, I also implemented a Python package[1] as well as an open-source web-based resource[2] to inspect and interrogate MOFA+ models with the Shiny R package (Wickham, 2021).

[1] https://pypi.org/project/mofax/
[2] https://www.ebi.ac.uk/shiny/mofa/

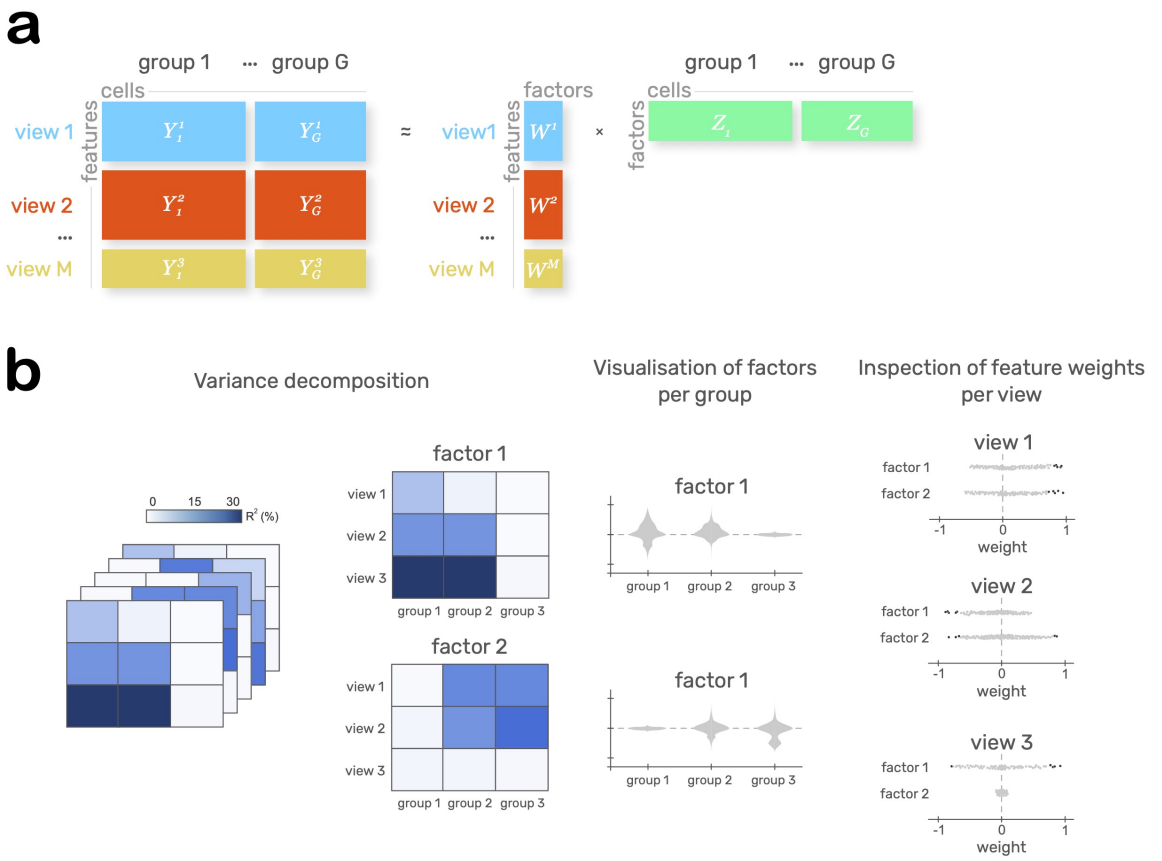

Fig. 3.2 MOFA+ framework for single-cell data integration across modalities and groups of cells

a: MOFA+ overview as a matrix factorisation problem. Data input consists of multiple matrices with features grouped into M views and cells grouped into G groups. Low-dimensional representation of the data is given by the inferred factors Z. Feature importance for each of the K factors is provided by the weight matrices W. **b:** Some of the downstream analyses to interpret the trained MOFA+ model include (from left to right) variance decomposition, factor values inspection and feature weights inspection.

3.2.1.1 MOFA+ illustration using simulated data

In order to assess if the formulation of group-wise sparsity in MOFA+ enables the detection of factors with differential activity in groups of cells, I simulated a dataset with differential factor activity and used MOFA models with and without respective sparsity to factorise it. For data simulations, I considered $M = 2$ views with $D = 800$ features in each of them and $G = 2$ groups with $N = 400$ cells in each of them. From $K = 8$ factors with differential activity, only the first two were simulated to be active in both views and groups, and overall, their activity pattern was simulated as outlined by the binary mask in Figure 3.3a. The results demonstrate that MOFA+ was able to reconstruct the true factor activity in different groups (Figure 3.3b).

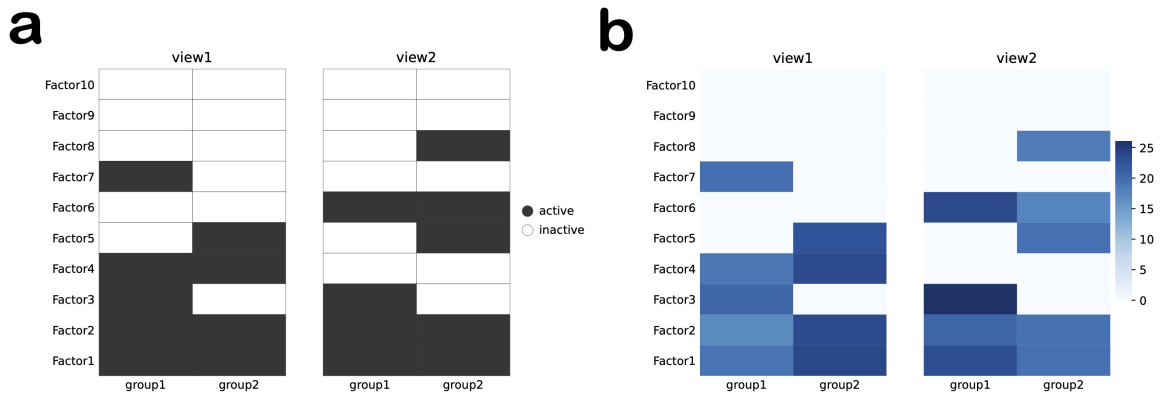

Fig. 3.3 MOFA+ recovers group-specific factor activity is simulated data
a: Representation of binary activity of factors in the simulated data with 2 views, 2 groups and 8 factors with differential activity. First column of each heatmap corresponds to the first group of cells and the second one corresponds to the second group of cells. Heatmaps correspond to respective views. **b:** Variance explained by each of the 10 factors in respective groups and views. Each heatmap corresponds to a respective view, with the first and the second columns corresponding to the first and the second group of cells, respectively. The colour denotes the percentage of variance explained by a given factor in a given view and group.

3.2.1.2 MOFA+ scalability

MOFA+ also provides two major technical advances that focus on performance and scalability as a result of work performed by Damien Arnol and Ricard Argelaguet. First, MOFA+ implementation comes with accelerated training on graphics processing units (GPU). Second, it comes with stochastic variational inference (SVI) that enables scalability to datasets with

the number of cells significantly larger than the number of features (M. D. Hoffman et al., 2013). The SVI framework uses a subset of the data (a batch of pre-defined size) to calculate an approximation of the ELBO gradient. The step size $\rho(t)$ at each iteration t is adjusted according to the hyperparameters of learning rate τ and forgetting rate κ:

$$\rho^{(t)} = \frac{\tau}{(1+\kappa t)^{3/4}}.$$

3.2.2 Model definition

MOFA+ can be defined as a following latent variable model:

$$\mathbf{Y}^{gm} = \mathbf{W}^{mT}\mathbf{Z}^{g} + \varepsilon^{gm},$$

where $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ is a weight matrix relating the original features in the view m to the latent representation and $\mathbf{Z}^g \in \mathbb{R}^{N^g \times K}$ is a factor matrix. The noise term ε^{gm} contains the unexplained variance for each feature in each group, with residuals assumed to be normally distributed and heteroskedastic:

$$p(\varepsilon_d^{gm}) = \mathcal{N}(\varepsilon_d^m \mid 0, 1/\tau_d^m).$$

Here and below, the following notation is used:

- k for a factor, with K factors in total,
- g for a group of cells, with G groups in total,
- n for a cell, with N_g cells in the group g,
- m for a view (modality), with M views in total,
- d for a feature, with D_m features in the view m.

Formulating the model in a Bayesian fashion, we introduce prior distributions on the unobserved variables. The two-level regularisation used on the prior distribution of the weights is similar to the MOFA model and can be expressed as a combination of ARD prior and spike-and-slab prior reparametrised as a product of a Gaussian and a Bernoulli random variables:

$$p(\hat{w}_{kd}^m, s_{kd}^m) = \mathcal{N}(\hat{w}_{kd}^m \mid 0, 1/\alpha_k^m)\,\mathrm{Ber}(s_{kd}^m \mid \theta_k^m).$$

Finally, hierarchical priors for θ and α are the following uninformative priors:

$$p(\theta_k^m) = \mathscr{B}(\theta_k^m \mid a_0^\theta, b_0^\theta)$$
$$p(\alpha_k^m) = \mathscr{G}(\alpha_k^m \mid a_0^\alpha, b_0^\alpha).$$

In MOFA+, the sparsity priors are also defined on factors, which makes the model symmetrical and accommodates the group structure:

$$p(\hat{z}_{nk}^g, s_{nk}^g) = \mathscr{N}(\hat{z}_{nk}^g \mid 0, 1/\alpha_k^g)\,\mathrm{Ber}(s_{nk}^g \mid \theta_k^g)$$
$$p(\theta_k^g) = \mathscr{B}(\theta_k^g \mid a_0^\theta, b_0^\theta)$$
$$p(\alpha_k^g) = \mathscr{B}(\alpha_k^g \mid a_0^\alpha, b_0^\alpha).$$

In addition to that, the noise level is also not just feature- and view- but also group-dependent:

$$p(\varepsilon_d^{m,g}) = \mathscr{N}(\varepsilon_d^{m,g} \mid 0, 1/\tau_d^{m,g})$$
$$p(\tau_d^{m,g}) = \mathscr{G}(\tau_d^{m,g} \mid a_0^\tau, b_0^\tau).$$

These modifications make the graphical model for MOFA+ symmetrical (Figure 3.4).

3.2.3 Inference

3.2.3.1 Update equations

Following the model definition, the update equations that are applied at every iteration of the variational inference algorithm match the ones for the original version of MOFA. Making the model symmetrical, I have derived update equations for factors (group sparsity). These equations are listed below while the equations for all the parts of the model are listed in Appendix A.1.

Sparse factors Prior distribution:

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

Variational distribution:

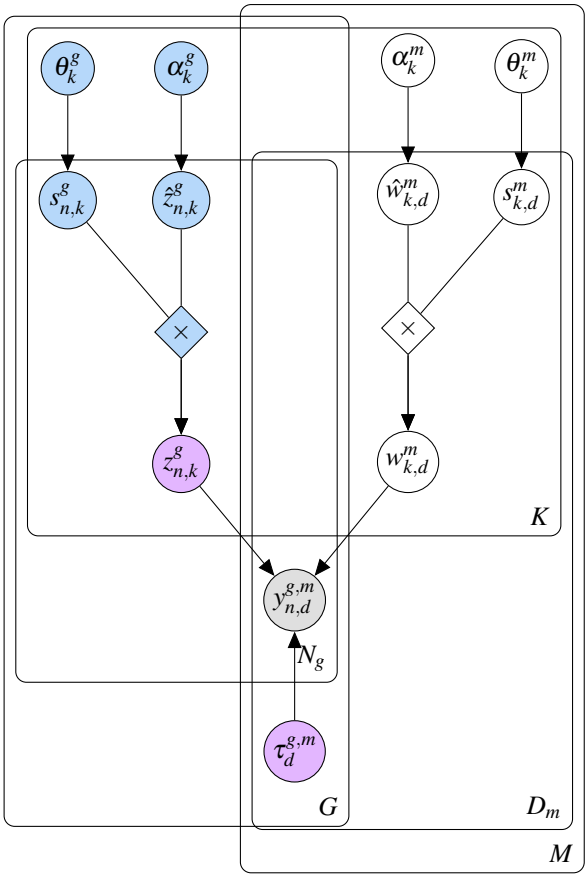

Fig. 3.4 MOFA+ graphical model
The grey circles represent observed variables. The white and coloured circles represent hidden variables inferred by the model. Elements added in MOFA+ are shown in light blue. Elements changed in MOFA+ compared to MOFA are shown in light purple. The plates represent dimensions of the model: M views, G groups, K factors, D_m features in view m and N_g samples in group g. Adopted from Argelaguet, Arnol, et al., 2020.

$$q(s_{nk}^g) = \text{Ber}(s_{nk}^g|\gamma_{nk}^g)$$

$$q(\hat{z}_{nk}^g|s_{nk}^g = 0) = \mathcal{N}\left(\hat{z}_{nk}^g\middle|0, 1/\alpha_k^g\right)$$

$$q(\hat{z}_{nk}^g|s_{nk}^g = 1) = \mathcal{N}\left(\hat{z}_{nk}^g\middle|\mu_{z_{nk}^g}, \sigma_{z_{nk}^g}^2\right),$$

where

$$\gamma_{nk}^g = \frac{1}{1+\exp(-\lambda_{nk}^g)}$$

$$\lambda_{nk}^g = \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle} - 0.5 \ln \left(\sum_{m=1}^M \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle} \right)$$

$$+ \frac{\langle \tau_d^{gm} \rangle}{2} \frac{\left(\sum_{m=1}^M \sum_{d=1}^{D_m} y_{nd}^{gm} \langle w_{kd}^m \rangle - \sum_{j \neq k} \langle s_{nj}^g \hat{z}_{nj}^g \rangle \sum_{m=1}^M \sum_{d=1}^{D_m} \langle w_{kd}^m \rangle \langle w_{jd}^m \rangle \right)^2}{\sum_{m=1}^M \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

$$\mu_{z_{nk}^g} = \frac{\sum_{m=1}^M \sum_{d=1}^{D_m} y_{nd}^{m,g} \langle w_{kd}^m \rangle - \sum_{j \neq k} \langle s_{nj}^g \hat{z}_{nj}^g \rangle \sum_{m=1}^M \sum_{d=1}^{D_m} \langle w_{kd}^m \rangle \langle w_{jd}^m \rangle}{\sum_{m=1}^M \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

$$\sigma_{z_{nk}^g}^2 = \frac{\langle \tau_d^{gm} \rangle^{-1}}{\sum_{m=1}^M \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

ARD for the factors Prior distribution:

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

Variational distribution:

$$q(\alpha_k^g) = \mathcal{G}(\alpha_k^g \mid \hat{a}_{gk}^\alpha, \hat{b}_{gk}^\alpha),$$

where

$$\hat{a}^{\alpha}_{gk} = a^{\alpha}_{0} + \frac{N_g}{2}$$

$$\hat{b}^{\alpha}_{gk} = b^{\alpha}_{0} + \frac{\sum_{n=1}^{N_g} \langle (z^g_{nk})^2 \rangle}{2}$$

Factor sparsity priors Prior distribution:

$$p(\theta^g_k) = \mathscr{B}(\theta^g_k \mid a^{\theta}_0, b^{\theta}_0)$$

Variational distribution:

$$q(\theta^g_k) = \mathscr{B}(\theta^g_k \mid \hat{a}^{\theta}_{gk}, \hat{b}^{\theta}_{gk}),$$

where

$$\hat{a}^{\theta}_{gk} = \sum_{n=1}^{N_g} \langle s^g_{nk} \rangle + a^{\theta}_0$$

$$\hat{b}^{\theta}_{gk} = b^{\theta}_0 - \sum_{n=1}^{N_g} \langle s^g_{nk} \rangle + N_g$$

3.2.3.2 Evidence Lower Bound

Evidence lower bound (ELBO) is used to track the algorithm convergence. ELBO can be decomposed into a sum of the expected log-likelihood and the Kullback-Leibler (KL) divergence between the prior and the variational distributions:

$$\mathscr{L}(X) = \mathbb{E}_q \log p(Y|X) - KL(q(X)||p(X \mid Y))$$

Log-likelihood term for Gaussian likelihood can be written as

$$\mathbb{E}_q[\ln P(Y|X)] = -\sum_{m=1}^{M} \frac{ND_m}{2} \ln(2\pi) + \sum_{g=1}^{G} \frac{N_g}{2} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \ln(\tau_d^{gm}) \rangle$$

$$- \sum_{g=1}^{G} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \frac{\langle \tau_d^{gm} \rangle}{2} \sum_{n=1}^{N_g} \left(y_{nd}^{m,g} - \sum_{k=1}^{K} \langle s_{kd}^m \hat{w}_{kd}^m \rangle \langle z_{nk}^g \rangle \right)^2$$

KL divergence between the true posterior distribution and an approximate posterior distribution can be expressed as

$$KL(q(X) \,||\, p(X \mid Y)) = -\int_x q(X) \ln \frac{p(X \mid Y)}{q(X)} = \mathbb{E}_q[\ln q(X)] - \mathbb{E}_q[\ln p(X \mid Y)]$$

Below are the analytical forms of these expectations for factors while the equations for all the parts of the model are listed in Appendix A.2

Sparse factors:

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

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

ARD for the factors:

$$\mathrm{E}_q[\ln p(\alpha)] = \sum_{g=1}^{G} \sum_{k=1}^{K} \left(a_0^\alpha \ln b_0^\alpha + \ (a_0^\alpha - 1) \langle \ln \alpha_k \rangle - b_0^\alpha \langle \alpha_k \rangle - \ln \Gamma(a_0^\alpha) \right)$$

$$\mathrm{E}_q[\ln q(\alpha)] = \sum_{g=1}^{G} \sum_{k=1}^{K} \left(\hat{a}_k^\alpha \ln \hat{b}_k^\alpha + (\hat{a}_k^\alpha - 1) \langle \ln \alpha_k \rangle - \hat{b}_k^\alpha \langle \alpha_k \rangle - \ln \Gamma(\hat{a}_k^\alpha) \right)$$

Factor sparsity priors:

$$\mathbb{E}_q\left[\ln p(\theta)\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\left((a_0-1)\times\langle\ln(\pi_{nk}^g)\rangle + (b_0-1)\langle\ln(1-\pi_{nk}^g)\rangle - \ln(\mathrm{B}(a_0,b_0))\right)$$

$$\mathbb{E}_q\left[\ln q(\theta)\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\sum_{n=1}^{N_g}\left((a_{kn}^g-1)\times\langle\ln(\pi_{nk}^g)\rangle + (b_{kn}^g-1)\langle\ln(1-\pi_{nk}^g)\rangle - \ln(\mathrm{B}(a_{kn}^g,b_{kn}^g))\right)$$

3.2.4 Model interpretation

3.2.4.1 Interpretation of the factors

Factors capture the sources of variability in the data and are analogous to the principal components in PCA. Each factor defines the location of cells along a one-dimensional axis. Higher absolute factor values correspond to a stronger effect. Different signs of factor values are expected to correspond to opposite properties along the inferred axis of variation, e.g. cell's phenotype such as cell cycle stage. Because of the zero-mean prior, cells with intermediate phenotype are expected to be located around zero along this axis (Figure 3.2b).

Strength of factor k in group g is controlled with ARD, with small posterior α_k^g values corresponding to active factors. Large α_k^g, in contrast, correspond to smaller variance around 0 and inactive factors. The fraction of non-zero factor values in a factor k is controlled by θ_k^g: as it is a parameter of the Bernoulli distribution, its posterior values close to 0 imply the shrinkage of most factor k values to 0 (sparse factor), while for dense factors, posterior θ_k^g values are close to 1.

3.2.4.2 Interpretation of the weights

The weights provide scores for feature importance for each factor. Features that are associated with the factor stronger have higher absolute weight values, and the sign corresponds to the direction of effect. For instance, positive weights for a gene means that this gene has higher expression in cells with positive factor values. When weights are compared between views,

they are to be scaled to $[-1,1]$ in order to account for different data distributions in different modalities (Figure 3.2b).

For ARD, large posterior α_k^m values correspond to low variance around 0 and and inactivity of factor k in view m. Conversely, small α_k^m implies an active factor. Posterior values of the Bernoulli parameter θ_k^m control the fraction of active weights in factor k, with values close to 0 or 1 indicating sparser or denser factors, respectively.

3.2.4.3 Variance decomposition

Variance explained factor k in each group g and view m can be calculated for a trained model as a coefficient of determination:

$$R^2_{gmk} = 1 - \frac{\left(\sum_{n=1}^{N_g}\sum_{d=1}^{D_m}(Y_{gm} - W_m Z_g)\right)^2}{\left(\sum_{n=1}^{N_g}\sum_{d=1}^{D_m} Y_{gm}\right)}.$$

Respective R^2 values can then be visualised in order to assess the activity of factors across groups of cells and views (Figure 3.3b, 3.2b).

3.2.5 Implementation

3.2.5.1 R and Python packages

The model described above has been implemented in Python[3] and has also been made accessible from R using the corresponding interface between the languages (Ushey et al., 2022).

The Python core has been built using the numerical stack including NumPy (Harris et al., 2020), Pandas (McKinney et al., 2011) and h5py (Collette, 2013). As part of this work, I have also implemented an interface for MOFA+ to be readily applied on data stored in AnnData (Virshup et al., 2021) as well as MuData (Bredikhin et al., 2022) objects and an interface to make MOFA+ models readily trained on Seurat objects (Satija et al., 2015).

[3] https://pypi.org/project/mofapy2/

3.2.5.2 Model storage

Practically, a MOFA+ factor model is trained on a dataset — a collection of matrices with feature values in views (modalities) and cell groups. The trained MOFA model is stored in an HDF5 file. The main result of the training are weight and factor matrices. As a file with the model frequently serves as an input for the downstream model interpretation, cell and feature names in corresponding groups and views are also stored in the model, alongside the original data and training parameters and statistics. To extend this, I implemented the storage of additional information for cells and features, which enables richer model interpretation, for instance by correlating factor values with additional cell covariates. Moreover, I generalised expectations storage so that not only $\mathbb{E}(\mathbf{Z})$ and $\mathbb{E}(\mathbf{W})$ but also expectations of all the other parameters of the model can be stored after the training.

3.2.5.3 Interactive model interrogation

The primary way to explore MOFA models has been through the MOFA R package[4], which is also distributed through the Bioconductor ecosystem (Huber et al., 2015). I improved this package in multiple ways so that MOFA2 R package allows to load, interrogate and visualise MOFA+ models with the group structure for cells. Leveraging this package, I then implemented web dashboards to allow interactive model interrogation for the user-provided models using Shiny[5] (Figure 3.5).

In order to integrate model interpretation into the Python ecosystem, I also implemented functionality[6] to read and inspect MOFA models in Python as well as generate plots for model interpretation.

[4] https://www.bioconductor.org/packages/release/bioc/html/MOFA2.html
[5] http://www.ebi.ac.uk/shiny/mofa/
[6] https://pypi.org/project/mofax/

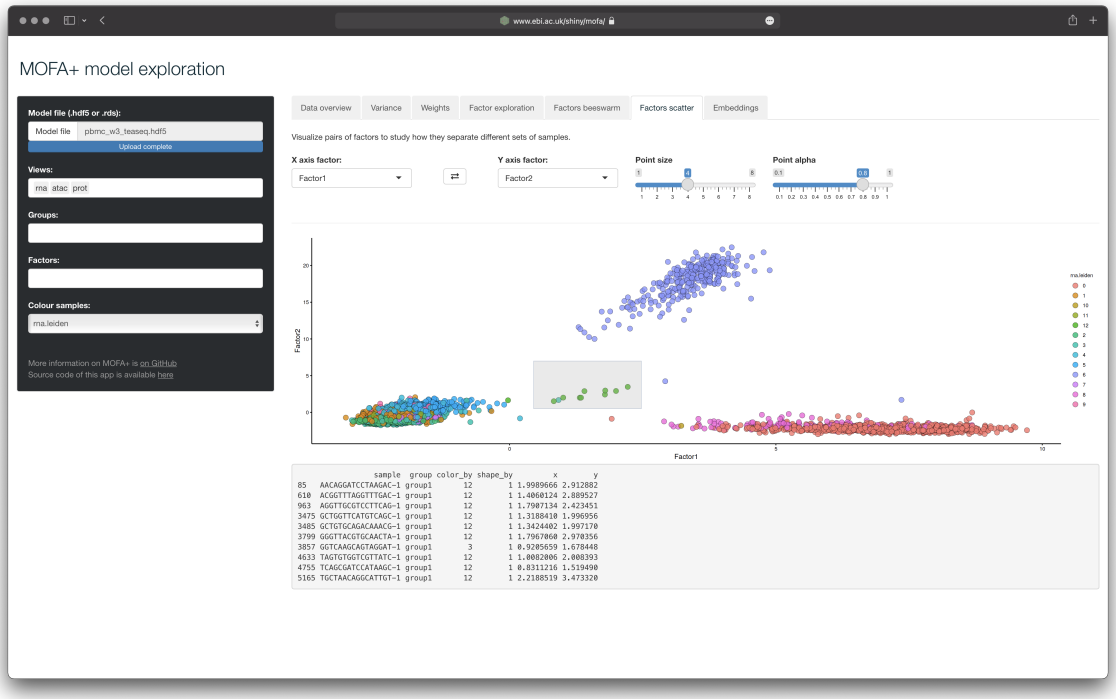

Fig. 3.5 Demonstration of interactive MOFA+ model exploration
The interface of the interactive model exploration includes (on the left) model file input fields as well as fields to select views, groups and factors of interest and a covariate to use for colouring cells on the corresponding plots. Here, cells are coloured by the cluster identity computed on the RNA modality. The major part of the interface is dedicated to plots with additional options to choose parameters and variables to display. A table with data including cell barcodes is displayed below the plot upon cell selection.

3.3 Discussion

3.3.1 Multimodal integration of single-cell data

Multimodal single-cell datasets have the potential to bring new biological insights, but they also present significant challenges in how the joint analysis of multiple modalities should be conducted and interpreted. New methods are required to address them, and as the size and the diversity of modalities in these datasets increases, such methods should be scalable and generalisable to combinations of various modalities.

Joint analysis of multimodal single-cell data can be approached from different directions. One of them described above formulates the integration problem as a latent variable model such that latent factors can act on multiple modalities. For instance, MOFA+ is a group factor analysis model that allows to infer a set of multimodal latent factors. The factors are defined by weight matrices, one per modality, which allow to describe the correspondence between factors and individual features such as genes, proteins or genomic loci, highlighting the interpretability of such models. Moreover, inferred latent factors can be conveniently used as input for other analytical methods such as computing cell neighbours and cell clustering.

Other approaches rely on latent factors computed for individual modalities and perform integration at the level of cell neighbours. For example, the weighted nearest neighbours (Hao et al., 2021) are computed based on the principal components for each modality. Cell neighbours can also be computed for each modality and then be fused together with methods like similarity network fusion (Bo Wang et al., 2014).

3.3.2 Challenges and perspectives

Integrating independent datasets across experiments and samples has been one of the major challenges in single-cell genomics (Lähnemann et al., 2020). Availability of multiple molecular layers adds a new dimension of complexity to it. Some methods such as NMF-based LIGER (Welch et al., 2019) and CCA-based Seurat (v3) (Stuart, Butler, et al., 2019) have been designed to integrate independent populations of cells by constructing a shared feature space (genes). MOFA+ is designed to integrate data with a shared sample space, i.e. with the same cells, while the features may come from different molecular layers. Using the terminology from Argelaguet, Cuomo, et al., 2021, this can be described as *vertical* integration.

Complex experimental designs that consist of unimodal and multimodal measurements motivate the development of mosaic integration strategies, which would combine *vertical* and *horizontal* integration approaches (Argelaguet, Cuomo, et al., 2021). Moreover, such methods would potentially be able to account for known dependencies in the data both in the cell and feature dimensions. Below, I outline some recent progress in these directions.

3.3.3 New models

3.3.3.1 MEFISTO

The increasing availability of omics datasets with spatial or temporal resolution (section **??**) motivates new computational methods to analyse them. In fact, factor analysis framework described in section 3.2 can be extended to infer smooth patterns of variation (Velten et al., 2022). Such a model named MEFISTO provides a toolbox for dimensionality reduction that also accounts for spatio-temporal dependencies in the data. To achieve this, MEFISTO combines MOFA+ framework with Gaussian processes by modelling each factor value as a realisation of a Gaussian process.

3.3.3.2 Encoding prior knowledge

Features such as genes can be viewed as parts of regulatory networks with complex interactions between them. While current models like MOFA+ assume independence between features, which is encoded in its prior distributions, more elaborate feature covariance structures can be incorporated to account for feature interactions. That would include interactions between features from different modalities such as promoters with their accessibility measured, respective genes with their expression profiled and respective proteins with their abundance quantified.

Prior domain information has been incorporated in factor analysis-based models (Qoku and Buettner, 2022) as well as in autoencoder-based frameworks in the form of curated gene sets (Lotfollahi et al., 2022).

3.3.4 Automatic differentiation variational inference

One of the large advances that arguably enabled the development of many of the aforementioned methods has been automatic differentiation variational inference (Kucukelbir et al., 2016). It greatly simplified probabilistic modelling by automatically deriving, as the name

suggests, a variational inference algorithm for the provided model and data. Importantly, it has been implemented and interfaced in commonly used numerical ecosystems including TensorFlow (Abadi et al., 2015), PyTorch (Paszke et al., 2019) and JAX (Bradbury et al., 2018).

New inferences schemes has also been recently proposed such as linear response variational inference for accurate covariate estimates from mean-field variational Bayes (Giordano et al., 2015) and Stein variational inference (Q. Liu and D. Wang, 2016).

3.3.5 Variational autoencoders

With explosive progress in deep neural networks, deep learning models have also found applications for computational biology (Angermueller, Pärnamaa, et al., 2016; Lopez et al., 2018). Deep neural networks provide a lot of benefits such as ability to learn directly on DNA sequences instead of pre-defining features based on prior knowledge, such as presence of variants, k-mer frequences, or conservation scores; ability to capture nonlinear dependencies and interaction effects, sequence context at multiple genomic scales (Avsec et al., 2021). However these models are still difficult to train, with overfitting as one of the key challenges: the model is too complex relative to the size of the training data. The benefits of such models often come at the cost of their limited interpretability, particularly when compared to linear methods such as PCA.

While PCA is constrained to a linear transformation, variational autoencodes (VAEs) can be viewed as a generalisation of the PCA approach (Rolinek et al., 2019) however allow to use non-linear functions. VAEs (Kingma and Welling, 2014) combined the ideas behind autoencoders and variational Bayes, were extended by beta-VAEs (Higgins et al., 2016) and later by VQ-VAEs (Razavi et al., 2019; van den Oord et al., 2017) and achieved state-of-the-art performance across different domains. The latest progress in VAEs for single-cell data has been largely incorporated in a framework for single-cell variational inference (Lopez et al., 2018).

Deep neural networks-based approaches such as single-cell variational inference (scVI) have recently made probabilistic models for single-cell genomics more accessible to a larger community of researchers (Gayoso, Lopez, et al., 2022; Lopez et al., 2018). Notably, such frameworks seem to provide necessary flexibility to implement strategies for integrating various modalities (Ashuach et al., 2021; Gayoso, Steier, et al., 2021) as well as to implement new methods on top of them. For instance, linearly decoded VAE replaces the neural network

that maps the latent dimensions back to the original space with a linear function providing greater interpretability (Svensson, Gayoso, et al., 2020).

Chapter 4

MOFA+ applications

MOFA+ has proved to be a useful matrix factorisation method to produce lower-dimensional representations of multi-omics dataset as shown in Chapters 2 and 3. In this chapter, I will reference a few applications of MOFA and will show how single-cell datasets can be visualised and interpreted with MOFA+. Such latent representations can also guide downstream analysis and further experiments as I will demonstrate in this chapter.

4.1 Introduction

The problems that have been addressed by multi-omics factor analysis can illustrate the utility of MOFA+ as a versatile method that can be applied to different questions. As a generic method for unsupervised data integration, multi-omics factor analysis can be used to gain insights in different biological contexts. MOFA has enabled identification of clinical markers of diseases in datasets with complex experimental designs such as multiple modalities combining continuous and discrete variables and missing data with samples profiled with some but not all the modalities. For instance, the original MOFA model was used to integrate and characterise multi-omics profiles in an application to chronic lymphocytic leukaemia (CLL) leveraging information across gene expression, methylation, mutation profiles as well as drug response data as described in the original publication (Argelaguet, Velten, et al., 2018). This has been followed by using MOFA to uncover an axis of heterogeneity associated with the disease outcome in CLL (Lu et al., 2021) as well as to show an enrichment of suppressive immune cells in IL4I1-high tumours (Sadik et al., 2020).

Besides disease markers in cancer, MOFA enabled integration of multimodal biomedical data such as the studies of SARS-CoV-2 infection (Rodriguez et al., 2020), cerebrospinal fluid to identify alterations in Alzheimer's disease (C. Clark et al., 2021), sepsis (Kwok et al., 2022), rheumatoid arthritis mouse model (L. Li et al., 2022) as well as of the microbiome co-variation patterns (Pattaroni et al., 2022).

The symmetric structure of MOFA+ and its improved scalability have enabled integration of single-cell multimodal omics datasets, which typically feature group structure on the cells. For instance, its utility has been demonstrated for learning factors based on gene expression, DNA methylation and chromatin accessibility in human oocytes and ovarian somatic cells (R. Yan et al., 2021). Most recently, MOFA+ has been used to integrate gene expression and chromatin accessibility in mouse embryos across multiple time points (Argelaguet, Lohoff, et al., 2022) extending MOFA's previous applications to studying mouse gastrulation (Argelaguet, S. J. Clark, et al., 2019). MOFA+ was also used to combine proteomic measurements with molecular and phenotypic datasets in a pan-cancer study of almost a thousand of human cell lines (Gonçalves et al., 2022).

4.2 MOFA+ model for CRISPRa screens

This section describes how MOFA+ can be particularly suited for a high-throughput screen analysis to quantify transcriptome-wide response to gene overexpression. In particular, I used MOFA+ to integrate expression of coding and non-coding transcriptomic changes in individual cells across hundreds of groups leveraging the MOFA+ advances described in Chapter 3. This allowed to identify new positive regulators of zygotic genome activation — a major developmental program.

The contents of this section represent the analysis that was performed by me and has been published as part of the publication *A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program* (Alda-Catalinas et al., 2020). The experiments described in this section were lead by Celia Alda-Catalinas while I performed all the described analyses and created the figures unless stated otherwise.

4.2.1 Primer on zygotic genome activation

Zygotic genome activation (ZGA) describes transcriptional changes that happen in the early embryogenesis. Initially, the development of embryos is driven by maternal gene products — RNAs and proteins, — however during the maternal-to-zygotic transition, the genome of the zygote (one-cell stage embryo) is being activated to produce functional transcripts (Vastenhouw et al., 2019). As part of this process, maternal RNAs undergo clearance. The timing of ZGA varies across species, and in mice, it starts with a minor wave of activation in the late zygote and then continues with the major wave in mid-two-cell and late-two-cell embryos (Melanie A. Eckersley-Maslin et al., 2018).

During the two waves of ZGA, global transcriptional and chromatin landscape changes occur. Chromatin remodellers seem to play a role at regulating these changes, together with transcription factors, histone modifiers and non-coding transcripts (Melanie A. Eckersley-Maslin et al., 2018). Notably, early two-cell embryos feature accessible chromatin in regions of repeat elements in the genome such as MERVL (mouse endogenous retrovirus with leucine tRNA primer) repeats. For instance, DUX (double homeobox), one of the ZGA-linked transcription factors, activates cleavage stage-specific genes as well as MERVL retrotransposons (Hendrickson et al., 2017). *Zscan4* (zinc finger and SCAN domain-containing 4) has been identified as a gene cluster specific to late two-cell stage embryos and mouse embryonic stem cells (Falco et al., 2007; M. S. H. Ko, 2016). Together with MERVL, it can be thus

considered as a marker of the transcriptional landscape of preimplantation embryos at the two-cell stage (Mélanie A. Eckersley-Maslin et al., 2016).

Notably, activation of the MERVL/Zscan4 network also occurs in mouse embryonic stem cells (mESCs), which cycle in and out this 2C-like state. Such cells resemble two-cell stage embryos both transcriptionally and epigenetically with global DNA demethylation and chromatin decondensation (Mélanie A. Eckersley-Maslin et al., 2016). Using mESCs as a model for studying ZGA enabled the identification of *Dppa2* (developmental pluripotency-associated 2) and *Dppa4* as positive regulators of ZGA genes transcription. Moreover, these genes were also shown to directly regulate *Dux* in order to initiate the 2C-like transcriptional signature while *Zscan4* reinforces this activation (De Iaco et al., 2019; M. Eckersley-Maslin et al., 2019).

4.2.2 Primer on CRISPR activation screening

Combination of single-cell RNA sequencing described in section 1.1.1.2 and clustered regularly interspaced short palindromic repeats (CRISPR)-based perturbations enables interrogation of gene regulation at scale (Datlinger et al., 2017; Dixit et al., 2016). In CRISPR screens, targeting a DNA endonuclease Cad9 to a specific genomic locus is achieved with single guide RNAs (sgRNAs) that contain a $5'$-terminal 20 nucleotides-long spacer sequence complementary to the DNA locus of interest (Jinek et al., 2012). Such screens can focus on loss-of-function perturbations through gene knockout (Datlinger et al., 2017; Dixit et al., 2016) or interference (CRISPRi) (Adamson et al., 2016; Gilbert et al., 2013). In contrast, CRISPR activation (CRISPRa) can be employed to selectively induce gene expression upregulation. This is achieved by using an engineered CRISPR-Cas9 complex where a catalytically inactive Cas9 (dCas9) protein fused to transcription activation domains is used as an RNA-guided transcription activator (Konermann et al., 2015).

One of the potent CRISPRa systems includes synergistic activation mediator (SAM) comprised of multiple components (Konermann et al., 2015). As part of SAM, dCas9-VP64, dCas9 fusion with a VP64 transcriptional activation domain (Maeder et al., 2013), is complemented with p65 and heat shock factor 1 (HSF1) that recruit distinct subsets of transcription factors and chromatin remodellers thus increasing the potential transcriptional activation. Both p65 and HSF1 are part of the MS2–p65–HSF1 fusion protein, in which the bacteriophage MS2 coat protein provides an RNA-binding function (Peabody, 1993). Single

guide RNAs (sgRNAs) in SAM are modified to include hairpin aptamers recruiting MS2 fusion.

While the use of preimplantation mouse embryos in high-throughput screens is limited due to the scarcity of raw material and experimental complexity, mESCs were shown to exhibit a ZGA-like state and provide a system for scalable *in vitro* screening (Mélanie A. Eckersley-Maslin et al., 2016). These cells were previously used to identify and to study ZGA regulators (M. Eckersley-Maslin et al., 2019; Mélanie A. Eckersley-Maslin et al., 2016). Combining CRISPRa with scRNA-seq in mESCs thus provides a setup to interrogate inducers and regulators of ZGA in a high-throughput manner.

A more detailed review of ZGA and its regulation has been provided in M. S. H. Ko, 2016 and Melanie A. Eckersley-Maslin et al., 2018.

4.2.3 CRISPRa screen for the regulators of zygotic genome activation

In order to systematically interrogate potential regulators of ZGA, a high-throughput screening method was developed combining CRISPRa with scRNA-seq. For the transcriptional activation, this method builds on the SAM CRISPRa described above with the screening performed in mESCs constitutively expressing dCas9-VP64 and MS2-p65-HSF1. The transcriptome of these mESCs is largely unchanged when compared to the parental ESC line E14 (Figure 4.2.3). A ZGA-like transcriptional response potentially induced by the upregulation of expression of ZGA regulators candidates is then captured with scRNA-seq.

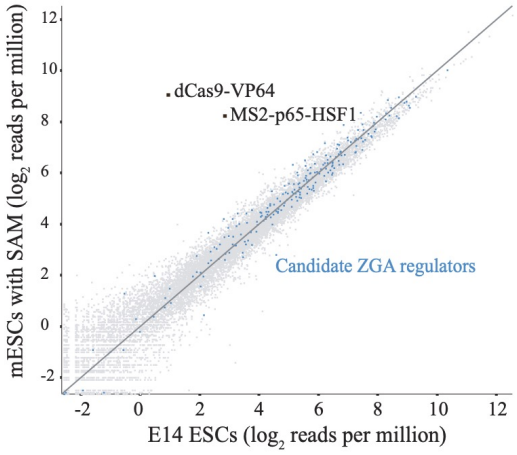

Fig. 4.1 Gene expression comparison for mESCs with SAM and parental line E14

Scatterplot shows normalised expression values of mESCs with SAM (constitutively expressing dCas9-VP64 and MS2-p65-HSF1) against their parental cell line E14. Highlighted are dCas9-VP64 and MS2-p65-HSF1 transcripts (in black) and screening candidates (in blue).

4.2.3.1 Pilot screen

To validate this screening system and optimise the approach, a pilot experiment was conducted with sgRNAs designed to target ZGA-linked MERVL long terminal repeats (LTRs) or *Zscan4* promoters (Table B.1). Both targets show the upregulation of their transcription upon CRISPR activation, which can be quantified using the proportion of cells expressing MERVL or *Zscan4* when compared to mESCs cells with a non-targeting sgRNA control (Figure 4.2). This also provides evidence for synergistic regulation in the MERVL/Zscan4 network (Mélanie A. Eckersley-Maslin et al., 2016) as *Zscan4* activation led to MERVL upregulation and vice versa. In order to assess global transcriptomic response to MERVL or *Zscan4* upregulation, I computed aggregated expression of 2115 genes (see Appendix B.2) described to be expressed during ZGA or in ZGA-like mESCs (Mélanie A. Eckersley-Maslin et al., 2016; Hendrickson et al., 2017). Such signature has been upregulated in cells with MERVL or *Zscan4* sgRNAs (Figure 4.2c). Overall, these results demonstrate that scRNA-seq can be used as a readout of ZGA-like transcriptional changes following CRISPR activation of relevant regulators.

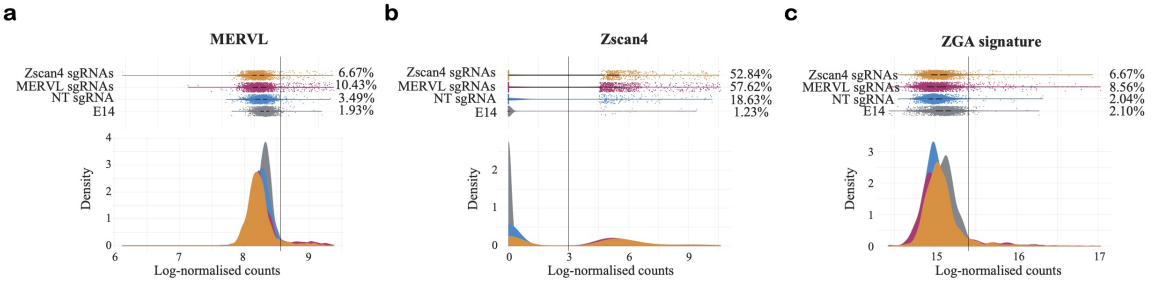

Fig. 4.2 Gene expression upon activation in the pilot screen
a: MERVL repeats expression in untransduced E14 ESCs (in grey) and SAM ESCs transduced with a non-targeting sgRNA (in blue), with sgRNAs targeting MERVL repeats (pink) or Zscan4 gene promoters (orange). Upper panel displays log-normalised counts in individual cells, lower panel provides density representation as a summary. Reported percentages were calculated for the cells with larger expression values than the denoted percentile in E14 ESCs (2%) marked the vertical black line with 100% corresponding to the total number of cells in each sample. Using this metric, MERVL activation resulted in a 3-fold increase of the number of cells expressing it from 3.49% to 10.43%. **b:** Same as **a** but for aggregated *Zscan4b/c/d/e/f* gene expression. *Zscan4* activation led to a 2.8-fold increase of the number of cells expressing it from 18.63% to 52.84%. **c:** Same as **a** but for aggregated ZGA genes expression. The number of cells with high ZGA gene expression increased 4.28-fold and 3.34-fold upon MERVL and *Zscan4* activation, respectively.

4.2.3.2 Main screen

Next, an extensive list of potential ZGA regulators was devised from a collection of proteins with nucleic acid binding and transcription factors activities detected in mouse oocytes and zygotes (see Section 4.2.5.1 below). As a result, pooled sgRNA library contained two sgRNAs per each of the 230 candidate regulators targeting the 180 nucleotide window upstream of these genes' transcriptional start sites in addition to 15 non-targeting sgRNA controls (Table B.1).

These 475 sgRNAs cloned into a lentiviral vector akin CRISPR droplet sequencing (CROP-seq) (Datlinger et al., 2017) were used to transduce mESCs at a <0.1 multiplicity of infection within three transduction replicates.

Transcriptome profiles of 341103 individual cells were obtained with scRNA-seq. I performed scRNA-seq quality control of the data (Figure 4.3a-c) and assigned sgRNAs to individual cells (see section 4.2.5.5 below). This resulted in 203894 cells that express unique sgRNAs (Figure 4.3d). In the dataset combined across three transduction replicates, each sgRNA is present on average in 437 cells, with sgRNAs captured consistently across replicates (Figure 4.3e).

I employed principal component analysis (PCA, see Section 3.1.1.1) to investigate the main sources of variation in the data. When components are sorted by the proportion of variance they explain, the second component is notably driven by the genes highly expressed in mid-to-late two-cell embryos during ZGA such as *Zscan4*, *Zscan4d*, *Gm8300* and *Tmem92* (Figure 4.4a-b). This is consistent between replicates highlighting the robustness of the screen (Figure 4.4c). Overall, this shows that a substantial proportion of cells exhibit ZGA-like transcriptional changes in response to activation.

4.2.3.3 Identification of zygotic genome activation-like signature with MOFA+

To characterise the ZGA-like transcriptional response with greater detail, I considered transposable or repeat elements expression in individual cells as an additional source of information (4.2.5.4) (Figure 4.5b-c). I then used MOFA+ (see Chapter 3) to combine the signal from the expression of coding genes and of repeat elements in a single model. I treated gene and repeat element expression as distinct views to disentangle sgRNA-specific activation of ZGA-like response while also accounting for different groups of cells defined by the sgRNA they bear (Figure 4.5a). With this joint model, I identified sources of variation (latent factors) the explain transcriptional variability across cells in both views. Having sorted

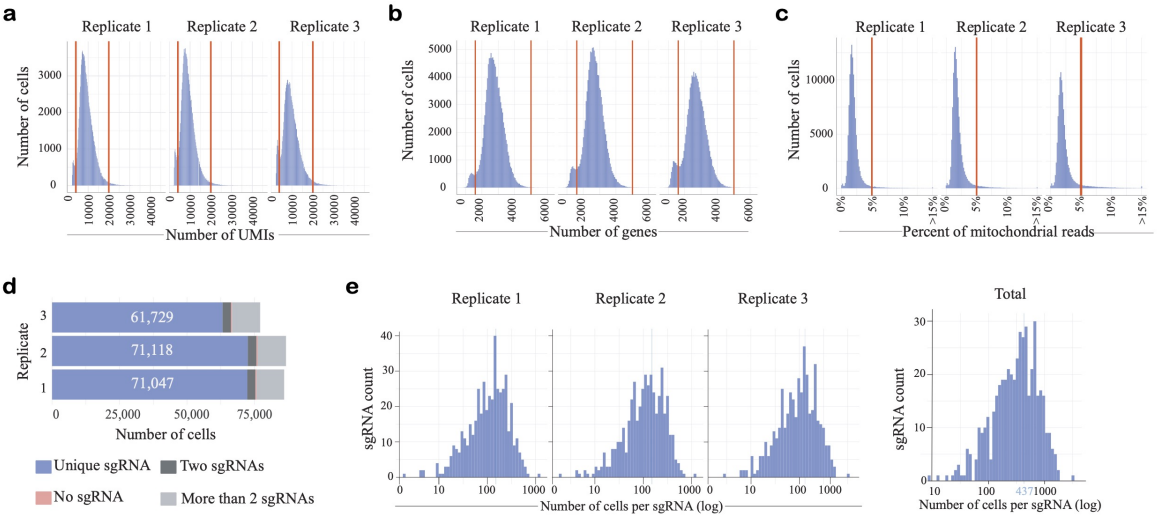

Fig. 4.3 Quality control for scRNA-seq and sgRNA assignment
a: Number of unique molecular identifiers (UMIs) for each transduction replicate. Cells with >4,000 UMIs and <20,000 UMIs were retained. **b:** Same as **a** for the number of detected genes. Cells with >1,600 and <5,000 genes were retained **c:** Same as **a** for the percentage of counts for mitochondrial genes. Cells with <5**d:** Number of cells with a unique sgRNA assigned (in blue), with two sgRNAs assigned (dark gray), with more than two sgRNAs assigned (in light gray), or no sgRNA assigned (in pink) in each of the three transduction replicates. For each replicate, the number of cells with a unique sgRNA assigned is shown. **e:** Number of cells that a corresponding sgRNA is uniquely assigned to for the 475 sgRNAs across three transduction replicates and in total. On average, there are 437 cells per sgRNA in the dataset.

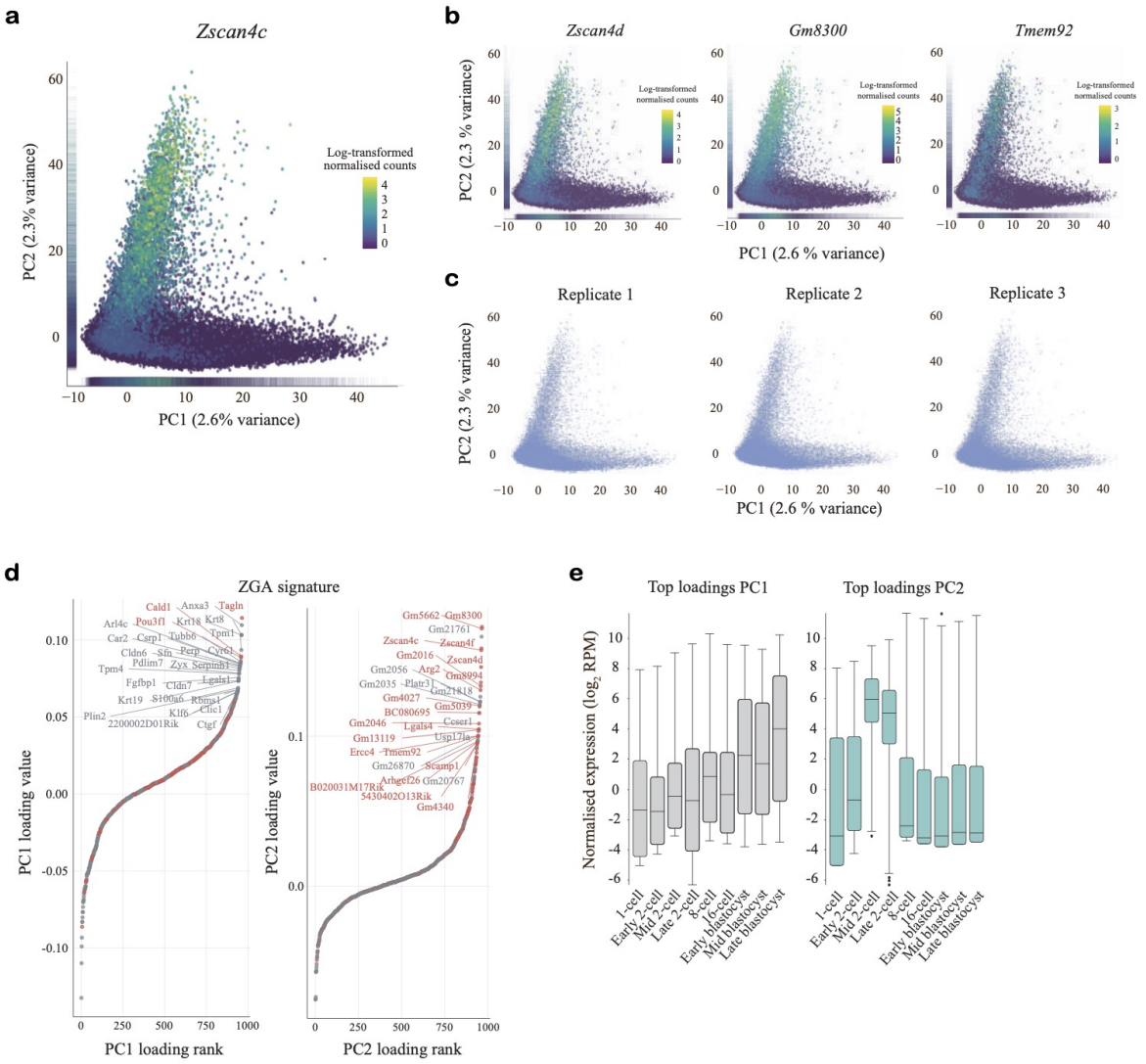

Fig. 4.4 Main sources of variation in gene expression
a,b: Scatterplot of the first two principal components (PC1 and PC2). Cells are coloured by the expression of ZGA markers *Zscan4c*, *Zscan4d*, *Gm8300*, *Tmem92*. Marginal distributions for principal components are displayed as rug plots along the respective axes. **c:** Scatterplots of the first two principal components with the cells in each of the three transduction replicates. **d:** Genes ranked by their PC1 (left) and PC2 (right) loadings. Previously know ZGA genes are highlighted in red. **e:** Aggregated expression (log_2 reads per million) for the top 50 PC1 (left) and PC2 (right) loadings during preimplantation development (Q. Deng et al., 2014). PC2 loadings peak at mid and late two-cell embryo, identifying this component as ZGA-like.

the factors by the proportion of variance they explain, I noted the factor 3 captured a ZGA-like signature (Figure 4.5d-f). In the coding genes view, the highest loadings for this factor are enriched in genes that are linked to ZGA (Figure 4.5d,f) and are highly expressed in mid-to-late two-cell embryos (Figure 4.5e). In the repeat elements view, MERVL repeats, which were linked to ZGA (Melanie A. Eckersley-Maslin et al., 2018; Mélanie A. Eckersley-Maslin et al., 2016), are the most prominent driver of factor 3 (Figure 4.5d). They are followed by major satellites notably linked to two-cell embryos where their active transcription is required for chromatin reorganisation and heterochromatin establishment (Casanova et al., 2013; M. S. H. Ko, 2016; Probst and Almouzni, 2011). I will use this interpretation of the factor 3 to refer to it as a ZGA-like MOFA+ factor below. Other factors captured technical and biological variability associated with mESCs rather than any gene expression programs in preimplantation development (Figure 4.6c).

I also applied MOFA+ to identify a ZGA-like signature in an *in vivo* mouse preimplantation dataset (Q. Deng et al., 2014). This dataset provides necessary temporal resolution to assess transcriptional changes during ZGA as it includes cells from various stages including zygotes, early, mid and late two-cell and four-cell embryos. The first MOFA+ factor ordered the cells according to their developmental stage, from zygotes to four-cell embryo while the second factor distinctively separates ZGA stages (mid and late two-cell embryos) from the rest (Figure 4.7a,b). I then compared that *in vivo* ZGA factor to the *in vitro* ZGA-like factor described above to note that both of them capture ZGA-linked genes among the top gene loadings (Figure 4.7c). Overall, these results point to the presence of ZGA-like transcriptional changes upon CRISPRa in mESCs detected using the robust screening strategy and MOFA+, which enabled unsupervised signature identification across multiple data views.

4.2.3.4 Identification of activators of zygotic genome activation-like signature

A pooled screening strategy combined with multi-group functionality of MOFA+ provides opportunity to assess the association of each sgRNA with the identified ZGA-like transcriptional changes. For this, I confined the analysis to 228 sgRNA that provided evidence of target gene upregulation and fitted a linear model for each sgRNA considering MOFA+ factor 3 activity for each cell as a function of presence or absence (non-targeting controls) of a targeting guide (see section 4.2.5.8). The regression coefficient δ in this model corresponds to the effect size of each sgRNA. This way I identified 25 sgRNAs that induce ZGA-like

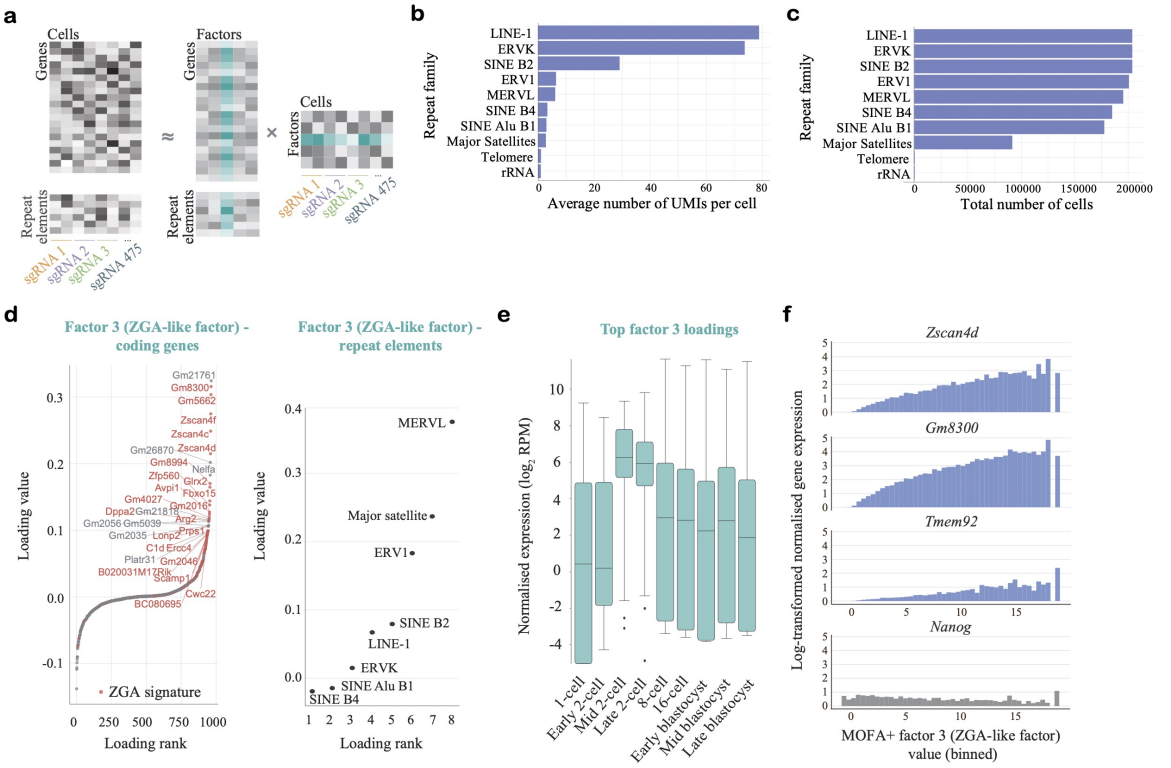

Fig. 4.5 Identification of a ZGA-like transcriptional signature with MOFA+
a: Schematic representation of the MOFA+ used for joint dimentionality reduction of expression matrices with coding genes and repeat elements. Data matrices with values across two respective views and 475 groups of cells (according to their sgRNAs) are decomposed into the product of weights (or loadings) and factors. Factor 3 of the trained model was interpreted as a ZGA-like factor and is highlighted in green. **b:** Average number of UMIs per cell after quality control. **c:** Total number of cells, after quality control, for which each repeat family is detected (number of UMIs > 0), out of 203,894 cells. **d:** Coding genes (left) and repeat element familites (right) ranked by their loadings of MOFA+ factor 3. Previously known ZGA genes are highlighted in red and indicate this factor captures a ZGA-like response. **e:** Aggregated expression (log_2 reads per million) for the top 50 MOFA+ factor 3 loadings during preimplantation development (Q. Deng et al., 2014). Factor 3 loadings peak at mid and late two-cell embryo, identifying this factor as ZGA-like. **f:** Log-transformed normalised expression levels for the ZGA markers *Zscan4d*, *Gm8300* and *Tmem92* (in blue) and the unrelated gene Nanog (in grey), for different quantiles of MOFA+ factor 3 values. This shows ZGA markers contribute to the variance explained by MOFA+ factor 3.

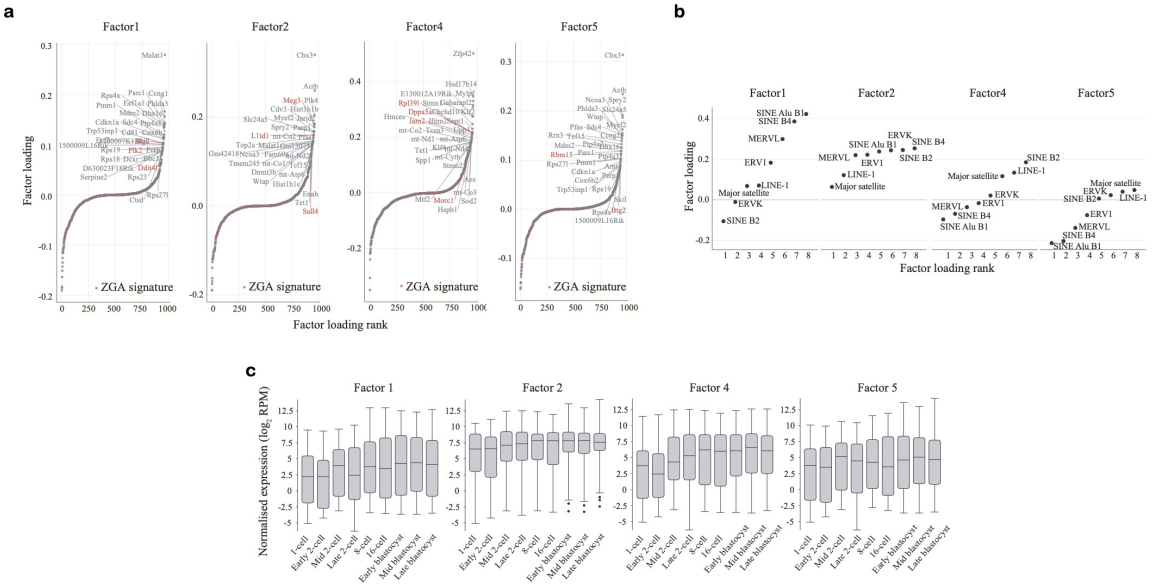

Fig. 4.6 MOFA+ factors that don't capture a ZGA-like response
a: Coding genes ranked by their loadings of MOFA+ factors 1, 2, 4 and 5. Previously know ZGA genes are highlighted in red. **b:** Repeat element families ranked by their loadings of MOFA+ factors 1, 2, 4 and 5. **c:** Normalised expression levels for the top 50 gene loadings of MOFA+ factors 1, 2, 4 and 5 during preimplantation development (Q. Deng et al., 2014).

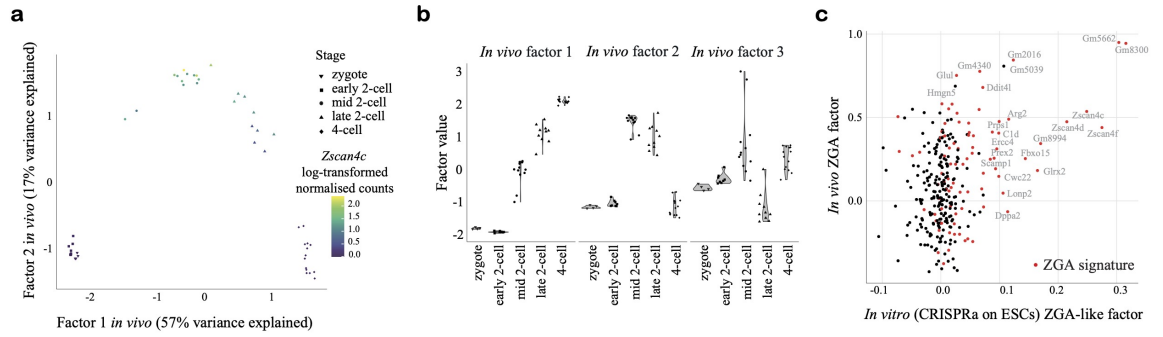

Fig. 4.7 MOFA+ captures ZGA response in vivo
a: Scatterplot for MOFA+ factor 1 and factor 2 values. MOFA+ model was trained on the scRNA-seq data for zygotes, early 2-cell, mid 2-cell, late 2-cell and four-cell stage embryos (Q. Deng et al., 2014). Cells are coloured by *Zscan4c* gene expression. **b:** MOFA+ factors 1, 2 and 3 values. Data as in **a**. **c:** Gene loadings for factor 2 from **a** and gene loadings for factor 3 from the MOFA+ model trained on the CRISPRa-perturbed ESCs scRNA-seq dataset. Genes with gene loadings for both factors are labelled. Previously known ZGA genes are coloured in red.

transcriptional changes upon CRISPRa of their target gene (Figure 4.8a). These 25 sgRNAs target 24 genes with two of them targeting *Dppa2* (Figure 4.8b).

As 25 sgRNAs were found to be associated with the ZGA-like MOFA+ factor, I confirmed the cells with these guides exhibit upregulated expression of coding genes driving this factor (Figure 4.8d). At the same time, the expression of genes linked to other factors was largely unchanged (Figure 4.9a). This highlights the specificity of these sgRNAs in promoting a ZGA-like response. Similarly, these sgRNAs upregulate mainly MERVL repeats (Figure 4.8d) as well as major satellites rather than other repeat families (Figure 4.9b), consistent with these repeat families having the top loadings for the ZGA-like MOFA+ factor (Figure 4.5d).

For each of the 25 sgRNAs hits I then investigated individual genes which expression was induced with CRISPRa and its downstream effects. With a transcriptome-wide differential expression test comparing cells with targeting and non-targeting cells sgRNAs, only a small set of genes was significantly differentially expressed. Ranking the top 400 upregulated genes by statistical significance, however, helped to identify enrichment of known ZGA-associated genes for 23 of 25 sgRNAs (Figure 4.10). As a complementary strategy, this analysis provides additional confidence in the selected sgRNAs identified with MOFA+. Moreover, it shows that MOFA+ allows to identify screen hits that otherwise might be missed with conventional differential expression analysis due to the lack of power to detect the effects on individual genes.

Some of the 24 identified genes which activation elicits ZGA-like transcriptional changes are previously known ZGA regulators such as *Dppa2* (M. Eckersley-Maslin et al., 2019; Hernandez et al., 2018), heat-shock factor-1 *Hsf1* (Christians et al., 2000) and yes-associated protein *Yap1* (Jukam et al., 2017; Yu et al., 2016).

Genes that are newly linked to ZGA include transcription factors *Patz1* (PATZ1 POZ/BTB and AT hook containing zinc finger 1), *Pou2f2* (POU class 2 homeobox 2, also known as *Oct-2*), *Foxo3* (forkhead box O-3), *Tsc22d4* (TSC22 domain family member 4), DNA demethylase *Tet3*, component of histone acetyltransferase complexes *Ing5* (inhibitor of growth family member 5), histone demethylase *Phf2* (PHD finger protein 2), heterochromatin component *Cbx5* (chromobox 5), components of the SWI/SNF chromatin remodelling complex *Arid1b* (AT-Rich interaction domain 1B), *Smarca5* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily A, member 5; also known as *Snf2h*), component of the the histone-deacetylase multiprotein complex (NuRD) *Mta1* (metastasis associated

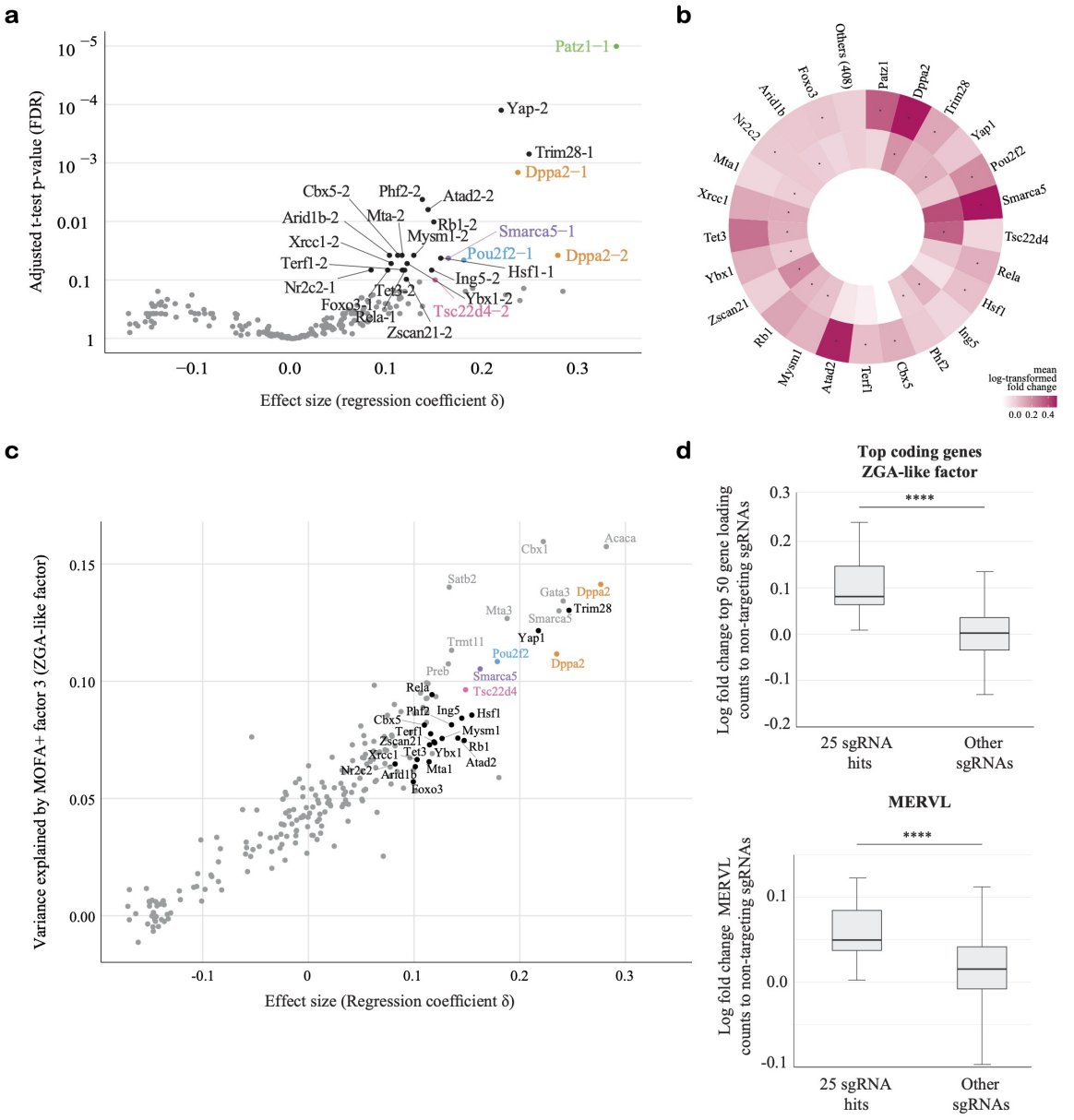

Fig. 4.8 Identification of MOFA+ ZGA-like signature activators
a: Effect size (regression coefficient δ) and the adjusted t-test p-value adjusted with the Benjamini-Hochberg adjustment. 25 sgRNAs tgat were considered to be potent positive regulators at FDR 10**b:** Target gene activation for the 25 sgRNAs measured as log fold change between gene expression in cells with the corresponding sgRNA and cells with non-targeting sgRNAs. Both sgRNAs per gene are shown (inner and outer circles) with sgRNAs identified as screen hits marked with dots. **c:** Fraction of explained variance for individual sgRNAs is consistent with the regression coefficient δ. **d:** Log fold change of expression of the top 50 MOFA+ factor 3 absolute loadings (top) and MERVL repeats (bottom) in cells expressing the 25 sgRNA and in cells expressing other targeting sgRNAs. Comparison was performed to cells expressing non-targeting sgRNA controls. $^{****}p\text{-}value = 3.7 \cdot 10^{-10}$ for coding genes and $8.2 \cdot 10^{-7}$ for MERVL repeats, Mann-Whitney two-tailed test.

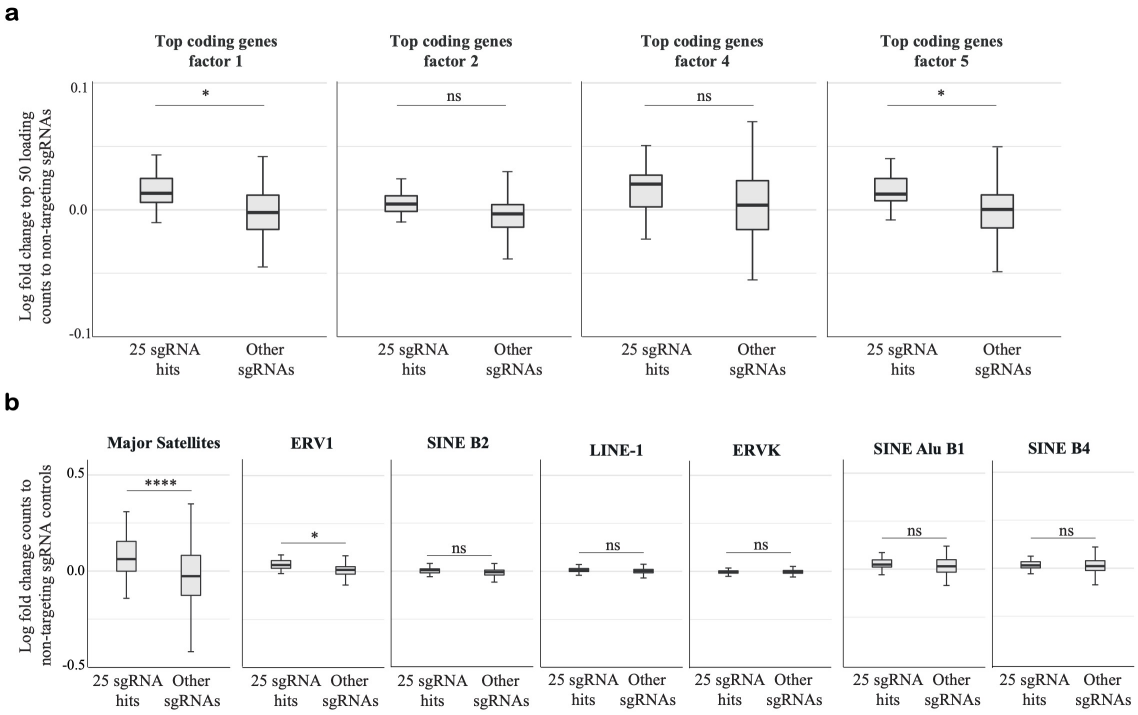

Fig. 4.9 Potent positive ZGA regulators and factors that don't capture ZGA-like response **a:** Log fold change of expression of the top 50 absolute gene loadings associated with MOFA+ factors 1, 2, 4 and 5 in cells expressing the 25 sgRNA hits and cells expressing other targeting sgRNAs, compared to cells expressing non-targeting sgRNAs. * : *p-value* < 0.05, ns: non-significant, Mann-Whitney two-tailed t-test. **b:** Log fold change of expression of different repeat families in cells expressing the 25 sgRNA hits and cells expressing other targeting sgRNAs, compared to cells expressing non-targeting sgRNAs. **** : *p-value* < 0.0001, * : *p-value* < 0.05, ns: non-significant, Mann-Whitney two-tailed t-test.

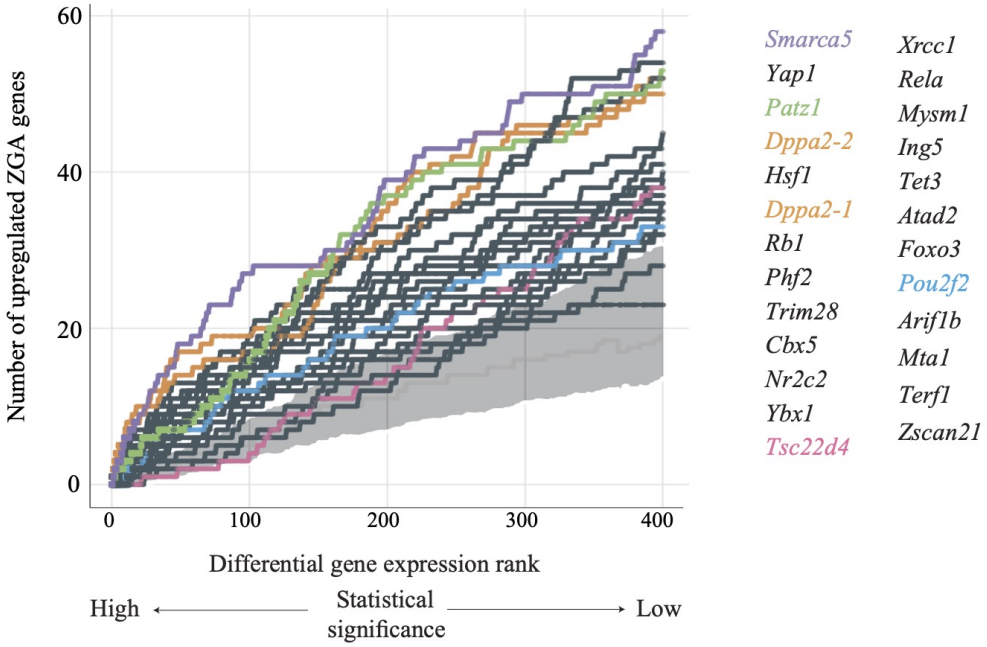

Fig. 4.10 ZGA genes upregulation by potent positive regulators
Number of ZGA-signature genes upregulated by each of the 25 sgRNAs compared to non-targeting sgRNA controls among the top 400 genes ranked by statistical significance of differential gene expression test. In grey, an empirical background distribution estimated based on differential gene expression between cells with non-targeting sgRNAs is shown (one standard deviation around the mean). 24 potent positive regulators are labelled, with some genes highlighted: *Patz1* in green, *Dppa2* in orange, *Smarca5* in purple, *Pou2f2* in blue, and *Tsc22d4* in pink, overlaping with the non-targeting sgRNAs in grey.

1), DNA-repair proteins *Xrcc1* (X-ray repair cross complementing 1), nuclear receptor *Nr2c2* (nuclear receptor subfamily 2 group C member 2), facilitator of chromatin modifiers recruitment *Atad2* (ATPase family AAA domain-containing), telomere-specific protein *Terf1* (telomeric repeat-binding factor 1).

4.2.3.5 *Patz1*, *Dppa2* and *Smarca5* as potent inducers of ZGA-like transcription

Some of these genes were later validated to promote ZGA-like transcriptional changes upon activation using complementary approaches. For instance, ten potential ZGA regulators were targeted with CRISPRa followed by the bulk RNA sequencing (Alda-Catalinas et al., 2020). This has confirmed the effectiveness of the CRISPRa screening strategy followed by scRNA-seq and also showed that weaker regulators such as *Arnt* (aryl hydrocarbon receptor nuclear translocator), *Sirt1* (sirtuin 1) or *Smad1* (SMAD family member 1), which were also included in the validation screen, can be identified with methods such as bulk RNA-seq even though they did not meet the significance criteria in the primary analysis (Figure 4.10).

I used this validation bulk RNA-seq dataset to find downstream effects of CRISPRa and compare them to the effects of the same sgRNAs estimated with scRNA-seq data. Namely, I compared fold changes from the differential expression analysis in bulk and single-cell readouts. While bulk RNA-seq provided more power in calling differentially expressed genes, the patterns of the downstream transcriptional changes were largely consistent between two datasets (Figure 4.11). Notably, both the scRNA-seq screen and validation bulk RNA-seq identified *Patz1*, *Dppa2* and *Smarca5* as strong positive regulators of ZGA-like transcription.

Patz1, *Dppa2*, and *Smarca5* were later confirmed to be positive regulators of ZGA-like transcription in mESCs using complementary experimental strategies. Alternative methods of gene overexpression such as transfecting cDNA-eGFP (enhanced green fluorescent protein) into mESCs and using bulk RNA-seq to quantify transcriptional response confirmed that there three potential regulators trigger similar transcriptome-wide changes in gene expression (Alda-Catalinas et al., 2020). This is in sharp contrast with the effects of the *Carhsp1*-targeting sgRNA used as a negative control, which displayed target gene activation but no ZGA-like response downstream. Immunofluorescence techniques confirmed that *Patz1*, *Dppa2* and *Smarca5* are expressed in a zygote with SMARCA5 localised in the nucleus and PATZ1 and DPPA2 localised both in the nucleus and cytoplasm. Overall, these results point to these three genes as positive regulators of ZGA-like transcription confirming the findings of the high-throughput screen.

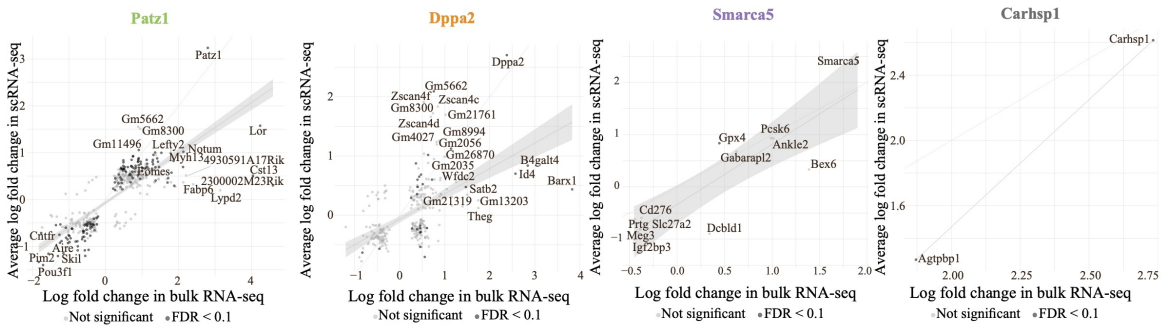

Fig. 4.11 Differentially expressed genes in scRNA-seq and in bulk
Log fold change values of differentially expressed genes estimated based on bulk CRISPRa RNA-sequencing data (FDR 10%) (x axis) versus log fold change estimates for the corresponding genes in cells expressing the same sgRNA based on the CRISPRa scRNA-seq data (y axis) for the target genes *Patz1*, *Dppa2*, *Smarca5*, and *Carhsp1*. Genes that were also differentially expressed in scRNA-seq data (FDR 10%) are labelled in dark grey whereas those genes differentially expressed in bulk RNA-sequencing data but not in scRNA-seq are labelled in light grey (not significant). For each target gene, a regression line was fitted to highlight the trend. Dashed lines mark the $y = x$ line.

Earlier, DPPA2 was shown to induce the ZGA gene expression network (De Iaco et al., 2019; M. Eckersley-Maslin et al., 2019), and there's evidence it interacts with SMARCA5 in mESCs (Hernandez et al., 2018) and colocalises with it in two-cell embryos (Alda-Catalinas et al., 2020). SMARCA5 is an adenosine triphosphatase (ATPase) subunit of ISWI (imitation switch) chromatin remodelling complex, one of the four mammalian remodeller families that orchestrate the nucleosome organisation. *Smarca5* knockdown in zygotes compromises development suggesting its importance for zygotic genome activation (Torres-Padilla and Zernicka-Goetz, 2006), and the recent *Smarca5* knockout data (Barisic et al., 2019) revealed that SMARCA5 regulates ZGA through its ATPase activity (Alda-Catalinas et al., 2020). Further work shows that *Dppa2* is required for *Smarca5*-mediated ZGA regulation. In particular, *Smarca5* knockout mESCs display downregulated ZGA genes, and their expression can be partially rescued by *Dppa2* overexpresison. At the same time, *Smarca5* overexpression doesn't induce ZGA-like gene expression in *Dppa2* knockout mESCs suggesting the latter acts downstream of *Smarca5* to regulate ZGA-like transcriptional changes (Alda-Catalinas et al., 2020).

4.2.4 Discussion

ZGA is a major developmental transition, yet the understanding of its timing and master regulators has remained limited (Jukam et al., 2017). This work demonstrates how a newly developed high-throughput CRISPRa screen coupled with scRNA-seq readout and a scalable factorisation framework can be used to identify potent ZGA regulators. I applied MOFA+ to the dataset comprised of more than 200 thousand mESCs with 460 sgRNAs targeting potential ZGA regulators. Joint analysis of coding gene and repeat elements expression revealed 24 genes that induced ZGA-like transcriptional response upon activation (Figure 4.8). Some of the candidate regulators were then experimentally validated and shown to induce ZGA-like transcription in mESCs with CRISPRa and complementary approaches (Alda-Catalinas et al., 2020). In particular, *Dppa2*-dependent regulation of ZGA was uncovered for *Smarca5*.

This CRISPRa-based screening method enables a systematic functional interrogation of a large number of genes and has a potential to be adapted to various biological contexts. Gene overexpression method is of notable importance here as CRISPRa provides advantages over traditional cDNA overexpression including physiologically relevant level of gene upregulation (J. Yang et al., 2019) and ease of scaling to more genes. Of 460 sgRNAs, only 228 sgRNAs had evidence for target gene upregulation, which might be explained by lack of activation due to sub-optimal sgRNA design or by lack of detection due to technical dropouts and data sparsity for the remaining sgRNAs. With an unsupervised factorisation framework, MOFA+, I inferred a ZGA-like transcriptional signature leveraging the correlation structure across genes. This signature has been defined in terms of expression of coding gene as well as of repeat elements. I selected the most potent positive regulators candidates by estimating the ZGA-like signature activation effect for each sgRNA and comparing the downstream effect of gene upregulation with list of genes in the ZGA network. Some of the 24 identified regulators (*Dppa2*, *Yap1*, *Hsf1*) were previously described as ZGA regulators, which validates the screening and analytical approach. Several new regulators including *Smarca5* and *Patz1* were further validated to confirm their ability to promote the ZGA-like signature upon upregulation.

In summary, the high-throughput CRISPRa-based screening strategy together with the unsupervised multimodal factorisation framework unraveled positive regulators of transcriptional changes in mESCs that resemble ZGA in mouse embryos. The results of this work motivate further functional studies of the potent ZGA regulators as well as investigations of

their interdependencies. Moreover, this method can be applied in a broad range of biological contexts for systematic transcriptional regulation understanding.

4.2.5 Methods

4.2.5.1 Candidate regulators selection

A list of potential ZGA regulators was constructed based on published data and resources. Namely, the intersection of the following gene sets was considered. First, mouse proteins associated with transcription factor (PC00218) and nucleic acid binding (PC00171) activities from the PANTHER[1] database (Mi et al., 2019). Second, proteins detected in metaphase II (MII) mouse oocytes (Pfeiffer, Siatkowski, et al., 2011; Pfeiffer, Taher, et al., 2015) and both MII oocytes and zygotes (Bingyuan Wang et al., 2016). Intersecting there four gene lists resulted in a list of 230 candidates (Table B.1).

4.2.5.2 Experimental procedures

Detailed description of experimental procedures has been provided in Alda-Catalinas et al., 2020. Briefly, the sgRNAs were designed according to Joung et al., 2017 to target the 180bp region upstream of the transcription start site (TSS) of the target gene. The lentiviral backbone included a fluorescent mCherry marker and a puromycin resistance cassette to enable fluorescence-activated cell sorting and antibiotic selection. Cloning, lentiviral packaging and transductions were performed as described in Alda-Catalinas et al., 2020. Following sorting and selection, the cells were used to prepare scRNA-seq libraries using 10x Genomics Single Cell 3′ Library & Gel Bead Kit v2 (Zheng et al., 2017). Each library was sequenced on Illumina HiSeq 4000. For each of the full length 10x cDNA samples, amplicon sgRNA PCRs were performed as described in Hill et al., 2018.

4.2.5.3 Analysis of scRNA-seq data

All the scRNA-seq data was processed wit the default CellRanger pipeline v2.1 (Zheng et al., 2017) using the mm10 mouse genome assembly. Data management, filtering, quality control and downstream analyses were performed with SCANPY (Wolf et al., 2018).

[1] https://pantherdb.org

For the pilot screen, cells with less than 15000 UMI counts, more than 40000 UMI counts, less than 4000 detected genes, more than 6500 detected genes or with more than 5% of counts coming from mitochondrial genes were discarded.

For the main screen, cells with less than 4000 UMI counts, more than 20000 UMI counts, less than 1600 detected genes, more than 5000 detected genes or with more than 5% of counts coming from mitochondrial genes were discarded (Figure 4.3a-c). This resulted in 109061, 118646 and 107591 cells in each of the three transduction replicates. After a sgRNA was assigned to each cell (see section 4.2.5.5), only cells with a unique sgRNA assignment were retained, resulting in 71047, 71188 and 61729 cells in three replicates, 203894 cells in total.

Genes that were detected in at least 10 cells after quality control were considered for downstream analysis. UMI counts for each gene were normalised by the library size, scaled by a factor of 10000 and log-transformed. Principal components were calculated with 965 highly variables genes that were selected using the SCANPY implementation (Wolf et al., 2018) with minimum mean of 0.01, maximum mean of 5 and minimum dispersion of 0.5.

4.2.5.4 Repeat Element Quantification

Repeat sequences from 12 families (LINE-1, LINE-2, ERV1, ERVK, MERVL, Major satellites, Minor satellites, SINE Alu B1, SINE B2, SINE B4, Ribosomal RNA, Telomeric repeats) with their genomic locations were downloaded from the UCSC Table Browser (Karolchik et al., 2004), concatenated and treated as a reference genome. Reads that had not been mapped to the mm10 reference by the CellRanger pipeline were extracted: 253330874 reads in replicate 1, 276401843 reads in replicate 2, 242863617 in replicate 3. These reads were mapped to the repeat elements reference using SAMtools (H. Li et al., 2009) and BWA v0.7.17-r1188 (Heng Li and Durbin, 2009) with default parameters. As a result, 15.31%, 13.49% and 10.64% of those reads were mapped to repeat elements. Minor satellites and LINE-2 elements were discarded from downstream analyses due to low mapping rates (Table 4.1). To obtain counts for each repeat family in each cell, reads sharing the same UMI and cell barcode were collapsed.

Table 4.1 Repeat element quantification

Repeat family	Number of mapped reads
LINE-1	40,490,996
LINE-2	185
ERV1	3,161,474
ERVK	36,949,603
MERVL	3,707,419
Major Satellites	798,309
Minor Satellites	9
Ribosomal RNA	2170
SINE Alu B1	1,227,023
SINE B2	14,151,376
SINE B4	1,425,154
Telomeric repeats	1348

4.2.5.5 Assignment of sgRNAs to Cells

In the sequenced amplicon sgRNA libraries, I extracted the potential sgRNA sequence, which corresponded to nucleotides 24 to 43 of a read. When considering exact matches to the list of sgRNAs, 16% of reads on average were not assigned to any sgRNA (15.3% – 16.8% for different libraries). The majority of 475 sgRNAs were recovered (470 – 474 in different libraries). In order to correct for sequencing errors, I allowed for a Levenshtein distance of

2 edits when comparing the expected sgRNA sequence with a list of possible sgRNAs and of 5 edits when comparing upstream and downstream 23bp-long context surrounding the sgRNA in the CROP-sgRNA-MS2. This correction procedure reduced the number of reads unassigned to sgRNAs to 2%. Cell barcodes that were detected in the amplicon libraries were matches with the cell barcodes from the scRNA-seq libraries. Among the latter, out of 317847 cells across three transduction replicates passing quality control, 249767 cell barcodes were detected in the amplicon sgRNA libraries. An sgRNA was assigned to a cell if more than 90% of the amplicon reads with that cell barcode shared the same sgRNA and if the standard error of binomial proportion was less than 10% (more than 8 reads if there's only one potential sgRNA, more than 13 reads if there are several). Table 4.2 summarises sgRNA assignment for each replicate.

Table 4.2 sgRNA assignment to cells across three replicates

Replicate	**Total number of cells**	**Assignment**	**Number of cells**	**Percentage**
1	85,933	No sgRNA	397	0.46
1	85,933	Unique sgRNA	71,047	82.62
1	85,933	Two sgRNAs	3,028	3.52
1	85,933	Multiple sgRNAs	11,521	13.40
2	86,671	No sgRNA	400	0.46
2	86,671	Unique sgRNA	71,118	82.14
2	86,671	Two sgRNAs	3,210	3.70
2	86,671	Multiple sgRNAs	11,873	13.70
3	77,103	No sgRNA	381	0.49

3	77,103	Unique sgRNA	61,729	80.06
3	77,103	Two sgRNAs	3,084	4.00
3	77,103	Multiple sgRNAs	11,909	15.45

4.2.5.6 MOFA+ application to the primary screen

I trained a MOFA+ model on two views of the data corresponding to coding genes and repeat elements. For the coding genes view, I considered a set of 965 highly variable genes as described in 4.2.5.3. For the repeat elements, I used expression levels of eight repeat elements (LINE-1, ERV1, ERVK, MERVL, SINE Alu B1, SINE B2, SINE B4 and Major satellites) estimated as described in 4.2.5.4. Telomeric repeats and ribosomal RNA were excluded due to low detection rate. Cells were grouped according to the assign sgRNA as described in 4.2.5.5. For the model training, I used 25 factors and Gaussian likelihoods for both view. Model training and interpretation were conducted with use of MOFA libraries in Python and in R.

4.2.5.7 MOFA+ application to the in vivo data

Single-cell RNA-seq data from five developmental stages including zygotes, early, mid and late two-cell and four-cell embryos (Q. Deng et al., 2014) were processed using SCANPY (Wolf et al., 2018). I applied MOFA+ on the top 5000 highly variables genes. Using the gene loadings for each factor, I concluded that factor 2 corresponded to a ZGA signature (Figure 4.7c). This was corroborated by high factor 2 activity during ZGA stages, i.e. mid and late two-cell embryos (Figure 4.7a,b).

4.2.5.8 Identification of potent positive regulators

For each of the 228 sgRNAs that prompted any target gene activation, I considered the ZGA-like MOFA+ factor values Z for the corresponding cells. In the model below, they were compared to the ZGA-like MOFA+ factor values Z with non-targeting sgRNAs *NT sgRNAs*. I fitted a linear model for each of the targeting sgRNAs *T sgRNA*:

$$Z_{[T\ sgRNA, NT\ sgRNAs]} \sim I_T,$$

where I_T is a binary indicator with $I_T = 1$ for cells with a targeting sgRNA and $I_T = 0$ for cells with non-targeting sgRNAs. Having fitted this model for each targeting sgRNA, I obtained effect size δ values that characterised guides' potency to induce the ZGA-like gene expression program. I then used a likelihood ratio test to test that δ is different from 0, obtained p-values and used the Benjamini-Hochberg procedure for multiple testing correction. Positive regulators were reported (Figure 4.8a) for 10% false discovery rate (FDR). I also demonstrated that δ values are correlated with higher expression of genes linked to ZGA (Figure 4.8d) and with the proportion of variance explained by the ZGA-like MOFA+ factor (Figure 4.8c).

4.2.5.9 Differential gene expression

To estimate downstream effects of the targeted gene upregulation, I compared cells with targeting and non-targeting sgRNAs using a generalised linear model (glm) and a likelihood ratio test as implemented in edgeR (Robinson et al., 2010). Using the top 400 genes ranked by statistical significance for each of the targeting sgRNA, I ranked the sgRNAs based on the proportion of ZGA-linked genes (see Appendix B.2) among them. Such cumulative ranking highlighted *Smarca5*, *Dppa2* and *Patz1* as most potent regulators. Empirical background distribution was estimated based on differential expression between cells with non-targeting sgRNAs.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems.

Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., & Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, *167*(7), 1867–1882.e21.

Ahlmann-Eltze, C., & Huber, W. (2021). Transformation and Preprocessing of Single-Cell RNA-Seq Data, 2021.06.24.449781.

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212.

Alda-Catalinas, C., Bredikhin, D., Hernando-Herraez, I., Santos, F., Kubinyecz, O., Eckersley-Maslin, M. A., Stegle, O., & Reik, W. (2020). A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program. *Cell Systems*, *11*(1), 25–41.e9.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, *3*(1), 246–259.

Almanzar, N., Antony, J., Baghel, A. S., Bakerman, I., Bansal, I., Barres, B. A., Beachy, P. A., Berdnik, D., Bilen, B., Brownfield, D., Cain, C., Chan, C. K. F., Chen, M. B., Clarke, M. F., Conley, S. D., Darmanis, S., Demers, A., Demir, K., de Morree, A., … The Tabula Muris Consortium. (2020). A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, *583*(7817), 590–595.

Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, *97*(18), 10101–10106.

Amezquita, R. A., Lun, A. T. L., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C., Waldron, L., Pagès, H., Smith, M. L., Huber, W., Morgan, M., Gottardo, R., & Hicks, S. C. (2020). Orchestrating single-cell analysis with Bioconductor. *Nature Methods*, *17*(2), 137–145.

Amir, E.-a. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., & Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology*, *31*(6), 545–552.

Androvic, P., Schifferer, M., Anderson, K. P., Cantuti-Castelvetri, L., Ji, H., Liu, L., Besson-Girard, S., Knoferle, J., Simons, M., & Gokce, O. (2022). Spatial Transcriptomics-correlated Electron Microscopy.

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O., & Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, *13*(3), 229–232.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, *12*(7), 878.

Apache Parquet. (2022). Apache Parquet. Retrieved June 19, 2022, from https://parquet.apache.org/

Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D., & Wagner, G. P. (2016). The origin and evolution of cell types. *Nature Reviews Genetics*, *17*(12), 744–757.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., & Stegle, O. (2020). MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, *21*(1), 111.

Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., … Reik, W. (2019). Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, *576*(7787), 487–491.

Argelaguet, R., Cuomo, A. S. E., Stegle, O., & Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, *39*(10), 1202–1215.

Argelaguet, R., Lohoff, T., Li, J. G., Nakhuda, A., Drage, D., Krueger, F., Velten, L., Clark, S. J., & Reik, W. (2022). Decoding gene regulation in the mouse embryo using single-cell multi-omics.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, *14*(6), e8124.

Armand, E. J., Li, J., Xie, F., Luo, C., & Mukamel, E. A. (2021). Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron*, *109*(1), 11–26.

Arzalluz-Luque, Á., & Conesa, A. (2018). Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biology*, *19*(1), 110.

Ashuach, T., Gabitto, M. I., Jordan, M. I., & Yosef, N. (2021). MultiVI: Deep generative model for the integration of multi-modal data, 2021.08.20.457057.

Attaf, M., Huseby, E., & Sewell, A. K. (2015). Alpha-beta T cell receptors as predictors of health and disease. *Cellular & Molecular Immunology*, *12*(4), 391–399.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, *18*(10), 1196–1203.

Baccin, C., Al-Sabah, J., Velten, L., Helbling, P. M., Grünschläger, F., Hernández-Malmierca, P., Nombela-Arrieta, C., Steinmetz, L. M., Trumpp, A., & Haas, S. (2020). Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nature Cell Biology*, *22*(1), 38–48.

Bachireddy, P., Azizi, E., Burdziak, C., Nguyen, V. N., Ennis, C. S., Maurer, K., Park, C. Y., Choo, Z.-N., Li, S., Gohil, S. H., Ruthen, N. G., Ge, Z., Keskin, D. B., Cieri, N., Livak, K. J., Kim, H. T., Neuberg, D. S., Soiffer, R. J., Ritz, J., … Wu, C. J. (2021). Mapping the evolution of T cell states during response and resistance to adoptive cellular therapy. *Cell Reports*, *37*(6), 109992.

Bakhtiari, S., Safavi-Naini, R., & Pieprzyk, J. (1995). *Cryptographic Hash Functions: A Survey*.

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., & Tanay, A. (2019). MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, *20*(1), 206.

Barisic, D., Stadler, M. B., Iurlaro, M., & Schübeler, D. (2019). Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature*, *569*(7754), 136–140.

Bellman, R. E. (2015). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Ben-Kiki, O., Bercovich, A., Lifshitz, A., & Tanay, A. (2022). Metacell-2: A divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biology*, *23*(1), 100.

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, *38*(12), 1408–1414.

Bergen, V., Soldatov, R. A., Kharchenko, P. V., & Theis, F. J. (2021). RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, *17*(8).

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98.

Bilous, M., Tran, L., Cianciaruso, C., Gabriel, A., Michel, H., Carmona, S. J., Pittet, M. J., & Gfeller, D. (2022). Metacells untangle large and complex single-cell transcriptome networks, 2021.06.07.447430.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., & Radvanyi, F. (2014). Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes. *Cell Reports*, *9*(4), 1235–1245.

Blanas, S., Wu, K., Byna, S., Dong, B., & Shoshani, A. (2014). Parallel data analysis directly on scientific file formats. *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, 385–396.

Blei, D. M. [David M.], Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Bloom, J. D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, *6*, e5578.

Bodenmiller, B. (2016). Multiplexed Epitope-Based Tissue Imaging for Discovery and Healthcare Applications. *Cell Systems*, *2*(4), 225–238.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., & Zhang, Q. (2018). *JAX: Composable transformations of Python+NumPy programs* (Version 0.3.13).

Bredikhin, D., Kats, I., & Stegle, O. (2022). MUON: Multimodal omics analysis framework. *Genome Biology*, *23*(1), 42.

Brüning, R. S., Tombor, L., Schulz, M. H., Dimmeler, S., & John, D. (2022). Comparative analysis of common alignment tools for single-cell RNA sequencing. *GigaScience*, *11*, giac001.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., & Stegle, O. (2017). F-scLVM: Scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, *18*(1), 212.

Burdziak, C., Azizi, E., Prabhakaran, S., & Pe'er, D. (2019). *A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene Regulatory Networks* (arXiv:1902.08138).

Burgess, D. J. (2019). Spatial transcriptomics coming of age. *Nature Reviews Genetics*, *20*(6), 317–317.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420.

Büttner, M. [M.], Ostner, J., Müller, C. L., Theis, F. J., & Schubert, B. (2021). scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications*, *12*(1), 6876.

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, *361*(6409), 1380–1385.

Cao, J., O'Day, D. R., Pliner, H. A., Kingsley, P. D., Deng, M., Daza, R. M., Zager, M. A., Aldinger, K. A., Blecher-Gonen, R., Zhang, F., Spielmann, M., Palis, J., Doherty, D., Steemers, F. J., Glass, I. A., Trapnell, C., & Shendure, J. (2020). A human cell atlas of fetal gene expression. *Science*, *370*(6518), eaba7721.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

Cao, Y., Lin, Y., Ormerod, J. T., Yang, P., Yang, J. Y., & Lo, K. K. (2019). scDC: Single cell differential composition analysis. *BMC Bioinformatics*, *20*(19), 721.

Casanova, M., Pasternak, M., El Marjou, F., Le Baccon, P., Probst, A. V., & Almouzni, G. (2013). Heterochromatin Reorganization during Early Mouse Development Requires a Single-Stranded Noncoding Transcript. *Cell Reports*, *4*(6), 1156–1167.

Cemgil, A. T. [Ali Taylan]. (2009). Bayesian Inference for Nonnegative Matrix Factorisation Models. *Computational Intelligence and Neuroscience*, *2009*, e785152.

Chari, T., Banerjee, J., & Pachter, L. (2021). The Specious Art of Single-Cell Genomics, 2021.08.25.457696.

Chari, T., Weissbourd, B., Gehring, J., Ferraioli, A., Leclère, L., Herl, M., Gao, F., Chevalier, S., Copley, R. R., Houliston, E., Anderson, D. J., & Pachter, L. (2021). Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across Clytia medusa cell types. *Science Advances*, *7*(48), eabh1683.

Chen, A. F., Parks, B., Kathiria, A. S., Ober-Reynolds, B., Goronzy, J. J., & Greenleaf, W. J. (2022). NEAT-seq: Simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nature Methods*, *19*(5), 547–553.

Chen, H., Ye, F., & Guo, G. (2019). Revolutionizing immunology with single-cell RNA sequencing. *Cellular & Molecular Immunology*, *16*(3), 242–249.

Chen, S., Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, *37*(12), 1452–1457.

Christians, E., Davis, A. A., Thomas, S. D., & Benjamin, I. J. (2000). Maternal effect of Hsf1 on reproductive success. *Nature*, *407*(6805), 693–694.

Clark, C., Dayon, L., Masoodi, M., Bowman, G. L., & Popp, J. (2021). An integrative multi-omics approach reveals new central nervous system pathway alterations in Alzheimer's disease. *Alzheimer's Research & Therapy*, *13*(1), 71.

Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O., & Reik, W. (2018). scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, *9*(1), 781.

Collette, A. (2013). *Python and HDF5: Unlocking scientific data*. " O'Reilly Media, Inc."

Comon, P., & Jutten, C. (2010). *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press.

Conesa, A., & Beck, S. (2019). Making multi-omics data accessible to researchers. *Scientific Data*, *6*(1), 251.

Consortium*, T. S., Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P. et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (New York, N.Y.)*, *376*(6594), eabl4896.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563.

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., & Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, *11*(1), 6077.

Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., & Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, *348*(6237), 910–914.

Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D., & Marioni, J. C. (2022). Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, *40*(2), 245–253.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, *14*(3), 297–301.

De Carlo, F., Gürsoy, D., Marone, F., Rivers, M., Parkinson, D. Y., Khan, F., Schwarz, N., Vine, D. J., Vogt, S., Gleber, S.-C., Narayanan, S., Newville, M., Lanzirotti, T., Sun, Y., Hong, Y. P., & Jacobsen, C. (2014). Scientific data exchange: A schema for HDF5-

based storage of raw and analyzed data. *Journal of Synchrotron Radiation*, *21*(6), 1224–1230.

De Iaco, A., Coudray, A., Duc, J., & Trono, D. (2019). DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. *EMBO reports*, *20*(5), e47382.

Deng, Q., Ramsköld, D., Reinius, B., & Sandberg, R. (2014). Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science*, *343*(6167), 193–196.

Deng, Y., Bartosovic, M., Ma, S., Zhang, D., Liu, Y., Qin, X., Su, G., Xu, M. L., Halene, S., Craft, J. E., Castelo-Branco, G., & Fan, R. (2021). Spatial-ATAC-seq: Spatially resolved chromatin accessibility profiling of tissues at genome scale and cellular level, 2021.06.06.447244.

Ding, J., Sharon, N., & Bar-Joseph, Z. (2022). Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics*, *23*(6), 355–368.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853–1866.e17.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21.

Domínguez Conde, C., Xu, C., Jarvis, L. B., Rainbow, D. B., Wells, S. B., Gomes, T., Howlett, S. K., Suchanek, O., Polanski, K., King, H. W., Mamanova, L., Huang, N., Szabo, P. A., Richardson, L., Bolt, L., Fasouli, E. S., Mahbubani, K. T., Prete, M., Tuck, L., … Teichmann, S. A. (2022). Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, *376*(6594), eabl5197.

Dougherty, M. T., Folk, M. J., Zadok, E., Bernstein, H. J., Bernstein, F. C., Eliceiri, K. W., Benger, W., & Best, C. (2009). Unifying biological image formats with HDF5. *Communications of the ACM*, *52*(10), 42–47.

Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C., & Reik, W. (2019). Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. *Genes & Development*, *33*(3-4), 194–208.

Eckersley-Maslin, M. A. [Melanie A.], Alda-Catalinas, C., & Reik, W. (2018). Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, *19*(7), 436–450.

Eckersley-Maslin, M. A. [Mélanie A.], Svensson, V., Krueger, C., Stubbs, T. M., Giehr, P., Krueger, F., Miragaia, R. J., Kyriakopoulos, C., Berrens, R. V., Milagre, I., Walter, J., Teichmann, S. A., & Reik, W. (2016). MERVL/Zscan4 Network Activation Results in Transient Genome-wide DNA Demethylation of mESCs. *Cell Reports*, *17*(1), 179–192.

Efremova, M., & Teichmann, S. A. (2020). Computational methods for single-cell omics across modalities. *Nature Methods*, *17*(1), 14–17.

Eggert, J., & Korner, E. (2004). Sparse coding and NMF. *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, *4*, 2529–2533 vol.4.

Eng, C.-H. L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., & Cai, L. (2019). Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, *568*(7751), 235–239.

Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., & Dölken, L. (2019). scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*, *571*(7765), 419–423.

Everaert, C., Luypaert, M., Maag, J. L. V., Cheng, Q. X., Dinger, M. E., Hellemans, J., & Mestdagh, P. (2017). Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Scientific Reports*, *7*(1), 1559.

Falco, G., Lee, S.-L., Stanghellini, I., Bassey, U. C., Hamatani, T., & Ko, M. S. H. (2007). Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. *Developmental Biology*, *307*(2), 539–550.

Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A., & Schier, A. F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, *360*(6392), eaar3131.

Févotte, C., & Cemgil, A. T. [A. Taylan]. (2009). Nonnegative matrix factorizations as probabilistic inference in composite models. *2009 17th European Signal Processing Conference*, 1913–1917.

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo González-Blas, C., Kalender Atak, Z., & Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, *17*(4), 246–254.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., Slichter, C. K., Miller, H. W., McElrath, M. J., Prlic, M., Linsley, P. S., & Gottardo, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, *16*(1), 278.

Flash-Frozen Human Healthy Brain Tissue (3k), 10x Genomics. (2022). 10x Genomics. Retrieved July 21, 2022, from https://www.10xgenomics.com/resources/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0

Francesconi, M., Di Stefano, B., Berenguer, C., de Andrés-Aguayo, L., Plana-Carmona, M., Mendez-Lago, M., Guillaumet-Adkins, A., Rodriguez-Esteban, G., Gut, M., Gut, I. G., Heyn, H., Lehner, B., & Graf, T. (2019). Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming (C. P. Ponting, D. Weigel, & C. P. Ponting, Eds.). *eLife*, *8*, e41627.

Gao, C., Liu, J., Kriebel, A. R., Preissl, S., Luo, C., Castanon, R., Sandoval, J., Rivkin, A., Nery, J. R., Behrens, M. M., Ecker, J. R., Ren, B., & Welch, J. D. (2021). Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*, *39*(8), 1000–1007.

Gao, M., Qiao, C., & Huang, Y. (2022). UniTVelo: Temporally unified RNA velocity reinforces single-cell trajectory inference, 2022.04.27.489808.

Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., … Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, *15*(3), 201–206.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., … Yosef, N. (2022). A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, *40*(2), 163–166.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., & Yosef, N. (2021). Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, *18*(3), 272–282.

Gerlach, J. P., van Buggenum, J. A. G., Tanis, S. E. J., Hogeweg, M., Heuts, B. M. H., Muraro, M. J., Elze, L., Rivello, F., Rakszewska, A., van Oudenaarden, A., Huck, W. T. S., Stunnenberg, H. G., & Mulder, K. W. (2019). Combined quantification of intracellular (phospho-)proteins and transcriptomics from fixed single cells. *Scientific Reports*, *9*(1), 1469.

Giladi, A., & Amit, I. (2018). Single-Cell Genomics: A Stepping Stone for Future Immunology Discoveries. *Cell*, *172*(1), 14–21.

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., Lim, W. A., Weissman, J. S., & Qi, L. S. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell*, *154*(2), 442–451.

Giordano, R., Broderick, T., & Jordan, M. (2015). Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes.

Gonçalves, E., Poulos, R. C., Cai, Z., Barthorpe, S., Manda, S. S., Lucas, N., Beck, A., Bucio-Noble, D., Dausmann, M., Hall, C., Hecker, M., Koh, J., Mahboob, S., Mali, I., Morris, J., Richardson, L., Seneviratne, A. J., Sykes, E., Thomas, F., … Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines reveals principles of cancer vulnerabilities.

Gopalan, P., Hofman, J. M., & Blei, D. M. (2014). Scalable Recommendation with Poisson Factorization.

Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled, 2022.02.12.480214.

Gorin, G., Svensson, V., & Pachter, L. (2020). Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, *21*(1), 39.

Gouwens, N. W., Sorensen, S. A., Baftizadeh, F., Budzillo, A., Lee, B. R., Jarsky, T., Alfiler, L., Baker, K., Barkan, E., Berry, K., Bertagnolli, D., Bickley, K., Bomben, J., Braun, T., Brouner, K., Casper, T., Crichton, K., Daigle, T. L., Dalley, R., … Zeng, H. (2020). Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells. *Cell*, *183*(4), 935–953.e19.

Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature Protocols*, 1–35.

Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y., & Greenleaf, W. J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, *53*(3), 403–411.

Griffiths, J., Scialdone, Antonio, & Marioni, John. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular Systems Biology*, *14*(4), e8046.

Gupta, I., Collier, P. G., Haase, B., Mahfouz, A., Joglekar, A., Floyd, T., Koopmans, F., Barres, B., Smit, A. B., Sloan, S. A., Luo, W., Fedrigo, O., Ross, M. E., & Tilgner, H. U. (2018). Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature Biotechnology*, *36*(12), 1197–1202.

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1), 296.

Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R., & Sandberg, R. (2020). Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature Biotechnology*, *38*(6), 708–714.

Hagemann-Jensen, M., Ziegenhain, C., & Sandberg, R. (2022). Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nature Biotechnology*, 1–6.

Haghverdi, L., Buettner, F., & Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, *31*(18), 2989–2998.

Haghverdi, L., Büttner, M. [Maren], Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427.

Hahaut, V., Pavlinic, D., Carbone, W., Schuierer, S., Balmer, P., Quinodoz, M., Renner, M., Roma, G., Cowan, C. S., & Picelli, S. (2022). Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. *Nature Biotechnology*, 1–5.

Halko, N., Martinsson, P.-G., & Tropp, J. A. (2010). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.

Ham, J., Lee, D. D., Mika, S., & Schölkopf, B. (2004). A kernel view of the dimensionality reduction of manifolds. *Proceedings of the Twenty-First International Conference on Machine Learning*, 47.

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091–1107.e17.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., Zhou, Y., Ye, F., Jiang, M., Wu, J., Xiao, Y., Jia, X., Zhang, T., Ma, X., Zhang, Q., … Guo, G. (2020). Construction of a human cell landscape at single-cell level. *Nature*, *581*(7808), 303–309.

Hanahan, D., & Coussens, L. M. (2012). Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment. *Cancer Cell*, *21*(3), 309–322.

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., Camara, P. G., Camp, J. G., Chédotal, A., Copp, A., Etchevers, H. C., Giacobini, P., Göttgens, B., Guo, G., Hupalowska, A., James, K. R., Kirby, E., Kriegstein, A., Lundeberg, J., … Webb, S. (2021). A roadmap for the Human Developmental Cell Atlas. *Nature*, *597*(7875), 196–205.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Harrison, P. W., Wright, A. E., & Mank, J. E. (2012). The evolution of gene expression and the transcriptome–phenotype relationship. *Seminars in Cell & Developmental Biology*, *23*(2), 222–229.

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666–673.

He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A., & Patro, R. (2022). Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nature Methods*, *19*(3), 316–322.

Heimberg, G., Bhatnagar, R., El-Samad, H., & Thomson, M. (2016). Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Systems*, *2*(4), 239–250.

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., Weaver, B. D., Pflueger, C., Emery, B. R., Wilcox, A. L., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T., & Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, *49*(6), 925–934.

Hernandez, C., Wang, Z. [Zheng], Ramazanov, B., Tang, Y., Mehta, S., Dambrot, C., Lee, Y.-W., Tessema, K., Kumar, I., Astudillo, M. et al. (2018). Dppa2/4 facilitate epigenetic remodeling during reprogramming to pluripotency. *Cell Stem Cell*, *23*(3), 396–411.

Hernando-Herraez, I., Evano, B., Stubbs, T., Commere, P.-H., Jan Bonder, M., Clark, S., Andrews, S., Tajbakhsh, S., & Reik, W. (2019). Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nature Communications*, *10*(1), 4361.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2016). Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Hill, A. J., McFaline-Figueroa, J. L., Starita, L. M., Gasperini, M. J., Matreyek, K. A., Packer, J., Jackson, D., Shendure, J., & Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. *Nature Methods*, *15*(4), 271–274.

Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., … Lein, E. S. (2019). Conserved cell types with divergent features in human versus mouse cortex. *Nature*, *573*(7772), 61–68.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*.

Hoffman, M. M., Buske, O. J., & Noble, W. S. (2010). The Genomedata format for storing large-scale functional genomics data. *Bioinformatics*, *26*(11), 1458–1459.

Hoyer, S., & Hamman, J. (2017). Xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1), 10.

Huang, Y., & Sanguinetti, G. (2021). BRIE2: Computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biology*, *22*(1), 251.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Hyvarinen, A., & Morioka, H. (2016). Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. *Advances in Neural Information Processing Systems*, *29*.

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4), 411–430.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166.

Jiang, R., Sun, T., Song, D., & Li, J. J. (2022). Statistics or biology: The zero-inflation controversy about scRNA-seq data. *Genome Biology*, *23*(1), 31.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821.

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127.

Joung, J., Konermann, S., Gootenberg, J. S., Abudayyeh, O. O., Platt, R. J., Brigham, M. D., Sanjana, N. E., & Zhang, F. (2017). Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. *Nature Protocols*, *12*(4), 828–863.

Jukam, D., Shariati, S. A. M., & Skotheim, J. M. (2017). Zygotic Genome Activation in Vertebrates. *Developmental Cell*, *42*(4), 316–332.

Kærn, M., Elston, T. C., Blake, W. J., & Collins, J. J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nature Reviews Genetics*, *6*(6), 451–464.

Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: Accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data, 2021.05.05.442755.

Kanton, S., Boyle, M. J., He, Z., Santel, M., Weigert, A., Sanchís-Calleja, F., Guijarro, P., Sidow, L., Fleck, J. S., Han, D., Qian, Z., Heide, M., Huttner, W. B., Khaitovich, P., Pääbo, S., Treutlein, B., & Camp, J. G. (2019). Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, *574*(7778), 418–422.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC table browser data retrieval tool. *Nucleic acids research*, *32*, D493–D496.

Kashima, Y., Sakamoto, Y., Kaneko, K., Seki, M., Suzuki, Y., & Suzuki, A. (2020). Single-cell sequencing techniques from individual to multiomics analyses. *Experimental & Molecular Medicine*, *52*(9), 1419–1427.

Katzenelenbogen, Y., Sheban, F., Yalin, A., Yofe, I., Svetlichnyy, D., Jaitin, D. A., Bornstein, C., Moshe, A., Keren-Shaul, H., Cohen, M., Wang, S.-Y., Li, B., David, E., Salame, T.-M., Weiner, A., & Amit, I. (2020). Coupled scRNA-Seq and Intracellular Protein Activity Reveal an Immunosuppressive Role of TREM2 in Cancer. *Cell*, *182*(4), 872–885.e19.

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740–742.

Kim, H. J., Lin, Y., Geddes, T. A., Yang, J. Y. H., & Yang, P. (2020). CiteFuse enables multi-modal analysis of CITE-seq data. *Bioinformatics*, *36*(14), 4137–4143.

Kim, T., Lee, I., & Lee, T.-W. (2006). Independent Vector Analysis: Definition and Algorithms. *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, 1393–1396.

Kingma, D. P., & Welling, M. (2014). *Auto-Encoding Variational Bayes* (arXiv:1312.6114).

Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, *20*(5), 273–282.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74.

Klami, A., Virtanen, S., Leppäaho, E., & Kaski, S. (2014). Group Factor Analysis.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, *161*(5), 1187–1201.

Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., Elmentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M. S., Park, J. S., Ramona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Stegle, O., & Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature Biotechnology*, *40*(5), 661–671.

Ko, J., Wilkovitsch, M., Oh, J., Kohler, R. H., Bolli, E., Pittet, M. J., Vinegoni, C., Sykes, D. B., Mikula, H., Weissleder, R., & Carlson, J. C. T. (2022). Spatiotemporal multiplexed immunofluorescence imaging of living cells and tissues with bioorthogonal cycling of fluorescent probes. *Nature Biotechnology*, 1–9.

Ko, M. S. H. (2016). Chapter Three - Zygotic Genome Activation Revisited: Looking Through the Expression and Function of Zscan4. In M. L. DePamphilis (Ed.), *Current Topics in Developmental Biology* (pp. 103–124). Academic Press.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, *58*(4), 610–620.

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., Nureki, O., & Zhang, F. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). Automatic Differentiation Variational Inference.

Kwok, A. J., Allcock, A., Ferreira, R. C., Smee, M., Cano-Gamez, E., Burnham, K. L., Zurke, Y.-X., Oxford acute medicine/ED research, McKechnie, S., Monaco, C., Udalova, I., Hinds, C. J., Davenport, E. E., Todd, J. A., & Knight, J. C. (2022). *Identification of deleterious neutrophil states and altered granulopoiesis in sepsis* (preprint). Genetic and Genomic Medicine.

L. Lun, A. T., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, *17*(1), 75.

La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., Vinsland, E., Mossi Albiach, A., Mattsson Langseth, C., Khven, I., Lederer, A. R., Dratva, L. M., Johnsson, A., Nilsson, M., Lönnerberg, P., & Linnarsson, S. (2021). Molecular architecture of the developing mouse brain. *Nature*, *596*(7870), 92–96.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498.

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., de

Barbanson, B., Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31.

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(5), 056117.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., & Theis, F. J. (2022). CellRank for directed single-cell fate mapping. *Nature Methods*, *19*(2), 159–170.

Lause, J., Berens, P., & Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, *22*(1), 258.

Lawrence, N. (2003). Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. *Advances in Neural Information Processing Systems*, *16*.

Leader, A. M., Grout, J. A., Maier, B. B., Nabet, B. Y., Park, M. D., Tabachnikova, A., Chang, C., Walker, L., Lansky, A., Berichel, J. L., Troncoso, L., Malissen, N., Davila, M., Martin, J. C., Magri, G., Tuballes, K., Zhao, Z., Petralia, F., Samstein, R., … Merad, M. (2021). Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer Cell*, *39*(12), 1594–1609.e12.

Lebrigand, K., Magnone, V., Barbry, P., & Waldmann, R. (2020). High throughput error corrected Nanopore single cell transcriptome sequencing. *Nature Communications*, *11*(1), 4025.

Lee, D., & Seung, H. S. (2000). Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, *13*.

Lee, D., & Seung, S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Lee, S., Hou, K.-y., Wang, K., Sehrish, S., Paterno, M., Kowalkowski, J., Koziol, Q., Ross, R. B., Agrawal, A., Choudhary, A., & Liao, W.-k. (2022). A case study on parallel HDF5 dataset concatenation for high energy physics data analysis. *Parallel Computing*, *110*, 102877.

Lein, E., Borm, L. E., & Linnarsson, S. (2017). The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*, *358*(6359), 64–69.

Levitin, H. M., Yuan, J., Cheng, Y. L., Ruiz, F. J., Bush, E. C., Bruce, J. N., Canoll, P., Iavarone, A., Lasorella, A., Blei, D. M. [David M] et al. (2019). De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Molecular systems biology*, *15*(2), e8557.

Li, C., Virgilio, M., Collins, K. L., & Welch, J. D. (2021). Single-cell multi-omic velocity infers dynamic and decoupled gene regulation, 2021.12.13.472472.

Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., & Ren, B. (2019). Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, *16*(10), 991–993.

Li, H. [H.], Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, H. [Heng], & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760.

Li, L., Freitag, J., Asbrand, C., Munteanu, B., Wang, B.-T., Zezina, E., Didier, M., Thill, G., Rocher, C., Herrmann, M., & Biesemann, N. (2022). *Multi-omics profiling of collagen-induced arthritis mouse model reveals early metabolic dysregulation via SIRT1 axis* (preprint). Cell Biology.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, *18*(1), 51–60.

Liu, Q., & Wang, D. (2016). Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. *Advances in Neural Information Processing Systems*, *29*.

Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C. C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., Norris, E., Pan, A., Li, J., Xiao, Y., Halene, S., & Fan, R. (2020). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, *183*(6), 1665–1681.e18.

Lock, E. F., Hoadley, K. A., Marron, J. S., & Nobel, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, *7*(1), 523–542.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058.

Lotfollahi, M., Rybakov, S., Hrovatin, K., Hediyeh-zadeh, S., Talavera-López, C., Misharin, A. V., & Theis, F. J. (2022). Biologically informed deep learning to infer gene program activity in single cells.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lu, J., Cannizzaro, E., Meier-Abt, F., Scheinost, S., Bruch, P.-M., Giles, H. A. R., Lütge, A., Hüllein, J., Wagner, L., Giacopelli, B., Nadeu, F., Delgado, J., Campo, E., Mangolini, M., Ringshausen, I., Böttcher, M., Mougiakakos, D., Jacobs, A., Bodenmiller, B., … Huber, W. (2021). Multi-omics reveals clinically relevant proliferative drive associated with mTOR-MYC-OXPHOS activity in chronic lymphocytic leukemia. *Nature Cancer*, *2*(8), 853–864.

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular systems biology*, *15*(6), e8746.

Lun, A. T. L., & Marioni, J. C. (2017). Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data. *Biostatistics*, *18*(3), 451–464.

Lun, A. T. L., Riesenfeld, S., Andrews, T., Dao, T. P., Gomes, T., Marioni, J. C., & participants in the 1st Human Cell Atlas Jamboree. (2019). EmptyDrops: Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biology*, *20*(1), 63.

Ma, A., Xin, G., & Ma, Q. (2022). The use of single-cell multi-omics in immuno-oncology. *Nature Communications*, *13*(1), 2728.

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A., & Buenrostro, J. D. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, *183*(4), 1103–1116.e20.

Macaulay, I. C., Haerty, W., Kumar, P., Li, Y. I., Hu, T. X., Teng, M. J., Goolam, M., Saurat, N., Coupland, P., Shirley, L. M., Smith, M., Van der Aa, N., Banerjee, R., Ellis, P. D., Quail, M. A., Swerdlow, H. P., Zernicka-Goetz, M., Livesey, F. J., Ponting, C. P., & Voet, T. (2015). G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, *12*(6), 519–522.

Macaulay, I. C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S. A., & Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. *Cell Reports*, *14*(4), 966–977.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes,

J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.

Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., & Joung, J. K. (2013). CRISPR RNA–guided activation of endogenous human genes. *Nature Methods*, *10*(10), 977–979.

Manning, K. S., & Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. *Nature Reviews Molecular Cell Biology*, *18*(2), 102–114.

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., & Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, *24*(3), 496–510.

Martens, L. D., Fischer, D. S., Theis, F. J., & Gagneur, J. (2022). Modeling fragment counts improves single-cell ATAC-seq analysis, 2022.05.04.490536.

Mattanovich, D., & Borth, N. (2006). Applications of cell sorting in biotechnology. *Microbial Cell Factories*, *5*(1), 12.

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *33*(8), 1179–1186.

McGinnis, C. S., Murrow, L. M., & Gartner, Z. J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, *8*(4), 329–337.e4.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

McKinney, W. (2011). Pandas: A Foundational Python Library for Data Analysis and Statistics, 9.

McKinney, W. et al. (2011). Pandas: A foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, *14*(9), 1–9.

Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H. (, da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., & Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, *39*(7), 813–818.

Meng, C., Kuster, B., Culhane, A. C., & Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, *15*(1), 162.

Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, *9*(5), 989–1009.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419–D426.

Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., & Rätsch, G. (1998). Kernel PCA and De-Noising in Feature Spaces. *Advances in Neural Information Processing Systems*, *11*.

Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B., Thakore, P. I., Wing, J. B., Nazor, K. L., Sakaguchi, S., Ludwig, L. S., Sankaran, V. G., Regev, A., & Smibert, P. (2020). Scalable, multimodal profiling of chromatin accessibility and protein levels in single cells, 2020.09.08.286914.

Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences*, *113*(39), 11046–11051.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628.

Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., & Onnela, J.-P. (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, *328*(5980), 876–878.

Mulè, M. P., Martins, A. J., & Tsang, J. S. (2022). Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature Communications*, *13*(1), 2099.

Narayan, A., Berger, B., & Cho, H. (2021). Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, *39*(6), 765–774.

Nehar-Belaid, D., Hong, S., Marches, R., Chen, G., Bolisetty, M., Baisch, J., Walters, L., Punaro, M., Rossi, R. J., Chung, C.-H., Huynh, R. P., Singh, P., Flynn, W. F., Tabanor-Gayle, J.-A., Kuchipudi, N., Mejias, A., Collet, M. A., Lucido, A. L., Palucka, K., … Banchereau, J. F. (2020). Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nature Immunology*, *21*(9), 1094–1106.

Ortiz, C., Navarro, J. F., Jurek, A., Märtin, A., Lundeberg, J., & Meletis, K. (2020). Molecular atlas of the adult mouse brain. *Science Advances*, *6*(26), eabb3446.

Papalexi, E., & Satija, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology*, *18*(1), 35–45.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*, *6*(1), 25533.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pattaroni, C., Macowan, M., Chatzis, R., Daunt, C., Custovic, A., Shields, M. D., Power, U. F., Grigg, J., Roberts, G., Ghazal, P., Schwarze, J., Gore, M., Turner, S., Bush, A., Saglani, S., Lloyd, C. M., & Marsland, B. J. (2022). Early life inter-kingdom interactions shape the immunological environment of the airways. *Microbiome*, *10*(1), 34.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K. B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., … Amit, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, *163*(7), 1663–1677.

Pbmc granulocyte sorted 10k, 10x Genomics. (2022). Retrieved July 21, 2022, from https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k?

Pbmc protein 5k v3, 10x Genomics. (2022). Retrieved July 23, 2022, from https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.2/5k_pbmc_protein_v3

Peabody, D. S. (1993). The RNA binding site of bacteriophage MS2 coat protein. *The EMBO journal*, *12*(2), 595–600.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., Thompson, M., Rappoport, N., Dahl, A., Lanata, C. M., Matloubian, M., Maliskova, L., Kwek, S. S., Li, T., Slyper, M., … Ye, C. J. (2022). Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, *376*(6589), eabf1970.

Perkel, J. M. (2019). Julia: Come for the syntax, stay for the speed. *Nature*, *572*(7767), 141–142
Bandiera_abtest: a Cg_type: Toolbox Subject_term: Computer science, Computational biology and bioinformatics, Software.

Perkel, J. M. (2020). Why scientists are turning to Rust. *Nature*, *588*(7836), 185–186
Bandiera_abtest: a Cg_type: Technology Feature Subject_term: Computer science, Computational biology and bioinformatics, Software.

Persad, S., Choo, Z.-N., Dien, C., Masilionis, I., Chaligné, R., Nawy, T., Brown, C. C., Pe'er, I., Setty, M., & Pe'er, D. (2022). SEACells: Inference of transcriptional and epigenomic cellular states from single-cell genomics data, 2022.04.02.486748.

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., Moore, R., McClanahan, T. K., Sadekova, S., & Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*(10), 936–939.

Pfeiffer, M. J., Siatkowski, M., Paudel, Y., Balbach, S. T., Baeumer, N., Crosetto, N., Drexler, H. C. A., Fuellen, G., & Boiani, M. (2011). Proteomic Analysis of Mouse Oocytes Reveals 28 Candidate Factors of the "Reprogrammome". *Journal of Proteome Research*, *10*(5), 2140–2153.

Pfeiffer, M. J., Taher, L., Drexler, H., Suzuki, Y., Makałowski, W., Schwarzer, C., Wang, B., Fuellen, G., & Boiani, M. (2015). Differences in embryo quality are associated with differences in oocyte composition: A proteomic study in inbred mice. *PROTEOMICS*, *15*(4), 675–687
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201400334.

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1098.

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181.

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L., Reik, W., Srinivas, S., Simons, B. D., Nichols, J., Marioni, J. C., & Göttgens, B. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, *566*(7745), 490–495.

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C., & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, *360*(6391), eaaq1723.

Pokhilko, A., Handel, A. E., Curion, F., Volpato, V., Whiteley, E. S., Bøstrand, S., Newey, S. E., Akerman, C. J., Webber, C., Clark, M. B., Bowden, R., & Cader, M. Z. (2021). Targeted single-cell RNA sequencing of transcription factors enhances the identification of cell types and trajectories. *Genome Research*, *31*(6), 1069–1081.

Probst, A. V., & Almouzni, G. (2011). Heterochromatin establishment in the context of genome-wide epigenetic reprogramming. *Trends in Genetics*, *27*(5), 177–185.

Qiu, Q., Hu, P., Qiu, X., Govek, K. W., Cámara, P. G., & Wu, H. (2020). Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nature Methods*, *17*(10), 991–1001.

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi (Joseph) Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., … Weissman, J. S. (2022). Mapping transcriptomic vector fields of single cells. *Cell*, *185*(4), 690–711.e45.

Qoku, A., & Buettner, F. (2022). Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity.

R Core Team. (2013). *R: A language and environment for statistical computing*. manual. R Foundation for Statistical Computing. Vienna, Austria.

Raasveldt, M., & Mühleisen, H. (2019). DuckDB: An Embeddable Analytical Database. *Proceedings of the 2019 International Conference on Management of Data*.

Ramos, M., Schiffer, L., Re, A., Azhar, R., Basunia, A., Rodriguez, C., Chan, T., Chapman, P., Davis, S. R., Gomez-Cabrero, D., Culhane, A. C., Haibe-Kains, B., Hansen, K. D., Kodali, H., Louis, M. S., Mer, A. S., Riester, M., Morgan, M., Carey, V., & Waldron, L. (2017). Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Research*, *77*(21), e39–e42.

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, *30*(8), 777–782.

Razavi, A., van den Oord, A., & Vinyals, O. (2019). *Generating Diverse High-Fidelity Images with VQ-VAE-2* (arXiv:1906.00446).

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology*, *26*(3), 303–304.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., & Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, *9*(1), 284.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Rodriguez, L., Pekkarinen, P. T., Lakshmikanth, T., Tan, Z., Consiglio, C. R., Pou, C., Chen, Y., Mugabo, C. H., Nguyen, N. A., Nowlan, K., Strandin, T., Levanov, L., Mikes, J., Wang, J., Kantele, A., Hepojoki, J., Vapalahti, O., Heinonen, S., Kekäläinen, E., & Brodin, P. (2020). Systems-Level Immunomonitoring from Acute to Recovery Phase of Severe COVID-19. *Cell Reports Medicine*, *1*(5), 100078.

Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, *363*(6434), 1463–1467.

Rogozhnikov, A. (2022). Einops: Clear and Reliable Tensor Manipulations with Einstein-like Notation.

Rolinek, M., Zietlow, D., & Martius, G. (2019). Variational Autoencoders Pursue PCA Directions (by Accident). *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12398–12407.

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182.

Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A., & Teichmann, S. A. (2017). The Human Cell Atlas: From vision to reality. *Nature*, *550*(7677), 451–453.

Sadik, A., Somarribas Patterson, L. F., Öztürk, S., Mohapatra, S. R., Panitz, V., Secker, P. F., Pfänder, P., Loth, S., Salem, H., Prentzell, M. T., Berdel, B., Iskar, M., Faessler, E., Reuter, F., Kirst, I., Kalter, V., Foerster, K. I., Jäger, E., Guevara, C. R., … Opitz, C. A. (2020). IL4I1 Is a Metabolic Immune Checkpoint that Activates the AHR and Promotes Tumor Progression. *Cell*, *182*(5), 1252–1270.e34.

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554.

Sarkar, H., Srivastava, A., Bravo, H. C., Love, M. I., & Patro, R. (2020). Terminus enables the discovery of data-driven, robust transcript groups from RNA-seq data. *Bioinformatics*, *36*, i102–i110.

Sastry, A. V., Hu, A., Heckmann, D., Poudel, S., Kavvas, E., & Palsson, B. O. (2021). Independent component analysis recovers consistent regulatory signals from disparate datasets. *PLOS Computational Biology*, *17*(2), e1008647.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502.

Saul, L. K., Jaakkola, T., & Jordan, M. I. (1996). Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research*, *4*, 61–76.

Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., Goeva, A., Nemesh, J., Kamitaki, N., Brumbaugh, S., Kulp, D., & McCarroll, S. A. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*, *174*(4), 1015–1030.e16.

Scala, F., Kobak, D., Bernabucci, M., Bernaerts, Y., Cadwell, C. R., Castro, J. R., Hartmanis, L., Jiang, X., Laturnus, S., Miranda, E., Mulherkar, S., Tan, Z. H., Yao, Z., Zeng, H., Sandberg, R., Berens, P., & Tolias, A. S. (2021). Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature*, *598*(7879), 144–150.

Schafflick, D., Xu, C. A., Hartlehnert, M., Cole, M., Schulte-Mecklenbeck, A., Lautwein, T., Wolbert, J., Heming, M., Meuth, S. G., Kuhlmann, T., Gross, C. C., Wiendl, H., Yosef, N., & Meyer zu Horste, G. (2020). Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nature Communications*, *11*(1), 247.

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, O., Chen, M. B., Chen, S., Green, F., Jones, R. C., Maynard,

A., Penland, L., Pisco, A. O., Sit, R. V., Stanley, G. M., Webber, J. T., … Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367–372.

Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature Methods*, *14*(10), 975–978.

Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., & Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nature Methods*, *17*(6), 629–635.

Seeger, M., & Bouchard, G. (2012). Fast Variational Bayesian Inference for Non-Conjugate Matrix Factorization Models. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 1012–1018.

Sepp, M., Leiss, K., Sarropoulos, I., Murat, F., Okonechnikov, K., Joshi, P., Leushkin, E., Mbengue, N., Schneider, C., Schmidt, J., Trost, N., Spänig, L., Giere, P., Khaitovich, P., Lisgo, S., Palkovits, M., Kutscher, L. M., Anders, S., Cardoso-Moreira, M., … Kaessmann, H. (2021). Cellular development and evolution of the mammalian cerebellum, 2021.12.20.473443.

Shah, S., Lubeck, E., Zhou, W., & Cai, L. (2016). In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron*, *92*(2), 342–357.

Shapiro, M., Preguiça, N., Baquero, C., & Zawirski, M. (2011). Conflict-Free Replicated Data Types. In X. Défago, F. Petit, & V. Villain (Eds.), *Stabilization, Safety, and Security of Distributed Systems* (pp. 386–400). Springer.

Simmons, S. K., Lithwick-Yanai, G., Adiconis, X., Oberstrass, F., Iremadze, N., Geiger-Schuller, K., Thakore, P. I., Frangieh, C. J., Barad, O., Almogy, G., Rozenblatt-Rosen, O., Regev, A., Lipson, D., & Levin, J. Z. (2022). Single cell RNA-seq by mostly-natural sequencing by synthesis, 2022.05.29.493705.

Single Cell Multiome ATAC + Gene Expression, 10x Genomics. (2022). 10x Genomics. Retrieved July 21, 2022, from https://www.10xgenomics.com/products/single-cell-multiome-atac-plus-gene-expression

Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., & Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, *11*(8), 817–820.

Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U., & Zinovyev, A. (2019). Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *International Journal of Molecular Sciences*, *20*(18), 4414.

Soneson, C., & Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, *15*(4), 255–261.

Spitz, F., & Furlong, E. E. M. (2006). Genomics and Development: Taking Developmental Biology to New Heights. *Developmental Cell*, *11*(4), 451–457.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, *12*(1), 5692.

Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S.,

Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78–82.

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews Genetics*, *20*(11), 631–656.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868.

Stovner, E. B., & Sætrom, P. (2020). PyRanges: Efficient comparison of genomic intervals in Python. *Bioinformatics*, *36*(3), 918–919.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., & Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888–1902.e21.

Stuart, T., Srivastava, A., Madad, S., Lareau, C. A., & Satija, R. (2021). Single-cell chromatin state analysis with Signac. *Nature Methods*, *18*(11), 1333–1341.

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R. A., Stephenson, E., Engelbert, J., Tuong, Z. K., Polanski, K., Yayon, N., Xu, C., Suchanek, O., Elmentaite, R., Domínguez Conde, C., He, P., Pritchard, S., Miah, M., … Teichmann, S. A. (2022). Mapping the developing human immune system across organs. *Science*, *376*(6597), eabo0510.

Svensson, V. (2020). Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, *38*(2), 147–150.

Svensson, V., da Veiga Beltrame, E., & Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. *Database*, *2020*, baaa073.

Svensson, V., Gayoso, A., Yosef, N., & Pachter, L. (2020). Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, *36*(11), 3418–3421.

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604.

Swanson, E., Lord, C., Reading, J., Heubeck, A. T., Genge, P. C., Thomson, Z., Weiss, M. D., Li, X.-j., Savage, A. K., Green, R. R., Torgerson, T. R., Bumol, T. F., Graybuck, L. T., & Skene, P. J. (2021). Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq (J. K. Tyler, H. Y. Chang, & A. C. Adey, Eds.). *eLife*, *10*, e63632.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382.

Tang, L. (2021). Multiomics sequencing goes spatial. *Nature Methods*, *18*(1), 31–31.

Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M. N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K. E., Lathia, K., Rimorin, C., Tieu, M., … Zeng, H. (2018). Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, *563*(7729), 72–78.

The HDF Group. (2000–2010). *Hierarchical data format version 5*. http://www.hdfgroup.org/HDF5

Thornton, C. A., Mulqueen, R. M., Torkenczy, K. A., Nishida, A., Lowenstein, E. G., Fields, A. J., Steemers, F. J., Zhang, W., McConnell, H. L., Woltjer, R. L., Mishra, A., Wright,

K. M., & Adey, A. C. (2021). Spatially mapped single-cell chromatin accessibility. *Nature Communications*, *12*(1), 1274.

Torres-Padilla, M. E., & Zernicka-Goetz, M. (2006). Role of TIF1 as a modulator of embryonic transcription in the mouse zygote. *Journal of Cell Biology*, *174*(3), 329–338.

Townes, F. W., & Engelhardt, B. E. (2021). Nonnegative spatial factorization.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, *20*(1), 295.

Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*(1), 5233.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Research*, *25*(10), 1491–1498.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386.

Trevino, A. E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H. Y., Pașca, A. M., Kundaje, A., Pașca, S. P., & Greenleaf, W. J. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell*, *184*(19), 5053–5069.e23.

Triana, S., Vonficht, D., Jopp-Saile, L., Raffel, S., Lutz, R., Leonce, D., Antes, M., Hernández-Malmierca, P., Ordoñez-Rueda, D., Ramasz, B., Boch, T., Jann, J.-C., Nowak, D., Hofmann, W.-K., Müller-Tidow, C., Hübschmann, D., Alexandrov, T., Benes, V., Trumpp, A., … Haas, S. (2021). Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nature Immunology*, *22*(12), 1577–1589.

Tung, P.-Y., Blischak, J. D., Hsiao, C. J., Knowles, D. A., Burnett, J. E., Pritchard, J. K., & Gilad, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, *7*(1), 39921.

Ushey, K., Allaire, J., & Tang, Y. (2022). *Reticulate: Interface to 'python'*. manual.

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., & Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nature Methods*, *14*(6), 565–571.

Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., & Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature Communications*, *11*(1), 1201.

van den Oord, A., Vinyals, O., & Kavukcuoglu, k. (2017). Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems*, *30*.

Van der Maaten L, & Hinton G. (2008). Visualizing Data using t-SNE. *Journal of machine learning research*, (9), 11.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial* (Vol. 620). Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H., & van Oudenaarden, A. (2021). Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature*, *597*(7877), 561–565.

Vastenhouw, N. L., Cao, W. X., & Lipshitz, H. D. (2019). The maternal-to-zygotic transition revisited. *Development*, *146*(11), dev161471.

Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., & Stegle, O. (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods*, *19*(2), 179–186.

Verma, A., & Engelhardt, B. E. (2020). A robust nonlinear low-dimensional manifold for single cell RNA-seq data. *BMC Bioinformatics*, *21*(1), 324.

Vickovic, S. [S.], Lötstedt, B., Klughammer, J., Mages, S., Segerstolpe, Å., Rozenblatt-Rosen, O., & Regev, A. (2022). SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nature Communications*, *13*(1), 795.

Vickovic, S. [Sanja], Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J. F., Gould, J., Griffin, G. K., Borg, Å., Ronaghi, M., Frisén, J., Lundeberg, J., Regev, A., & Ståhl, P. L. (2019). High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, *16*(10), 987–990.

Vieth, B., Parekh, S., Ziegenhain, C., Enard, W., & Hellmann, I. (2019). A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature Communications*, *10*(1), 4667.

Villani, A.-C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., Jardine, L., Dixon, D., Stephenson, E., Nilsson, E., Grundberg, I., McDonald, D., Filby, A., Li, W., De Jager, P. L., … Hacohen, N. (2017). Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, *356*(6335), eaah4573.

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., & Wolf, F. A. (2021). Anndata: Annotated data, 2021.12.16.473007.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Virtanen, S., Klami, A., Khan, S., & Kaski, S. (2012). Bayesian Group Factor Analysis. *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, 1269–1277.

Wan, Y., Wei, Z., Looger, L. L., Koyama, M., Druckmann, S., & Keller, P. J. (2019). Single-Cell Reconstruction of Emerging Population Activity in an Entire Developing Circuit. *Cell*, *179*(2), 355–372.e23.

Wang, B. [Bingyuan], Pfeiffer, M. J., Drexler, H. C. A., Fuellen, G., & Boiani, M. (2016). Proteomic Analysis of Mouse Oocytes Identifies PRMT7 as a Reprogramming Factor that Replaces SOX2 in the Induction of Pluripotent Stem Cells. *Journal of Proteome Research*, *15*(8), 2407–2421.

Wang, B. [Bo], Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., & Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*(3), 333–337.

Wang, W., Tan, H., Sun, M., Han, Y., Chen, W., Qiu, S., Zheng, K., Wei, G., & Ni, T. (2021). Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Research*, *49*(9), e54.

Wang, X., He, Y., Zhang, Q., Ren, X., & Zhang, Z. (2021). Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics, Proteomics & Bioinformatics*, *19*(2), 253–266.

Wang, Z. [Zhong], Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wattenberg, M., Viégas, F., & Johnson, I. (2016). How to Use t-SNE Effectively. *Distill*, *1*(10), 10.23915/distill.00002.

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., & Macosko, E. Z. (2019). Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, *177*(7), 1873–1887.e17.

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61).

Wickham, H. (2021). *Mastering shiny*. " O'Reilly Media, Inc."

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018.

Winther, O., & Petersen, K. B. (2007). Bayesian independent component analysis: Variational methods and non-negative decompositions. *Digital Signal Processing*, *17*(5), 858–872.

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15.

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., Mburu, F. M., Mantalas, G. L., Sim, S., Clarke, M. F., & Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, *11*(1), 41–46.

Wu, T. D., Madireddi, S., de Almeida, P. E., Banchereau, R., Chen, Y.-J. J., Chitre, A. S., Chiang, E. Y., Iftikhar, H., O'Gorman, W. E., Au-Yeung, A., Takahashi, C., Goldstein, L. D., Poon, C., Keerthivasan, S., de Almeida Nagata, D. E., Du, X., Lee, H.-M., Banta, K. L., Mariathasan, S., … Grogan, J. L. (2020). Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature*, *579*(7798), 274–278.

Xi, N. M., & Li, J. J. (2021). Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Systems*, *12*(2), 176–194.e6.

Xu, C., Tao, D., & Xu, C. (2013). *A Survey on Multi-view Learning* (arXiv:1304.5634).

Yan, F., Powell, D. R., Curtis, D. J., & Wong, N. C. (2020). From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biology*, *21*(1), 22.

Yan, R., Gu, C., You, D., Huang, Z., Qian, J., Yang, Q., Cheng, X., Zhang, L., Wang, H., Wang, P., & Guo, F. (2021). Decoding dynamic epigenetic landscapes in human oocytes using single-cell multi-omics sequencing. *Cell Stem Cell*, *28*(9), 1641–1656.e7.

Yang, J., Rajan, S. S., Friedrich, M. J., Lan, G., Zou, X., Ponstingl, H., Garyfallos, D. A., Liu, P., Bradley, A., & Metzakopian, E. (2019). Genome-Scale CRISPRa Screen Identifies Novel Factors for Cellular Reprogramming. *Stem Cell Reports*, *12*(4), 757–771.

Yang, Z., & Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, *32*(1), 1–8.

Yi, L., Pimentel, H., Bray, N. L., & Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology*, *19*(1), 53.

Young, M. D., & Behjati, S. (2020). SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, *9*(12), giaa151.

Yu, C., Ji, S.-Y., Dang, Y.-J., Sha, Q.-Q., Yuan, Y.-F., Zhou, J.-J., Yan, L.-Y., Qiao, J., Tang, F., & Fan, H.-Y. (2016). Oocyte-expressed yes-associated protein is a key activator of the early zygotic genome in mouse. *Cell Research*, *26*(3), 275–287.

Zappia, L., Phipson, B., & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, *14*(6), e1006245.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., Manno, G. L., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., & Linnarsson, S. (2018). Molecular Architecture of the Mouse Nervous System. *Cell*, *174*(4), 999–1014.e22.

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., & Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, *40*(19), 9379–9391.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137.

Zhang, Z., Zhou, J., Tan, P., Pang, Y., Rivkin, A. C., Kirchgessner, M. A., Williams, E., Lee, C.-T., Liu, H., Franklin, A. D., Miyazaki, P. A., Bartlett, A., Aldridge, A. I., Vu, M., Boggeman, L., Fitzpatrick, C., Nery, J. R., Castanon, R. G., Rashid, M., … Callaway, E. M. (2021). Epigenomic diversity of cortical projection neurons in the mouse brain. *Nature*, *598*(7879), 167–173.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049.

Zhu, C., Preissl, S., & Ren, B. (2020). Single-cell multimodal omics: The power of many. *Nature Methods*, *17*(1), 11–14.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., & Siebert, P. D. (2001). Reverse transcriptase template switching: A SMART approach for full-length cDNA library construction. *BioTechniques*, *30*(4), 892–897.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., & Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, *65*(4), 631–643.e4.

žurauskienė, J., & Yau, C. (2016). pcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, *17*(1), 140.

Appendix A

MOFA+

A.1 Update equations

The update equations are applied at every iteration of the variational inference algorithm. The update equations for all the parts of the MOFA+ model are listed below.

Non-sparse factors Prior distribution:

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

Variational distribution:

$$q(z_{nk}^g) = \mathcal{N}(z_{nk}^g|\mu_{z_{nk}^g}, \sigma^2_{z_{nk}^g}),$$

where

$$\sigma^2_{z_{nk}^g} = \left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_d^{gm} \rangle \langle (w_{kd}^m)^2 \rangle + 1 \right)^{-1}$$
$$\mu_{z_{nk}^g} = \sigma^2_{z_{nk}^g} \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_d^{gm} \rangle \langle w_{kd}^m \rangle \left(\sum_{g=1}^{G} y_{nd}^{gm} - \sum_{j \neq k} \langle w_{jd}^m \rangle \langle z_{nj}^g \rangle \right)$$

Sparse factors Prior distribution:

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

Variational distribution:

$$q(s_{nk}^g) = \mathrm{Ber}(s_{nk}^g|\gamma_{nk}^g)$$

$$q(\hat{z}_{nk}^g|s_{nk}^g = 0) = \mathcal{N}\left(\hat{z}_{nk}^g \middle| 0, 1/\alpha_k^g\right)$$

$$q(\hat{z}_{nk}^g|s_{nk}^g = 1) = \mathcal{N}\left(\hat{z}_{nk}^g \middle| \mu_{z_{nk}^g}, \sigma^2_{z_{nk}^g}\right),$$

where

$$\gamma_{nk}^{g} = \frac{1}{1+\exp(-\lambda_{nk}^{g})}$$

$$\lambda_{nk}^{g} = \langle \ln \frac{\theta}{1-\theta} \rangle + 0.5 \ln \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle} - 0.5 \ln \left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle} \right)$$

$$+ \frac{\langle \tau_d^{gm} \rangle}{2} \frac{\left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} y_{nd}^{gm} \langle w_{kd}^m \rangle - \sum_{j \neq k} \langle s_{nj}^g \hat{z}_{nj}^g \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w_{kd}^m \rangle \langle w_{jd}^m \rangle \right)^2}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

$$\mu_{z_{nk}^g} = \frac{\sum_{m=1}^{M} \sum_{d=1}^{D_m} y_{nd}^{m,g} \langle w_{kd}^m \rangle - \sum_{j \neq k} \langle s_{nj}^g \hat{z}_{nj}^g \rangle \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle w_{kd}^m \rangle \langle w_{jd}^m \rangle}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

$$\sigma_{z_{nk}^g}^2 = \frac{\langle \tau_d^{gm} \rangle^{-1}}{\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle (w_{kd}^m)^2 \rangle + \frac{\langle \alpha_k^g \rangle}{\langle \tau_d^{gm} \rangle}}$$

ARD for the factors Prior distribution:

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

Variational distribution:

$$q(\alpha_k^g) = \mathcal{G}(\alpha_k^g \mid \hat{a}_{gk}^\alpha, \hat{b}_{gk}^\alpha),$$

where

$$\hat{a}_{gk}^\alpha = a_0^\alpha + \frac{N_g}{2}$$

$$\hat{b}_{gk}^\alpha = b_0^\alpha + \frac{\sum_{n=1}^{N_g} \langle (\hat{z}_{nk}^g)^2 \rangle}{2}$$

Factor sparsity priors Prior distribution:

$$p(\theta_k^g) = \mathscr{B}(\theta_k^g \mid a_0^\theta, b_0^\theta)$$

Variational distribution:

$$q(\theta_k^g) = \mathscr{B}(\theta_k^g \mid \hat{a}_{gk}^\theta, \hat{b}_{gk}^\theta),$$

where

$$\hat{a}_{gk}^\theta = \sum_{n=1}^{N_g} \langle s_{nk}^g \rangle + a_0^\theta$$

$$\hat{b}_{gk}^\theta = b_0^\theta - \sum_{n=1}^{N_g} \langle s_{nk}^g \rangle + N_g$$

Non-sparse weights Prior distribution:

$$p(w_{kd}^m) = \mathscr{N}(w_{kd}^m | 0, 1)$$

Variational distribution:

$$q(w_{kd}^m) = \mathscr{N}(w_{kd}^m | \mu_{w_{kd}^m}, \sigma_{w_{kd}^m}^2),$$

where

$$\sigma_{w_{kd}^m}^2 = \left(\sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle \tau_d^{gm} \rangle \langle (z_{nk}^g)^2 \rangle + 1 \right)^{-1}$$

$$\mu_{w_{kd}^m} = \sigma_{w_{kd}^m}^2 \sum_{g=1}^{G} \sum_{n=1}^{N_g} \langle \tau_d^{gm} \rangle \langle z_{nk}^g \rangle \left(\sum_{m=1}^{M} y_{nd}^{gm} - \sum_{j \neq k} \langle z_{nj}^g \rangle \langle w_{jd}^m \rangle \right)$$

Sparse weights Prior distribution:

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

Variational distribution:

$$q(s_{kd}^m) = \mathrm{Ber}(s_{kd}^m|\gamma_{kd}^m)$$

$$q(\hat{w}_{kd}^m|s_{kd}^m = 0) = \mathcal{N}\left(\hat{w}_{kd}^m\middle|0, 1/\alpha_k^m\right)$$

$$q(\hat{w}_{kd}^m|s_{kd}^m = 1) = \mathcal{N}\left(\hat{w}_{kd}^m\middle|\mu_{w_{kd}^m}, \sigma_{w_{kd}^m}^2\right),$$

where

$$\gamma_{kd}^m = \frac{1}{1+\exp(-\lambda_{kd}^m)}$$

$$\lambda_{kd}^m = \left\langle \ln \frac{\theta}{1-\theta} \right\rangle + 0.5 \ln \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} - 0.5 \ln \left(\sum_{g=1}^G \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle} \right)$$

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

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

$$\sigma_{w_{kd}^m}^2 = \frac{\langle \tau_d^{gm} \rangle^{-1}}{\sum_{g=1}^G \sum_{n=1}^{N_g} \langle (z_{nk}^g)^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^{gm} \rangle}}$$

ARD for the weights Prior distribution:

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

Variational distribution:

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

where

$$\overset{\alpha}{}_{mk} = a_0^{\alpha} + \frac{D_m}{2}$$

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

Weight sparsity priors Prior distribution:

$$p(\theta_k^m) = \mathcal{B}(\theta_k^m \mid a_0^{\theta}, b_0^{\theta})$$

Variational distribution:

$$q(\theta_k^m) = \mathcal{B}(\theta_k^m \mid \hat{a}_{mk}^{\theta}, \hat{b}_{mk}^{\theta}),$$

where

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

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

Noise Prior distribution:

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

Variational distribution:

$$q(\tau_d^{gm}) = \mathscr{G}(\tau_d^{gm} \mid \hat{a}_d^{gm}, \hat{b}_d^{gm}),$$

where

$\hat{a}_d^{gm} = a_0^{\tau} + \frac{N_g}{2}$

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

A.2 Expectations equations

Non-sparse factors:

$$\mathbb{E}_q[\ln p(Z)] = -\tfrac{NK}{2}\ln(2\pi) - \tfrac{1}{2}\textstyle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\langle(z_{nk}^g)^2\rangle$$

$$\mathbb{E}_q[\ln q(Z)] = -\tfrac{NK}{2}(1+\ln(2\pi)) - \tfrac{1}{2}\textstyle\sum_{g=1}^{G}\sum_{n=1}^{N_g}\sum_{k=1}^{K}\ln(\sigma_{z_{nk}}^2)$$

Sparse factors:

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

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

ARD for the factors:

$$\mathrm{E}_q[\ln p(\alpha)] = \textstyle\sum_{g=1}^{G}\sum_{k=1}^{K}\left(a_0^\alpha\ln b_0^\alpha + \ (a_0^\alpha-1)\langle\ln\alpha_k\rangle - b_0^\alpha\langle\alpha_k\rangle - \ln\Gamma(a_0^\alpha)\right)$$

$$\mathrm{E}_q[\ln q(\alpha)] = \textstyle\sum_{g=1}^{G}\sum_{k=1}^{K}\left(\hat{a}_k^\alpha\ln\hat{b}_k^\alpha + (\hat{a}_k^\alpha-1)\langle\ln\alpha_k\rangle - \hat{b}_k^\alpha\langle\alpha_k\rangle - \ln\Gamma(\hat{a}_k^\alpha)\right)$$

Factor sparsity priors:

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

Non-sparse weights:

$$\mathbb{E}_q[\ln p(W)] = -\frac{DK}{2}\ln(2\pi) - \frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\langle(w_{kd}^m)^2\rangle$$
$$\mathbb{E}_q[\ln q(W)] = -\frac{DK}{2}(1+\ln(2\pi)) - \frac{1}{2}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\sum_{k=1}^{K}\ln(\sigma_{w_{kd}}^2)$$

Sparse weights:

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

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

ARD for the weights:

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

Weight sparsity priors:

$$\mathbb{E}_q\left[\ln p(\theta)\right] = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}((a_0-1)\times\langle\ln(\pi_{d,k}^m)\rangle + (b_0-1)\langle\ln(1-\pi_{d,k}^m)\rangle - \\ -\ln(\mathrm{B}(a_0,b_0)))$$

$$\mathbb{E}_q\left[\ln q(\theta)\right] = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{d=1}^{D_m}((a_{k,d}^m-1)\times\langle\ln(\pi_{d,k}^m)\rangle + (b_{k,d}^m-1)\langle\ln(1-\pi_{d,k}^m)\rangle - \\ -\ln(\mathrm{B}(a_{k,d}^m,b_{k,d}^m)))$$

Noise:

$$\mathbb{E}_q[\ln p(\tau)] = \sum_{m=1}^{M} D_m a_0^\tau \ln b_0^\tau + \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm}(a_0^\tau-1)\langle\ln\tau_d^{gm}\rangle - \\ -\sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{Dm} b_0^\tau\langle\tau_d^{gm}\rangle - \sum_{m=1}^{M} D_m\ln\Gamma(a_0^\tau)$$

$$\mathbb{E}_q[\ln q(\tau)] = \sum_{g=1}^{G}\sum_{m=1}^{M}\sum_{d=1}^{D_m}\left(\hat{a}_{dgm}^\tau\ln\hat{b}_{dgm}^\tau + (\hat{a}_{dgm}^\tau-1)\langle\ln\tau_d^{gm}\rangle - \hat{b}_{dgm}^\tau\langle\tau_d^{gm}\rangle - \ln\Gamma(\hat{a}_{dgm}^\tau)\right)$$

Appendix B

MOFA+ application for CRISPRa screens

B.1 sgRNA sequences and target genes

Nucleotide sequence and target gene for the 475 sgRNAs used in the study are listed below.
Table B.1 Sequence and target gene for 475 sgRNAs

1	Baz2b	CAACTTTCATTGTTAGAGAG	36	Arid4a	CTCATGACGCCGAATAGCTT
2	Baz2b	GCTGATAGGTCATTTATTAT	37	Tsc22d2	GCCCAGAGCAGCCCGTGACT
3	Fam170a	AGGAACTGTTCTTGTAAACA	38	Tsc22d2	GCCCTTCTCGGTTTCCGGTG
4	Fam170a	GAGAACTGAGTTCCTTAGAG	39	Pbrm1	CGGGATCACCACGACTTCTC
5	Mdc1	AGTTACCGGCGGCGGCTAGA	40	Pbrm1	GGTTCAACGTCACGCGCTCC
6	Mdc1	CAAACGCCTACCAGAGAGTC	41	Chd2	ATGTAGAGAGTGTCAATCAA
7	Ttf2	AGACTCTCCCGCTTGATTGG	42	Chd2	GCTCTCTTGGATGGAGAATC
8	Ttf2	GCGCGCTCTGAAATCAAACT	43	Dux	CAGAGGCAGAGGTATTTAAG
9	Crebbp	CGCGCTCGGGAGGCGGGCAG	44	Dux	GGTGCCAATGACGGCCTCCC
10	Crebbp	GGGCGCCGTCGCTGCCAGCC	45	Ubp1	GCACTGGAGGAAGTGGTTCC
11	Lin28b	ACAGAGTCACGTGTGCTCAG	46	Ubp1	GGCCAATGAGCTCTTACTTA
12	Lin28b	TTCTTGCTACTAATTCACCT	47	Arid1b	ACCCTCCTTGAGAGCCGCCT
13	Zfp407	AAGTGGCTCGACGGAGAGGC	48	Arid1b	GCGGCCGCTCTGCGCGCTGT
14	Zfp407	GAATGTTAGGTCATTAGCCA	49	Lin9	GAACTGCAGGCCGGTTCCCG
15	Lyg2	ACCTGATTATAGTGCTACTT	50	Lin9	GTGGCGCGGCGGACGAGACC
16	Lyg2	GAGGTCAATGCTTTAAAGAG	51	Fiz1	GGCTAGCGCCATCTTTGTTG
17	Arnt	CCGGTGACCCGAGGAGTGCT	52	Fiz1	TCCTTTCCTATTTCCAGCCG
18	Arnt	CTGACCGCGCCCATAGTTTG	53	Sin3a	GAAAGCCAATAGATACTGGG
19	Akap1	AGGCGGGCCCGTACAGCTCC	54	Sin3a	TTGGAGGCGACTCTTACTCG
20	Akap1	GGAGCTCTCCGCTGACATCC	55	Brd3	CGCGGACTACGACTCCCGCC
21	Zscan21	CAACGCTCGGATTCCTAGAG	56	Brd3	GACATGTAGTCCGTGCGCCT
22	Zscan21	GTCGCAGCCTCCGGTGAAAC	57	Tcf20	CTCAGCTGATGTCACCGCCC
23	Bahd1	GCGGCCGGACGGGCGAGCAG	58	Tcf20	GCAGAGGTGTGATCCGGCAG
24	Bahd1	GCGTAGCGGATCCCGGAGCC	59	Tnrc18	AAGTGTGAGGCGGCTGCGGG
25	Nfyc	GCTCTTCAGGTGCAGTCTTC	60	Tnrc18	GAGCTCGGCGGGAGGAAAGC
26	Nfyc	GTACCACCACACCTGAATCT	61	Sirt2	GATTGGCTGTGGTGCAGTAC
27	Cbx5	CGCTGGTCCCGCCACTACGC	62	Sirt2	TTGTAGTTTGATAACATGAG
28	Cbx5	GCGCAGGGAAGGATCCGTTT	63	Gmeb1	GTCCTTATTGCGGGCCGCGC
29	Trim33	GAAGACTTCTGTCTCACCGC	64	Gmeb1	TGCGCACAAAGCCGCCACGG
30	Trim33	TCACAGCCCTCTAGAGAACG	65	Trp63	GCTGAAAGGGAGGCAGAAGG
31	Arid1a	GAGAAGACGAAGACAGGGCC	66	Trp63	TGAAATGAAGAGTGAGTTCT
32	Arid1a	GAGCGAGCGCAGCGCAGCAG	67	Cecr2	GGGCCCGGCAGAAACCTGGG
33	Foxk2	CCGGCGTCGTCGCGCGACGC	68	Cecr2	GTTGTTGTTGTGGTGCGCGC
34	Foxk2	CGTGCGCCGCGCCCATTGGT	69	Max	ACTACAAGTCTCAGCTACCC
35	Arid4a	AGCTGGGAAACTTCCATGAT	70	Max	AGTCTGGCGAAGCCTGTGTG

71	Irf1	GGTGGCGCGCTGGCTCGGGA	111	Chd7	CTTCCCGTCCCGCCGGGCGC
72	Irf1	TGAGCTCGCCGGAGCTGCGC	112	Chd7	GGGCGCGCCGCGGAGAGAGC
73	Sirt1	AGGCGCCAGAGGCCGCTGAG	113	Terf2	AGCATTGCAGCAGACACAAG
74	Sirt1	GAGCCGAGCCGCCCTGGGTG	114	Terf2	TGAAGACTACGCTGCAGACT
75	Pou2f2	CCGGGAGGAAACCACAGGGC	115	Mier1	TCTCTGTCTCTTTGCTACTG
76	Pou2f2	TGACTTCCGGCAGGGAGGGC	116	Mier1	TGTGAGACTGAGTGAGCTTG
77	Satb1	CCCGCAAGCCAGCTTGTGTG	117	Nfx1	AAAGCAATTACGCAAGTTGG
78	Satb1	CCTCCTTCCCTTCAAACTGA	118	Nfx1	CGCCATTTCACTGTGTGCTC
79	Zfp512b	AGGCGGGCTGGGCACAGGGC	119	Cux1	CGTGAAAGTGACTGCGGAGC
80	Zfp512b	TTTCTGGAGGAGGAGGCGGG	120	Cux1	GACCCGCCACACCGGGTAGC
81	Hmga1	AACAGCAGGGCCGCCGTCAC	121	Tfdp1	CGAGAGGCGCGGGAGATGAG
82	Hmga1	GTAGCTCGCCACTGTGCGCG	122	Tfdp1	GGTGCTACCCGCCCTGGGAG
83	Phf14	CTAGTTCTTAGGCAGATGGA	123	Hmgb3	CAACCTCCTCGCCTGGCCAT
84	Phf14	TACTATGGAAATCAGAGTAC	124	Hmgb3	CATTCAAACAGAACAACCCG
85	Zfp57	CTAAGATAAGTATTTACTCC	125	Preb	GGCAGCTTCATCGTCATTCG
86	Zfp57	TTTCCGTAATTCCTTGTACT	126	Preb	TAGGACCCGGATCCGACGCC
87	Mta3	CGTCGCTCTGCCGGCGCTGG	127	Usp3	CCCACACTAGTCAGTGGGCG
88	Mta3	GGGTAGCACACCAGGGTCAG	128	Usp3	CGTAACCAGCAGGGTGGGCG
89	Yap1	AAGGAAGAGCGTCGAGGAGG	129	Phf1	CAGCTAACGGATGTGGCATG
90	Yap1	CGCGCTGCGTGCGCACCCAC	130	Phf1	GGAGCTTGGAGTGGCCACGG
91	Smarca4	AGTCCTTTGAGGGCGCGAGG	131	Ing1	GCGACACAAAGGGACGATGG
92	Smarca4	CCAGAGGCTGCACAAGATCC	132	Ing1	TCCCGGTGCTGGGTTTAGAA
93	Dnmt1	AACATCTCCAATACCCTGAT	133	Cic	CGCAGCCAATCAGCAACGGC
94	Dnmt1	AGCATGGCCCGCTCTGTGCC	134	Cic	TAAACCAGAGTAGAGAAGTT
95	Jarid2	ATTGTAGTTCTTCGCTTCCG	135	Ubtf	GCCGCCCGGAGGAGGCGGAG
96	Jarid2	GTTGAGTGACAGGAGGCGTG	136	Ubtf	GCCTTGGCGCCGCCGCCTCC
97	Eya1	CAATTGAGAGGAGCAGAGCT	137	Csnk2b	CGCGGGAAACTTGAACCTTA
98	Eya1	CTGTTTATTCAGATATCATG	138	Csnk2b	GCACGCGATGCTTTACTACG
99	Shmt2	AGCTTCAAGAAAGGACATGG	139	Cbx1	CTCCACTTTCAAGGATTGTA
100	Shmt2	TCCCTGCGCATGCGCAGTAC	140	Cbx1	TGGGCGGCGCGGGATTCGCG
101	Mllt10	GAAAGCCTGGAGGTAGGCAG	141	Cbx2	GCTGCGTGCTCCGGGAATCC
102	Mllt10	GCGGAGTCGGCTGCTCAGCG	142	Cbx2	GGGCGCTCTCTGCACTGTGT
103	Patz1	GTGCTACGCGGGTCCGGCGG	143	Cbx3	CCCGCCGGGAACGCGTCCTG
104	Patz1	GTTAAGGGCGGCAATGTGGG	144	Cbx3	GCGCGCTCCGAGACGCAGAG
105	Tdrd3	TGACCAAGCAGCCCGGTCCC	145	Dhx9	CGCGCGAGATCTGCTGGCCT
106	Tdrd3	TTCTAAGTTGCACACACATG	146	Dhx9	GTCGGCGACTCGCGCCTGCG
107	Hdac9	AAGCAGAAAGAAGTTAATGC	147	Dnajc1	AGAGAACAGCTTTCCGTGGC
108	Hdac9	GTTCTGCTCAGACAATAAGA	148	Dnajc1	AGCTGCACGTGTGTCAGCGC
109	Trim24	CCCGCCGGACGCGAGGACCG	149	E2f5	CGTTACTAGGCAACGCCGCG
110	Trim24	CTGCAGAAGCCCGCGAGGCT	150	E2f5	GGCAGTCGGGTTCCCTCAGG

151	Ezh2	AGGCGCTTGATAGTGCTGGG	191	Adnp	CCGAAGAGCTCTCCCTTGGG
152	Ezh2	TTCGGAGCGAGCTCCAGCCC	192	Adnp	GGACTCATTGTCCAGCTGGG
153	Gata3	CTGCAACCCAAACCCGCTCC	193	Brca1	ACAAGATCCAGACACGTCAG
154	Gata3	TGGCGCGACGCAACTTAAGG	194	Brca1	ACGGAAGAGAAAGAATTTCG
155	Nr3c1	CTTTGTGGATGGGAGCAGGG	195	Creb1	AGCTCGGCTGTTTCCGTGAG
156	Nr3c1	GCCTATTAAATAAACCCTTT	196	Creb1	GAAGGCCTTCGGCTACTTCC
157	Hdac2	ACGGGTAGTCACACACAGTC	197	Gbx2	GGAGCGCGATTTAAAGGTGC
158	Hdac2	AGTCCGCTTGTGCGCACCTC	198	Gbx2	TGGGCTACCGGCGCCACGGT
159	Hmgb2	ATGTACTGCTCGGCCCTGAT	199	Lhx2	AGCCACTGGGAGAATTGAAG
160	Hmgb2	CTCTCTAGACCGGTTCCAAC	200	Lhx2	GGCTCAGCTCGCAGGCTTCG
161	Hsf1	CTCTGATTGGTGAGCAGCCC	201	Smad2	GACGTTTCCCTTCCGATTCC
162	Hsf1	TGGGCCGGCAAAGTAGGCAG	202	Smad2	GGGTGTGAGCACCGCGTCGG
163	Smad1	CGCGGTGCGCGTGTGAGCGA	203	Msh6	CGGCCAATCCGAAGGCGTGT
164	Smad1	GAAGGCGCGGGCCGGTAATT	204	Msh6	TGGAGCTCGCGGTGGCTACC
165	Smad4	GGACACTGAGCATGCGCGCG	205	Nfatc3	GATAGGTCGGTGAGGAGGCG
166	Smad4	TGGCAGCAACAACACGGCCC	206	Nfatc3	GCCAGAGCCCTGCTCCCTCT
167	Mybl2	AGCCCGCGCGCCAGAACCCT	207	Nfya	AAGGCGCCTGCGTAGGCCTT
168	Mybl2	CTATCTCCCGCCAAGTGCGT	208	Nfya	GCGTGGTATATAATTCAGGG
169	Ncoa3	CGCCCAGCACTACGGGTCCC	209	Nfyb	AGCCTTTCCGCGTGGAGCTC
170	Ncoa3	GGAGGTGAAGAGGACGTTCT	210	Nfyb	GCTCTGGATGTTCGCTCTTC
171	Rb1	GAGACAGGCCCGGGCAGGCG	211	Uhrf1	CTATGAGGTTTACATGGTGG
172	Rb1	TCGCGGGCCGCGGGCGTCAG	212	Uhrf1	GAGGGCTCTTGATTCCTGAT
173	Rbbp4	ATTGGTCGGTGTTGGCGCGA	213	Nrf1	CCGCCCGGAGACCGCCATGA
174	Rbbp4	TGGGCGGAGCATCCGCCTCG	214	Nrf1	GCGAGGCCAGCACGGCTCTG
175	Rela	GATGTCACCCTGGCCGGGAC	215	Phf2	GGCGCAGGGCTGAGAGTGCG
176	Rela	TGCGGCCATTCGCCCTGCGC	216	Phf2	TTTGTGTCCCGCGCGCGGCC
177	Rfx1	GGAGCTCGGTGAAGTCGAGA	217	Pou2f1	GCCTCGGGAGAGGAAGGAAC
178	Rfx1	GGGCGCTTCCAGCATGCTAT	218	Pou2f1	TGCGCACCGCGCGCGCAGGG
179	Smarcc1	CCGCCCAGCCAACCCTCTTC	219	Rest	GACGGCGCGGGCCGGGTGCG
180	Smarcc1	CGGGCAGCCTCAGTCCGTAG	220	Rest	GGCGCGCGCGGACGGCGAGG
181	Terf1	CGCCTGGGCTCGGATCGACA	221	Sub1	AGTGAGATGCCACGCAGCTT
182	Terf1	TTTAGCAGTTCTCAGCCAAT	222	Sub1	CGTGCCATTGGGCGGAGCCA
183	Tfam	CTGTGGGAAGCCCGGACTTC	223	Smarca2	CCCGAGTTTAGGAAGAGGAG
184	Tfam	GGCGGAGCCTGAGGCTAGGG	224	Smarca2	GCGTCTTCCGGCGCCCGTGG
185	Usf1	TGGCTGGCGTTTCCCAGGCC	225	Sox2	CTGGGCTCGGGCGCAGGAGC
186	Usf1	TGGGCACGGAGCCTTTGTTT	226	Sox2	GGCGGCCAATCAGCGAGCGC
187	Xrcc1	CGCCGGGTTCCCTTAGCAAC	227	Tbx3	GAGTGATTGGGAGCTGGAGT
188	Xrcc1	GTTGGGCCTCTCCGCGAGCG	228	Tbx3	GCGAGGGAGGATCAAGAAGA
189	Dnajc2	ACGTAGGAACTGATGCGTCT	229	Trim28	GCCACGGCAGAGCAAAGCGC
190	Dnajc2	GCATTGGCGCGAGGCCGTGG	230	Trim28	GCGGCAGCCTGGCCTATAGT

231	Nr2c2	CCGCCGGAGCTGAGGCCCGC	271	Pias4	ATTGGTCAGATCCTCGGGCG
232	Nr2c2	CTCGGCGCGGCGCCAGGCAA	272	Pias4	CTCCCGCCTCTCAGACGCGG
233	Tulp3	CATGATGCTGGAGGGTTTCC	273	Carm1	CGGGCGAGCGCAAGCACGAC
234	Tulp3	TGGGCGTCGGCGACAGTGCG	274	Carm1	CTCCCTCAGGCGCCAGGAGG
235	Ybx1	AAAGATACCAATGAGAGCGC	275	Ing3	GGCTCCACTTTGGCGCCTGA
236	Ybx1	GCCTATTGGCTCACGCTCCG	276	Ing3	TCTGCAGCGATGACGTCAAC
237	Esrrb	ATGTATGATTTGGCTCCTAT	277	Mphosph8	GGGAGTTCACGGGAAGTGGA
238	Esrrb	TTAAAGTGACAGAATCGAGC	278	Mphosph8	GGGCGGGTCGACGGACTCTG
239	Mbd3	CGTCTGCGCAGGCGTGGAGT	279	Mgea5	CTCCTTCCGTGGAGACGCGG
240	Mbd3	TGATTGGCTTCACAGCCTTG	280	Mgea5	GACGAGCGAAGGCGCCTCCT
241	Pou5f1	ATCTGCCTGTGTCTTCCAGA	281	Phf10	CAGCGCGCTGGCGGCCATGG
242	Pou5f1	GTCTGGACAGGACAACCCTT	282	Phf10	GTCAGCCGGCCGCAGCTGCG
243	Mga	GCGGTCGGCTCCAAGCTGGT	283	Ing5	GCTCATGAATAGTGAGCGCG
244	Mga	TACGGCTTCGGCTCCGTGGG	284	Ing5	TTGCCCTGAGGGAACGCGGA
245	Prmt5	ACAAATGGACTGGGCATTCT	285	Hmg20a	TCCTCCGCAGTACTTTATTT
246	Prmt5	ATGAGAGTCTGTCATTGGTG	286	Hmg20a	TCTGTCAAGCTGTGAGACGA
247	Mtf2	GAATTAAGTCCTCTTGCTGC	287	Carhsp1	GCGCGGCCCGGACCTTGGAC
248	Mtf2	GCTCACTATTGTGTCTGCAC	288	Carhsp1	TCCAGCCGCTGCCGAGTCCC
249	Smad3	GGCCAGCAAAGTTTGGCTGG	289	Mina	AGGTTCATAGGCCAGAGCAG
250	Smad3	TCTGCGCACCAAAGCCATCC	290	Mina	GAAGGCCTGTAGCTGTAGGC
251	Ybx2	TCGGCTGAAGCGGCGCGCTC	291	Hat1	TAAGTCCTTGCGGTCACTTC
252	Ybx2	TGGCCTCGGCTCAGGGCCCG	292	Hat1	TTCTCCCTCTCATTGGTTCG
253	Rfx5	TCTCTAGGAGAAAGGAAGCT	293	Calcoco1	CCTGTGGAACTACATGTCCC
254	Rfx5	TTTAGAGAAGGCAAGTGGTG	294	Calcoco1	GGGCGGTGATGGACTCAGCG
255	Smarcal1	ATCACCCACAATTTATGACC	295	Rnmt	AGGAAGCCACGCTCCCAGTC
256	Smarcal1	ATTTGAGAGGCGTAACGTTG	296	Rnmt	CCGGGACTTAGTCAATTCCA
257	Nfat5	GGCCGCTGCTGATCCCGGGC	297	Smyd2	GAGCGAAGGCAGGTCTGCCG
258	Nfat5	TCCGGGAGCCCGGAAGAGCC	298	Smyd2	GCACGCGCGTCACGGGAGGG
259	Setdb1	AGAGTATGCATGTGTGAAGT	299	Spata3	AATTAGACCATGGAGGAACA
260	Setdb1	GCAATAAATTTGGTCCCTTC	300	Spata3	GTCTTTGGGCCTAAGAGAAC
261	Arid3b	AACTCATATGGGTTGGAGGG	301	Smyd3	CAGGTCCTGTGAGCATGCGC
262	Arid3b	TCATCCTGAAAGTCAGGGTT	302	Smyd3	TGTACGCAAGCGCAGGACGC
263	Foxo3	CGCACGCACAGCAGCCCTCC	303	Zmynd8	GGTGGGAGAAGTGCTTGTGG
264	Foxo3	TGTGCGCCCTCGCGCCTGCG	304	Zmynd8	TTTGCTGGAGCAGGTCAGGC
265	Snd1	CTGGAGCTCCGCACCCGTTT	305	Atad2	ACCGGAGCCGACGCGAACCG
266	Snd1	TTCATAAAGAGAGTGCCTAT	306	Atad2	CTCCTCCAATCCGAAGAGCG
267	Brd4	AGCGGCGGGCGCAACTGCCC	307	Prdm16	GCCGCCTCGGCGCGGGATTG
268	Brd4	ATGGCGGCGGCAGGAAGTGC	308	Prdm16	GGTGAAGACCGAGAAGGCCT
269	Smarce1	GAACTGTTTAAAGACAAGCG	309	Taf3	CTGCTCGTACTGCCTCGCCC
270	Smarce1	TGAACTGGCCCTCCTGGGAA	310	Taf3	GGCCGAGAGATAGGCGGCCC

311	Ints12	ACGTTTCCGAGTGAGACAGC	351	Aff4	CTGCGGCCCTGCTGCTGCTC
312	Ints12	GAAACAAAGGTTCCCGCAAC	352	Aff4	TGGCCGCGGCCCATGTGACC
313	Setd3	CCAGGGTGCTGATCAGCCAT	353	Smarca5	GCGTAGGTAGCTGTCCTGGT
314	Setd3	CTTCGGATACGGCTCCGCCC	354	Smarca5	TGGAGAAGAGAATCCACTCC
315	Zfp655	CCACTCTCTCGACTGCCAGA	355	Baz2a	CTCAGGGACGGCTGTCGCGA
316	Zfp655	GGCGACGGAAGTCCACGGGA	356	Baz2a	GACGGCGGGTGGCGTCTCCT
317	Tdrkh	CTGCGGCGGAGGAGCGCGAT	357	Mta1	CGCGGCCTCGGCGGCCTCGG
318	Tdrkh	TTGGTGAGGATGCGCTGCGG	358	Mta1	GGGACGGGAGGGCCGCGGCG
319	Trmt11	CTCCAAGTGGAGTATCCCTG	359	Ing4	CCTTACTAAACTATCTGCCT
320	Trmt11	TAAGGGCAAGAGTGAGATCC	360	Ing4	CGGAAGTTCAGGTACGGAAA
321	Dppa4	GTTGAGGGTGGGACCAGAAG	361	Acaca	AGTAAAGCATAAACCACAAG
322	Dppa4	ATCACAATTTTGTTGAGGGT	362	Acaca	TCTTGTGTGTTCTCATCTAT
323	Dppa2	ACACAGGAGGACCCTCCCTC	363	Prmt3	CCGAGGCGCCGGCGCTTACA
324	Dppa2	CCGGATGCAGAAATAGCAGC	364	Prmt3	CGACTTATCGCAGCGCCCTC
325	Elp3	GGATGTGCACAGTGATTTAG	365	Lbr	GCCCTCCTGCGGAGCGATCT
326	Elp3	TAGACGCTTCTCCAAGAAGC	366	Lbr	GGACGCCGGACGCTCACGAG
327	Cxxc1	ACCATACGTTCCGGGAAGCA	367	Nfxl1	ACATTTCCCAGAGTGTTCCG
328	Cxxc1	TCGCAACGTCGAAGCAAATG	368	Nfxl1	GGCGGTGGGCGCCGTCGGTA
329	Tdrd9	CCTGGACAGGCACGTGTTTA	369	Psip1	GCTACAGCCGCCGAGCGTCG
330	Tdrd9	GGCCAGGCCCTGCATGCACG	370	Psip1	TGGTACGGCCCAGCGCTAGC
331	Chd5	GCTGGCGTAACGATGGGATG	371	Mbtd1	CCGAGGGTCCCACGCTGGAG
332	Chd5	GTTAACAGCCGAGAACAAGT	372	Mbtd1	CTAATGTTGTTGATATCCTG
333	Ep400	CGGATGAGGCCTTTCGGAGG	373	Ybx3	GCTCCCGGAGCTGGTTAGTC
334	Ep400	GCTGTATGCAGGAGGTGCGT	374	Ybx3	GGCGCCCGCGCTACTCGGCA
335	Cdyl2	ATCGCGGCGACACGCGAGCG	375	Satb2	AAGACCGGTTCTGGAGAGAA
336	Cdyl2	TTTGCAGAATGACAGGCTTG	376	Satb2	GGAAAGGAGGACTCCTCCAA
337	Tsc22d4	CTGGCTCCGGGCCAGCTGGC	377	Plac8	ATTTGGTAAGAGATGGCTTT
338	Tsc22d4	GTGAGCAGGCCCGAGAAGGC	378	Plac8	TTTCCTCCTTACAATTCCAA
339	Stk31	GGAGCGCAGCCCTGGTCACG	379	Dppa3	GAACTGGCTGGGATTGCGCA
340	Stk31	GTGCTGCTGCGCTATAGCGG	380	Dppa3	TTAGCATTAAAGTAGCCTGG
341	Phf23	AAGGGATTTGTAGTCTGTGC	381	Gatad2b	CAGAATAAAGTTCCTCATGT
342	Phf23	CCTCCGGAACTTCAGTTCCC	382	Gatad2b	GTGAAGTGAAAGAAATGCTG
343	Brd8	AACTGCGCGCTGAGCTTGTC	383	Cbx7	CCCAATTTGGTCAGCAGGCG
344	Brd8	CTGCGCGGGTGGGTAGAAGC	384	Cbx7	GTGCCTCTGAAAGCGCCCAG
345	Zcchc4	GTGGGTTCCGCGTGCCGCGG	385	Csde1	GGTGAATTTAGGCGTTCACC
346	Zcchc4	TCTCACTCTGGGTCTGACGG	386	Csde1	TAGTGACTCTGCTGCGACGC
347	Trps1	CTTTAAGGGAGCTCGGTCTG	387	Prmt7	GCGATCGCCGCTATGGAGGC
348	Trps1	GTGCGGCCGGCCGGTCCCGC	388	Prmt7	TCCATCCAGTTCAGCGGGCG
349	Jmjd6	ACGCCAATTGGGCGCCATGT	389	Ehmt2	CGCGTGCAGCCGGGAGGAGG
350	Jmjd6	GCGTAGGCCAAAGCAGGGAG	390	Ehmt2	GGGCGCAGCGGGCGTTGCAG

391	Lin28a	CCTCCCTTTCCCTCTGGCTC	431	Ss18l1	CGACGGCCACCTCTTCCAAG
392	Lin28a	GTCAGAGACCAGAGCAGTGG	432	Ss18l1	TGCTGGCGTCTGGCGTCAGG
393	Chd4	GCCGGGAAATCCCACCGCAC	433	Sirt5	ATATAGGTACAAACATTTGC
394	Chd4	TATTATGGGCTGTCCGTGGG	434	Sirt5	CAGTGTTGTTACTTGTATAT
395	Chd3	CGCCGCCGAGGAGGAGGAGG	435	Ctcf	CTTATCAGCACCCGCGGCCC
396	Chd3	GGTGGTGGTGGTAGCGGTGG	436	Ctcf	GGGTTGGCGCAGGGCAGCAT
397	Tdrd7	GCGCACGCCGTCCTGACGCT	437	Npm2	AAGGGTCAGCTTCCCACAGC
398	Tdrd7	TCACACGCGGACCTCGAAGT	438	Npm2	TCTGCTGTTTCCCATTGAAC
399	E2f4	AGCAGCCTTGATAGCTCCGG	439	Ssrp1	CCCTGGTCGGCTAAAGCGAG
400	E2f4	GCTGCCCTAAGGAGTTGTTT	440	Ssrp1	CTTCAAAGACTAGATTCCGA
401	Cdc5l	CTCTCGCGAGATGCCAGAGT	441	Tet3	TGCCTGGGAAGCAGCCCTTG
402	Cdc5l	CTTGACCGGGAACGCAGTGA	442	Tet3	TGCTGGGCAGGTCCTTGGCG
403	Kdm5b	AGAGGTTAACTGAAGCATTA	443	Arid4b	CGCCTCTTCCAAGTTTCTAT
404	Kdm5b	TAGACTGTTAACTTCTTGAG	444	Arid4b	GGCATCCGGGAAAGAGGTAG
405	Pms1	GAATTGTCCTGTGACCAGCC	445	Rcor1	GCCGGCCGAGGCGGCGGCTG
406	Pms1	GCACCAGTAGCTCTAGCCAG	446	Rcor1	GGGCCCTTCTCCACCATGGC
407	Smndc1	CTCGGAAGGCGGAGCGGGTG	447	Gpr6	TAAGTGAAGCAGTTAGAGCG
408	Smndc1	GCAATCGCAGGCTGCTGAGA	448	Gpr6	TCTCACCTCAGGAACACGCA
409	Wdhd1	GGCAGCGGCAAGTCTGACTC	449	Sall4	GCCTTTGTCACATGTAGGGC
410	Wdhd1	GTGGGAGCCGAACCCGGAAG	450	Sall4	TTAAATGATCTCTGAGGTCT
411	Phf13	ATGACGAAGTCCAGTCAACC	451	Tsc22d1	CGAGAAATGCCCACCTTCTT
412	Phf13	TAGTCCCTAGCGTTGCGCCC	452	Tsc22d1	GGTTGAGCTGGCTCCGGAGT
413	Fbxl19	CCCGCCCACCGCCGGTTCCC	453	Stat3	TAAGGAATGGCCAGCTGGCT
414	Fbxl19	CTTCCCAGCTCCAACTGCCC	454	Stat3	TTATGCATGGAGGCGTGTCT
415	Tox3	CCGGGACGGCCGGCGAGCTC	455	Zscan29	GCTCCGCGGCGGGAGCGAGG
416	Tox3	GGTGCGCGGCGAAGCTGGGA	456	Zscan29	TCTCCTCTCCCGGTGGACTG
417	Hspbap1	AGACGCTAAGGGTGGAAGAA	457	Zscan4b/c/e/f	AATGGTAATCTGCCCCACCC
418	Hspbap1	GGCGCTGGGCGTGGCCACGC	458	Zscan4c/f	TGGTTCCCTTTTATAGCGCC
419	Arid2	GAAGGAGGGAGCGGGAGGCC	459	MERVL LTRs	AATGAACTACAATCCGGAATTGG
420	Arid2	GAGGCGGCGAGGCAGGGTGG	460	MERVL LTRs	GCTCAGCAGTGACCCTTATCTGG
421	Zfp513	AGCGCATCACAATCACTTCC	461	-	GCTTTCACGGAGGTTCGACG
422	Zfp513	TCCTAGTTCGTATATGGAGG	462	-	ATGTTGCAGTTCGGCTCGAT
423	Mysm1	GCAGCTTTCACGCAGTCTCC	463	-	CCGCGCCGTTAGGGAACGAG
424	Mysm1	GCCTTGGCGTGGCGCACTTC	464	-	ATTGTTCGACCGTCTACGGG
425	Ep300	CAGAGACACTCACCTCCTCC	465	-	ACCCATCGGGTGCGATATGG
426	Ep300	GAGGGCGGCTCTCAGGGTGG	466	-	GCTTCTACTCGCAACGTATT
427	Dido1	AGATGCGAACTGGCAACCAA	467	-	TACAGTTATACGTCGCGGTG
428	Dido1	GGGTTGGCGCGAGAGTCAAA	468	-	CCTTAGACCGGGTGTACCTC
429	Hmgxb4	CGGAGGAGTGTATTCTCAAA	469	-	AAGTCTATGCGGGGCTCGTA
430	Hmgxb4	TCGACGCTCTTTCGCCGGCG	470	-	TTGTCAACTTCGGCCAACGC

471	-	ATAGATGTCTACGCGCCGTT	474	-	CCCTATATGCGAGATCCATA
472	-	CTCGGGCTATTCAGCGATAG	475	-	TCTCAGTTCGTAGCGAACGA
473	-	GCGGTTACCGCGAAAACCAT			

B.2 List of ZGA signature gene names

A set of 2115 gene names was used for estimating the ZGA signature. These gene names are listed below and are also described in (Alda-Catalinas et al., 2020).

0610005C13Rik, 0610009B14Rik, 0610031J06Rik, 0610040J01Rik, 1110005A03Rik, 1110006O24Rik, 1110017F19Rik, 1110018J18Rik, 1110021L09Rik, 1110032F04Rik, 1110038B12Rik, 1190002H23Rik, 1300014I06Rik, 1500010J02Rik, 1500012F01Rik, 1600002K03Rik, 1600010M07Rik, 1600012H06Rik, 1600015I10Rik, 1600020E01Rik, 1600025M17Rik, 1700001G17Rik, 1700007K13Rik, 1700009P17Rik, 1700010D01Rik, 1700012B15Rik, 1700013H16Rik, 1700016D06Rik, 1700019B21Rik, 1700019E08Rik, 1700019N12Rik, 1700024F13Rik, 1700025E21Rik, 1700029I01Rik, 1700034F02Rik, 1700034H15Rik, 1700048O20Rik, 1700060J05Rik, 1700069L16Rik, 1700080O16Rik, 1700084E18Rik, 1700086L19Rik, 1700086O06Rik, 1700093K21Rik, 1700096K18Rik, 1700112E06Rik, 1700123O20Rik, 1810019D21Rik, 1810026B05Rik, 1810032O08Rik, 1810035L17Rik, 1810044D09Rik, 1810062G17Rik, 2010001M09Rik, 2010204K13Rik, 2010317E24Rik, 2010320M18Rik, 2210016L21Rik, 2210404O07Rik, 2210414B05Rik, 2310003L22Rik, 2310011J03Rik, 2310040G24Rik, 2310045N01Rik, 2310047M10Rik, 2410002F23Rik, 2410004N09Rik, 2410016O06Rik, 2410075B13Rik, 2510002D24Rik, 2610005L07Rik, 2610019E17Rik, 2610027L16Rik, 2610028H24Rik, 2610206C17Rik, 2610306M01Rik, 2610318N02Rik, 2700023E23Rik, 2700038G22Rik, 2700078E11Rik, 2810004N23Rik, 2810006K23Rik, 2810008D09Rik, 2810021B07Rik, 2810029C07Rik, 2810055F11Rik, 2810405K02Rik, 2810417H13Rik, 2810429I04Rik, 2810459M11Rik, 2900002K06Rik, 2900079G21Rik, 3010003L10Rik, 3110056K07Rik, 3830408C21Rik, 4632427E13Rik, 4732471D19Rik, 4831440E17Rik, 4921517L17Rik, 4921531C22Rik, 4930413G21Rik, 4930427A07Rik, 4930447C04Rik, 4930455C21Rik, 4930465K10Rik, 4930479M11Rik, 4930483J18Rik, 4930515G01Rik, 4930528A17Rik, 4930547E08Rik, 4930558C23Rik, 4930578G10Rik, 4930578I07Rik, 4930579G24Rik, 4930583H14Rik, 4932411N23Rik, 4933404O12Rik, 4933406J08Rik, 4933408N05Rik, 4933411G11Rik, 4933430I17Rik, 4933430M04Rik, 4933440M02Rik, 5330411J11Rik, 5330426P16Rik, 5430402O13Rik, 5430416N02Rik, 5730408K05Rik, 5730419I09Rik, 5730455P16Rik, 5730508B09Rik, 5730590G19Rik, 5830403L16Rik, 5830433M19Rik, 6030408B16Rik, 6230400D17Rik, 6230427J02Rik, 6430573F11Rik, 9130008F23Rik, 9330020H09Rik, 9330133O14Rik, 9330159M07Rik, 9430008C03Rik, 9430015G10Rik, 9430016H08Rik, 9430021M05Rik, 9430060I03Rik, 9530077C05Rik, A130040M12Rik, A2m, A330049M08Rik, A430035B10Rik, A430084P05Rik, A430089I19Rik, A530032D15Rik, A530040E14Rik, A630001G21Rik, A630066F11Rik, A630072M18Rik, A730018C14Rik, A730037C10Rik, A730085A09Rik, AA467197, Aacs, Abcb10, Abcb5, Abcc8, Abhd14b, Abhd3, Ablim1, Abo, Abpg, AC079644.1, AC079644.2, AC079644.3, AC091683.1, AC091683.2, AC121866.1, AC121888.1, AC122464.1, AC122464.2, AC124724.1, AC127274.2, AC127374.1, AC130840.1, AC132362.1, AC134841.1,

AC140240.1, AC140409.1, AC153539.1, AC153998.1, AC156643.1, AC158600.1, AC158600.2, AC158600.3, AC158600.4, AC158600.6, AC164627.1, AC165327.2, AC211878.1, AC231112.1, AC231112.2, AC238940.1, AC238940.3, AC239617.1, AC239678.2, AC240744.3, Acaa1a, Acad8, Acap3, Acn9, Acot1, Acrbp, Acrv1, Acss3, Actr2, Actr5, Adad1, Adam2, Adamts1, Adamts5, Adck4, Adcy2, Adh7, Adi1, Adm, Adnp, Adora2b, Adrb3, Adrbk1, AF067061, AF067063, Aff1, Aga, Agmat, Agpat9, Agrp, Agtrap, Ahnak, Ahnak2, AI427809, AI462493, AI464131, AI838599, Ajap1, Ak4, Akr1b7, Aldh18a1, Aldh1l1, Alg5, Alkbh4, Aloxe3, Alppl2, Amd2, Amhr2, Amica1, Ankdd1b, Ankrd17, Ankrd22, Ankrd35, Ankrd50, Ankrd54, Ankrd9, Antxr1, Ap2a2, Ap4m1, Apoa2, Apoc2, Apod, Apol7b, Apom, Aqp11, Aqp3, Aqr, Arap2, Arf2, Arfgap2, Arfip2, Arfrp1, Arg1, Arg2, Arhgap29, Arhgap36, Arhgap8, Arhgdib, Arhgdig, Arhgef26, Arid3c, Arid5a, Arid5b, Arih2, Arl13b, Arl14, Arl15, Arl16, Arl5c, Arnt, Arrdc3, Arrdc4, Arx, Ascl1, Ascl2, Asl, Asns, Ate1, Atf2, Atf3, Atf5, Atg4d, Atg9a, Atp5s, Atp6v1e1, Atp6v1g3, Atp8b1, Atxn7l3, AU019990, AU022252, AU041133, Auts2, Avpi1, AW822073, Axin2, AY761184, B020004C17Rik, B020004J07Rik, B020031M17Rik, B3gnt4, B3gntl1, B4galnt2, B4galt1, B4galt6, B930059L03Rik, Baat1, Bag3, Bambi, Bambi-ps1, BB287469, BB287469,Gm4027, BC002230, BC003965, BC016495, BC017647, BC025920, BC027231, BC028528, BC033916, BC037704, BC046404, BC048355, BC048507, BC049715, BC049762, BC057022, BC061212, BC080695, BC147527, Bcdin3d, Bcl2l14, Bcor, Bcorl1, Bcs1l, Bdh1, Bdnf, Bend3, Bend6, Best2, Bex1, Bex2, Bex6, Bhlhb9, Bicc1, Bid, Blnk, Blvrb, Bmp2, Bmyc, Bnc2, Bnip3, Bola2, Boll, Bop1, Brd2, Bre, Bri3bp, Bst1, Btbd19, Btg1, Btg2, Bud13, C030034I22Rik, C130026I21Rik, C130060C02Rik, C130074G19Rik, C1d, C1ql1, C1qtnf4, C230096C10Rik, C2cd2l, C2cd4b, C630043F03Rik, C86695, Cab39, Cacna1h, Cacna1s, Calcoco2, Cald1, Calml4, Cand1, Cap1, Capn5, Capsl, Car6, Caskin2, Catsperg1, Cbx7, Ccdc106, Ccdc110, Ccdc116, Ccdc125, Ccdc126, Ccdc134, Ccdc137, Ccdc160, Ccdc50, Ccdc60, Ccdc64b, Ccdc66, Ccdc83, Ccdc85b, Cchcr1, Ccl3, Ccng2, Ccnjl, Ccno, Ccrn4l, Ccs, Cct8l1, Cd200r2, Cd24a, Cd52, Cd63, Cd97, Cda, Cdan1, Cdc42ep3, Cdc42ep4, Cdk4, Cdk5r1, Cdk5rap3, Cdkn2aip, Cdkn3, Cdrt4, Cdsn, Cdyl2, Ceacam1, Ceacam19, Cebpa, Celf4, Cep57l1, Cep97, Cetn1, Chchd4, Chchd5, Chek2, Chga, Chit1, Chkb, Chrac1, Chrm3, Chrna10, Chrna5, Chrna9, Chst7, Chtf18, Chtf8, Churc1, Cirbp, Cited1, Cited4, Clcn5, Cldn18, Cldn3, Cldn5, Cldn6, Clec10a, Clic1, Clip2, Clk1, Clk3, Cln8, Clp1, Clu, Cml1, Cml2, Cmtm2a, Cndp2, Cnn2, Cnnm2, Cnnm3, Cnpy1, Cnpy3, Cntnap3, Cobl, Coch, Col12a1, Col13a1, Col28a1, Col4a1, Col4a2, Colq, Commd7, Coq4, Cotl1, Cox19, Cox8c, Cpb2, Cpe, Cpeb2, Cpt1a, Crabp1, Crebl2, Creld1, Crip1, Crtap, Crx, Crxos1, Csmd1, Csrnp1, Csrnp3, Cst6, Cst7, Cst9, Cstb, CT025616.1, CT030687.1, CT485612.1, CT485612.2, CT954323.2, Ctf2, Ctla2b, Ctr9, Ctrb1, Ctsa, Ctsc, Ctsl, Ctss, Ctsz, Ctu1, Cuedc2, Cwc22, Cwc25, Cxadr, Cxcl10, Cxcl11, Cxcl16, Cyba, Cyp2b23, Cyp2c44, Cyp2c67, Cyp2s1, Cypt1,Cypt8, Cypt12, Cypt2-ps, Cypt3, Cypt4,Cypt9, Cypt7, D17Ertd648e, D1Pas1, D4Wsu53e, D5Ertd605e, D7Ertd143e, Dact3, Dapl1, Dars2, Dazl, Dbc1, Dbndd2, Dbnl, Dbp, Dbr1, Dbx1, Dcbld1, Dcc, Dcdc5, Dclre1c, Dcp2, Ddhd1, Ddit4, Ddit4l,

Ddr2, Ddx26b, Ddx31, Ddx39, Ddx43, Decr2, Dedd2, Def8, Defb13, Defb23, Defb25, Dennd4c, Depdc1b, Depdc5, Depdc7, Dexi, Dgkk, Dhcr24, Dhcr7, Dhdds, Dhh, Dhrs7b, Dhtkd1, Dido1, Dkkl1, Dlgap2, Dlgap3, Dlx2, Dmrt1, Dnahc7b, Dnajb14, Dnajb3, Dnajb9, Dnajc12, Dnajc28, Dnajc6, Dnlz, Dnpep, Doc2a, Dock9, Dohh, Dolpp1, Dpagt1, Dpep1, Dpf3, Dph2, Dpp7, Dppa2, Dppa3, Dppa5a, Dpys, Dpysl3, Dpysl5, Drd3, Drr1, Dsg1b, Dst, Dub1, Dub1a, Dub2a, Dub3, Duox2, Duoxa2, Dusp1, Dusp28, Dusp4, Dux, Dyrk3, Dyrk4, Dzip1, E130309D02Rik, E2f7, Ears2, Ebf3, Ecel1, Echdc2, Echdc3, Ecm2, Ecsit, Efcab4b, Egflam, Egr1, Eid2, Eif2s3y, Eif4ebp3, Eif5a2, Elovl4, Elovl6, Emp3, Endod1, Enpp3, Entpd7, Epha10, Ephb1, Epm2a, Epm2aip1, Ercc4, Ero1l, Errfi1, Esp24, Esrrb, Esrrg, Esyt3, Etfdh, Etohd2, Etv3, Etv5, Exoc7, Exoc8, Exog, Exosc6, F11r, Fabp9, Fadd, Fah, Fahd1, Fam102b, Fam105a, Fam109a, Fam124a, Fam134a, Fam13c, Fam150a, Fam151a, Fam158a, Fam171b, Fam173b, Fam176a, Fam181b, Fam190a, Fam195b, Fam203a, Fam20b, Fam33a, Fam43b, Fam53c, Fam57b, Fam73a, Fam76a, Fam81a, Fam84a, Fam84b, Fam89b, Fan1, Fancf, Fat3, Fblim1, Fbll1, Fbxl15, Fbxo15, Fbxo31, Fbxo34, Fbxo6, Fbxw9, Fcgbp, Fcgr2b, Fcgrt, Fdxr, Fgf1, Fgf4, Filip1l, Fkbp11, Fkbp1b, Flrt2, Flt4, Flywch2, Folr1, Folr2, Fos, Foxi3, Foxo6, Foxp1, Foxred1, Frat2, Frrs1, Fscn1, Fstl4, Fthl17, Fundc1, Fv1, Fxyd6, Fzd4, Fzd5, Fzd7, Fzd9, G2e3, Gabarap, Gabpa, Gabra1, Gadd45g, Gal, Gal3st2, Galns, Galnt3, Galt, Gan, Gata1, Gba2, Gcdh, Gcnt1, Gcnt2, Gdap10, Gdpd4, Gfra3, Ggn, Gja1, Gjb3, Gla, Glipr2, Glod5, Glrx, Glrx2, Glrx3, Glt25d1, Glud1, Glul, Gm10188, Gm10264, Gm10354, Gm10394, Gm10424, Gm10505, Gm10553, Gm10639, Gm10668, Gm10696, Gm10718, Gm10800, Gm10801, Gm11052, Gm11232, Gm11236, Gm11237, Gm11238, Gm11239, Gm11487, Gm11517, Gm11544, Gm11564, Gm11602, Gm11640, Gm11756, Gm11757, Gm11974, Gm12088, Gm12114, Gm12531, Gm12702, Gm12714, Gm12724, Gm12730, Gm12789, Gm12790, Gm12794, Gm12800, Gm12823, Gm12824, Gm12953, Gm13040, Gm13043, Gm13057, Gm13078, Gm13083, Gm13101, Gm13109, Gm13119, Gm13128, Gm13139, Gm13335, Gm13498, Gm13693, Gm13694, Gm13695, Gm13696, Gm13698, Gm13718, Gm13871, Gm13962, Gm13964, Gm14322, Gm14325, Gm14326, Gm14393, Gm14403, Gm14634, Gm14742, Gm14798, Gm14929, Gm15023, Gm1527, Gm15421, Gm15455, Gm15787, Gm16008, Gm16023, Gm16028, Gm16062, Gm16119, Gm16211, Gm16239, Gm16243, Gm16381, Gm16429, Gm16513, Gm16517, Gm17019, Gm17026, Gm17611, Gm1995, Gm2016, Gm20199, Gm2022, Gm2027, Gm2042, Gm20431, Gm20440, Gm2046, Gm20580, Gm20625, Gm20631, Gm20634, Gm2075, Gm281, Gm2a, Gm3139, Gm3258, Gm4027, Gm428, Gm4301, Gm4302, Gm4303, Gm4305, Gm4307, Gm4312, Gm4340, Gm44, Gm4532, Gm4778, Gm4782, Gm4827, Gm4858, Gm4971, Gm498, Gm4981, Gm4984, Gm5077, Gm5127, Gm5148, Gm5286, Gm5577, Gm5590, Gm5612, Gm5635, Gm5647, Gm5662, Gm5698, Gm5699, Gm5773, Gm581, Gm6086, Gm6189, Gm6351, Gm6432, Gm6468, Gm6502, Gm6507, Gm6509, Gm6568, Gm6654, Gm6763, Gm6880, Gm6890, Gm6902, Gm7102, Gm749, Gm7647, Gm7682, Gm773, Gm7942, Gm7982, Gm8038, Gm8094, Gm8104, Gm8300, Gm8766, Gm8994, Gm9, Gm9116, Gm9125, Gm973, Gm9895, Gm9958, Gmeb2, Gmpr2, Gna15, Gnat1, Gnaz, Gnb4, Gnl1,

Gnl3l, Golga1, Gorasp1, Gpa33, Gpatch3, Gpbp1l1, Gpc5, Gpkow, Gpr126, Gpr137c, Gpr161, Gpr182, Gpr19, Gpr50, Gpr63, Gpr75, Gpr83, Gprc5c, Gpx7, Grb7, Gria1, Grifin, Grik3, Grk4, Grk6, Grp, Grwd1, Gsta4, Gstm1, Gstm6, Gsto1, Gt(ROSA)26Sor, Gtf2b, Gtf2i, Gtf3c4, Gtf3c5, Gtpbp3, Gtsf1l, Guca1a, Gulo, Gusb, H1f0, H2-Bl, H2-D1, H2-DMa, H2-M10.4, H2-Q7,H2-Q9, H2-Q8, H2-T22,H2-T9, H2-T9, H2afx, H47, Hadha, Hars2, Hba-a2, Hcrt, Hdc, Hddc3, Hdhd3, Herpud1, Hes1, Hesx1, Hexa, Hexb, Hexdc, Hexim1, Hinfp, Hip1, Hipk1, Hist1h1a, Hist1h1c, Hist1h2aa, Hist1h2ab, Hist1h2ac, Hist1h2ad, Hist1h2ag, Hist1h2ah, Hist1h2bh, Hist1h2bj, Hist1h2bk, Hist1h2bl, Hist1h2bp, Hist1h3c, Hist1h3d, Hist1h3e, Hist1h3g, Hist1h3h, Hist1h4b, Hist1h4f, Hist1h4i, Hist1h4j, Hist1h4n, Hist2h3c2, Hist3h2ba, Hlx, Hmgb4, Hmgcl, Hmgn3, Hmx1, Hoxa1, Hoxa9, Hoxb2, Hoxd10, Hoxd11, Hoxd13, Hpdl, Hrh2, Hs6st2, Hsd17b10, Hsd17b14, Hspa1a, Hspa1b, Hspa1l, Hspa2, Hspa8, Hspb3, Hspb8, Hspbp1, Htra1, I0C0044D17Rik, Iah1, Iars, Id1, Id3, Id4, Idh2, Idh3b, Ier2, Ier3, Ier3ip1, Ier5, Iffo1, Ifi35, Ifitm1, Ifitm3, Ifltd1, Ifng, Igf2bp1, Igfbp2, Igfbp3, Ighe, Igtp, Ikbip, Ikzf5, Il12rb2, Il18rap, Il2rg, Il6ra, Il6st, Impa1, Impdh1, Ing4, Inpp4b, Inpp5j, Insl6, Insrr, Ints7, Ipo4, Iqcf1, Iqch, Iqub, Irak2, Irak3, Irf2bp1, Irf3, Irf7, Irf9, Irgq, Irx1, Irx2, Irx4, Isca2, Isg15, Isg20, Isg20l2, Itfg2, Itgae, Itpkc, Jam2, Jmjd4, Jmjd8, Jmy, Jub, Jun, Junb, Kank4, Kat2b, Katnb1, Kbtbd10, Kbtbd2, Kcna1, Kcna3, Kcne1, Kcnj13, Kcnk6, Kcnn2, Kdelc2, Kdm4c, Kdm5a, Kif14, Kif18b, Kifc2, Kirrel, Klb, Klf3, Klf5, Klf9, Klhl13, Klhl7, Klrg2, Kpna1, Kpna4, Kri1, Krt18, Krt28, Krtap4-13, L1td1, L2hgdh, Laptm5, Lass5, Lass6, Lat2, Lck, Lctl, Ldhc, Lefty2, Lemd3, Leng8, Lgals1, Lgals12, Lgals2, Lgals4, Lgals9, Lhfpl1, Lias, Limch1, Lime1, Limk1, Lin7a, Lin7b, Lmbr1l, Lmo4, Lmx1a, Lnp, LOC100503496, Lonp2, Lonrf3, Loxl1, Lpar6, Lpcat3, Lpcat4, Lphn2, Lphn3, Lrig1, Lrp12, Lrrc2, Lrrc38, Lrrc3b, Lrrn4, Lst1, Lta, Lxn, Ly6g5b, Lyrm1, Lyrm2, Lyrm5, Macrod1, Mad2l1bp, Mafb, Mafk, Magel2, Mak16, Malt1, Man1b1, Man2a2, Man2b1, Maneal, Map2k3, Map3k3, Mapkapk2, March1, March3, Marcksl1, Mars2, Marveld1, Maz, Mb, Mbd4, Mbd5, Mbd6, Mbnl3, Mccc2, Mcm6, Mcm9, Mctp2, Mcts2, Mdn1, Mdp1, Mecom, Med26, Med29, Mef2a, Mef2d, Meg3, Meox2, Mep1b, Mesdc2, Mest, Mettl13, Mettl2, Mfap4, Mfap5, Mfsd12, Mfsd5, Mfsd7b, Mfsd7c, Mga, Mgat2, Mgll, Micall1, Micu1, Mir17hg, Mlh1, Mlh3, Mlxip, Mmp19, Mmrn2, Mn1, Mob3a, Mob3b, Mob3c, Morc1, Mpdu1, Mpp1, Mppe1, Mpv17, Mreg, Mrgprb1, Mrgprx2, Mrm1, Mrpl17, Mrpl33, Mrps34, Msra, Msx1, Mt1, Mt2, Mt3, Mta2, Mtap6, Mtap7d3, Mtch2, Mterfd3, Mthfr, Mtrf1l, Mtss1, Mttp, Muc13, Mum1l1, Mvd, Mvk, Myadml2, Myc, Mycs, Myg1, Myl3, Mylpf, Myo1e, Myof, N4bp3, N6amt2, Naalad2, Nanog, Nanos3, Nanp, Narfl, Ncbp1, Ncf1, Ndrg1, Ndufaf1, Ndufaf3, Ndufb7, Ndufc1, Nedd4, Nedd4l, Nefl, Nek10, Neto2, Neu4, Neurl2, Neurog2, Nfam1, Nfat5, Nfatc3, Ngfrap1, Nid1, Nid2, Ninj2, Nipsnap1, Nkapl, Nkx2-5, Nlrp4f, Nme4, Nmnat2, Nop16, Nop2, Npc1l1, Npl, Nr2c2, Nrarp, Nrip1, Nrip3, Nrp, Nrxn3, Nsmaf, Nt5dc2, Nudt16l1, Nudt22, Nup62cl, Nupr1, Nxf2, Oaz3, Obox1, Obox2, Obox3, Obox6, Olfm3, Olfr118, Olfr119, Olfr120, Olfr1277, Olfr161, Olfr18, Olfr214, Olfr293, Olfr328, Olfr376, Olfr450, Olfr697, Olfr699, Olfr787, Olfr788, Olfr815, Olfr847, Olfr881, Oog4, Oraov1, Ormdl3, Os9, Osbpl6, Osgep, Osgin2, Osm, Osr2, Otud6a, Otx1,

Ovol1, Ovol2, Oxsr1, P2ry1, P4ha2, Paip2b, Pank3, Papolb, Papolg, Parp10, Pax9, Pcdh10, Pcdh17, Pcdh8, Pcdhgb7, Pcf11, Pclo, Pcyt2, Pdcd1, Pde3b, Pde6a, Pde9a, Pdgfrl, Pdk3, Pdk4, Pdlim3, Pdp2, Pdxk, Pemt, Per1, Pgap1, Pgk2, Pgm3, Phc2, Phf11, Phf16, Phf17, Phf21b, Phldb2, Pi4k2b, Pias3, Pif1, Pigl, Pigo, Pin1, Pisd-ps1, Pitpnb, Pitpnc1, Piwil2, Pla2g16, Pla2g1b, Plcb1, Plcd4, Plcl1, Pld4, Plekhf1, Plekhm1, Plekho2, Plk2, Plk3, Plod3, Plp2, Plxdc2, Pmaip1, Pmm2, Pmpca, Pnp, Pnp2, Polg2, Polh, Polq, Polr2a, Pomc, Popdc3, Porcn, Pou2f3, Pou3f1, Pou6f2, Ppat, Ppcs, Ppig, Ppil1, Ppp1r12c, Ppp1r15a, Ppp1r2-ps7, Ppp1r27, Ppp1r8, Ppp2cb, Ppp2r3a, Pramef6, Pramef8, Pramel1, Pramel3, Pramel4, Pramel5, Pramel6, Pramel7, Prdm15, Prdm9, Prdx4, Prex2, Prkch, Prkcz, Prkg1, Prmt10, Prmt6, Prnp, Prps1, Prr14, Prr15l, Prr19, Prrg2, Prss23, Prss8, Prtg, Prtn3, Psat1, Psmb10, Psmd13, Psme1, Psmg2, Ptch1, Ptpmt1, Ptpn18, Ptpn4, Ptprg, Ptrf, Ptrh1, Pura, Purb, Purg, Pusl1, Pygb, Pygm, Qpctl, Qrsl1, Qtrt1, Rab11fip4, Rab19, Rab20, Rab24, Rab30, Rab34, Rab39, Rab40c, Rab42-ps, Rab43, Rab4b, Rab7l1, Rabep2, Rabepk, Rad9, Rad9b, Radil, Ramp1, Ranbp6, Rapgef2, Rarg, Rasd1, Rasgef1c, Rasl11a, Rasl2-9, Rbm10, Rbm15, Rbm41, Rbms3, Rbmx, Rbp7, Rc3h2, Rccd1, Rcn3, Rdh1, Rdh16, Rdh9, Rdm1, Recql4, Reep6, Renbp, Rfc5, Rfpl4b, Rfx4, Rg9mtd2, Rgl2, Rgn, Rgs11, Rgs16, Rgs2, Rhbdl2, Rhox5, Ric3, Rilpl2, Rimbp2, Rimklb, Ripk4, Rln1, Rnase4, Rnaseh2a, Rnf103, Rnf113a1, Rnf121, Rnf128, Rnf19b, Rnmtl1, Rnpc3, Robo1, Rpgrip1, Rpia, Rpl10l, Rpl39l, Rpp25, Rpp40, Rprd1b, Rps26, Rps27, Rpusd1, Rpusd2, Rrp9, Rrs1, Rsph1, Rtn2, Rundc1, Rusc1, Rxra, Rxrg, Sall4, Sat2, Satb2, Scamp1, Sccpdh, Scd1, Scd2, Scg5, Sco1, Sco2, Sdr39u1, Sdr9c7, Sec1, Sec61a2, Sel1l2, Sephs2, Sept1, Sept11, Sept2, Sept6, Sept9, Sepx1, Serpinb1c, Serpinf1, Sertad1, Sesn1, Sfta2, Sfxn2, Sfxn4, Sgk3, Sh2d3c, Sh3bp4, Sh3kbp1, Shb, Shroom1, Siah2, Sik1, Six1, Six4, Skil, Slc10a3, Slc14a2, Slc16a3, Slc16a6, Slc1a4, Slc1a5, Slc20a1, Slc22a20, Slc22a28, Slc25a14, Slc25a43, Slc26a1, Slc27a2, Slc27a5, Slc29a2, Slc2a1, Slc2a8, Slc30a2, Slc34a2, Slc35e2, Slc35e3, Slc38a2, Slc38a4, Slc39a1, Slc39a14, Slc39a4, Slc4a4, Slc4a5, Slc52a2, Slc5a4b, Slc5a6, Slc6a6, Slc6a8, Slc7a3, Slc7a9, Slc9a9, Slfn10-ps, Smad7, Smg5, Smpd3, Smtn, Smyd5, Snai1, Snai2, Snai3, Snhg11, Snhg12, Snhg3, Snhg4, Snhg7, Snhg8, Snrnp35, Snrpc, Snta1, Snw1, Snx8, Socs2, Socs3, Sord, Sowaha, Sowahc, Sox10, Sox11, Sox2, Sox3, Sox30, Sox8, Sp110, Sp140, Sp4, Sp6, Spag9, Spdya, Specc1, Speer4c, Speer4d, Speer4e, Speer7-ps1, Speer8-ps1, Spesp1, Spic, Spin2, Spns1, Spock3, Spp1, Sprr2d, Sprr2e, Spry2, Spryd4, Spty2d1, Spz1, Srd5a1, Srgap3, Srl, Stac2, Stag3, Stard6, Steap1, Stim1, Stk16, Stk17b, Stk19, Stk38, Stmn3, Ston2, Stox1, Sult2b1, Sult5a1, Sun1, Surf2, Sva, Sycp1-ps1, Syde2, Syne2, Syngr1, Synm, Synpo2, Sytl2, Tacc3, Tacr3, Taf1d, Tagln, Tagln2, Tapbp, Tatdn3, Tbc1d12, Tbl1x, Tbl2, Tbl3, Tbrg3, Tbrg4, Tbx1, Tbx20, Tbx3, Tbxa2r, Tceal8, Tchh, Tcirg1, Tcstv1, Tcstv3, Tdh, Tdpoz1, Tdpoz2, Tdpoz3, Tdpoz4, Tdpoz5, Tef, Tekt2, Terc, Tex101, Tex13, Tex19.1, Tfap2a, Tfcp2, Tfcp2l1, Tfpi2, Tfrc, Thbs1, Thnsl2, Thsd7b, Thy1, Ticam2, Tie1, Timd2, Timm22, Tktl1, Tle3, Tlk2, Tm4sf1, Tmco2, Tmed6, Tmeff1, Tmem101, Tmem106a, Tmem119, Tmem126b, Tmem129, Tmem146, Tmem167, Tmem167b, Tmem177, Tmem189, Tmem191c, Tmem20, Tmem208, Tmem215, Tmem229b, Tmem35, Tmem37, Tmem39a, Tmem56, Tmem79, Tmem92,

Tmppe, Tmsb15b1, Tmsb15b2, Tmsb15l, Tmx1, Tmx2, Tnfaip6, Tnfaip8l2, Tnfrsf13b, Tnfrsf17, Tnip2, Tnnc2, Tob1, Toe1, Tomm5, Tor1b, Tor3a, Tpo, Traf2, Trak2, Trap1a, Trh, Trib3, Trim25, Trim32, Trim43a, Trim43b, Trim43c, Trim52, Trim8, Triml2, Trmt1, Trmt61a, Trmt61b, Trp53bp2, Trub2, Tsc22d3, Tsen2, Tsen54, Tsga8, Tspan1, Tspan4, Tspan6, Tssk6, Tsx, Ttc30b, Ttc39d, Ttll12, Tuba1a, Tubb2b, Txnip, Uap1, Uap1l1, Ubap1, Ubc, Ube2o, Ube2t, Ube2w, Ubtd2, Ubtfl1, Ubxn2a, Ugt1a1, Uhrf1bp1, Uhrf2, Uimc1, Unc119b, Upk1a, Upk3a, Uqcrc1, Ush1g, Usp-ps, Usp17l5, Usp25, Usp26, Usp50, Usp9y, Utf1, Utp14b, Utp23, Uty, Vcam1, Vcan, Vegfa, Vgll1, Vhl, Vim, Vmn1r227, Vmn1r53, Vmn1r88, Vmn2r1, Vmn2r94, Vps33a, Vps39, Wbscr27, Wdr16, Wdr18, Wdr27, Wdr4, Wdr45, Wdr62, Wdr77, Wdtc1, Whsc2, Wipf2, Wnt5a, Wrap53, Wt1, Wwtr1, Xab2, Xaf1, Xkr9, Xlr, Xpnpep1, Xpot, Xylb, Yars, Ydjc, Yeats2, Yif1a, Yod1, Zbed6, Zbtb17, Zbtb5, Zbtb7a, Zbtb8a, Zc3h10, Zc3h12a, Zc3h7a, Zc3hc1, Zcchc11, Zcchc12, Zcchc13, Zcchc17, Zcchc24, Zfand2a, Zfp1, Zfp119b, Zfp142, Zfp184, Zfp187, Zfp217, Zfp239, Zfp280c, Zfp292, Zfp296, Zfp30, Zfp317, Zfp335, Zfp352, Zfp353, Zfp367, Zfp382, Zfp386, Zfp42, Zfp446, Zfp51, Zfp516, Zfp53, Zfp54, Zfp560, Zfp566, Zfp57, Zfp572, Zfp574, Zfp593, Zfp599, Zfp608, Zfp622, Zfp623, Zfp637, Zfp647, Zfp689, Zfp704, Zfp706, Zfp707, Zfp719, Zfp771, Zfp775, Zfp800, Zfp808, Zfp809, Zfp810, Zfp825, Zfp867, Zfp871, Zfp874b, Zfp882, Zfp941, Zfp948, Zfpm2, Zfx, Zfy1, Zfy2, Zfyve26, Zgpat, Zic3, Zkscan1, Zmiz1, Zmym2, Znhit2, Znhit3, Zrsr1, Zscan4a, Zscan4b, Zscan4c, Zscan4d, Zscan4e, Zscan4f, Zswim2, Zswim3, Zswim5, Zswim6, Zyg11a, Zyg11b.