Unravelling cell fate decisions through single cell methods and mathematical models

Mircea, M.

1

How a cell decides its own fate: A single-cell view of molecular mechanisms and dynamics of cell type specification

On its path from a fertilized egg to one of the many cell types in a multicellular organism, a cell turns the blank canvas of its early embryonic state into a molecular profile fine-tuned to achieve a vital organismal function. This remarkable transformation emerges from the interplay between dynamically changing external signals, the cell's internal, variable state and a tremendously complex molecular machinery we are only beginning to understand. Recently developed single-cell omics techniques have started to provide an unprecedented, comprehensive view of the molecular changes during cell type specification and promise to reveal the underlying gene regulatory mechanism. The exponentially increasing amount of quantitative molecular data being created at the moment is slated to inform predictive, mathematical models. Such models can suggest novel ways to manipulate cell types experimentally, which has important biomedical applications. This review is meant to give the reader a starting point to participate in this exciting phase of molecular developmental biology. We first introduce some of the principal molecular players involved in cell type specification and discuss the important organizing ability of biomolecular condensates, which has been discovered recently. We then review some of the most important single-cell omics methods and relevant findings they produced. We devote special attention to the dynamics of the molecular changes and discuss methods to measure them, most importantly lineage tracing. Finally, we introduce a conceptual framework that connects all molecular agents in a mathematical model and helps us make sense of the experimental data.

1

1.1 Introduction

What is a cell type, anyway? Traditionally, cell types have been defined by their function within an organism: Neurons process and transmit information, macrophages remove harmful microorganisms and podocytes are crucial for blood filtration in the kidney. As function can be difficult to ascertain, especially for subtle variants of cell types, cell morphology and the presence of certain marker genes are often used as proxies [2, 3]. With the advent of single-cell **omics technologies**, cell types have increasingly come to be identified with their molecular profiles. While most cell types persist over long periods of time,

> **Single-cell omics technologies**:
> Experimental methods to measure the entire genome, epigenome, transcriptome, proteome etc. of a cell in high-throughput.
> **Pluripotency**:
> Ability of a cell to give rise to multiple cell types.
> **Histone**:
> Proteins that are crucial for the organization of DNA in the nucleus. DNA is tightly wound around nucleosome core particles which consist of 8 histones.

often the entire life span of an adult organism, cells are found in short-lived, transient states such as different phases of the cell cycle, different metabolic states or multiple forms of stress response. Here, we are only concerned with the specification of cell types, which occurs during embryonic development or regeneration of adult tissues. During development, **pluripotent** embryonic cells differentiate into progenitors with diminishing developmental potential, and eventually fully specified cell types [2, 4–6]. In adult tissue, long-lived adult stem cells give rise to multiple types of descendants. These processes are collectively termed differentiation. Differentiation involves changes in gene expression (i.e., messenger RNA and protein levels), which are accompanied and guided by epigenetic changes. Broadly, the epigenetic profile of a cell encompasses any heritable molecular mark, with the exceptions of changes in the DNA sequence [7, 8]. Two of the most important epigenetic marks are DNA methylation and **histone** modifications. These marks are tightly linked to the accessibility of the DNA and thus influence the expression of specific genes. A comprehensive introduction to epigenetics can be found in [9].

Here, we will review a few of the many cell-autonomous molecular mechanism that make differentiation a reproducible process and ensure the long-term stability of cell types. Importantly, cells do not develop in isolation. Their communication with neighboring cells via chemical and mechanical signaling is an integral part of embryonic development and tissue regeneration, which we will not discuss here (for recent reviews see [10, 11]). Equally important, but also outside the scope of this review, is the role of molecular noise, which can drive cell type decisions but must also be controlled to ensure the stability of the fully differentiated state (for recent reviews see [12, 13]). In this review, we will first introduce some of the most important molecular players, which can be used to define a cell type, and discuss omics techniques that can measure molecular profiles comprehensively in single cells. We will then focus on the dynamics of differentiation and novel methods that allow the inference of the developmental lineage tree. Finally, we will discuss challenges arising in the analysis of data sets comprising multiple modalities and a conceptual framework that enables a quantitative understanding of differentiation (Figure 1.1).
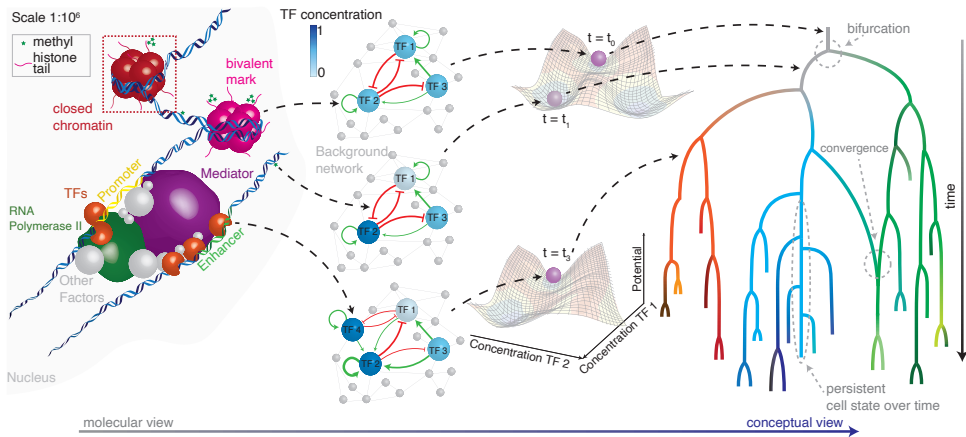
Figure 1.1: Conceptual framework for cell types and differentiation dynamics. Left: TFs bind to enhancer and promoter regions depending on the chromatin state. All molecules are drawn at a scale of 1:106. Center left: A simplified gene regulatory network consisting of TFs (nodes of the network graph) and their interactions (edges of the network graph). Center right: A gene regulatory network can be modeled by a dynamical system, which can be represented by a potential energy landscape. The position in the landscape is determined by TF abundance. The shape of the landscape depends on the TF interactions. Cells follow the path of steepest descent to stable states, which correspond to cell types. Right: The collection of trajectories in the potential landscape form a cell type decision tree, which highlights the hierarchical nature of differentiation.

1.2 Molecular embodiment of a cell type

Most, if not all, cell type decisions involve specific **transcription factors** (TFs) [2, 3, 5, 6, 14, 15]. These DNA binding proteins control a gene's transcription level by binding to **cis-regulatory elements** (CREs) in the DNA. Enhancers, CREs that can be found at large distances from the regulated gene, play a particularly important role for cell type determination. Enhancers work in concert and physically interact with promoters, another type of CRE that is usually found near the regulated gene. TF binding of CREs not only depends on the presence of specific DNA sequence motifs but is also strongly modulated by the configuration of the **chromatin** (the complex of DNA,

> **Transcription factor**:
> A protein that binds to specific DNA sequences and regulates transcription.
> **Cis-regulatory elements**:
> Sequences of non-coding DNA which regulate the transcription of genes.
> **Chromatin**:
> The complex of nucleosomes, DNA and other associated proteins.
> **Nucleosome**:
> Smallest unit of DNA organization. Consists of DNA wound around 8 histones.
> **Pioneer factor**:
> A transcription factor that can bind to nucleosome-bound DNA.

nucleosomes and other associated proteins, see Figure 1, left). With the exception of so-called **pioneer factors** [16], TFs only bind accessible, nucleosome-free DNA [17]. Chromatin configuration and TF binding are affected by chemical modifications of histones (the components of a nucleosome), as well as the DNA [17]. Different histone modifications, or marks, are associated with different functions, broadly categorized as active and repressive, and the effect of DNA modifications strongly depends on the genomic context [18, 19]. For example, DNA methylation at the enhancer regions of the pluripotency gene Sox2 results in silencing of its expression in embryonic stem cells [19]. Importantly, the interaction between TFs and chromatin configuration is reciprocal: TFs recruit enzymes that locally

1

change the molecular make-up of the chromatin [17]. Both histone marks and DNA methylation are heritable molecular marks, as they are copied to the newly synthesized DNA during cell division [8, 9, 20]. They can therefore function as long-term memory of a cell's molecular profile and hence cell type. The pattern of chromatin accessibility and epigenetic marks can thus be used to identify a cell type and reveal relevant CREs [17, 21]. Importantly, cell type specification cannot be understood by studying individual TFs or epigenetic features in isolation. Cell types rather emerge from the complex interactions of several TFs. The presence of particular subsets of TFs has therefore been used to define a periodic table of cell types [5]. Together with their target genes, TFs form **gene regulatory networks** that establish and maintain cell identity [2], see Figure 1, middle. Regulatory interactions between TFs, in particular negative feedback loops, are crucial for the stability of molecular states. Due to the presence of fluctuations in the environment as well as the internal state of the cell, robustness is an important requirement for regulatory networks. At the same time, they need to be dynamic and react appropriately to external signaling

> **Regulatory network**:
> A system of interacting molecules that regulate each other's gene expression as well as a set of target genes.
> **Master transcription factor / master regulator**:
> A transcription factor that effects the transcription of multiple downstream genes and is essential for cell type specification.
> **Super-enhancer**:
> A group of multiple enhancers in close proximity characterized by high levels of Mediator complex, which strongly drives gene expression of its target genes.
> **Mediator**:
> A multiprotein complex that coactivates transcription by interacting with TFs and RNA polymerase II.
> **Biomolecular condensates**:
> Droplets of a condensed liquid phase formed in cells by homotypic, multivalent interactions (i.e., interactions between identical molecules that involve multiple binding sites). One example are membrane-less organelles.
> **Liquid-liquid phase separation**:
> De-mixing of a homogeneous liquid into two distinct liquid phases.

inputs [2]. Mutual repression of TFs is one mechanism by which multiple, alternative cell types can be created. A prominent example is the interaction between the TFs GATA6 and NANOG, which governs the lineage decision between two of the earliest cell types in the mammalian embryo [22–24]. The conceptual framework discussed in the final section of this review explains how various stable cell types and unidirectional differentiation dynamics emerge from gene regulatory networks, see Figure 1.1, right.

Despite the fact that TFs always work in concert, some have a particularly large impact on lineage decisions: Overexpression of certain TFs can revert a differentiated cell back to a pluripotent state (reprogramming) or convert one cell type into another (transdifferentiation) [25]. The remarkable power of these TFs, termed **master TFs** or **master regulators**, can be rationalized by their DNA binding patterns. Master TFs have been shown to bind clusters of enhancers, or **super-enhancers**, which drive high levels of key cell type-specific genes [26, 27], see Figure 1.2. Super-enhancers owe their special role to a high density of co-localized **Mediator** complex [26, 27], a protein complex that links TF binding to the recruitment of the transcription machinery and therefore gene expression. A well-studied example of master TFs that bind cell-type specific super-enhancers are regulators of the pluripotent state in embryonic stem cells: NANOG, SOX2 and OCT4 [27]. TFs, CREs, epigenetic marks and enzymes that modify chromatin state are just a small, albeit important, subset of the many molecular species that are involved in cell type specification. It has long been unclear, how all of these mobile molecules, some of which are freely diffusing in the nucleus, can interact in an efficient manner. Recently, **biomolecular condensates**, which form through **liquid-liquid phase separation** (LLPS) [28–30], have been suggested as a possible answer to this question. Biomolecular condensates form

according to well-known thermodynamic principles as a result of multivalent, homotypic interactions between molecules [31]. The high concentration of several molecular species in the condensed phase leads to increased interaction rates [30]. Examples of biomolecular condensates are the well-known membrane-less organelles, such as the nucleolus or Cajal bodies, as well as **paraspeckles**

> **Paraspeckle**:
> A biomolecular condensate that forms in the presence of the long non-coding RNA NEAT1 and several RNA binding proteins.
> **Intrinsically disordered regions**:
> Segments of a protein that do not form a stable three-dimensional structure.

and many more [28, 29, 32]. It has been found that **intrinsically disordered regions** (IDRs) of proteins can lead to the multivalent interactions that can cause condensates to form [33–35]. Interestingly, MED1, a member of the Mediator complex, and BRD4, a coactivator of transcription, have large IDRs and form condensates at super-enhancers [33], see Figure 1.2. Thus, phase-separated condensates likely concentrate components of the transcription apparatus and thereby ensure robust transcription of key cell type-specific genes. Additionally, the large size of the Mediator cluster at super-enhancers enables the contact with multiple promoter sites [36]. Therefore, biomolecular condensates are likely of crucial importance for establishing a cell type.
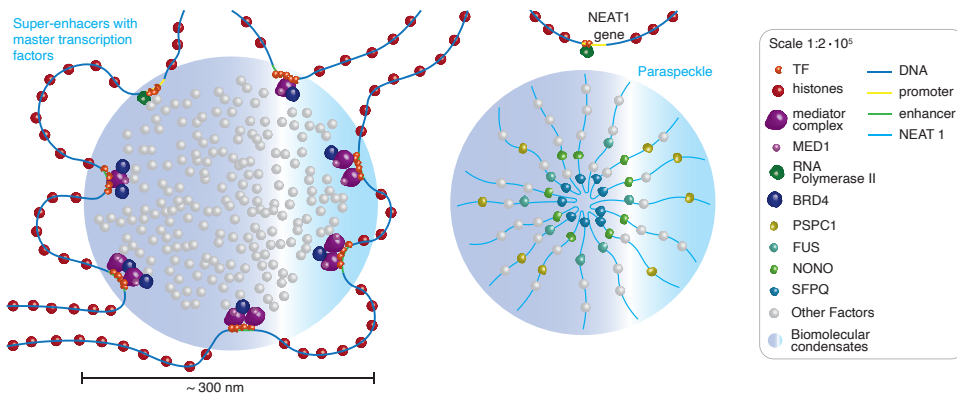

Figure 1.2: Master transcription factors and super-enhancers play a major role in guiding cell type specification and are compartmentalized by biomolecular condensates. Left: super-enhancer with bound master transcription factors in a large biomolecular condensate. Right: A paraspeckle that forms in the presence of the lncRNA NEAT1. All molecules are drawn at a scale of $1 : 2 \cdot 10^5$.

1.3 Molecular profiling

In recent years, omics technologies have emerged that measure one or multiple molecular species comprehensively in single cells (see Box 3 and Figure 1.3 for a selection of common methods). These technologies can reveal cell type-specific molecular profiles in high throughput. With single-cell RNA-sequencing (scRNA-seq) the transcriptomes of individual cells can be obtained [2, 5, 6], which enables the identification of new cell types and cell states in complex tissues [37]. Multiple large consortia are currently generating transcriptional atlases of entire organisms (reviewed in [38]). The human cell atlas [39]

and Tabula Muris [40] are two prominent examples. Notwithstanding the great value of scRNA-seq measurements, gene expression should ideally be measured at the protein level. Numerous regulatory mechanisms at the translational and post-translational level make mRNA abundance just a proxy for protein abundance. Protein measurements have indeed revealed phenotypic features that could not be discerned with scRNA-seq alone [41, 42]. As
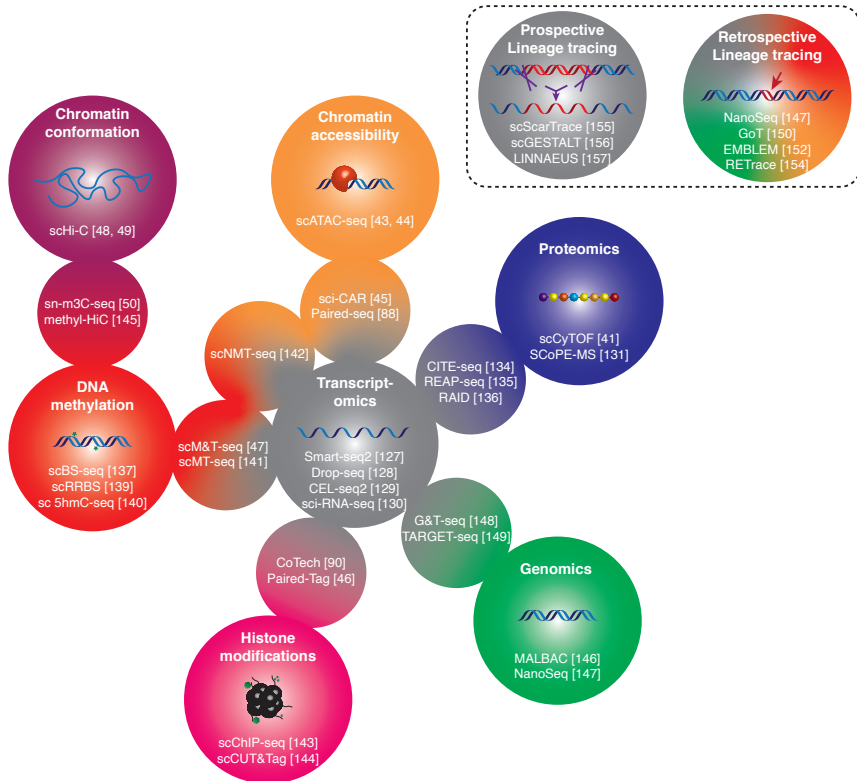

Figure 1.3: A selection of single-cell omics and multi-omics techniques useful for studying cell type specification. Colors indicate the measurement of different molecular species in a cell (green: DNA sequence, grey: RNA abundance, blue: protein abundance , orange: chromatin accessibility, red: DNA methylation, pink: histone modifications, violet: chromatin conformation). Circles with color gradients contain techniques that measure several types of molecules at the same time. Dashed lines envelop techniques that aim at understanding similar concepts.

mentioned before, chromatin state is an important factor in gene regulation. Knowledge of the chromatin landscape can therefore improve the identification of cell types [43]. There is a growing variety of single-cell methods that measure chromatin features. For example, by using scATAC-seq [44, 45], which reveals accessible chromatin regions in single cells, it is possible to identify cell type-specific regulatory elements and candidate master TFs. By combining scATAC-seq and scRNA-seq, open chromatin regions can be associated with active transcription, which improves the identification of TFs and target genes compared

to pure RNA measurements. It was also shown that considering the chromatin state of distal CREs significantly increased the power to predict cell-type specific gene expression, compared to using promoter chromatin state alone [46].

Box 1: Common single-cell omics and multi-omics techniques

Transcriptomics: The currently most prevalent single-cell omics method is single-cell RNA sequencing, which measures RNA abundance. Different experimental implementations of this method include: Smart-seq2 [47], Drop-seq [48], CEL-seq2 [49] and Sci-RNA-seq [50].

Proteomics: It is not yet feasible to measure every protein in a single cell. Antibody-based methods, for example scCyTOF [42], can measure hundreds of proteins, but cannot easily be scaled to the whole proteome and rely on the existence of highly specific antibodies. Mass spectrometry-based proteomics methods, which recently became available, might soon produce high-quality proteomes of single cells. One example is SCoPE-MS [51], which detects around 1000 proteins per cell. Improvements to this method have been recently made in SCoPE2 [52] and another method [53]. By sequencing of DNA-tagged antibodies, the quantification of hundreds of proteins together with the transcriptome in single cells is possible with CITE-seq [54] and REAP-seq [55]. RAID [136] uses RNA-tagged antibodies for the same purpose.

Epigenomics: To gain insights into chromatin accessibility, one of the most prominent techniques is scATAC-seq [44, 45], which uses transposons to barcode accessible DNA. scATAC-seq can be performed simultaneously with scRNA-seq, which was implemented, for example, by sci-CAR [46] and Paired-seq [56]. DNA methylation is measured by scBS-seq [57, 58] and scRRBS [59]. Single-cell 5hmC-seq measures DNA hydroxymethylation [60]. Joint measurements of DNA methylation and the transcriptome is possible with, for example, scM&T-seq [61] and scMT-seq [62]. A method to measure all three molecular profiles (DNA methylation, transcriptome and chromatin accessibility) is scNMT-seq [63]. Histone modifications can be measured, for example, with scChIP-seq [64] and scCUT&Tag [65]. New methods to measure the transcriptome jointly with histone modifications are CoTech [66] and PairedTag [67]. It is now also possible to study the chromatin conformation in every single cell with scHi-C [68, 69]. Recent methods have allowed the capture of both chromatin conformation and DNA methylation (sn-m3C-seq [70] and methyl-HiC [71]).

Genomics: The DNA sequence of single cells can be measured by methods such as MALBAC [72] and NanoSeq [73]. NanoSeq which has been designed to detect even small somatic mutations in single DNA molecules. Measurement methods that combine DNA sequencing with transcriptomics are, for example, G&T-seq [74] and TARGET-seq [75].

1

Similarly, a simultaneous measurement of histone modifications and transcriptome showed that active enhancers are epigenetically more variable across cell types than promoter regions [67]. A related finding resulted from the simultaneous measurement of DNA methylation and transcriptome (scM& T-seq [61]). The authors confirmed that promoter DNA methylation in mouse ESCs is typically correlated with reduced gene expression. By contrast, DNA methylation of distal enhancers is more often correlated with increased gene expression, compared to promoters. Since active transcription typically requires the physical proximity of enhancers and promoters, knowledge of chromatin organization can be helpful to understand cell type decision making.

scHi-C is a high-throughput method to reveal chromatin interactions throughout the genome in single cells [68, 69]. In combination with DNA methylation measurements, cell type specific chromatin conformations can be obtained [70], which might help to clarify the role of biomolecular condensates [76, 77]. In a recent study, a new variant of Hi-C [77] was used to determine the stability of chromatin interactions, which were revealed to vary substantially between organelles. Approaches to measure the spatial distribution of transcripts [78, 79] and proteins [80] with sub-cellular resolution might lead to an even better understanding of cellular compartmentalization through biomolecular condensates. Multi-omics single-cell methods, like those presented here, promise to enable an improved mechanistic understanding of cell type-specific gene regulation [81]. A more comprehensive discussion of these methods can be found in [82].

1.4 Dynamics of differentiation

During differentiation, the molecular profile of a cell is remodeled substantially. TFs are, unsurprisingly, important drivers of this transformation. As the majority of TFs binds to accessible chromatin regions, differentiation is accompanied by pervasive changes in chromatin accessibility [13, 16, 17]. One underlying mechanism involves pioneer TFs, which bind to nucleosome-associated DNA and create an open chromatin state [16, 17]. These TFs can explore nucleosomal DNA through non-specific and transient binding, which in turn allows partial opening of

> **Chromatin remodeler**:
> Protein complexe that catalyze molecular changes of the chromosome, such as nucleosome removal.
> **Bivalent domain**:
> Chromatin domain that carries both activating and repressing histone marks.
> **RNA Polymerase II**:
> A multiprotein complex that transcribes DNA into messenger RNA.

the chromatin and other, non-pioneering factors to bind [16]. This mechanism has been recently validated, for example, for the pioneer factor PAX7 [83]. Another mechanism is passive competition of TFs for DNA binding during short periods of local chromatin opening, which increases and stabilizes with higher TF concentrations [17, 84].

Chromatin state is also influenced by **chromatin remodelers** that are recruited by TFs [16, 18, 85] and bind to different histone marks [85, 86]. This is one mechanism by which epigenetic marks strongly impact chromatin accessibility. Importantly, activating and repressing histone marks can also occur simultaneously, on the same nucleosome. These **bivalent domains** play a particularly important role in cell type decisions [87] and are more abundant in embryonic stem cells (ESCs) than in adult tissues. A prominent example is the combination of H3K4me3 (Trimethylation of histone H3 on lysine 4) which is associated with active transcription and H3K27me3 (Trimethylation of histone H3 on lysine

27), which causes chromatin compaction and is thus a repressive mark. It has been shown that bivalent domains are positioned at key TF genes that are important for development [87–89]. Enhancers and promoters with bivalent marks are thought to be in a poised state, that can be quickly resolved to either activation or repression. This effect can be mediated by multiprotein complexes composed of polycomb group (PcG) proteins. These proteins cause gene silencing by, for example, catalyzing methylation of H3K27 [90, 91]. In ESCs, the occupancy of PcG proteins at bivalent histone marks can change during differentiation, which results in altered gene expression [89, 92]. Poised enhancers have been found to be necessary, for example, for the differentiation into specific neural cell types [93]. The examples mentioned here are only few of many epigenetic mechanisms that drive dynamic chromatin remodeling during stem cell differentiation (reviewed in [94]).

Epigenetic marks are not homogeneously distributed in the nucleus, but rather need to be localized at important regulatory sites, which might be promoted by biomolecular condensates. The formation of biomolecular condensates during differentiation has been linked to different long non-coding RNAs (lncRNA) and RNA-binding proteins (RBPs) [32, 95]. For example, the lncRNA DIGIT forms biomolecular condensates together with the RNA binding protein BRD3, which contains an IDR [95]. BRD3 is recruited to sites of the activating histone mark H3K18ac (Acetylation of histone H3 on lysine 18). Paraspeckles are another important class of biomolecular condensates defined by the presence of the lncRNA NEAT1, which recruits several RBPs [96], see Figure 1.2. These condensates can, for example, influence transcriptional regulation via associated RBPs [96–98]. NEAT1 has been found to physically interact with EZH2, a PcG protein, which is involved in catalyzing histone methylation [99]. Interestingly, paraspeckles were found to be involved in slowing down the differentiation process and their number changes dynamically during differentiation to several lineages [32, 97]. Another example of dynamic transcriptional regulation through biomolecular condensates is the association of RNA polymerase II with Mediator condensates (see Figure 1.2). It has been found that, upon phosphorylation, **RNA polymerase II** transitions from condensates involved in transcription initiation to condensates involved in RNA splicing at genes associated with super-enhancers [35].

1.5 Measurement of differentiation dynamics

In simple organisms the entire lineage tree can be assembled using a microscope [100]. In larger organisms that becomes unfeasible and scRNA-seq data of developing tissues has been used instead to infer lineage relationships [101]. Due to asynchrony in embryonic development or regeneration of adult tissues, a single scRNA-seq measurement can capture cells in different stages of differentiation [43, 102] and developmental order, or pseudotime, can be inferred by computational methods (reviewed in [103], see also the section on data analysis below). If the developmental process is sufficiently accessible for repeated sampling, scRNA-seq measurements at several time points can be used to resolve developmental dynamics [14, 104–107]. This approach improves the temporal resolution and revealed that cells with different lineage histories can converge to globally similar cells [108]. However, combining multiple data sets to infer the correct developmental trajectory is challenging.

Lineage reconstruction has also been performed based on protein measurements in single

1

cells at different time points. In a recent study [108], 27 proteins, of which 16 were TFs, were measured over a time course of 22 days during hematopoietic differentiation. This study showed that, at the protein level, cell type decisions are accompanied by gradual changes in lineage specific TFs, as no abrupt switches in TF levels were observed.

Box 2: Lineage tracing

There are two, conceptually distinct approaches to lineage tracing: retrospective and prospective. In retrospective lineage tracing, lineage relationships are inferred from naturally occurring somatic mutations. These mutations can be traced using DNA sequencing methods [73]. In a recent study, such mutations were linked with scRNA-seq data to investigate clonal relationships and cell types in human [109]. Mitochondrial DNA has a 10 fold higher mutation rate than nuclear DNA [110, 111], which makes it a good candidate for retrospective lineage tracing. Interestingly, these mutations can be tracked with ATAC-seq measurements because mitochondrial DNA is accessible [111]. DNA methylation also undergoes stochastic changes during cell division known as epimutations, which allows tracking of lineage histories through measurements of DNA methylation [110, 112]. Coupling genomics to DNA methylation measurements allows both lineage tracing and the study of cell type specific methylation patterns [113]. However, naturally occurring mutations are rare, which requires highly accurate and sensitive measurement techniques and computational methods. In prospective lineage tracing, heritable markers are introduced that are read out at a later time point. The most recently developed dynamic lineage tracing methods insert 'scars' into the DNA at random or pre-determined locations, resulting in a large variety of different markers [114, 115]. In some cases, these markers, or barcodes, are also transcribed, so that scRNA-seq is able to capture transcriptomes and lineage information simultaneously. Different omics technologies have been used in the context of lineage tracing (see Box 1 for a list of omics techniques). For retrospective lineage tracing, NanoSeq [73] has been used to track even small somatic mutations and GoT [109] linked transcriptomics to genotyping. scATAC-seq has been used to track mutations in mitochondrial DNA [111] and scRRBS has been used together with DNA sequencing to track DNA mutations together with DNA methylation [113]. Examples of prospective lineage tracing techniques that use transcriptomics measurements are scScarTrace [116], scGESTALT [117] and LINNAEUS [118].

To reveal the gene regulatory programs that cause gene expression changes, chromatin conformation measurements during development can be used. Bulk methods have been used extensively to measure epigenetic changes and chromatin accessibility of cell populations [119], which produced many important insights. However, cell-to-cell variability and rare cell populations can only be distinguished with single-cell methods. Therefore, time-resolved single cell chromatin accessibility measurements can be very informative [120], in particular in combination with transcriptomics [46, 56, 121, 122]. One study found a class of genes with a high number of putative enhancers whose chromatin accessibility is predictive of gene expression [122]. These genes are enriched in TFs that regulate cell type-specific gene expression. These findings suggest participation in super-enhancers and

a central role in cell-type specification. Additionally, it was observed that the expression of TFs precedes the accessibility of their target sites, which might indicate a causal role of TFs in chromatin remodeling, possibly through additional epigenetic mechanisms [121]. Another interesting case is the simultaneous measurement of chromatin accessibility, DNA methylation and transcriptome (scNMT [123]) at several timepoints in mouse development. The authors were able to study the dynamic changes of all three profiles in time and confirmed ectoderm, one of the three embryonic germ layers, as the default developmental pathway. Specific histone marks have also been measured during differentiation and development. For example, the co-occurrence of H3K4me3 and H3K27me3 (bivalent mark) was measured in mouse ESCs together with scRNA-seq. The authors calculated a bivalency score along an RNA based pseudotime trajectory and were able to classify genes by trends in bivalency dynamics [66]. A similar method found a significant overlap between H3K27ac (Acetylation of histone H3 on lysine 27) and H3K27me3 in the adult mouse brain at CREs related to forebrain development [67].

An entirely different approach to study developmental dynamics is used in lineage tracing techniques [114, 115, 124] (see Box 2, which aim to find the correct phylogenetic tree [125, 126] from pluripotent cells to fully specified cell types. Lineage tracing methods have produced a large number of valuable insights. A recent study used lineage tracing to reveal early biases towards particular cell types [127] that are not resolved with transcriptomics: Transcriptionally similar cells were found to be committed to particular cell types prior to the divergence of their transcriptional profiles [114, 128]. Importantly, such cells can easily be mistaken for multipotent progenitors. Coupling lineage tracing with epigenomics or proteomics measurements might help to avoid some of these biases and pinpoint the correct sequence of transcriptional and epigenetic changes during development. Lineage tracing experiments also seem to indicate that cell fate decisions occur in a more continuous manner rather than abruptly, as previously believed [129]. Finally, lineage tracing made it possible to observe the convergence of differentiation trajectories from distinct developmental origins [128].

1

1.6 Data analysis

Many single-cell methods involve advanced data analysis (see Box 3 and Figure 1.4 for a selection of computational methods). In scRNA-seq data, cell types can in principle be identified by clustering similar transcriptomes [130] and the underlying gene regulatory networks can be inferred [131–133]. However, both cell type identification and network inference are improved by integrating multiple omics data sets [41, 134]. Integration methods typically aim to extract variations common to all measured modalities [135–137]. That is even possible if molecular species are not measured simultaneously in the same cell, as shown, for example, for DNA methylation and transcriptome measurements [138]. Trajec-
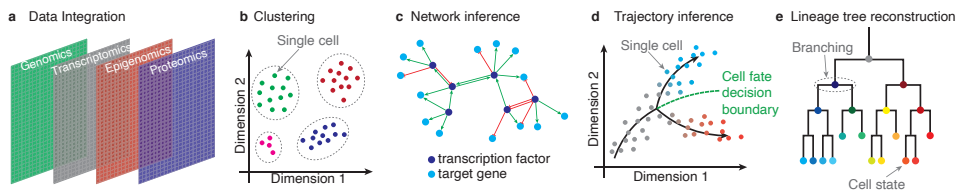

Figure 1.4: Common elements of single-cell omics data analysis. a) Each rectangle represents a data matrix from an omics technology with cells in columns and features in rows. Data Integration methods can be used to combine these data sets. b) Each circle represents a cell projected into a two-dimensional space. Cells that are closer have more similar molecular profiles. Different colors indicate different clusters of cells, which have been determined through a clustering algorithm. Clusters are usually identified with cell types. c) Inferred gene regulatory network. Circles show transcription factors and their target genes. Edges correspond to the interactions between them. d) Each circle represents a cell in a two-dimensional space. Developmental dynamics, indicated by a black line, are revealed with trajectory inference methods. The cell fate decision boundary separates molecular profiles that develop into different cell types. e) Each node represents a cell type. The leaves of the tree (lowest nodes) are the observed final cell types, whereas the other cell types are transient. Lines indicate the lineage relationships.

tory inference algorithms seek to reconstruct gene expression dynamics from scRNA-seq measurements of developing tissues. Many of these methods use similarity of transcriptomes to estimate temporal proximity, which comes with many challenges and limitations [102, 127, 139, 140]: For example, the starting point of a differentiation trajectory has to be provided by the user, because most methods cannot infer directionality. One exception is RNA velocity [141, 142], which exploits RNA splicing dynamics to infer gene expression dynamics and directionality. Time-resolved measurements can be analyzed with optimal transport theory to infer probabilities for the transitions between the observed cell types [14, 106].

Prospective lineage tracing presents a completely different set of challenges for data analysis. The increase in data complexity caused by randomly inserted barcodes, necessitates the development of novel algorithms to infer the underlying phylogenetic tree [127, 143], which captures the hierarchy and relationship of cells during differentiation. As barcoding is often limited to a short period of time, it becomes difficult to infer the lineage tree beyond the point where barcoding has stopped. However, a new method [127] leverages covariances between barcodes to transcend this limitation. An interesting concept in this regard is phylodynamics [125], which studies how the cell type distribution changes over

time, given an observed lineage tree. For example, it has been shown that a model with constant cell division rate, can result in a skewed lineage tree that appears like earlier generations were dividing more rapidly [125].

The algorithms mentioned here are just a small selection of the many tools that have been developed specifically to deal with the challenges arising in single-cell methods. We refer the reader to [144–146] for a much more comprehensive overview.

Box 3: A selection of computational methods for the analysis of single-cell omics data

Clustering methods have been used extensively for transcriptomics data (reviewed in [99]), where they partition cells based on the similarity of their transcriptomes. Clustering is now also applied to a combination of different omics data sets (reviewed in [147–149]). Clusters can be first obtained separately for each modality and then combined, or the different data sets are integrated prior to clustering. Popular examples of integration methods are WNN [41], totalVI [137], MOFA+ [135] and LIGER [138].

Inference of gene regulatory networks has been frequently performed using scRNA-seq data [150]. Examples of existing algorithms are GENIE3 [131], SCENIC [132], SINCERITIES [133] and Scribe [151]. A new method, CellOracle [134], allows the identification of gene regulatory networks from a combination of scRNA-seq and scATAC-seq data. Symphony [152] provides multi-omics clustering as well as gene regulatory network inference. Importantly, these methods often rely on the proper identification of TFs and CREs.

Inference of differentiation trajectories was first introduced for transcriptomics data. These methods make use of the asynchrony during differentiation and order cells by developmental progress (pseudotime). Examples of trajectory inference methods are PAGA [153], DPT [154], Monocle3 [155], FateID [156] and Palantir [157] (reviewed in [103]). A pseudotime method that makes use of spliced and unspliced RNA is RNA velocity [141, 142], which has also been expanded to include protein dynamics [158]. In order to combine several transcriptomics data sets and recreate the differentiation trajectory, optimal transport theory has been applied [14, 106]. A new, interesting method is MATCHER [159], which infers pseudotime based on multi-omics assays.

Reconstruction of lineage trees is the goal of dynamic lineage tracing techniques, where barcodes are introduced randomly during a short period of time. Classic reconstruction methods, like neighbor joining [160] are not robust enough for this purpose. Several studies therefore designed custom made methods [118, 161] and additionally, a new, more robust inference method has been proposed, Cassiopeia [143]. Building on the neighbor joining algorithm, CLiNC [127] tries to discover inconsistencies within the phylogenic tree.

1.7 Conceptual framework

Even with appropriate data analysis algorithms in place, we still need a conceptual framework for the quantitative understanding of cell types and their formation. The challenge is to reveal, how gene regulatory networks with certain topologies give rise to the observed cell types and molecular dynamics during differentiation. Dynamical systems theory has been used extensively to model gene regulatory networks quantitatively. In this framework, cell types can be understood as **stable states** in a system of **coupled differential equations** [162]. Number, position and robustness of these stable states all depend on parameters that reflect the interactions between TFs and other members of the regulatory network. These parameters can be difficult to infer from experiments, except for (unrealis-

> **Coupled differential equations**:
> Differential equations describe the temporal evolution of a system. They are coupled, if variables appear in several equations. Such equations can have multiple stable solutions, which do not evolve in time, unless perturbed.
> **Critical point**:
> A point in parameter space where the number or stability of solutions to a dynamical system change abruptly.
> Intrinsically disordered regions: Segments of a protein that do not form a stable three-dimensional structure.
> **Stable states**:
> A solution of a dynamical system that is a local minimum of the corresponding potential landscape.

tically) small networks. Nevertheless, dynamical systems describe key properties of the differentiation process. They explain how the interactions between several TFs jointly give rise to cell types that are robust up to a certain level of perturbation [163]. They also explain how a change in TF interactions causes cell types to destabilize [164]. Finally, unstable, intermediate cell states can be found, depending on the parameters of the system [165, 166].

A dynamical systems model can be represented by a potential energy landscape, where a cell follows the path of steepest decent into locally stable states, that correspond to cell types [166–168], see Figure 1.1, middle right. This potential energy landscape is closely related to Waddington's epigenetic landscape [169], a pioneering metaphor that abstracted from molecular details to conceptualize embryonic development. Importantly, the shape of Waddington's landscape is constant in time and a location in the landscape corresponds to the complete molecular profile of a cell. By contrast, most dynamical systems models identify the state of a cell by its transcriptome or even just the expression levels of the TFs in a gene regulatory network (see Figure 1.1, middle left). The shape of the potential landscape is then defined by the gene regulatory network, most importantly the interactions between TFs and their target genes [162]. Changes in the epigenetic state and other gene regulatory molecules can modulate the strength of those interactions (i.e. the parameters of the dynamic system) and thereby cause different stable and unstable states to appear or disappear [166, 167]. Defining the gene expression profile as the state of the cell and modeling the epigenetic profile as parameters of the gene regulatory network has certain conceptual advantages. For example, at **critical points**, which have been studied extensively by catastrophe theory, small changes of the parameters can cause large changes in the stable states of a dynamical system [162, 166]. Lineage decisions might thus be driven by dynamic epigenetic changes around critical points. Importantly, Waddington's landscape implies a strict hierarchy of differentiation, leading from multipotent to more and more specified, unipotent states (see Figure 1.1, right).

Despite its many advantages, the landscape model also has clear drawbacks, including its inability to describe periodic trajectories, e.g. caused by the cell cycle [102, 167]. Therefore,

many other ways to conceptualize differentiation have been devised. For example, the spin glass, a model that originated in physics, describes a system of interacting particles that can have stable low energy states corresponding to different cell types [162, 170]. It accommodates different strengths of interactions between TFs, can describe symmetry breaking events and is scalable to larger numbers of TFs. However, it is often simplified by the usage of binary TF expression (on/off) and symmetric interactions for mathematical tractability. The concepts discussed here are just a few examples of the many models that are currently being developed. More comprehensive overviews can be found in [162, 171].

1.8 Perspectives

- To discover the molecular underpinnings of cell types and their formation is of fundamental interest in developmental and stem cell biology. It is equally important for the understanding of diseases such as cancer, where cell types lose their stability and are transformed to malignant states.

- New single-cell measurement techniques have given us unprecedented insights into the interactions and dynamics of the relevant molecular agents. In the current paradigm, transcription factors, regulatory DNA elements and other classes of molecules form a regulatory network from which cell types emerge.

- In the future, lineage tracing and other quantitative methods will be leveraged to reveal the complete lineage tree and infer a predictive mathematical model of the underlying gene regulatory network. Such a model would allow us to manipulate cell types at will, which has numerous medical applications.

Author Contributions
S.S. and M.M. wrote the manuscript. M.M. created the figures.

Competing Interests
The authors declare no competing interests.

1

1.9 Outline of Thesis

Despite being the object of intense study, embryonic development has been difficult to model for several reasons. One primary reason is that complex tissues can comprise many cell types, of which we probably only know a subset. Thus, we first focus on discovering cell types with single-cell RNA sequencing (scRNA-seq), which has proven very successful. Throughout this thesis, we will use scRNA-seq to uncover different cell types and transcriptional changes during cellular differentiation.

Additionally, many signaling processes and morphogenic events co-occur in a developing tissue, making it hard to isolate the cell's individual contributions. For this purpose, we look at stem cell-derived in vitro systems, in which a small number of specific cell types can be studied in isolation. In this thesis, we mainly focus on stem-cell-derived endothelial cells, which line the inside of vessels, and gastruloids which mimic the process of gastrulation. We use scRNA-seq to investigate changes in gene expression patterns of master transcription factors and their target genes. Cellular differentiation, particularly, depends on extensive communication between cells and must lead to the formation of non-trivial spatial patterns in a robust and reproducible way. For this purpose, we analyze different in vitro model systems showing that cellular communication causes morphogenic and transcriptional changes in the developing cells. Specifically, we show in one chapter that cellular communication can overrule the developmental origin and in another that it can cause the formation of epithelial structures.

Lastly, we want to use measurements of developmental processes to reveal the underlying regulatory mechanisms. To that end, we use a neural network to infer the parameters of a model for gene regulation and cellular communication.

In **Chapter 2** we develop a new method to assess clusterability in single-cell transcriptomics data. Single-cell transcriptomics data has revolutionized biology by its ability to find new cell phenotypes through clustering. However, all clustering methods have adjustable parameters, making it challenging to find the correct number of clusters. There was no principled method to decide whether a cluster of cells contains meaningful sub-populations that can be further resolved. We developed phiclust (ϕ_{clust}) a clusterability measure derived from random matrix theory and low-rank perturbation theory that can identify the presence of non-random substructures. We showed that, by using this method, we could identify previously overlooked subtypes.

In **Chapter 3** we analyze the co-differentiation of endothelial cells and cardiomyocytes from human-induced pluripotent stem cells (hiPSCs). Both cell types are building blocks of the heart: Cardiomyocytes generate contractile forces, and endothelial cells line the inside of blood vessels. During embryonic development, these two cell types arise from a common hematoendothelial lineage. A known master regulator of this specification in mice is ETV2. We used scRNA-seq and a reporter cell line to uncover the role of ETV2 in the differentiation from human iPSCs. We showed that a transient expression of an ETV2-high state initiates the specification of endothelial cells. Functional cardiomyocytes can arise from cells that do not express ETV2 or, surprisingly, have sub-threshold expression.

In **Chapter 4** we investigate the importance of cellular communication compared to developmental origin in a 3D model of endothelial cell differentiation from hiPSCs. Endothelial cells appear across different organs where, on the one hand, their common function is to create a barrier between a liquid, such as blood, and the surrounding tissue. On the other hand, each vessel has specific requirements for each environment and liquid. We lacked an understanding of how endothelial cells acquire their organ-specific function. We investigated the hypothesis that developmental origin plays a role in the specification. For this purpose, endothelial cells were differentiated from paraxial or cardiac mesoderm. We showed that upon integration into a 3D microtissue, including cardiomyocytes and fibroblasts, the transcriptomic signature of the developmental origin is partially removed. This finding suggests that environmental cues might be more critical in function specification than developmental origin.

In **Chapter 5** we characterize the effect of cellular communication between embryonic and extra-embryonic cell types in an enhanced gastruloid system. In particular, we combined in vitro mouse embryonic stem cells (mESCs), used in regular gastruloid protocols, with extraembryonic-endoderm (XEN) cells. With this addition, we observed the formation of a neural epithelium absent in gastruloids derived only from mESCs. We characterized the neural epithelia with scRNA-seq, imaging, and differentiation protocols and observed similarities with the formation of a neural tube with dorsal characteristics. The XEN cells influence morphogenic changes in mESCs, and we also showed that mESCs induce differentiation in the XEN cells. We analyzed candidates of signaling pathways between these cell types and found that local inhibition of WNT signaling is one of the processes.

In **Chapter 6** we explore the use of physics-informed neural networks (PINNs) for the inference of gene regulatory networks (GRNs). GRNs regulate, among many other things, the robust differentiation of a progenitor into specified cell types. Existing methods to infer GRNs mostly use correlation or other similarity-based metrics which limits the predictive power of the resulting model. In order to incorporate prior biological knowledge and thereby make the inference of mechanistic relationships feasible, we chose to use PINNs. We analyze two relevant experimental scenarios in detail: In the first scenario single-cell trajectories are available and cells communicate with each other. In the other scenario, cells do not communicate and the provided data is only a snapshot, which corresponds to a destructive single-cell RNA-sequencing measurement. In both cases, a PINN is used to infer the strengths of gene interactions. We show the benefits of using PINNs compared to regular feedforward NNs in this context and determine the performance level of PINNs for different experimental designs. This analysis will serve as the starting point for exploring the great potential of PINNs for GRN inference.

1

References

[1] M. Mircea and S. Semrau. How a cell decides its own fate: a single-cell view of molecular mechanisms and dynamics of cell-type specification. *Biochemical Society Transactions*, dec 2021.

[2] S. A. Morris. The evolving concept of cell identity in the single cell era. *Development (Cambridge)*, 146(12), jun 2019.

[3] D. Arendt et al. The origin and evolution of cell types, dec 2016.

[4] A. F. Savulescu et al. Pinpointing Cell Identity in Time and Space. *Frontiers in Molecular Biosciences*, 7:209, aug 2020.

[5] B. Xia and I. Yanai. A periodic table of cell types. *Development (Cambridge)*, 146(12), jun 2019.

[6] N. Almeida et al. Employing core regulatory circuits to define cell identity. *The EMBO Journal*, 40(10):e106785, may 2021.

[7] C. Dupont, D. R. Armant, and C. A. Brenner. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Seminars in reproductive medicine*, 27(5):351, sep 2009.

[8] C. D. Allis and T. Jenuwein. The molecular hallmarks of epigenetic control, aug 2016.

[9] R. Paro, U. Grossniklaus, R. Santoro, and A. Wutz. *Introduction to Epigenetics*. Learning Materials in Biosciences. Springer International Publishing, Cham, 2021.

[10] E. Armingol, A. Officer, O. Harismendy, and N. E. Lewis. Deciphering cell–cell interactions and communication from gene expression, feb 2021.

[11] A. A. Almet, Z. Cang, S. Jin, and Q. Nie. The landscape of cell–cell communication through single-cell transcriptomics. *Current Opinion in Systems Biology*, 26:12–23, jun 2021.

[12] N. Eling, M. D. Morgan, and J. C. Marioni. Challenges in measuring and understanding biological noise. *Nature Reviews Genetics*, 20(9):536–548, may 2019.

[13] A. Guillemin and M. P. Stumpf. Noise and the molecular processes underlying cell fate decision-making. *Physical Biology*, 18(1):11002, jan 2021.

[14] M. Mittnenzweig et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell*, 184(0):1–18, apr 2021.

[15] R. Stadhouders, G. J. Filion, and T. Graf. Transcription factors and 3D genome conformation in cell-fate decisions, may 2019.

[16] K. S. Zaret. Pioneer Transcription Factors Initiating Gene Network Changes, nov 2020.

[17] S. L. Klemm, Z. Shipony, and W. J. Greenleaf. Chromatin accessibility and the regulatory epigenome, apr 2019.

[18] Y. Stelzer et al. Tracing Dynamic Changes of DNA Methylation at Single-Cell Resolution. *Cell*, 163(1):218–229, sep 2015.

[19] D. Song, D. Yang, C. A. Powell, and X. Wang. Cell–cell communication: old mystery and new opportunity, apr 2019.

[20] A. D. Riggs, R. A. Martienssen, and V. E. Russo. *Epigenetic Mechanisms of Gene Regulation.* Cold Spring Harbor Laboratory Press, 1996.

[21] C. H. Ludwig and L. Bintu. Mapping chromatin modifications at the single cell level. *Development (Cambridge)*, 146(12), jun 2019.

[22] Z. Cang et al. A multiscale model via single-cell transcriptomics reveals robust patterning mechanisms during early mammalian embryo development. *PLOS Computational Biology*, 17(3):e1008571, mar 2021.

[23] S. Bessonnard et al. Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development (Cambridge)*, 141(19):3637–3648, oct 2014.

[24] N. Schrode, N. Saiz, S. D. Talia, and A.-K. Hadjantonakis. GATA6 Levels Modulate Primitive Endoderm Cell Fate Choice and Timing in the Mouse Blastocyst. *Developmental Cell*, 29(4):454–467, may 2014.

[25] T. Graf and T. Enver. Forcing cells to change lineages, dec 2009.

[26] D. Hnisz et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, 155(4):934, nov 2013.

[27] W. A. Whyte et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, apr 2013.

[28] S. V. Razin and A. A. Gavrilov. The Role of Liquid–Liquid Phase Separation in the Compartmentalization of Cell Nucleus and Spatial Genome Organization, jun 2020.

[29] A. S. Lyon, W. B. Peeples, and M. K. Rosen. A framework for understanding the functions of biomolecular condensates across scales, mar 2021.

[30] S. F. Banani, H. O. Lee, A. A. Hyman, and M. K. Rosen. Biomolecular condensates: Organizers of cellular biochemistry, may 2017.

[31] E. Gomes and J. Shorter. The molecular language of membraneless organelles, may 2019.

[32] M. Grosch, S. Ittermann, D. Shaposhnikov, and M. Drukker. Chromatin-Associated Membraneless Organelles in Regulation of Cellular Differentiation, dec 2020.

[33] B. R. Sabari et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400):eaar3958, jul 2018.

1

[34] M. Esposito et al. TGF-β-induced DACT1 biomolecular condensates repress Wnt signalling to promote bone metastasis. *Nature Cell Biology*, 23(3):257–267, mar 2021.

[35] Y. E. Guo et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770):543–548, aug 2019.

[36] W. K. Cho et al. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415, jul 2018.

[37] M. Hochane et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biology*, 17(2):e3000152, feb 2019.

[38] Y. Ando, A. T. J. Kwon, and J. W. Shin. An era of single-cell genomics consortia, sep 2020.

[39] A. Regev et al. The human cell atlas. *eLife*, 6, dec 2017.

[40] N. Schaum et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, oct 2018.

[41] Y. Hao et al. Integrated analysis of multimodal single-cell data. *Cell*, 0(0), may 2021.

[42] S. C. Bendall et al. Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum. *Science*, 332(6030):687–696, may 2011.

[43] A. Tanay and A. Regev. Scaling single-cell genomics from phenomenology to mechanism, jan 2017.

[44] J. D. Buenrostro et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, jul 2015.

[45] D. A. Cusanovich et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, may 2015.

[46] J. Cao et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, sep 2018.

[47] S. Picelli et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098, sep 2013.

[48] E. Z. Macosko et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, may 2015.

[49] T. Hashimshony et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1):1–7, apr 2016.

[50] J. Cao et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, aug 2017.

[51] B. Budnik, E. Levy, G. Harmange, and N. Slavov. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biology*, 19(1):1–12, oct 2018.

[52] H. Specht et al. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biology*, 22(1):1–27, jan 2021.

[53] E. M. Schoof et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nature Communications*, 12(1):1–15, jun 2021.

[54] M. Stoeckius et al. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14(9):865–868, aug 2017.

[55] V. M. Peterson et al. Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, 35(10):936–939, aug 2017.

[56] C. Zhu et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural and Molecular Biology*, 26(11):1063–1070, nov 2019.

[57] S. A. Smallwood et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820, jul 2014.

[58] M. Farlik, N. C. Sheffield, J. Klughammer, and C. Bock Correspondence. Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *CellReports*, 10:1386–1397, 2015.

[59] H. Guo et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Research*, 23(12):2126–2135, dec 2013.

[60] D. Mooijman et al. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature Biotechnology*, 34(8):852–856, jan 2016.

[61] C. Angermueller et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, 13(3):229–232, feb 2016.

[62] Y. Hu et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biology*, 17(1):1–11, may 2016.

[63] S. J. Clark et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature Communications*, 9(1):1–9, feb 2018.

[64] A. Rotem et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11):1165–1172, nov 2015.

[65] H. S. Kaya-Okur et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1–10, dec 2019.

1

1

[66] H. Xiong et al. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nature Methods*, pages 1–9, may 2021.

[67] C. Zhu et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature Methods*, 18(3):283–292, mar 2021.

[68] V. Ramani et al. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3):263–266, jan 2017.

[69] T. Nagano et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, sep 2013.

[70] D.-S. Lee et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature Methods*, 16(10):999–1006, sep 2019.

[71] G. Li et al. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods*, 16(10):991–993, aug 2019.

[72] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science*, 338(6114):1622–1626, dec 2012.

[73] F. Abascal et al. Somatic mutation landscapes at single-molecule resolution. *Nature*, 593(7859):405–410, may 2021.

[74] I. C. Macaulay et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522, apr 2015.

[75] A. Rodriguez-Meira et al. Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell*, 73(6):1292–1305.e8, mar 2019.

[76] S. V. Ulianov et al. Suppression of liquid–liquid phase separation by 1,6-hexanediol partially compromises the 3D genome organization in living cells. *Nucleic Acids Research*, 49(18):10524–10541, oct 2021.

[77] H. Belaghzal et al. Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. *Nature Genetics*, 53(3):367–378, feb 2021.

[78] K. Holler et al. Spatio-temporal mRNA tracking in the early zebrafish embryo. *Nature Communications*, 12(1):3358, dec 2021.

[79] C.-H. L. Eng et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, 568(7751):235–239, mar 2019.

[80] E. Lundberg and G. H. Borner. Spatial proteomics: a powerful discovery tool for cell biology, may 2019.

[81] E. Shema, B. E. Bernstein, and J. D. Buenrostro. Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution, jan 2019.

[82] J. Lee, D. Y. Hyeon, and D. Hwang. Single-cell multiomics: technologies and data analysis methods, sep 2020.

[83] A. Mayran et al. Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nature Genetics*, 50(2):259–269, jan 2018.

[84] S. R. Joseph et al. Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *eLife*, 6, apr 2017.

[85] C. R. Clapier and B. R. Cairns. The Biology of Chromatin Remodeling Complexes. *Annual Review Biochem*, 78:273–304, jun 2009.

[86] M. Tyagi, N. Imam, K. Verma, and A. K. Patel. Chromatin remodelers: We are the drivers!! *Nucleus*, 7(4):388, jul 2016.

[87] V. Azuara et al. Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8(5):532–538, mar 2006.

[88] B. E. Bernstein et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2):315–326, apr 2006.

[89] E. Blanco et al. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends in Genetics*, 36(2):118–131, feb 2020.

[90] R. Margueron et al. Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature*, 461(7265):762–767, sep 2009.

[91] L. Di Croce and K. Helin. Transcriptional regulation by Polycomb group proteins. *Nature Structural and Molecular Biology*, 20(10):1147–1155, oct 2013.

[92] S. Kundu et al. Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. *Molecular Cell*, 65(3):432–446.e5, feb 2017.

[93] S. Cruz-Molina et al. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell*, 20(5):689–705.e9, may 2017.

[94] Y. Atlasi and H. G. Stunnenberg. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, 18(11):643–658, aug 2017.

[95] K. Daneshvar et al. lncRNA DIGIT and BRD3 protein form phase-separated condensates to regulate endoderm differentiation. *Nature Cell Biology*, 22(10):1211–1222, oct 2020.

[96] M. Modic et al. Cross-Regulation between TDP-43 and Paraspeckles Promotes Pluripotency-Differentiation Transition. *Molecular Cell*, 74(5):951–965.e13, jun 2019.

[97] M. Grosch et al. Nucleus size and DNA accessibility are linked to the regulation of paraspeckle formation in cellular differentiation. *BMC Biology*, 18(1):1–19, apr 2020.

1

[98] G. J. Knott, C. S. Bond, and A. H. Fox. The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. *Nucleic Acids Research*, 44(9):3989–4004, may 2016.

[99] S. Wang et al. Long noncoding RNA Neat1 modulates myogenesis by recruiting Ezh2. *Cell Death and Disease*, 10(7):1–15, jun 2019.

[100] J. E. Sulston and H. R. Horvitz. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental Biology*, 56(1):110–156, mar 1977.

[101] Sagar and D. Grün. Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis. *Annual Review of Biomedical Data Science*, 3(1):1–22, jul 2020.

[102] C. Weinreb et al. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*, 115(10):E2467–E2476, mar 2018.

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

[104] B. Pijuan-Sala et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, feb 2019.

[105] J. Delile et al. Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord. *Development*, 146(12):dev173807, jun 2019.

[106] G. Schiebinger et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.e22, feb 2019.

[107] S. Nowotschin et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, 569(7756):361–367, apr 2019.

[108] C. G. Palii et al. Single-Cell Proteomics Reveal that Quantitative Changes in Co-expressed Lineage-Specific Transcription Factors Determine Cell Fate. *Cell Stem Cell*, 24(5):812–820.e5, may 2019.

[109] A. S. Nam et al. Somatic mutations and cell identity linked by Genotyping of Transcriptomes. *Nature*, 571(7765):355–360, jul 2019.

[110] A. Abyzov and F. M. Vaccarino. Cell Lineage Tracing and Cellular Diversity in Humans. *Annual Review of Genomics and Human Genetics*, 21:101–116, sep 2020.

[111] J. Xu et al. Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *eLife*, 8, 2019.

[112] F. Gaiti et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, may 2019.

[113] C. J.-Y. Wei and K. Zhang. RETrace: simultaneous retrospective lineage tracing and methylation profiling of single cells. *Genome Research*, 30(4):gr.255851.119, mar 2020.

[114] D. E. Wagner and A. M. Klein. Lineage tracing meets single-cell omics: opportunities and challenges, jul 2020.

[115] C. S. Baron and A. van Oudenaarden. Unravelling cellular relationships during development and regeneration using genetic lineage tracing, dec 2019.

[116] A. Alemany et al. Whole-organism clone tracing using single-cell sequencing. *Nature*, 556(7699):108–112, apr 2018.

[117] B. Raj, J. A. Gagnon, and A. F. Schier. Large-scale reconstruction of cell lineages using single-cell readout of transcriptomes and CRISPR–Cas9 barcodes by scGESTALT. *Nature Protocols*, 13(11):2685–2713, nov 2018.

[118] B. Spanjaard et al. Simultaneous lineage tracing and cell-type identification using CrIsPr-Cas9-induced genetic scars. *Nature Biotechnology*, 36(5):469–473, jun 2018.

[119] B. Carter and K. Zhao. The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics*, 22(4):235–250, nov 2020.

[120] S. Preissl et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature Neuroscience*, 21(3):432–439, mar 2018.

[121] G. Jia et al. Single cell RNA-seq and ATAC-seq analysis of cardiac progenitor cell transition states and lineage settlement. *Nature Communications*, 9(1):1–17, dec 2018.

[122] A. E. Trevino et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution, dec 2020.

[123] R. Argelaguet et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491, dec 2019.

[124] A. McKenna and J. A. Gagnon. Recording development with single cell dynamic lineage tracing. *Development (Cambridge)*, 146(12), jun 2019.

[125] T. Stadler, O. G. Pybus, and M. P. Stumpf. Phylodynamics for cell biologists, jan 2021.

[126] C. Weinreb and A. M. Klein. Lineage reconstruction from clonal correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 117(29):17041–17048, jul 2020.

[127] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), feb 2020.

[128] F. Wagner, Y. Yan, and I. Yanai. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*, page 217737, 2018.

1

1

[129] A. E. Rodriguez-Fraticelli et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature*, 553(7687):212–216, jan 2018.

[130] M. Krzak et al. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. *Frontiers in Genetics*, 10:1253, dec 2019.

[131] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):12776, 2010.

[132] S. Aibar et al. SCENIC: Single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, oct 2017.

[133] N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, and R. Gunawan. SINCERITIES: Inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics*, 34(2):258–266, jan 2018.

[134] K. Kamimoto, C. M. Hoffmann, and S. A. Morris. CellOracle: Dissecting cell identity via network inference and in silico gene perturbation, feb 2020.

[135] R. Argelaguet et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):1–17, may 2020.

[136] A. Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, apr 2018.

[137] A. Gayoso et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nature Methods*, 18(3):272–282, feb 2021.

[138] J. D. Welch et al. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell*, 177(7):1873–1887.e17, jun 2019.

[139] S. Tritschler et al. Concepts and limitations for learning developmental trajectories from single cell genomics, jun 2019.

[140] C. Weinreb, S. Wolock, and A. M. Klein. SPRING: A kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, 2018.

[141] V. Bergen et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, dec 2020.

[142] G. La Manno et al. RNA velocity of single cells. *Nature*, 560(7719):494–498, aug 2018.

[143] M. G. Jones et al. Inference of single-cell phylogenies from lineage tracing data using Cassiopeia. *Genome Biology*, 21(1):1–27, apr 2020.

[144] H. Nguyen et al. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data. *Briefings in Bioinformatics*, 22(3):1–15, may 2021.

1

[145] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols*, 16(1):1–9, dec 2020.

[146] T. Stuart and R. Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, jan 2019.

[147] N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546, nov 2018.

[148] M. Colomé-Tatché and F. J. Theis. Statistical single cell multi-omics integration. *Current Opinion in Systems Biology*, 7:54–59, feb 2018.

[149] M. Efremova and S. A. Teichmann. Computational methods for single-cell omics across modalities. *Nature Methods*, 17(1):14–17, jan 2020.

[150] A. Pratapa et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, feb 2020.

[151] X. Qiu et al. Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe. *Cell Systems*, 10(3):265–274.e11, mar 2020.

[152] C. Burdziak, E. Azizi, S. Prabhakaran, and D. Pe'er. A Nonparametric Multi-view Model for Estimating Cell Type-Specific Gene Regulatory Networks. *arXiv*, feb 2019.

[153] F. A. Wolf et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):1–9, mar 2019.

[154] L. Haghverdi et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods 2016 13:10*, 13(10):845–848, aug 2016.

[155] J. Cao et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, feb 2019.

[156] J. S. Herman, Sagar, and D. Grün. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386, apr 2018.

[157] M. Setty et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37(4):451–460, mar 2019.

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

[159] J. D. Welch, A. J. Hartemink, and J. F. Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, 18(1):1–19, jul 2017.

[160] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, jul 1987.

1

[161] M. M. Chan et al. Molecular recording of mammalian embryogenesis. *Nature*, 570(7759):77–82, jun 2019.

[162] A. E. Teschendorff and A. P. Feinberg. Statistical mechanics meets single-cell biology. *Nature reviews. Genetics*, pages 1–18, apr 2021.

[163] C. Furusawa and K. Kaneko. A Dynamical-Systems View of Stem Cell Biology. *Science*, 338(6104):215–217, oct 2012.

[164] J. J. Tyson and B. Novak. A Dynamical Paradigm for Molecular Cell Biology. *Trends in Cell Biology*, 30(7):504–515, jul 2020.

[165] A. L. MacLean, T. Hong, and Q. Nie. Exploring intermediate cell states through the lens of single cells, jun 2018.

[166] A. Guillemin and M. P. H. Stumpf. Non-equilibrium statistical physics, transitory epigenetic landscapes, and cell fate decision dynamics. *Mathematical Biosciences and Engineering*, 17(6):7916–7930, nov 2020.

[167] P. Greulich, R. Smith, and B. D. MacArthur. The physics of cell fate. In *Phenotypic Switching*, pages 189–206. Elsevier, jan 2020.

[168] L. Xu and J. Wang. Quantifying Waddington landscapes, paths, and kinetics of cell fate decision making of differentiation/development. In *Phenotypic Switching*, pages 157–187. Elsevier, jan 2020.

[169] C. Waddington. *The strategy of the genes : a discussion of some aspects of theoretical biology / by C.H. Waddington. | Wellcome Collection.* Allen and Unwin, London, 1975.

[170] A. H. Lang, H. Li, J. J. Collins, and P. Mehta. Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. *PLOS Computational Biology*, 10(8):e1003734, aug 2014.

[171] P. Cahan et al. Computational Stem Cell Biology: Open Questions and Guiding Principles. *Cell Stem Cell*, 28(1):20–32, jan 2021.