ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis

Sagar[1] and Dominic Grün[1,2]

[1]Max Planck Institute of Immunobiology and Epigenetics, D-79108 Freiburg, Germany; email: gruen@ie-freiburg.mpg.de

[2]CIBSS (Centre for Integrative Biological Signaling Studies), University of Freiburg, D-79104 Freiburg, Germany

Annu. Rev. Biomed. Data Sci. 2020. 3:1–22

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-111419-091750

Keywords

cell fate, lineage specification, lineage tree, single-cell RNA-seq, differentiation trajectory, dimensionality reduction, lineage tracing

Abstract

Cellular differentiation is a common underlying feature of all multicellular organisms through which naïve cells progressively become fate restricted and develop into mature cells with specialized functions. A comprehensive understanding of the regulatory mechanisms of cell fate choices during development, regeneration, homeostasis, and disease is a central goal of modern biology. Ongoing rapid advances in single-cell biology are enabling the exploration of cell fate specification at unprecedented resolution. Here, we review single-cell RNA sequencing and sequencing of other modalities as methods to elucidate the molecular underpinnings of lineage specification. We specifically discuss how the computational tools available to reconstruct lineage trajectories, quantify cell fate bias, and perform dimensionality reduction for data visualization are providing new mechanistic insights into the process of cell fate decision. Studying cellular differentiation using single-cell genomic tools is paving the way for a detailed understanding of cellular behavior in health and disease.

INTRODUCTION

The multitude of complex functions performed by various tissues and organs of higher multicellular organisms require a large number of specialized cell types. The cell types eventually arise from a totipotent zygote in a process termed cellular differentiation. Besides cellular differentiation during ontogeny, many tissues host multipotent progenitors continuously differentiating into specialized cell types during homeostasis, as well as upon regeneration and during other challenges. Differentiation is a multilayered process through which immature or naïve cell states transform into increasingly specialized progenitors and eventually into mature cell types requiring sequential epigenetic and transcriptional changes affecting the biochemical properties of a cell. Although differentiation is frequently irreversible, cellular plasticity (i.e., trans- and dedifferentiation of mature cell types) can be observed, e.g., in the regenerating liver and limbs (1, 2). Importantly, cellular differentiation does not occur in isolation and is tightly coupled to paracrine and autocrine signals from the microenvironment in which stem and progenitor cells reside. Moreover, differentiation is contingent on the stochastic interactions of thousands of molecules in a cell and is affected by biological variability or noise (3). Thus, understanding differentiation requires insights into the process of cellular decision making and fate specification. Such insights are not only fundamental to understand embryonic development and tissue homeostasis but also important to uncover the mechanisms of cellular transformation in cancer, as well as in cellular reprogramming and dedifferentiation during regeneration.

Advances in profiling molecular features of cells such as messenger RNA (mRNA) (4, 5) and chromatin accessibility (6) at single-cell resolution are providing novel insights into the process of cell fate specification, e.g., during embryonic development (7–11), tissue turnover (12), and bone marrow hematopoiesis (13, 14). Such studies are challenging the classical view of cell fate commitment as a discrete binary decision process where immature multipotent progenitors become lineage restricted in a stepwise fashion and instead suggest that cells differentiate in a continuous transcriptional and chromatin landscape where cell fates are progressively specified in a probabilistic process (**Figure 1*a***).

Historically, cell fate specification studies mainly focused on embryonic development where fate maps were created using simple imaging-based tools, such as chick-quail transplantation experiments and lipophilic dye–labeling techniques (15–17). With advances in imaging and recombinant DNA technology, fluorescent reporter–based and genetic labeling techniques were developed and became popular alternatives to these classical methods (18). Although these pioneering techniques allowed for the successful reconstruction of cell lineage trees, the genome-wide molecular features of the cells undergoing fate specification could not be profiled simultaneously. Current state of the art allows for the successful integration of genetic labeling techniques with single-cell genomics, enabling us to understand the cellular differentiation landscape at unprecedented resolution (19).

In this review, we provide an overview of recent advances in single-cell biology that empower us to understand the mechanisms of cell fate priming and specification during development and tissue turnover in health and disease. Our focus is on single-cell transcriptomics as a method to interrogate cell fate decisions in a variety of biological systems. We discuss various computational frameworks developed to explore and visualize differentiation trajectories of cells profiled with single-cell RNA sequencing (scRNA-seq). We also provide an overview of state-of-the-art single-cell lineage tracing to reveal cellular ancestry. Furthermore, we discuss the prospects of further breakthroughs in single-cell biology that will help produce a comprehensive multilayered molecular map of cell fate choices at a high-throughput scale with unparalleled resolution.

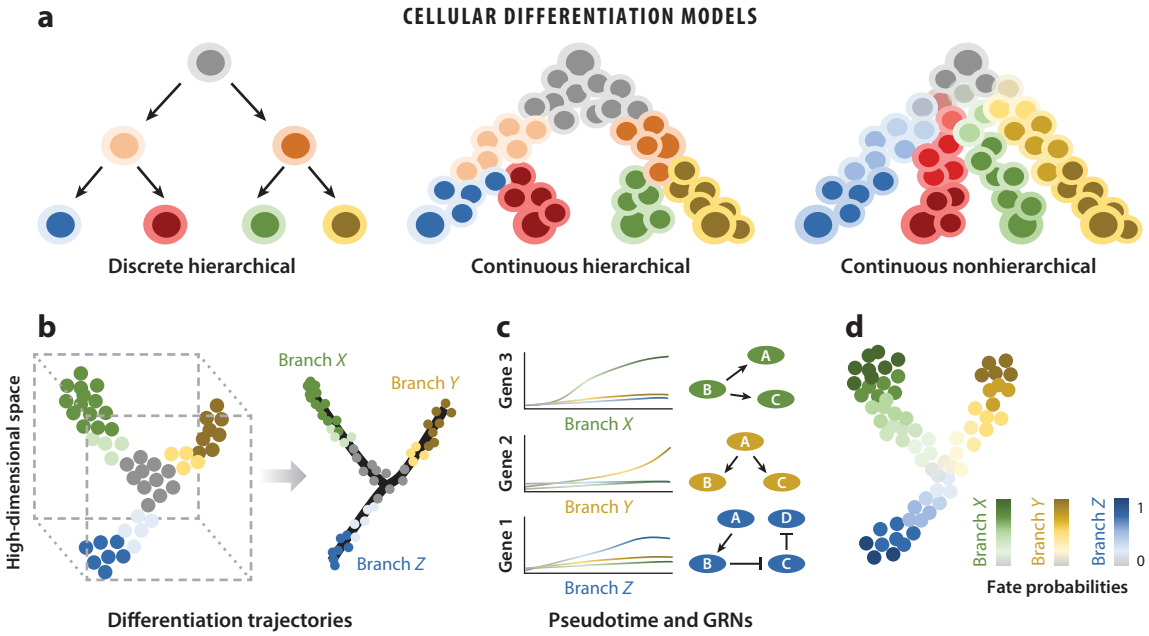

Figure 1

Understanding cell fate choices using scRNA-seq. (*a*) The prevalent models of cellular differentiation. Traditionally, cellular differentiation is considered as a discrete hierarchical process where progenitors become lineage restricted in a series of stepwise bifurcation events. Advances in single-cell technologies are challenging this classical view, suggesting that differentiation is rather a continuous process where progenitors progressively become fate restricted. These continuous processes can be strictly hierarchical or nonhierarchical (i.e., the progenitor compartment is heterogeneous and consists of lineage-primed subpopulations giving rise to differentiated cell types), or a combination of both. (*b*) To elucidate the mechanisms of cell fate decisions, one can reconstruct differentiation trajectories from snapshot single-cell transcriptome data to characterize the differentiation of progenitors (*gray*) into terminal cell states (branches *X*, *Y*, and *Z*). (*c*) Mechanistic insights can be gained by performing pseudotemporal ordering and identifying the accompanying gene expression changes. Gene regulatory networks (GRNs) can also be reconstructed from these data to characterize the interactions among the different regulators. (*d*) Single-cell transcriptomic and epigenomic studies support a probabilistic view of differentiation. Consequently, cell fate commitment can be modeled as a probabilistic process where fate probabilities of progenitor cells differentiating into different terminal states can be predicted, providing insights into lineage commitment.

REVEALING THE MECHANISMS OF CELL FATE SPECIFICATION USING SINGLE-CELL TRANSCRIPTOMICS

Since cell fate decisions occur at a single-cell level, it is critical to study differentiating systems using technologies and methods that can provide this resolution to avoid the misinterpretation associated with bulk analysis yielding average readouts. Indeed, in vitro culture, in vivo transplantation, and lineage-tracing studies at single-cell resolution have brought about many key insights into cell fate specification in various biological systems (20, 21). However, a major limitation of such pioneering studies was their inability to profile genome-wide molecular features of the individual cells under investigation.

Molecular profiling of mRNA species, chromatin accessibility, and epigenetic modifications in single cells at a genome-wide level holds the key to understanding the molecular mechanisms governing how cell fate choices are executed within a single cell. During the last decade, significant advances in single-cell transcriptomics have enabled individual cells to be profiled at a high-throughput scale (22–24). These technological advances parallel the emergence of novel

computational tools utilizing machine learning and statistical modeling approaches to analyze these datasets from different perspectives. Initially, application of various clustering approaches to scRNA-seq data revealed cell type heterogeneity within conventionally defined populations of cells considered homogeneous, e.g., those purified based on the expression of cell surface markers by fluorescence-activated cell sorting (FACS). During the last few years many computational strategies have emerged for deriving differentiation trajectories and modeling cellular fate decisions, revealing the topology of lineage trees and the continuous changes of gene expression along the branches of these trees (19, 25).

In the following, we discuss state-of-the-art computational methods to study cell fate choices and their underlying concepts. We also discuss several visualization methods for reducing the high-dimensional gene expression space of cells to two or three dimensions, enabling us to view and interpret the manifolds populated by observed cell states. We discuss examples of biological systems where these methods have been applied to study cell fate decisions during differentiation. Lastly, we describe the limitations of scRNA-seq alone as a technique to study differentiation processes, and we give our perspective on further advances that are required to understand cell fate choices more comprehensively.

Reconstructing Differentiation Trajectories to Characterize Cell Fate Specification

The reconstruction of differentiation trajectories from scRNA-seq data relies on the assumption that single-cell transcriptomes encompass all naïve, intermediate, and mature cell states with sufficient sampling coverage. As homeostatic adult tissues frequently undergo permanent turnover with a lack of synchrony, it is feasible to sample many (if not all) differentiation stages at any time point from the tissue of interest for scRNA-seq library preparation. However, in certain cases such as early embryonic development, where lineage decision happens in a relatively short period of time, it is essential to sample cells across several stages of development in order to explore the control of lineage specification extending across different developmental time points (8–10).

Once a suitable scRNA-seq dataset is obtained encompassing a single snapshot from the tissue of interest or several time points from a developing tissue, various computational methods can be utilized to reconstruct the differentiation trajectories. In case of a common progenitor differentiating into multiple lineages, reconstructing developmental trajectories implies the identification of branching points characterized by the gradual emergence of transcriptionally distinct cellular states corresponding to alternative fates. Deciphering changes in the transcription profiles of these cells at the verge of lineage commitment may provide novel insights into the molecular mechanisms of cell fate specification. It is a common assumption of lineage reconstruction methods that similarity in gene expression profiles reflects developmental proximity. The general goal of these methods is to identify a low-dimensional continuous manifold capturing all observed cell states. The topology of this manifold is expected to correspond to the lineage tree, and positions along the branches are interpreted as progressive stages of differentiation, recapitulating actual differentiation processes occurring in high-dimensional gene expression space (**Figure 1*b***). Depending on the methods, such manifolds can be linear trajectories, bi- or multifurcations, or even more complex topologies including the presence of circular graphs for modeling periodic processes such as the cell cycle (26). **Table 1** provides an overview of currently available computational methods discussed below for the reconstruction of differentiation trajectories and quantification of cell fate probabilities from scRNA-seq data.

Since scRNA-seq datasets are snapshot measurements at one or multiple time points, it is important to note that the ordering of the cells inferred along the differentiation trajectories does

Table 1 Overview of methods available for differentiation trajectory reconstruction and cell fate probability quantification from scRNA-seq data

Method name	Implementation	Availability	Reference(s)
Tree-based methods			
Monocle2 and Monocle3	R	Bioconductor and GitHub	11, 28
SCUBA	Matlab	GitHub	31
SLICE	R	Webpage	32
TSCAN	R	Bioconductor and GitHub	33
Waterfall	R	Webpage	34
Graph-based methods			
Wanderlust	Python	Not available anymore	35
Wishbone	Matlab and Python	GitHub	36
PAGA	Python	GitHub	37
SLICER	R	CRAN	39
p-Creode	Python	GitHub	40
STITCH	Matlab	GitHub	9
Cluster partition–based methods			
StemID	R	CRAN	41
Slingshot	R	Bioconductor and GitHub	43
Methods utilizing transcriptional dynamics			
RNA velocity	Python and R	Webpage	47
Pseudodynamics	Matlab	GitHub	50
GRAND-SLAM	Java	Available upon request	51
sci-fate	Python and R	GitHub	52
PBA	Python	GitHub	60
Methods quantifying cell fate bias			
GPfates	Python	GitHub	57
STEMNET	R	Webpage	53
FateID	R	CRAN	42
Palantir	Python	GitHub	45
Others			
DPT	Python and R	Bioconductor and GitHub	44
Waddington-OT	Python	GitHub	49

Abbreviations: DPT, diffusion pseudotime; PAGA, partition-based graph abstraction; PBA, population balance analysis.

not reflect the actual dynamics of any individual cell in real time of the differentiation process and is thus rather suitably termed "pseudotime." The first method to introduce the concept of pseudotime was Monocle, which allows linear differentiation trajectories to be reconstructed (27). The method involves an initial dimensionality reduction step followed by the construction of a minimum spanning tree (MST). A trajectory is then created by identifying the longest path through the MST. The subsequently developed Monocle2 algorithm utilizes a reverse graph embedding technique to construct lineage trees with multiple branching points without requiring the prior knowledge of the number of branches (28). Reverse graph embedding is a graph-based representation that learns a set of low-dimensional latent variables and functions on a weighted undirected graph such that these latent variables faithfully represent the high-dimensional data points. Using this approach, given a high-dimensional gene expression dataset, Monocle2 learns

the corresponding latent variables for each cell in a low-dimensional space, and a graph connecting these variables represents the inferred differentiation trajectory. Trajectory analysis of mouse kidney collecting duct cells using Monocle2 revealed a plastic novel cell type that transitions from intercalated cells to principal cells through the activation of Notch signaling (29). Monocle2 was also successfully applied to derive the developmental trajectories of the human prefrontal cortex (30). Other methods that involve dimensionality reduction and subsequent MST construction or curve fitting include SCUBA (31), SLICE (32), TSCAN (33), and Waterfall (34).

Another class of computational methods utilizes *k*-nearest neighbor (kNN) graph–based approaches to derive low-dimensional manifolds capturing cell state transitions during differentiation. The earliest methods in this category were Wanderlust (35) and Wishbone (36). A new method of this kind that accounts for continuous and discrete structure in the dataset is partition-based graph abstraction (PAGA) (37). PAGA derives a kNN graph in a low-dimensional representation obtained using, e.g., principal component analysis (PCA) and a given metric for measuring neighborhood relations such as Euclidean distance. The core step of PAGA is the partitioning of the graph at a desirable resolution to obtain cell clusters or partitions using the Louvain density clustering algorithm. Subsequently, a so-called PAGA graph is created with nodes representing the identified cell clusters and weighted edges reflecting the PAGA connectivity between these cell clusters. PAGA connectivity is a test statistic ranging between 0 and 1 that quantifies the degree of connectivity between cell clusters and is defined as the ratio of the observed inter-edge number and the inter-edge number expected under random assignment. Importantly, PAGA graphs can be created at multiple resolutions. Cell ordering along the high-confidence paths is performed according to the distance of a cell from its root (a progenitor) based on random walks on the single-cell graph. PAGA has been shown to work on complex datasets and has successfully inferred the complex lineage tree of the freshwater planarian flatworm *Schmidtea mediterranea*, encompassing all somatic lineages arising from an adult pluripotent stem cell (38). The pseudotemporal ordering of the cells from all lineages identified 48 gene sets activated during the differentiation of various cell types in *S. mediterranea*. Such lineage trees are instrumental to understanding the lineage specification of specialized cell types and cell fate decisions of progenitor cells during differentiation. A similar strategy to PAGA has been implemented in the latest scalable version of Monocle, Monocle3, which has been used to derive differentiation trajectories during mouse embryonic development (11). Other kNN graph–based methods include SLICER (39) and p-Creode (40).

A third class of methods, including StemID (41, 42) and Slingshot (43), infers lineage trees utilizing a given clustering partition as input. StemID derives the topology of the lineage tree by identifying links between clusters representing cell states in the dataset based on transcriptome similarities between clusters. Each cell is then assigned to one of these inter-cluster links to populate the tree. Finally, the significantly populated inter-cluster links are retained, resulting in a differentiation tree (41). Another method is diffusion pseudotime (DPT), which applies the concept of diffusion maps (discussed below) to reconstruct differentiation trajectories from single-cell data (44). DPT involves convolving Gaussians that are centered at proximal cells to construct a nearest neighbor graph. Subsequently, the transition probabilities between cells are derived from random walks on these graphs. DPT between two cells is defined as the Euclidean distance between the vectors containing the transition probabilities between a cell and all other cells. DPT is robust to noise and scalable to larger datasets but not suitable to study complex differentiation trajectories such as those with several branching points unless multiple diffusion components are considered (45).

Importantly, most of the methods described above, except for SLICE, SLICER, and StemID, do not predict the starting and the end points of the inferred trajectories. In order to infer the directionality of a lineage tree, one must characterize the developmental potential of the cells,

thereby assigning them either to the root (a putative stem cell) or to the tip (a differentiated cell) of the trajectory. Usually, such information is obtained by identifying the cells expressing the marker genes attributed to the respective cell types based on prior knowledge. However, for novel, understudied systems, such knowledge may not be available and needs to be inferred de novo. StemID addresses this problem by utilizing the concept of entropy to identify the putative stem cells in the dataset (41). To determine the root of the tree, i.e., the putative stem cell cluster, the method calculates the transcriptome entropy of each cluster and assumes that immature cells display higher uniformity in their transcriptome and thus have higher entropy, with a decrease in entropy accompanying differentiation toward more specialized mature cell types. In this way, the directionality on the branches of the tree is inferred. StemID correctly identified stem cells in mouse intestinal and bone marrow single-cell datasets and predicted a putative multipotent ductal stem cell population in mouse pancreas (41). Other methods using the concept of entropy to infer the directionality in a lineage tree have been developed, including SLICE (32), SLICER (39), and SCENT (46).

A recently introduced approach known as RNA velocity infers differentiation trajectories based on the kinetics of the mRNA lifecycle (47). Focusing on RNA splicing, the method profiles the velocity of the mRNAs of a cell, i.e., the time derivative of spliced mRNA molecules determined by the balance between the production of spliced mRNA from its unspliced counterpart and its degradation. Therefore, in steady state, when a cell is not undergoing transcriptional change coupled to differentiation, these derivatives are zero. However, cells undergoing differentiation yield a nonzero RNA velocity vector that can be used to predict the future state of the cell. Hence, these velocity vectors translate into a vector field indicating the directionality of differentiation for each cell in cell state space. Finally, cell fates are predicted by modeling cellular trajectories as a Markov process with transition probabilities determined by the local velocity field. RNA velocity successfully recapitulated the direction of differentiation of chromaffin cells from Schwann cell precursors in the mouse brain and confirmed the PAGA-predicted lineage relationships in *S. mediterranea* (38). A recent study characterizing cell fate decisions and lineage relationships in murine neural crest cells successfully used RNA velocity to identify the directions of cell state progression during neural crest migration and differentiation (48). A unique aspect of RNA velocity is the prediction of a future cell state based on information solely obtained from an individual cell, i.e., the ratio of spliced versus unspliced reads, enabling a more reliable prediction of cell state dynamics. For instance, modulations of RNA velocity along a trajectory could help discriminate between a continuous differentiation process and a stepwise process connecting longer-lived metastable states by fast transitions.

Notably, the earliest scRNA-seq studies mainly dealt with asynchronously differentiating systems where datasets sampled at once already contained naïve, intermediate, and mature cell types. Consequently, the computational tools described above were developed for analyzing snapshot data captured at a single time point. However, to discern the mechanisms of lineage specification in nonhomeostatic systems, e.g., during early embryonic development or reprogramming, researchers need to sample cells at several subsequent time points. Leveraging the power of these temporally resolved datasets, various computational tools are now available to infer differentiation and reprogramming trajectories.

The first algorithm of this kind, STITCH (9), utilizes a kNN graph–based strategy to account for the increasing transcriptional complexities in the developing zebrafish embryo during the first 24 hours where several lineage decisions are made. The scRNA-seq data consisted of seven developmental time points and continuous developmental trajectories were reconstructed using STITCH. Instead of projecting all cells exhibiting complex gene expression patterns onto a single low-dimensional manifold, STITCH constructs kNN graphs separately at each time point

in a locally defined low-dimensional subspace obtained using, e.g., PCA. These graphs are then stitched together in a stepwise manner to generate a complete single-cell graph manifold visualized using a force-directed layout (described below). Furthermore, a coarse-grained graph can be constructed to abstract the main features of the single-cell graph (9).

A method called Waddington-OT was recently developed as a means of learning the relationship between cells during reprogramming (49). The method utilizes the information present in the temporally resolved scRNA-seq datasets to model cells as time-varying probability distributions and infers how these probability distributions change over time using optimal transport theory. Specifically, differentiation or reprogramming of a set of cells between two time points on short timescales in high-dimensional gene expression space is defined as the change in mass distribution of all the cells at the destination point given by transporting these cells according to a temporal coupling calculated using optimal transport. Temporal couplings over the longer timescale are inferred by composing the transport maps between every pair of consecutive intermediate time points. Waddington-OT accommodates growth and death rates of cells while computing transport maps. These rates are calculated based on the gene expression signature of cells related to proliferation and cell death. The model was applied to a time course dataset of induced pluripotent stem cell reprogramming, identified previously uncharacterized developmental programs, and validated the role of two candidates in enhancing the reprogramming efficiency (49).

Utilizing the possibility of gaining information on the population dynamics from time course scRNA-seq data to accommodate for the changes in cell type frequencies during developmental trajectory reconstruction, the Theis laboratory has recently developed a framework called pseudodynamics (50). Pseudodynamics models the rate of change of the population distribution across continuous cellular states in a low-dimensional space obtained from, e.g., diffusion maps as a reaction-diffusion-drift model. The reaction parameter in this partial differential equation describes cell proliferation and death, while drift and diffusion parameters encode directed and stochastic movements of cells along the differentiation trajectory, respectively. These parameters are defined as continuous functions of cell states and time and are estimated from input scRNA-seq data at different time points. The model requires the total cell numbers at each time point as an input to estimate the birth-death parameter and imputes the probability distributions of cell states at unsampled time points. Applied to early T cell development, it successfully mapped beta-selection and characterized birth-death rates along the T cell maturation trajectory. For mouse pancreatic beta cell maturation, pseudodynamics revealed that beta cell proliferation during the early stages of life is determined by the molecular cell state and not affected by extracellular regulators (50).

As mentioned earlier, current experimental protocols to perform scRNA-seq capture a snapshot of the transcriptome of each individual cell (even if conducted as a time course experiment). Consequently, the computational tools to reconstruct differentiation trajectories do not account for the underlying dynamics of transcription. RNA velocity is a notable exception, as the ratio of spliced and unspliced mRNA is used to predict the future state of the cell. Nevertheless, instead of inferring such vectors with the help of computational tools, it is desirable for the experimental protocols to capture the temporal dynamics of transcription and robustly characterize the cell fate decisions during differentiation. A proof-of-principle study recently developed single-cell, thiol-(SH)-linked alkylation of RNA for metabolic labeling sequencing (scSLAM-seq) to profile transcriptional dynamics and stochastic gene expression by labeling newly synthesized mRNA with 4-thiouridine (4sU) (51). Cells were incubated with 4sU for two hours to incorporate it into nascent RNA, and then an alkylation reaction with iodoacetamide was performed to convert 4sU into a cytosine analog. After mRNA of these cells is profiled using scRNA-seq, newly synthesized mRNA can be distinguished by U-to-C conversions from the sequencing reads. Furthermore,

a Bayesian method called GRAND-SLAM was developed to estimate the expression of old and new RNA and the ratio of new to total RNA (NTR). scSLAM-seq was used to study the transcriptional dynamics of cytomegalovirus infection in mouse fibroblasts. Importantly, infected cells did not exhibit any difference from the uninfected cells in terms of old and total RNA. However, NTR clearly separated infected and uninfected cells. scSLAM-seq enabled a detailed characterization of the gene expression changes related to newly transcribed RNA upon infection, ON-OFF switches, transcriptional bursting dynamics, and promoter structure analysis (51).

Another recent high-throughput method akin to scSLAM-seq combining 4sU labeling and single-cell combinatorial indexing RNA sequencing (sci-RNA-seq) is sci-fate (52). sci-fate was used to characterize the transcriptional dynamics of cortisol response in A549 cells in vitro by dexamethasone treatment for different periods of time. A computational framework was developed to reconstruct single-cell transition trajectories, where the past transcriptional state of each cell was estimated to link the cells at different time points. In the future, application of these experimental techniques in vivo and the development of dedicated computational tools will be essential to elucidate transcriptional dynamics underlying cell fate decisions.

In summary, the computational methods described in this section can be used to reconstruct lineage trees and thereby identify the manifolds of cell fate specification in cell state space. By characterizing the cellular states in regions within these manifolds where lineages diverge and the accompanying gene expression changes, one can gain novel insights into the process of lineage specification, e.g., by identifying novel genes, regulatory networks, and signaling pathways activated in the emerging cell lineages (**Figure 1*c***). Furthermore, new experimental techniques and computational methods now make it possible to go beyond snapshot profiling and pseudotemporal ordering and to decipher the dynamics of transcriptional programs activated during differentiation, enabling improved reconstruction of differentiation trajectories.

Modeling Differentiation as a Probabilistic Process to Quantify Cell Fate Bias

Most of the differentiation trajectory reconstruction algorithms described above assign individual cells to fixed states or fixed positions on the derived cell state manifolds without accounting for the probabilistic nature of cell fate decision, i.e., the possibility of switching fate to another lineage with a certain probability. However, single-cell studies in various biological systems suggest that differentiation proceed not in discrete stages but rather as a continuous process (23, 53–55). This indicates that transcriptional bias toward a mature cell fate emerges gradually and implies that the probability to commit to a particular fate could be modulated in an equally continuous fashion. Importantly, such conclusions can be drawn not only from the transcriptional state of the sampled cells profiled by scRNA-seq. Single-cell assay for transposase-accessible chromatin and sequencing (scATAC-seq) studies profiling the genome-wide chromatin accessibility landscape also suggest the continuous nature of the epigenomic changes during differentiation (56).

Hence, to explore the possibility of cell fate priming in progenitor and stem cells, one should consider modeling differentiation processes as probabilistic events where each progenitor cell can be assigned a probability of differentiation into, or cell fate bias toward, each of the mature cell lineages in the system (**Figure 1*d***). Several algorithms have been introduced that consider differentiation as a probabilistic process and quantify such fate probabilities in single progenitor cells.

The first two methods that attempted to quantify cell fate bias were GPfates (57) and STEMNET (53). GPfates uses the Bayesian Gaussian process latent variable model to perform dimensionality reduction and pseudotime inference of scRNA-seq data. Subsequently, a bifurcation

of cell fates is considered as a mixture problem to be solved by fitting the overlapping mixtures of Gaussian processes (OMGP) model. The Gaussian process–assigned probabilities for the different trajectories of the OMGP model are used to quantify cell fate bias. This method was used to model the Th1 and Tfh bifurcation of $CD4^+$ T cells during *Plasmodium* infection in mice and to characterize the transcriptional changes associated with this bifurcation (57).

The development of STEMNET was inspired by the idea that fate specification of multipotent hematopoietic progenitors could be a continuous process (53). In order to quantify the extent of lineage priming in stem cells, the algorithm relies on the prior identification of mature cell states based on, e.g., known marker genes in the dataset, which are then used to fit an elastic net–regularized generalized linear model for regressing the transcriptome of the multipotent progenitors on the mature states. The regression coefficients are then used to estimate the probability of each cell resulting in a particular mature cell state, i.e., to quantify cell fate bias of immature cells. STEMNET was used to predict the transcriptional lineage priming among human hematopoietic stem cells (HSCs) toward six lineages (53), suggesting that these lineages may emerge directly from lowly primed progenitors without passing through a series of discrete metastable progenitors in a stepwise fashion.

The FateID algorithm (42) utilizes a random forests–based approach to predict cell fate probabilities toward a priori defined mature cell types. It requires the input of a clustering partition defining these mature states or, alternatively, a set of marker genes specific to these states. These so-called target cell states are used to train an iterative random forests classifier to infer the fate probabilities of the remaining cells included in the analysis based on their transcriptome. Unlike STEMNET, which only regresses on the mature cell types to infer cell fate probabilities of the progenitors, FateID classifies with a dynamic training set by iteratively moving backward along the differentiation trajectory. At each iteration the training set is updated by replacing more mature cells with more immature cells closer to the cells to be classified in the current iteration. This strategy ensures that the classification is not performed solely based on the transcriptional state of the most mature cell types. FateID was able to resolve domains of predominant cell fate bias toward distinct lineages in mouse hematopoietic progenitors and identified a common lymphoid progenitor for B cells and plasmacytoid dendritic cells, which was experimentally validated (42) and confirmed by subsequent in-depth studies (58, 59).

A very different approach was implemented by the Palantir method, which models differentiation as a stochastic process using Markov chains (45). The method first constructs a nearest neighbor graph using diffusion maps (discussed below) to identify an optimal low-dimensional manifold recapitulating the differentiation landscape as a basis for fitting an absorbing Markov chain. This enables the calculation of differentiation probabilities toward the terminal states. Palantir was applied on human bone marrow data, where it recapitulates the known trends in human hematopoiesis and supports a continuous and hierarchical differentiation model.

To address the limitations of many existing lineage tree reconstruction algorithms in inferring the underlying dynamics of differentiation, the Klein group has developed an approach termed population balance analysis (PBA) that attempts to reconstruct trajectories from single-cell snapshot data by formulating a population balance equation (60). Under the assumption that cell trajectories are Markovian and that there is no oscillating gene expression, a diffusion–drift equation modeling population balance is solved by utilizing special graph theory under steady state conditions. The output of PBA is a Markov chain that allows for the derivation of cell fate probabilities and pseudotemporal ordering of progenitors along differentiation trajectories. PBA was applied to mouse bone marrow single-cell data (13) and supports a continuous and hierarchical model for murine hematopoiesis.

Using Dimensionality Reduction to Visualize and Interpret Cell Fate Specification

With the advances in throughput and sensitivity of scRNA-seq protocols, it is now feasible to measure thousands of genes across tens to hundreds of thousands of cells (7, 11, 61) generating large datasets with sufficient sampling of complex lineage trees in high-dimensional cell state space. A dimensionality reduction to two or three dimensions is a useful approach to visualizing and interpreting such datasets that can preserve key topological features. The core idea is that observable cell states only populate a manifold of much lower dimensionality than the full gene expression space, in which each dimension corresponds to a particular gene, due to the fact that genes are not regulated independently of each other. An ideal dimensionality reduction method should not only preserve the local and global structure of the data to permit meaningful interpretation but also be scalable to a large number of cells. In the context of visualization and interpretation of cellular differentiation trajectories, it is important that the selected method for dimensionality reduction preserves the continuity of the populated manifold in gene expression space.

Current dimensionality reduction methods can be broadly categorized into two types: matrix factorization–based and neighbor graph–based methods (**Figure 2*a***). The most common dimensionality reduction methods used in single-cell biology are PCA and *t*-distributed stochastic neighbor embedding (t-SNE). PCA-based dimensionality reduction utilizes matrix factorization to project the data into a space defined by the two or three major principal components, capturing the major directions of variability in the data. Since the principal components correspond to eigenvectors obtained from a linear transformation of the covariance matrix, complex nonlinear trajectories cannot be mapped to a low-dimensional space using PCA (**Figure 2*b***). In contrast, nonlinear dimensionality reduction methods like t-SNE are more suitable for preserving local structure of the data beyond the major axis of variability (62). In contrast to PCA, t-SNE relies on neighborhood graphs to infer a dimensionally reduced representation of the data. Indeed, t-SNE is currently a method of choice for dimensionality reduction and visualization of scRNA-seq data and preserves salient features related to local data structure. This method is based on a nonlinear transformation of local normal density distributions in original space into local Student *t*-distributions, measuring the density in the low-dimensional space, which minimizes the relative entropy, termed Kullback–Leibler divergence, of these distributions. The output and performance of t-SNE heavily depends on the perplexity parameter determining the degree of locality of the transformed distribution. This parameter controls the number of neighbors captured by the local density distributions in the original space and implies a local adjustment of the bandwidth. Values of the perplexity parameter that are too large or too small may lead to an artifactual local structure or to loss of structure, respectively. In particular, the t-SNE approach by definition does not preserve global relations between distant cells and is thus insensitive to the global topology (**Figure 2*b***). Therefore, relative distances of remotely related cell types might not be informative in low-dimensional space obtained by t-SNE, potentially disrupting undersampled differentiation trajectories.

To circumvent this problem with t-SNE, researchers are increasingly applying another neighborhood graph–based approach, unique manifold approximation (UMAP), to visualize scRNA-seq data (63). The algorithm approximates the high-dimensional data manifold by assuming it to be uniformly distributed and reconstructs the manifold by gluing together simple local topological elements connecting nearest neighbors, termed simplices, into simplicial complexes to obtain a combinatorial representation of the topology of the populated cell state manifold. A local distance metric reflecting the local density is introduced to account for nonuniform coverage. Edges connecting cells within simplices arising from different local distance metrics are combined in

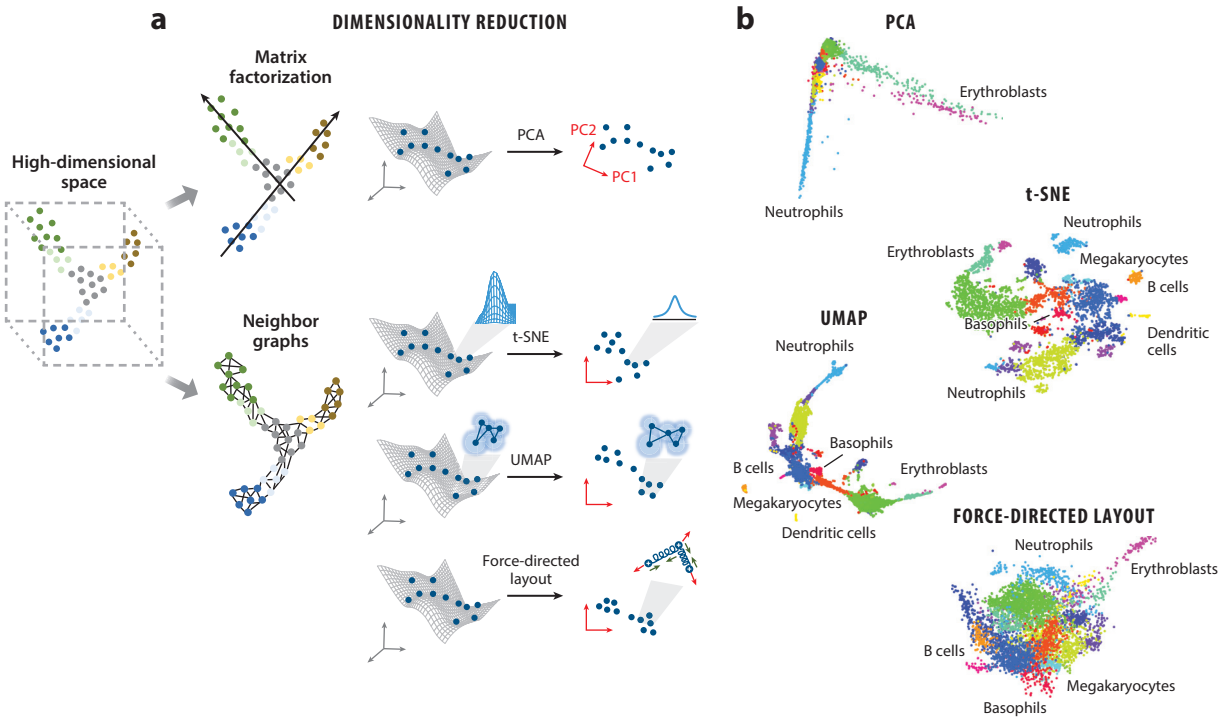

Figure 2

Dimensionality reduction to visualize and interpret scRNA-seq data. (*a*) Since scRNA-seq profiles thousands of genes in single cells, the data are high dimensional and dimensionality reduction is necessary for visualization and meaningful interpretation. Dimensionality reduction methods can be broadly divided into two major categories: matrix factorization and neighbor graphs. A commonly used matrix factorization method for dimensionality reduction of scRNA-seq data is principal component analysis (PCA), a linear transformation identifying the major axes of variability. In contrast, a neighbor graph–based approach is more suitable for preserving the local structure of the data. Such methods include *t*-distributed stochastic neighbor embedding (t-SNE), unique manifold approximation (UMAP), and force-directed layout. t-SNE transforms local Gaussian distributions measuring the density of data points in high-dimensional space into local Student's *t*-distributions. The optimization of the low-dimensional space is performed by minimizing the Kullback–Leibler divergence between these distributions. UMAP constructs a topological representation of the high-dimensional space by patching together local topological elements called simplices into simplicial complexes. A similar process is used to construct an equivalent low-dimensional topological representation of the data. The cross-entropy between the two representations is minimized to optimize the layout in low-dimensional space. Force-directed layouts visualize *k*-nearest neighbor graphs by assigning attractive forces (depicted as springs) to the edges and repulsive forces (e.g., positive charges) to the nodes. The net forces are minimized until equilibrium is achieved. (*b*) Application of different dimensionality reduction techniques on the scRNA-seq data of hematopoietic progenitors from Reference 13. Dimensionality reduction using PCA only resolves neutrophil and erythroblast differentiation trajectories. t-SNE, UMAP, and force-directed Fruchterman–Reingold layout representations allow the visualization of all cell types including the underrepresented cell populations such as megakaryocytes, basophils, dendritic cells, and B cells. However, UMAP provides a smoother manifold and better resolves the global and continuous structure of the differentiation manifold in low-dimensional space.

a probabilistic fashion to make these local metrics compatible, leading to so-called fuzzy simplicial sets for each cell, which are glued together into a fuzzy simplicial union. The resulting fuzzy graph is mapped onto a fuzzy graph in two-dimensional Euclidean space with a nearest neighbor distance given as a hyperparameter. This hyperparameter is inferred by optimizing the cross-entropy between the two graphs, yielding a low-dimensional topology-conserving representation of the manifold. In contrast to t-SNE, the mapping of global graphs ensures that UMAP preserves not only local but also—to some extent—global structure and thus reveals the relationship and

the continuity of the cell clusters (**Figure 2*b***). As another advantage over t-SNE, UMAP allows embedding of new data points into the low-dimensional layout. Finally, UMAP has a strongly reduced runtime compared to t-SNE and scales much better with increasing dataset size (63).

Two other dimensionality reduction methods developed to visualize cell clusters or differentiation trajectories are diffusion maps (64) and scvis (65). DPT, the method described above to reconstruct lineage trees, is an extension of the application of diffusion maps. A diffusion map is a nonlinear dimensionality reduction method where each cell is represented by a Gaussian distribution in high-dimensional space. The overlap of these Gaussians gives rise to diffusion paths along the nonlinear data manifold. Formally, a Markovian transition probability matrix is calculated whose first n eigenvectors, termed diffusion components, are used for the visualization of the data in n-dimensional space with $n = 2$ or 3. scvis is a latent variable model, and unlike t-SNE, it is a parametric dimensionality reduction method that is based on the assumption that the gene expression vectors of cells are governed by low-dimensional latent vectors captured by normal distributions parameterized by mean and standard deviation as functions of the position of a cell in gene expression space that are determined by a model neural network. A variational inference feedforward neural network infers the parameterization of the distribution of the latent vectors in order to obtain a low-dimensional representation defined by these vectors. scvis is a probabilistic generative model that also provides a log-likelihood to measure the quality of the embedding. This method has been shown to preserve both local and global structure of the data (65).

Another set of visualization techniques suitable for preserving the global architecture of the differentiation manifold are force-directed layout algorithms. They solve the problem of visualizing kNN graphs in the most meaningful and aesthetical way by assigning forces to the edges and nodes of the graphs and simulating them as a physical system. Typically, attractive forces are assigned to the edges, repulsive forces are assigned to the nodes, and the net forces on the graph are minimized until an equilibrium is achieved (**Figure 2*a***). This can be done in low dimensions in order to obtain a dimensionally reduced representation of the data. There are various layout models available such as spring–electric and Fruchterman–Reingold (FR) layouts (66). In the spring–electric layout model, nodes and edges are considered as charged particles and springs, respectively, and all edges have a uniform length. FR layouts replace nodes and edges by steel rings and springs, respectively, and the repulsive force inversely scales with the distance between nodes. SPRING is one of the first implementations of such a force-directed layout for scRNA-seq data, and it is publicly available as a web tool (67). Lastly, embeddings of PAGA, a kNN graph–based manifold learning approach described in the previous section, can also be used with UMAP and force-directed layout algorithms to visualize scRNA-seq data (37). Advantages include faster convergence and preservation of global and local data manifold structure.

Limitations of Single-Cell Transcriptomics to Study Cell Fate Decisions

There are experimental and computational limitations to using scRNA-seq as a technique to study cell fate decisions. Since single-cell experiments destroy the cells sampled from a population, the static snapshot data may not contain all cell states occurring during differentiation, e.g., fast transitioning cell states. This problem may be circumvented by sampling large numbers of cells using high-throughput scRNA-seq protocols such as sci-RNA-seq (68) or SPLiT-seq (69) that use combinatorial barcoding, but these approaches currently suffer from relatively low sensitivity. Moreover, it becomes cost prohibitive to maintain a high sequencing depth while massively increasing cell numbers to ensure the capture of lowly expressed genes such as transcription factors, which may play important roles during cell fate commitment. An alternative solution is the enrichment

of such intermediate states using FACS, provided that the cell surface markers characterizing these cellular states are previously known. However, this results in skewed frequencies of cellular states in the scRNA-seq dataset, leading to the distortion of the original high-dimensional differentiation landscape and the inferred low-dimensional layout. Furthermore, as discussed above, since current scRNA-seq protocols generate only snapshot data, they do not reveal the actual dynamics of differentiation processes. There are multiple conceivable dynamic processes that can in theory give rise to the same high-dimensional differentiation manifold captured by the snapshot scRNA-seq data. Therefore, the question of how and when exactly cell fate decisions happen in the biological system is difficult to answer.

Computationally, all lineage reconstruction algorithms are based on the assumption that cells with similar transcriptional profiles are developmentally closer to each other, as the actual lineage relationships among the profiled cells are not known. Thus, more complex dynamics such as molecular oscillations and asymmetric cell divisions during differentiation are difficult to decipher. Of note, molecular oscillations and asymmetric cell divisions have been shown to play important roles in many developmental processes, homeostasis, and cancer (70–72). Furthermore, there are other confounding factors that can prevent the successful reconstruction of lineage trees, such as cell cycle states and technical batch effects in the data. For example, irrespective of the stage in the differentiation process, similar expression of cell cycle–associated genes may reduce the distance of otherwise unrelated cell states in gene expression space, thereby distorting the inferred differentiation trajectories. Moreover, technical batch effects due to, e.g., integrating libraries generated with different protocols or in separate experiments may complicate the problem of lineage inference. However, there are several batch correction methods that can be applied to remove these technical artifacts, such as matching the mutual nearest neighbors between batches (73) or identifying the common correlation structure across batches (74–76). Yet the application of such batch correction methods could lead to unwanted removal of actual biological variability.

Since scRNA-seq profiles mRNA at single-cell resolution to study lineage specification, which sheds lights on the transcriptional basis of cell fate decisions, it suffers from the limitation that it provides a readout on only one of the many molecular layers involved in this multimodal process. Therefore, profiling of other molecular features is crucial to comprehensively elucidate the mechanisms of lineage choices and fully understand the process of cell fate decision. For example, readouts on chromatin accessibility using scATAC-seq can reveal the involvement of epigenetic regulation and binding of transcription factors in cell fate decisions. Other features such as the proteome, histone modifications, DNA methylation, and posttranslational modifications are important in controlling the commitment of a cell toward a particular lineage. There has been a significant improvement in methods profiling some of these features (6, 77–85), but they still suffer from limitations in applicability and power compared to scRNA-seq. Furthermore, since scRNA-seq requires dissociation of tissues and organs into a single-cell suspension, the role of the microenvironment and the spatial context of the profiled cells in which cell fate decisions are executed cannot be characterized.

Current efforts are focused on the development of experimental protocols enabling the simultaneous profiling of genomic, transcriptomic, and epigenomic features of cells (86–89), as well as retaining their spatial context (**Figure 3*b***) (90–93). Computationally, tools to integrate single-cell datasets measuring different modalities are also helping to elucidate the mechanisms of cell fate decisions more comprehensively (75, 94). Moreover, various genetic lineage tracing techniques have been developed to causally infer the lineage relationship between cells at single-cell resolution, and experimental approaches combining them with single-cell molecular profiling techniques such as scRNA-seq are emerging. In the next section, we discuss the advances of these techniques and their role in understanding the mechanisms of cell fate decisions.

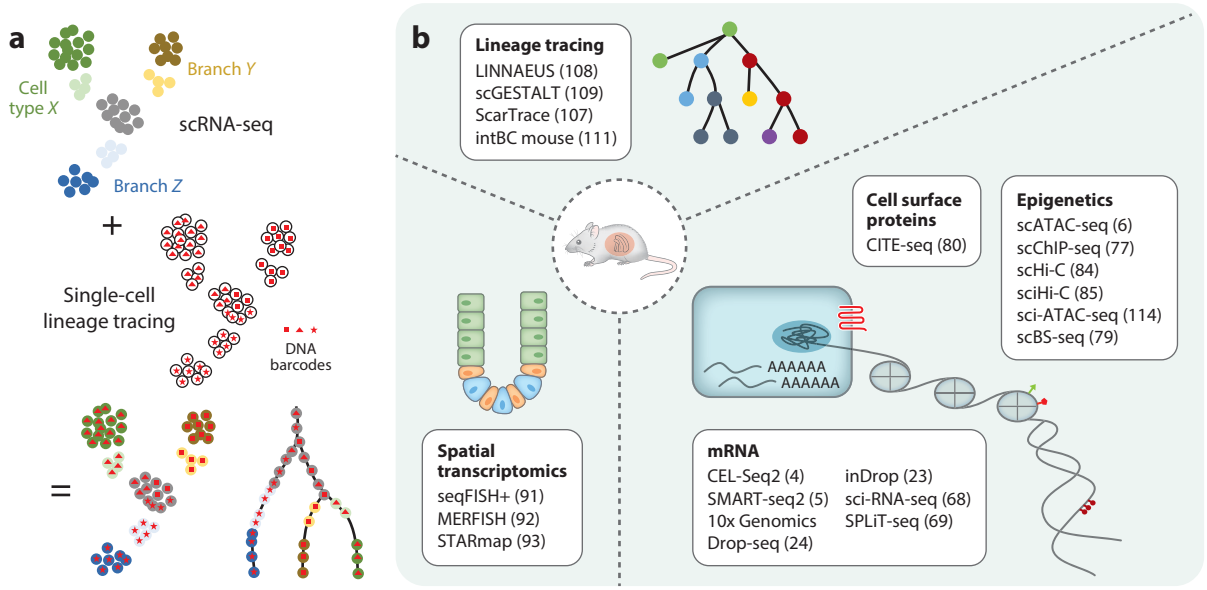

Figure 3

Multimodal single-cell analysis to comprehensively understand the mechanisms of cell fate decisions. (*a*) Lineage tree inferences from scRNA-seq data are based on the assumption that cells with similar transcriptomes are developmentally closer to each other, but the real progenitor–progeny relationship between the cells cannot be known. Advances in single-cell lineage tracing allow for sequencing of the genetic labels and profiling of the whole transcriptome simultaneously, making it feasible to reconstruct refined cell lineage trees with the information of genome-wide transcriptomes. (*b*) The currently available methods to study cell fate choices and lineage commitment can be broadly divided into three different categories: single-cell genetic lineage tracing to reconstruct lineage trees, methods to profile mRNA in situ to characterize cell types in their original spatial location and their microenvironment, and methods profiling different molecular features of a cell (e.g., mRNA, histone modifications, DNA methylation, chromatin accessibilities and architecture, cell surface proteins). Future advances will enable simultaneous application of these methods or their computational integration for an unprecedented resolution of cellular differentiation.

INTEGRATING LINEAGE TRACING AND SINGLE-CELL TRANSCRIPTOMICS TO EXPLORE THE UNDERPINNIGS OF CELL FATE DECISIONS

Contemporary lineage tracing techniques allow cells to be marked with fluorescent proteins or nucleic acids, which are heritable and, hence, can be found in the progeny of the marked cells, enabling us to reconstruct lineage trees and understand cell fate decisions (18, 21). Owing to the limited number of fluorophores that can be simultaneously imaged in an imaging experiment, cellular barcoding using DNA-based tags is an attractive alternative as, by controlling the number of nucleotides in such tags, a large number of unique sequences can be generated to exhaustively mark the entirety of cells in a tissue or even a whole organism. Current DNA-based barcoding approaches can be broadly divided into three different categories: viral transduction–based, recombinase-based, and CRISPR-based techniques (19, 95).

Viral barcoding involves the generation of viral vector libraries carrying a large number of unique nucleotide sequences that, upon transduction, integrate into the genomic DNA of the host cells, thereby labeling them uniquely. These integrated unique DNA barcodes are then sequenced from single cells to identify the clones and reconstruct a lineage tree. A viral barcoding approach was first used in the early 1990s to reveal that clonally related neurons are not clustered in the specific functional areas of the cerebral cortex but are widely dispersed across the whole cortex

during development (96). Recently, the use of viral barcoding techniques in the hematopoietic system identified two fate-biased subpopulations of HSCs (97), discovered the existence of fate-restricted lymphoid-primed multipotent progenitors (98), and revealed that the megakaryocyte lineage branches off early and independently of other hematopoietic lineages (99). It has also been used to study the clonal dynamics of $CD8^+$ T cell differentiation during immune response (100, 101). Although many short-length unique barcodes can be generated using the viral barcoding approach, it is not feasible to use this technique to study lineage relationships in many tissues owing to the practical constraints in viral delivery.

Alternatives to viral barcoding include recombinase-based and CRISPR-based barcoding systems where transgenic model organisms can be generated to mark the cells of any tissue of interest or even the entire organism during the desired temporal windows to study cell fate decisions in vivo. Importantly, unlike viral barcoding-based strategies, the DNA tags in these two approaches are not static and evolve over time, resulting in increased barcode diversity. Recombinase-based barcoding relies on the activity of Cre recombinase on an array of nucleotide sequences flanked by several loxP sites. Once Cre is induced, the nucleotide sequences undergo excision or inversion, creating a diverse array of unique DNA barcodes. However, since Cre favors deletion over inversion, the nucleotide array will exhibit a gradual decrease in sequence diversity over time. The alternative solution is to use a recombinase that is unable to perform deletion of the flanked sequences, e.g., *Rci* DNA recombinase (102). Notably, the other limitation of recombinase-based approaches is the requirement of long-read sequences to read the barcodes, e.g., generated using low-throughput single-molecule real-time sequencing, as the array used for marking the cells contains several DNA sequences flanked by loxP sites to generate high sequence diversity. A Cre recombinase–based barcoding system utilizing multiple loxP sites called Polylox has been recently used to characterize HSC fates in vivo (103).

As an alternative to recombinase-based barcoding, CRISPR-Cas9-based barcode generation strategies have been employed in various studies. CRISPR-based systems rely on the activity of the Cas9 nuclease on multiple transgenes or DNA arrays containing CRISPR target sites to induce a double-strand break that is repaired by nonhomologous end joining (NHEJ), leading to random insertions and deletions (scars) at the target sites that are used to lineage-trace the cells undergoing differentiation. Like Cre recombinase, NHEJ favors deletions over insertions, thereby shrinking the diversity of scars over time. Alternative approaches include the use of multiple independently evolving target sites to increase the complexity or self-targeting CRISPR guide RNAs (104, 105). Transgenic mouse and zebrafish lines have been generated using CRISPR-based barcoding approaches to study the mechanisms of lineage decisions at the whole-organism level (106–108).

Although single-cell genetic lineage tracing using the abovementioned techniques can elucidate the clonal relationships between the different cell types, the underlying molecular mechanisms of linage commitment and differentiation cannot be established. Moreover, single-cell transcriptomics enables the identification of cell types but cannot establish the lineage relationships between them. Therefore, combining single-cell genetic lineage tracing with scRNA-seq represents a powerful approach to establish the causal link between clones, as well as to map their molecular identities in order to understand the regulatory features of fate choices during differentiation (**Figure 3*a***). Simultaneous profiling of cellular barcodes and mRNA using scRNA-seq is an active area of research. Such combinatorial approaches enabling the profiling of both readouts have already been applied to in vitro differentiation systems, as well as to developing vertebrate embryos, e.g., zebrafish and mouse (107–111). Future studies combining single-cell lineage tracing with scRNA-seq and, e.g., multiplexed single-molecule RNA fluorescence in situ hybridization (112) in mammalian systems will pave the way to recovering cell type information, developmental

relationships between cell types, and their spatial context, thereby adding an extra layer of information to help decipher the mechanisms of cell fate decisions during differentiation.

DISCUSSION AND OUTLOOK

Traditionally, cells have been considered the fundamental units of life in biology. Therefore, the molecular programs operating in single cells during homeostasis and the changes they undergo during dynamic processes such as development and disease progression are central themes of contemporary biology. Single-cell multi-omics technologies represent extremely promising tools to shed light on these molecular programs at unprecedented resolution. In recent years, major developments in single-cell experimental and computational methods have been mainly restricted to scRNA-seq. The application of scRNA-seq to elucidate the mechanisms of cell fate decisions has undoubtedly provided many novel insights challenging the classical and longstanding models of cellular differentiation, e.g., for hematopoiesis (113). However, we are still far from a comprehensive understanding of cellular differentiation and cell fate decision even in well-characterized systems such as hematopoiesis. In order to achieve the goal of a more holistic view on cell fate specification, researchers need to develop methods to measure other molecular features within single cells. Owing to the rapid ongoing developments in single-cell biology, it is now possible to profile chromatin accessibility, DNA methylation, histone modifications, and cell surface proteins in single cells, as well as spatial arrangements of cells in situ (**Figure 3*b***) (6, 77–80, 90, 91, 114). With the growing availability of experimental techniques to simultaneously profile these features within individual cells and the development of computational tools to enable the integration of datasets representing distinct modalities (75, 86–89, 94, 115), we expect to see the emergence of a more integrative framework to answer longstanding questions of cell fate choices and lineage specification.

Retrospectively, since the publication of the first scRNA-seq method in 2009 (116), single-cell technologies have seen tremendous growth, enabling various molecular features to be profiled in large numbers of cells (**Figure 3*b***). Consequently, leveraging this growth and continuously reducing sequencing costs, global and large-scale collaborative efforts such as Human Cell Atlas (117) and the LifeTime initiative have been initiated to produce comprehensive maps comprising millions of cells sampled from the human body in health and disease using single-cell multi-omics technologies. These atlases will be the basis for a much-improved understanding of human tissues and for the development of better disease treatment options. Naturally, the generation of such a large amount of data will require the development of computational methods scalable to millions of cells in a memory- and time-efficient way. The application of deep learning tools to achieve a scalable solution for the analysis of very large scRNA-seq datasets is an active area of ongoing research (118).

In summary, the answer to how cells undergo fate commitment and lineage specification is instrumental to comprehending how complex multicellular organisms are formed from a totipotent zygote during embryonic development. Additionally, after birth, cell fate decisions are critically important in maintaining homeostasis during organisms' entire lifespans. With the advent of and ongoing accelerated progress in single-cell multi-omics, many new insights in these processes are gained at a large scale with remarkable resolution. Such groundbreaking insights will allow us to dig deeper and acquire an improved understanding of the mechanisms of cellular dysfunction and molecular deregulation occurring in human pathologies such as cancer.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Development of our own methods and ideas was supported by the Max Planck Society, the German Research Foundation (DFG) (grants SPP1937 GR4980/1-1, GR4980/3-1, and GRK2344 MeInBio), the DFG under Germany's Excellence Strategy (Centre for Integrative Biology Signalling Studies grant EXC-2189, project ID 390939984), the European Research Council (Consolidator Grant ImmuNiche, project ID 818846), and the Behrens-Weise Foundation. We would like to apologize to all our colleagues whose work could not be mentioned due to space limitations.

LITERATURE CITED

1. Kopp JL, Grompe M, Sander M. 2016. Stem cells versus plasticity in liver and pancreas regeneration. *Nat. Cell Biol.* 18:238–45
2. Brockes JP, Kumar A. 2005. Appendage regeneration in adult vertebrates and implications for regenerative medicine. *Science* 310:1919–23
3. Raj A, van Oudenaarden A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135:216–26
4. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, et al. 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* 17:77
5. Picelli S, Bjorklund AK, Faridani OR, Sagasser S, Winberg G, Sandberg R. 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* 10:1096–98
6. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523:486–90
7. Pijuan-Sala B, Griffiths JA, Guibentif C, Hiscock TW, Jawaid W, et al. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* 566:490–95
8. Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF. 2018. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* 360:eaar3131
9. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360:981–87
10. Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, et al. 2018. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360:eaar5780
11. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, et al. 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 566:496–502
12. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, et al. 2017. A single-cell survey of the small intestinal epithelium. *Nature* 551:333–39
13. Tusi BK, Wolock SL, Weinreb C, Hwang Y, Hidalgo D, et al. 2018. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* 555:54–60
14. Dahlin JS, Hamey FK, Pijuan-Sala B, Shepherd M, Lau WWY, et al. 2018. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice. *Blood* 131:e1–11
15. Le Douarin NM. 1980. The ontogeny of the neural crest in avian embryo chimaeras. *Nature* 286:663–69
16. Serbedzija GN, Bronner-Fraser M, Fraser SE. 1989. A vital dye analysis of the timing and pathways of avian trunk neural crest cell migration. *Development* 106:809–16
17. Selleck MA, Stern CD. 1991. Fate mapping and cell lineage analysis of Hensen's node in the chick embryo. *Development* 112:615–26
18. Kretzschmar K, Watt FM. 2012. Lineage tracing. *Cell* 148:33–45
19. Kester L, van Oudenaarden A. 2018. Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* 23:166–79
20. Etzrodt M, Endele M, Schroeder T. 2014. Quantitative single-cell approaches to stem cell research. *Cell Stem Cell* 15:546–58
21. McKenna A, Gagnon JA. 2019. Recording development with single cell dynamic lineage tracing. *Development* 146:dev169730

22. Papalexi E, Satija R. 2018. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* 18:35–45
23. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, et al. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* 161:1187–201
24. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161:1202–14
25. Tritschler S, Buttner M, Fischer DS, Lange M, Bergen V, et al. 2019. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development* 146:dev170506
26. Saelens W, Cannoodt R, Todorov H, Saeys Y. 2019. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* 37:547–54
27. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32:381–86
28. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, et al. 2017. Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* 14:979–82
29. Park J, Shrestha R, Qiu C, Kondo A, Huang S, et al. 2018. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* 360:758–63
30. Zhong S, Zhang S, Fan X, Wu Q, Yan L, et al. 2018. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 555:524–28
31. Marco E, Karp RL, Guo G, Robson P, Hart AH, et al. 2014. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *PNAS* 111:E5643–50
32. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. 2017. SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res.* 45:e54
33. Ji Z, Ji H. 2016. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 44:e117
34. Shin J, Berg DA, Zhu Y, Shin JY, Song J, et al. 2015. Single-cell RNA-seq with Waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell* 17:360–72
35. Bendall SC, Davis KL, Amir ED, Tadmor MD, Simonds EF, et al. 2014. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157:714–25
36. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, et al. 2016. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* 34:637–45
37. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, et al. 2019. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* 1535 20:59
38. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, et al. 2018. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360:eaaq1723
39. Welch JD, Hartemink AJ, Prins JF. 2016. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* 17:106
40. Herring CA, Banerjee A, McKinley ET, Simmons AJ, Ping J, et al. 2018. Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell Syst.* 6:37–51.e9
41. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, et al. 2016. De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* 19:266–77
42. Herman JS, Sagar, Grün D. 2018. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat. Methods* 15:379–86
43. Street K, Risso D, Fletcher RB, Das D, Ngai J, et al. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genom.* 19:477
44. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ. 2016. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13:845–48
45. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D. 2019. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* 37:451–60
46. Teschendorff AE, Enver T. 2017. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* 8:15599

47. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, et al. 2018. RNA velocity of single cells. *Nature* 560:494–98
48. Soldatov R, Kaucka M, Kastriti ME, Petersen J, Chontorotzea T, et al. 2019. Spatiotemporal structure of cell fate decisions in murine neural crest. *Science* 364:eaas9536
49. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, et al. 2019. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 176:928–43.e22
50. Fischer DS, Fiedler AK, Kernfeld EM, Genga RMJ, Bastidas-Ponce A, et al. 2019. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* 37:461–68
51. Erhard F, Baptista MAP, Krammer T, Hennig T, Lange M, et al. 2019. scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* 571:419–23
52. Cao J, Zhou W, Steemers F, Trapnell C, Shendure J. 2019. Characterizing the temporal dynamics of gene expression in single cells with sci-fate. bioRxiv 666081. **https://doi.org/10.1101/666081**
53. Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, et al. 2017. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* 19:271–81
54. Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, et al. 2016. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* 128:e20–31
55. Macaulay IC, Svensson V, Labalette C, Ferreira L, Hamey F, et al. 2016. Single-Cell RNA-sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep*. 14:966–77
56. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, et al. 2018. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* 173:1535–48.e16
57. Lonnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, et al. 2017. Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves T_H1/T_{FH} fate bifurcation in malaria. *Sci. Immunol.* 2:eaal2192
58. Rodrigues PF, Alberti-Servera L, Eremin A, Grajales-Reyes GE, Ivanek R, Tussiwand R. 2018. Distinct progenitor lineages contribute to the heterogeneity of plasmacytoid dendritic cells. *Nat. Immunol.* 19:711–22
59. Dress RJ, Dutertre CA, Giladi A, Schlitzer A, Low I, et al. 2019. Plasmacytoid dendritic cells develop from $Ly6D^+$ lymphoid progenitors distinct from the myeloid lineage. *Nat. Immunol.* 20:852–64
60. Weinreb C, Wolock S, Tusi BK, Socolovsky M, Klein AM. 2018. Fundamental limits on dynamic inference from single-cell snapshots. *PNAS* 115:E2467–76
61. Tabula Muris Consort. 2018. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. *Nature* 562:367–72
62. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579–605
63. Becht E, McInnes L, Healy J, Dutertre CA, Kwok IWH, et al. 2018. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37:38–44
64. Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31:2989–98
65. Ding J, Condon A, Shah SP. 2018. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002
66. Fruchterman TMJ, Reingold EM. 1991. Graph drawing by force-directed placement. *Softw. Pract. Exp.* 21:1129–64
67. Weinreb C, Wolock S, Klein AM. 2018. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* 34:1246–68
68. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, et al. 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357:661–67
69. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, et al. 2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360:176–82
70. Harmer SL, Panda S, Kay SA. 2001. Molecular bases of circadian rhythms. *Annu. Rev. Cell Dev. Biol.* 17:215–53
71. Oates AC, Morelli LG, Ares S. 2012. Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development* 139:625–39

72. Morrison SJ, Kimble J. 2006. Asymmetric and symmetric stem-cell divisions in development and cancer. *Nature* 441:1068–74
73. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36:421–27
74. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36:411–20
75. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, et al. 2019. Comprehensive integration of single-cell data. *Cell* 177:1888–902.e21
76. Hie B, Bryson B, Berger B. 2019. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* 37:685–91
77. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, et al. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33:1165–72
78. Luo C, Keown CL, Kurihara L, Zhou J, He Y, et al. 2017. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* 357:600–4
79. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, et al. 2014. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* 11:817–20
80. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, et al. 2017. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14:865–68
81. Hainer SJ, Boskovic A, McCannell KN, Rando OJ, Fazzio TG. 2019. Profiling of pluripotency factors in single cells and early embryos. *Cell* 177:1319–29.e11
82. Ai S, Xiong H, Li CC, Luo Y, Shi Q, et al. 2019. Profiling chromatin states using single-cell itChIP-seq. *Nat. Cell Biol.* 21:1164–72
83. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, et al. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* 10:1930
84. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, et al. 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502:59–64
85. Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, et al. 2017. Massively multiplex single-cell Hi-C. *Nat. Methods* 14:263–66
86. Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, et al. 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 361:1380–85
87. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, et al. 2016. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* 13:229–32
88. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, et al. 2015. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* 12:519–22
89. Dey SS, Kester L, Spanjaard B, Bienko M, van Oudenaarden A. 2015. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* 33:285–89
90. Shah S, Lubeck E, Zhou W, Cai L. 2016. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* 92:342–57
91. Eng CL, Lawson M, Zhu Q, Dries R, Koulena N, et al. 2019. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* 568:235–39
92. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348:aaa6090
93. Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, et al. 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* 361:eaat5691
94. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. 2019. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177:1873–87.e17
95. Kebschull JM, Zador AM. 2018. Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods* 15:871–79
96. Walsh C, Cepko CL. 1992. Widespread dispersion of neuronal clones across functional regions of the cerebral cortex. *Science* 255:434–40
97. Lu R, Neff NF, Quake SR, Weissman IL. 2011. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* 29:928–33

98. Naik SH, Perie L, Swart E, Gerlach C, van Rooij N, et al. 2013. Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* 496:229–32
99. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, et al. 2018. Clonal analysis of lineage fate in native haematopoiesis. *Nature* 553:212–16
100. Gerlach C, Rohr JC, Perie L, van Rooij N, van Heijst JW, et al. 2013. Heterogeneous differentiation patterns of individual $CD8^+$ T cells. *Science* 340:635–39
101. van Heijst JW, Gerlach C, Swart E, Sie D, Nunes-Alves C, et al. 2009. Recruitment of antigen-specific $CD8^+$ T cells in response to infection is markedly efficient. *Science* 325:1265–69
102. Peikon ID, Gizatullina DI, Zador AM. 2014. In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res.* 42:e127
103. Pei W, Feyerabend TB, Rossler J, Wang X, Postrach D, et al. 2017. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* 548:456–60
104. Kalhor R, Mali P, Church GM. 2017. Rapidly evolving homing CRISPR barcodes. *Nat. Methods* 14:195–200
105. Perli SD, Cui CH, Lu TK. 2016. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* 353:aag0511
106. Kalhor R, Kalhor K, Mejia L, Leeper K, Graveline A, et al. 2018. Developmental barcoding of whole mouse via homing CRISPR. *Science* 361:eaat9804
107. Alemany A, Florescu M, Baron CS, Peterson-Maduro J, van Oudenaarden A. 2018. Whole-organism clone tracing using single-cell sequencing. *Nature* 556:108–12
108. Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, et al. 2018. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* 36:469–73
109. Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, et al. 2018. Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* 36:442–50
110. Weinreb C, Rodriguez-Fraticelli A, Camargo F, Klein AM. 2018. Lineage tracing on transcriptional landscapes links state to fate during differentiation. bioRxiv 467886. **https://doi.org/10.1101/467886**
111. Chan MM, Smith ZD, Grosswendt S, Kretzmer H, Norman TM, et al. 2019. Molecular recording of mammalian embryogenesis. *Nature* 570:77–82
112. Frieda KL, Linton JM, Hormoz S, Choi J, Chow KHK, et al. 2017. Synthetic recording and in situ readout of lineage information in single cells. *Nature* 541:107–11
113. Laurenti E, Gottgens B. 2018. From haematopoietic stem cells to complex differentiation landscapes. *Nature* 553:418–26
114. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, et al. 2015. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348:910–14
115. Stuart T, Satija R. 2019. Integrative single-cell analysis. *Nat. Rev. Genet.* 20:257–72
116. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6:377–82
117. Regev A, Teichmann SA, Lander ES, Amt I, Benoist C, et al. 2017. The Human Cell Atlas. *eLife* 6:e27041
118. Eraslan G, Avsec Z, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20:389–403