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 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.

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 [1,2]. 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, 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 the 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 [1,3–5]. 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 exception of changes in the DNA sequence [6,7]. 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 [8].

Here, we will review few of the many cell-autonomous molecular mechanisms 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 [9,10]). 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 [11,12]). 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).

**Molecular embodiment of a cell type**

Most, if not all, cell-type decisions involve specific transcription factors (TFs) [1,2,4,5,13,14]. 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 in 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, nucleosomes, and other associated proteins, see Figure 1, left). With the exception of so-called pioneer factors [15], TFs only bind accessible, nucleosome-free DNA [16]. Chromatin configuration

---

**Figure 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 1:10^6. 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.

---

© 2021 The Author(s). This is an open access article published by Portland Press Limited on behalf of the Biochemical Society and distributed under the Creative Commons Attribution License 4.0 (CC BY-NC-ND).
and TF binding are affected by chemical modifications of histones (the components of a nucleosome), as well as the DNA [16]. 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 [17,18]. For example, DNA methylation at the enhancer regions of the pluripotency gene Sox2 results in silencing of its expression in embryonic stem cells [18]. Importantly, the interaction between TFs and chromatin configuration is reciprocal: TFs recruit enzymes that locally change the molecular make-up of the chromatin [16]. Both histone marks and DNA methylation are heritable molecular marks, as they are copied to the newly synthesized DNA during cell division [7,8,19]. 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 [16,20].

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 [4]. Together with their target genes, TFs form gene regulatory networks that establish and maintain cell identity [2,4], 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 inputs [1]. Mutual repression of TFs is one mechanism by which multiple, alternative cell types can be created. A prominent example is an interaction between the TFs GATA6 and NANOG, which governs the lineage decision between two of the earliest cell types in the mammalian embryo [21–23]. 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, 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) [24]. 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 [25,26], see Figure 2. Super-enhancers owe their special role to a high density of co-localized Mediator complex [25,26], 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 [26].

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) [27–29], 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 [30]. The high concentration of several molecular species in the condensed phase leads to increased interaction rates [29]. Examples of biomolecular condensates are the well-known membrane-less organelles, such as the nucleolus or Cajal bodies, as well as paraspeckles and many more [27,28,31]. It has been found that intrinsically disordered regions (IDRs) of proteins can lead to multivalent interactions that can cause condensates to form [32–34]. Interestingly, MED1, a member of the Mediator complex, and BRD4, a coactivator of transcription, have large IDRs and form condensates at super-enhancers [32], see Figure 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 contact with multiple promoter sites [35]. Therefore, biomolecular condensates are likely of crucial importance for establishing a cell type.

**Molecular profiling**

In recent years, omics technologies have emerged that measure one or multiple molecular species comprehensively in single cells (see Box 1 and Figure 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 [1,4,5], which enables the identification of new cell types and cell states in complex tissues [36]. Multiple large consortia are currently generating transcriptional atlases of entire organisms (reviewed in [37]). The human cell atlas [38] and Tabula Muris [39] 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 [40,41].

As mentioned before, chromatin state is an important factor in gene regulation. Knowledge of the chromatin landscape can therefore improve the identification of cell types [42]. There is a growing variety of single-cell methods that measure chromatin features. For example, by using scATAC-seq [43,44], 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 with 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 with using promoter chromatin state alone [45].

Similarly, a simultaneous measurement of histone modifications and transcriptome showed that active enhancers are epigenetically more variable across cell types than promoter regions [46]. A related finding resulted from the simultaneous measurement of DNA methylation and transcriptome (scmK&T-seq [47]). The authors confirmed that promoter DNA methylation in mouse ESCs is typically correlated with reduced gene expression. In contrast, DNA methylation of distal enhancers is more often correlated with increased gene expression, compared with 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 [48,49]. In combination with DNA methylation measurements, cell-type-specific chromatin conformations can be obtained [50], which might help to clarify the role of biomolecular condensates [51,52]. In a recent study, a new variant of Hi-C [52] 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 [53,54] and proteins [55] 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 [56]. A more comprehensive discussion of these methods can be found in [57].

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 bind to accessible chromatin regions, differentiation is accompanied by pervasive changes in chromatin accessibility [12,15,16]. One underlying mechanism involves pioneer TFs, which bind to nucleosome-associated DNA and create an open chromatin state [15,16]. These TFs can explore nucleosomal DNA through non-specific and transient binding, which in turn allows partial opening of the chromatin and other, non-pioneering factors to bind [15]. This mechanism has been recently validated, for example, for the pioneer factor PAX7 [58]. Another mechanism is a passive competition of TFs for DNA binding during short periods of local chromatin opening, which increases and stabilizes with higher TF concentrations [16,59].

Chromatin state is also influenced by chromatin remodelers that are recruited by TFs [15,16,60] and bind to different histone marks [60,61]. 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 [62] and are more

---

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 [127], Drop-seq [128], CEL-seq2 [129], and Sci-RNA-seq [130].

Proteomics: It is not yet feasible to measure every protein in a single cell. Antibody-based methods, for example, scCyTOF [41], 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 [131], which detects around 1000 proteins per cell. Improvements to this method have been recently made in SCoPE2 [132] and another method [133]. By sequencing of DNA-tagged antibodies, the quantification of hundreds of proteins together with the transcriptome in single cells is possible with CITE-seq [134] and REAP-seq [135]. 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 [43,44], which uses transposons to barcode accessible DNA. scATAC-seq can be performed simultaneously with scRNA-seq, which was implemented, for example, by sci-CAR [45] and Paired-seq [88]. DNA methylation is measured by scBS-seq [137,138] and scRRBS [139]. Single-cell 5hmC-seq measures DNA hydroxymethylation [140]. Joint measurements of DNA methylation and the transcriptome are possible with, for example, scM&T-seq [47] and scMT-seq [141]. A method to measure all three molecular profiles (DNA methylation, transcriptome, and chromatin accessibility) is scNMT-seq [142]. Histone modifications can be measured, for example, with scChIP-seq [143] and scCUT&Tag [144]. New methods to measure the transcriptome jointly with histone modifications are CoTech [90] and PairedTag [46]. It is now also possible to study the chromatin conformation in every single cell with scHi-C [48,49]. Recent methods have allowed the capture of both chromatin conformation and DNA methylation (sn-m3C-seq [50] and methyl-HiC [145]).

Genomics: The DNA sequence of single cells can be measured by methods such as MALBAC [146] and NanoSeq [147]. NanoSeq 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 [148] and TARGET-seq [149].
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 [62–64]. 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 the methylation of H3K27 [65,66]. In ESCs, the occupancy of PcG proteins at bivalent histone marks can change during differentiation, which results in altered gene expression [64,67]. Poised enhancers have been found to be necessary, for example, for the differentiation into specific neural cell types [68]. The examples mentioned here are only a few of many epigenetic mechanisms that drive dynamic chromatin remodeling during stem cell differentiation (reviewed in [69]).

Figure 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.
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) [31,70]. For example, the lncRNA DIGIT forms biomolecular condensates together with the RNA binding protein BRD3, which contains an IDR [70]. 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 [71], see Figure 2. These condensates can, for example, influence transcriptional regulation via associated RBPs [71–73]. NEAT1 has been found to physically interact with EZH2, a PcG protein, which is involved in catalyzing histone methylation [74]. Interestingly, paraspeckles were found to be involved in slowing down the differentiation process and their number changes dynamically during differentiation to several lineages [31,72]. Another example of dynamic transcriptional regulation through biomolecular condensates is the association of RNA splicing at genes associated with super-enhancers [34].

**Measurement of differentiation dynamics**

In simple organisms, the entire lineage tree can be assembled using a microscope [75]. In larger organisms that becomes unfeasible and scRNA-seq data of developing tissues have been used instead to infer lineage relationships [76]. Due to asynchrony in embryonic development or regeneration of adult tissues, a single scRNA-seq measurement can capture cells in different stages of differentiation [42,77] and developmental order, or pseudotime, can be inferred by computational methods (reviewed in [78], 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 [13,79–82]. This approach improves the temporal resolution and revealed that cells with different lineage histories can converge to globally similar cells [82]. 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 cells at different time points. In a recent study [83], 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.

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 [84], 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 [85], in particular in combination with transcriptomics [45,86–88]. One study found a class of genes with a high number of putative enhancers whose chromatin accessibility is predictive of gene expression [87]. 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 [86].

Another interesting case is the simultaneous measurement of chromatin accessibility, DNA methylation, and transcriptome (scNMT [89]) at several time points 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 [90]. 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 [46].

An entirely different approach to study developmental dynamics is used in lineage tracing techniques [91–93] (see Box 2), which aim to find the correct phylogenetic tree [94,95] from pluri-potent cells to fully specified cell
Lineage tracing methods have produced a large number of valuable insights. A recent study used lineage tracing to reveal early biases toward particular cell types [96] 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 [91,97]. 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 [98]. Finally, lineage tracing made it possible to observe the convergence of differentiation trajectories from distinct developmental origins [97].

Data analysis

Many single-cell methods involve advanced data analysis (see Box 3 and Figure 4 for a selection of computational methods). In scRNA-seq data, cell types can in principle be identified by clustering similar transcriptional profiles [99] and the underlying gene regulatory networks can be inferred [100–103]. However, both cell-type identification and network inference are improved by integrating multiple omics data sets [40,104]. Integration methods typically aim to extract variations common to all measured modalities [105–107]. 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 [108].

Trajectory inference algorithms seek to reconstruct gene expression dynamics from scRNA-seq measurements of developing tissues. Many of these methods use the similarity of transcriptomes to estimate temporal proximity, which comes with many challenges and limitations [77,96,109,110]: 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 [111,112], which exploits RNA splicing dynamics to infer gene expression
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 [158–160]). 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 [40], totalVI [107], MOFA+ [105] and LIGER [108].

Inference of gene regulatory networks has been frequently performed using scRNA-seq data [102]. Examples of existing algorithms are GENIE3 [100], SCENIC [101], SINCERITIES [103], and Scribe [161]. A new method, CellOracle [104], allows the identification of gene regulatory networks from a combination of scRNA-seq and scATAC-seq data. Symphony [162] 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 [163], DPT [164], Monocle3 [165], FateID [166] and Palantir [167] (reviewed in [78]). A pseudotime method that makes use of spliced and unspliced RNA is RNA velocity [111,112], which has also been expanded to include protein dynamics [168]. To combine several transcriptomics data sets and recreate the differentiation trajectory, optimal transport theory has been applied [13,81]. A new, interesting method is MATCHER [169], 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 [170] are not robust enough for this purpose. Several studies therefore designed custom-made methods [157,171] and additionally, a new, more robust inference method has been proposed, Cassiopeia [113]. Building on the neighbor-joining algorithm, CLiNC [95] tries to discover inconsistencies within the phylogenetic tree.

Figure 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.
dynamics and directionality. Time-resolved measurements can be analyzed with optimal transport theory to infer probabilities for the transitions between the observed cell types [13,81].

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 [96,113], 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 [95] leverages covariances between barcodes to transcend this limitation. An interesting concept in this regard is phylodynamics [94], 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 a constant cell division rate can result in a skewed lineage tree that appears like earlier generations were dividing more rapidly [94].

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 [114–116] for a much more comprehensive overview.

**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 [117]. The 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 (unrealistically) 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 [118]. They also explain how a change in TF interactions causes cell types to destabilize [119]. Finally, unstable, intermediate cell states can be found, depending on the parameters of the system [120,121].

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 [121–123], see Figure 1, middle right. This potential energy landscape is closely related to Waddington’s epigenetic landscape [124], 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. In contrast, most dynamical system 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, 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 [117]. 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 [121,122]. 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 [117,121]. 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, right).

Despite its many advantages, the landscape model also has clear drawbacks, including its inability to describe periodic trajectories, for example, caused by the cell cycle [77,122]. Therefore, many other ways to conceptualize differentiation have been devised. For example, 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 [117,125]. 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 [117,126].
Perspectives

- Importance of the field: 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.

- Summary of current thinking: 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.

- Comment on future directions: 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.

Glossary

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 is membrane-less organelles.

Bivalent domain: Chromatin domain that carries both activating and repressing histone marks.

Chromatin: The complex of nucleosomes, DNA, and other associated proteins.

Chromatin remodeler: Protein complexes that catalyze molecular changes of the chromosome, such as nucleosome removal.

Cis-regulatory elements: Sequences of non-coding DNA, which regulate the transcription of genes.

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.

Histone: Proteins that are crucial for the organization of DNA in the nucleus. DNA is tightly wound around nucleosome core particles which consist of eight histones.

Intrinsically disordered regions: Segments of a protein that do not form a stable three-dimensional structure.

Liquid–liquid phase separation: De-mixing of a homogeneous liquid into two distinct liquid phases.

Master transcription factor/master regulator: A transcription factor that affects the transcription of multiple downstream genes and is essential for cell-type specification.

Mediator: A multiprotein complex that coactivates transcription by interacting with TFs and RNA polymerase II.

Nucleosome: Smallest unit of DNA organization. Consists of a DNA wound around eight histones.

Paraspeckle: A biomolecular condensate that forms in the presence of the long non-coding RNA NEAT1 and several RNA binding proteins.

Pioneer factor: A transcription factor that can bind to nucleosome-bound DNA.

Pluripotency: Ability of a cell to give rise to multiple cell types.

Regulatory network: A system of interacting molecules that regulate each other’s gene expression as well as a set of target genes.

RNA Polymerase II: A multiprotein complex that transcribes DNA into messenger RNA.

Single-cell omics technologies: Experimental methods to measure the entire genome, epigenome, transcriptome, proteome, etc. of a cell in high throughput.
Stable states: A solution of a dynamical system that is a local minimum of the corresponding potential landscape.
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.
Transcription factor: A protein that binds to specific DNA sequences and regulates transcription.

Competing Interests
The authors declare that there are no competing interests associated with the manuscript.

Funding
M.M. and S.S. were supported by the Netherlands Organization for Scientific Research (NWO/OCW, www.nwo.nl), as part of the Frontiers of Nanoscience (NanoFront) program. We acknowledge funding by an NWO/OCW Vidi grant (016.Vidi.189.007) for S.S.

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

Abbreviations
CREs, cis-regulatory elements; ESCs, embryonic stem cells; IDRs, intrinsically disordered regions; RBPs, RNA-binding proteins; TFs, transcription factors.

References
1. Morris, S.A. (2019) The evolving concept of cell identity in the single cell era. Development 146, e169748 https://doi.org/10.1242/dev.169748
2. Arendt, D., Musser, J.M., Baker, C.V.H., Bergman, A., Cepko, C., Erwin, D.H. et al. (2016) The origin and evolution of cell types. Nat. Rev. Genet. 17, 744–757 https://doi.org/10.1038/nrg.2016.127
3. Savulescu, A.F., Jacobs, C., Negishi, Y., Davignon, L. and Mhlanga, M.M. (2020) Pinpointing cell identity in time and space. Development 146, 401–412 https://doi.org/10.1242/dev.146145
4. Dupont, C., Armant, D.R. and Brenner, C.A. (2009) Epigenetics: definition, mechanisms and clinical perspective. Semin. Reprod. Med. 27, 351–363 https://doi.org/10.1055/S-0029-1237423
5. Allis, C.D. and Jenuwein, T. (2016) The molecular hallmarks of epigenetic control. Nat. Rev. Genet. 17, 487–500 https://doi.org/10.1038/nrg.2016.59
6. Paro, R., Grossniklaus, U., Santoro, R. and Wutz, A. (2021) Epigenetics: definition, mechanisms and clinical perspective. Semin. Reprod. Med. 27, 351–363 https://doi.org/10.1055/S-0029-1237423
7. Almet, A.A., Cang, Z., Jin, S. and Nie, Q. (2021) The landscape of cell–cell communication through single-cell transcriptomics. Curr. Opin. Biol. 56, 12–23 https://doi.org/10.1016/j.cub.2021.03.007
8. Zaret, K.S. (2020) Pioneer transcription factors initiating gene network changes. Annu. Rev. Genet. 54, 367–385 https://doi.org/10.1146/annurev-genet-030220-015007
9. Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. 20, 207–220 https://doi.org/10.1038/s41576-019-0089-8
10. Riggs, A.D., Martienssen, R.A. and Russo, V.E. (1996) Epigenetic Mechanisms of Gene Regulation, Cold Spring Harbor Laboratory Press
11. Song, Y., van den Berg, P.R., Markoulaki, S., Soldner, F., Dal’Agne, A., Henninger, J.E. et al. (2019) Dynamic enhancer DNA methylation as basis for transcriptional and cellular heterogeneity of ESCs. Mol. Cell 75, 905–920 e6 https://doi.org/10.1016/J.MOLCEL.2019.06.045
12. Ludwig, C.H. and Bintu, L. (2019) Mapping chromatin modifications at the single cell level. Development 146, e170217 https://doi.org/10.1242/dev.170217
