## REVIEW

# Concepts and limitations for learning developmental trajectories from single cell genomics

Sophie Tritschler[1,2,3,*], Maren Büttner[1,4,*], David S. Fischer[1,3], Marius Lange[1,4], Volker Bergen[1,4], Heiko Lickert[2,5,6,‡] and Fabian J. Theis[1,4,‡]

## ABSTRACT

Single cell genomics has become a popular approach to uncover the cellular heterogeneity of progenitor and terminally differentiated cell types with great precision. This approach can also delineate lineage hierarchies and identify molecular programmes of cell-fate acquisition and segregation. Nowadays, tens of thousands of cells are routinely sequenced in single cell-based methods and even more are expected to be analysed in the future. However, interpretation of the resulting data is challenging and requires computational models at multiple levels of abstraction. In contrast to other applications of single cell sequencing, where clustering approaches dominate, developmental systems are generally modelled using continuous structures, trajectories and trees. These trajectory models carry the promise of elucidating mechanisms of development, disease and stimulation response at very high molecular resolution. However, their reliable analysis and biological interpretation requires an understanding of their underlying assumptions and limitations. Here, we review the basic concepts of such computational approaches and discuss the characteristics of developmental processes that can be learnt from trajectory models.

KEY WORDS: Computational approaches, Developmental trajectories, Single cell genomics, Trajectory inference, Pseudotime

## Introduction

Cell specification, proliferation, differentiation and morphogenesis are processes that are essential for the development of an adult organism from a single cell, the zygote. Individual progenitor cells commit and differentiate into lineage-restricted cell types during development and thereby follow paths to distinct fates with perfect precision. Coordinated temporal and spatial gene regulation and multiple signalling pathways orchestrate these cell fate decisions and are key for the development of a zygote into a healthy individual (McGinnis and Krumlauf, 1992). A central endeavour in biomedical sciences, therefore, is to understand these developmental processes and identify how they are mis-regulated in disease. This knowledge can also be harnessed for *in vitro* generation of cells or tissues from stem cells for cell-replacement therapy and drug screens.

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [2]Institute of Diabetes and Regeneration Research, Helmholtz Zentrum München, 85764 Neuherberg, Germany. [3]TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85353 Freising, Germany. [4]Department of Mathematics, Technische Universität München, 85748 Garching, Germany. [5]German Center for Diabetes Research, 85764 Neuherberg, Germany. [6]Institute of Stem Cell Research, Helmholtz Zentrum München, 85764 Neuherberg, Germany.
*These authors contributed equally to this work

‡Authors for correspondence (heiko.lickert@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de)

H.L., 0000-0002-4597-8825; F.J.T., 0000-0002-2419-1943

The advent of single cell-based sequencing now allows us to study how cellular heterogeneity in previously assumed homogeneous cell populations leads to different cell fate decisions (Griffiths et al., 2018; Pijuan-Sala et al., 2018). With single cell resolution, we can monitor an arising population from its earliest emergence onwards. Thus, we can gain detailed insight into the differentiation path that individual cells follow and the underlying molecular programmes executed, which is impossible on the bulk level. However, given that we can generally only measure each cell once, we need computational models to deduce these cellular trajectories and changes in molecular programmes from static snapshot data. Once we have inferred such a trajectory, we can potentially identify the (transcriptional) cell states at which fate decisions are made and the factors driving these decisions. However, a good understanding of the concepts, limitations and scope of such computational models is crucial for drawing robust and reliable biological conclusions.

In this Review, we describe the mathematical concepts of pseudotemporal ordering and trajectory inference as well as recent advances in computational biology that increase the robustness and flexibility with which these concepts can be applied. In particular, we focus on single cell transcriptomic data because this type of data is abundantly available and has profited from the commercialisation of experimental platforms; other important single cell data types, which mirror multiple levels of cellular information, are summarised in Box 1. Furthermore, we highlight current applications of pseudotime algorithms and point out their promises and limitations for studying developmental processes. Finally, we summarise important aspects in the application of pseudotime to disease modelling.

## The concepts of pseudotime and trajectory inference

Current single cell protocols require cell lysis before sequencing, which means that we cannot follow cells over time but can only obtain static snapshots of the developmental process we wish to investigate. One approach to overcome this limitation is to infer a pseudotemporal ordering (see Box 2) of the cells according to their similarity in gene expression, such that each cell is assigned a one-dimensional coordinate that reflects its state in the developmental process. It is assumed that this transcriptomic similarity is dominated by the process of interest, and that the more similar two cells are the closer is their stage of progression along this process. First, a similarity measure is defined that quantifies the similarity between the gene expression profiles of two cells. Then, by ordering the cells according to this similarity measure, a hypothetical trajectory – the pseudotime – of the cells' development can be inferred. In other words, each cell is projected to one point along a one-dimensional process using the information available from its gene expression profile. Thus, although we only obtained a static snapshot, i.e. one measurement of each cell at one time point,

we can investigate developmental processes such as cellular differentiation or maturation.

In such applications of pseudotemporal ordering, the experimental time is the same for all measurements, whereas the internal time, or state, is different for all cells. Pseudotime inference also assumes sampling of a continuum of cellular states independent of sampling time, which implies that all cells following the same (undirected) process will be collapsed onto one trajectory. This means that even if cells from consecutive (developmental) time points are measured, this time information is lost. In addition, it is important to note that the inferred ordering does not convey any information about real elapsed time nor the progression rate of cells. The pseudotime should therefore rather be interpreted as a sequence of cellular states. Moreover, because the ordering is only based on similarity, we need prior knowledge or additional data to assign a direction to the process, i.e. to define which end of the process is 'early or immature' and which one is 'late or mature'.

During development, cells do not follow one single trajectory but commit and differentiate into multiple lineages (Fig. 1A). A simplistic view is that such cell fate decisions occur when transcriptional states diverge, which is represented by branching events. Such diverging lineages can be modelled by tree-like structures. A commonly used analogy for this is Waddington's landscape, where cells are marbles that start at the peak of this landscape and roll down paths into multiple valleys representing different fates (Waddington, 1957). In such tree-like structures and models, cells with different cell fates can be assigned to different paths that segregate at different time points in development (Fig. 1B). To include branching in these models, early trajectory

inference methods (see Box 3) have been extended such that they can embed cells not only along linear paths but also along bi- or multi-furcating, or even more complex, paths.

Multiple methods exist for inferring one-dimensional orderings or complex branching trees. They all include two basic steps. First, cells are projected to a simplified embedding through dimensionality reduction (Fig. 1C), graph construction or clustering. Second, cells are ordered within this simplified embedding (Fig. 1D) and can be further abstracted to a graph (Fig. 1E). The best choice of method heavily depends on the topology of the trajectory in question (e.g. whether it is a linear, bi- or multi-furcated process, see Fig. 1F) (Saelens et al., 2019). For a comprehensive overview and guidelines, we refer the reader to Saelens et al. (2019), who benchmarked the performance of 45 trajectory inference algorithms in different situations. The authors found PAGA (Wolf et al., 2019), Slingshot (Street et al., 2018) and SCORPIUS (Cannoodt et al., 2016) to perform best across all tested datasets. They also propose to both confirm and complement the inferred results by using multiple methods. As a support for end-users to easily apply trajectory inference to their data, and as a resource for developers to evaluate their method, a comprehensive and user-friendly website was compiled (https://dynverse.org/).

## Modelling undirected cell trajectories and lineage relationships in development

Pseudotime and lineage tree inference from single cell transcriptomic data can be broadly applied in the study of developing organisms. For example, pseudotime can be applied to reduce the complex molecular interplay in development to a one-dimensional trajectory that cells follow and can be examined separately by cell fate (Fig. 1G). On Waddington's landscape, this corresponds to the path from a stem cell to just one of its differentiated progenies. Along this path, a continuous spectrum of cellular states is observed, instead of the distinct cell clusters observed in most mature tissues. Thus, the cell types derived from pseudotemporal ordering are fuzzy and overlap. Nonetheless, pseudotime allows the mapping of cell subtypes such as stem cells, differentiating progenitor cells and terminally differentiated cells (i.e. root cells, intermediate states and final states). Using this approach, lineage hierarchies and cellular differentiation trajectories
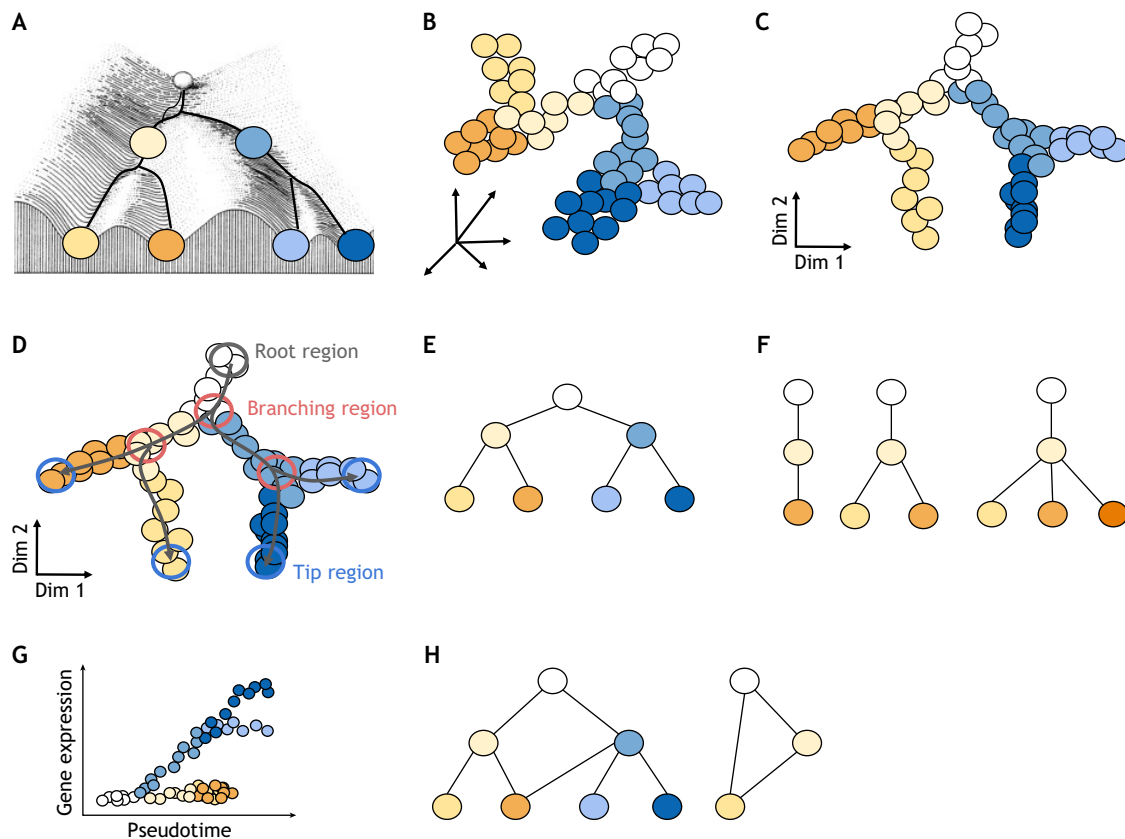
DEVELOPMENT

Fig. 1. Pseudotemporal ordering can reconstruct lineage trees. (A) Waddington's landscape (Waddington, 1957) illustrates the state-space of differentiation and cell fates, where likely cell states are depicted as valleys and unlikely cell states as hills, similar to the free energy landscape. We highlight the positions of four different cell fates and two intermediate cell stages, indicated by coloured circles. Lines connecting all cell stages highlight potential differentiation paths. (B) In a high-dimensional space, cell states correspond to specific expression profiles, with all cell samples lying on the high-dimensional data manifold. (C) The high-dimensional data manifold can be reconstructed via low-dimensional embedding of the data to represent the data manifold in two dimensions (Dim 1 and Dim 2). (D) Important regions (e.g. root, branching and tip regions; indicated by coloured rings) and pseudotime inference (indicated by arrows) can be identified and visualised on the low-dimensional embedding. Pseudotime direction depends on the choice of the root population. (E) An abstracted graph (without directionality) can be constructed from the data based on pseudotime inference and clustering. This is a discretised representation of the data manifold and indicates lineage relationships. (F) Different graph topologies can be inferred from the data. These topologies include (from left to right) continuous transition, bifurcation and trifurcation. (G) Pseudotime inference and lineage tree reconstruction are essential for the analysis of gene expression changes on several cell differentiation paths. The scatter plot of gene expression versus pseudotemporal ordering illustrates how gene expression changes over pseudotime. The scaling of pseudotime depends on the similarity measure (i.e. the amount of changes from root to tip cells) and differs between each branch. (H) Multiple branching events with alternative loops (shortcuts) and a cyclic architecture can also be observed and are special cases of differentiation, reflecting convergent differentiation paths.

have been catalogued in the mouse pancreas (Byrnes et al., 2018; Scavuzzo et al., 2018), in zebrafish embryos (Wagner et al., 2018), and even in whole complex animals such as planaria (Plass et al., 2018). These studies revealed that pseudotime as a proxy for a developmental process and real time can be closely connected.

For example, Byrnes et al. and Scavuzzo et al. studied mouse pancreatic development, which is characterised by a successive increase in complexity (Byrnes et al., 2018; Scavuzzo et al., 2018). By simply aggregating data from several time points, they were able to connect early, undifferentiated cell states with more mature cell populations. Importantly, the authors identified several lineage segregation events that lead to multi-lineage formation and endocrine/exocrine compartmentalisation of the pancreas. Similarly, Wagner et al. sequenced full zebrafish embryos at seven different time points within the first 24 h post-fertilisation (Wagner et al., 2018). Here, the authors reconstructed a coarse-grained developmental lineage tree from the expression profiles of 92,000 cells. The inferred lineage tree immediately suggests that lineage decision making is a tightly timed process. Strikingly, by tracing cell clones using a transposon-based barcode, they found

converging/diverging behaviours undiscovered by pseudotime. Therefore, the developmental routes in zebrafish could not be fully represented by a tree structure. Converging behaviour of cell populations can be described as alternative paths in differentiation processes or as shortcuts in a lineage hierarchy. These shortcuts represent a generalisation of the tree structure, called a directed acyclic graph (Fig. 1H), as reported in a study of hematopoiesis (Moignard et al., 2013). Directed acyclic graphs and convergent developments represent a particular challenge for trajectory inference (see Box 3 for details). Most trajectory inference methods expect fixed starting and end points, which are difficult to define in the loop structure. Furthermore, the detection of branching points is similarly difficult. Both challenges are addressed by specific cyclic methods for trajectory inference or graph-based methods (Saelens et al., 2019).

In contrast to mice and zebrafish, adult planaria are immortal and consist of a considerable number of pluripotent stem cells that are able to regenerate the whole organism. In other words, an adult planaria is approximately in steady-state, i.e. all cell types remain in the same proportions. Therefore, studying the whole organism over

---

**Box 3. Lineage trees and graph abstraction**
For the inference of branching trajectories, the one-dimensional representation (discussed in Box 2) is extended in order to incorporate more complex topologies. A variety of methods exist to uncover these often tree-like structures. Most methods start with embedding the cells in a K-nearest neighbour graph (KNN; see Glossary, Box 4), where nodes represent cells and edges indicate that the expression profile of the two connected cells are similar. A subgraph with a tree-like structure, which represents the developmental process, can then be inferred (Qiu et al., 2017b; Setty et al., 2016). Another approach is to cluster cells based on this graph into relevant cell subtypes or 'metacells', and to infer lineage relationships between these subtypes based on a statistical measure of cluster connectivity (Baran et al., 2018 preprint; Wolf et al., 2019). The resulting 'abstracted graph' only includes the detected cell clusters and their relatedness. An advantage is that this graph uncovers cellular lineage hierarchies without assuming a tree-like structure *a priori*. Thus, it does not exclude the possibility of loops or alternative paths leading to the same cell type and also detects disconnected cell types that are not close to other parts of the data. An alternative approach to clustering the KNN graph directly is to simulate random walks on this graph, i.e. to simulate possible trajectories cells can follow, and to infer the branching tree by clustering a large ensemble of these trajectories (Farrell et al., 2018). Furthermore, dimensionality reduction techniques based on neighbourhood relationships (Becht et al., 2018) or autoencoders (Ding, Condon, and Shah, 2018) can be used to reveal the global branching structure in the embedding.

---

time does not yield any new insight into cell differentiation because we cannot observe the emergence of new cell types and structures as in other developmental model organisms such as mice and zebrafish. In a couple of recent studies, the full, complex lineage tree of planaria was reconstructed in an unbiased way using pseudotime and partition-based graph abstraction (PAGA) (Plass et al., 2018; Wolf et al., 2019). The estimated lineage tree, rooted in pluripotent stem cells, described the transitions into all mature tissues of planaria. Here, pseudotemporal ordering along one lineage path in the abstracted graph is interpreted as a differentiation coordinate, which can be used to examine cell fate decision processes in the fully developed, adult organism.

## Limitations of pseudotemporal ordering

The examples discussed above demonstrate that pseudotemporal ordering is a versatile tool that can be used to connect developmental regimes and to describe cell fate decision making. However, there are several important aspects to be aware of when trying to extrapolate the correct biological interpretation. Firstly, pseudotemporal ordering requires sufficient sampling from the underlying biological process, i.e. the more cells sequenced, the higher the resolution of the continuous differentiation path. Unbiased sampling, such as sequencing a complete organ, yields the most abundant cell types. The less frequent a cell type, the more cells should be sequenced to ensure that all cell types are covered. Enrichment for known markers by fluorescence-activated cell sorting (FACS) is a common approach to obtain sufficient amounts of rare cell types. For example, Scavuzzo et al. enriched the $Neurog3^+$ endocrine lineage in the pancreas to study cell fate priming towards α and β cells (Scavuzzo et al., 2018). However, to distort the developmental landscape as little as possible, prior sorting should be minimised, or could be ameliorated by resampling, and the combination of several sorted subpopulations avoided (Weinreb et al., 2018b). Hence, there is a trade-off between sampling a sufficient amount of (rare) cells for continuous trajectory inference and deforming the developmental landscape.

An important aspect to capture cell fate decisions and the emergence of new populations via pseudotemporal ordering is to decide how often to sample a developmental process. For example, Wagner et al. sampled zebrafish embryos every 2-4 h to monitor their fast development (Wagner et al., 2018). Fischer et al. resolved T-cell maturation in the mouse thymus based on daily samples (Fischer et al., 2019). Scavuzzo et al. sampled mouse pancreas cells from embryonic day (E) 14.5 and E16.5, i.e. used a 48 h interval, to capture the lineage allocation of α and β cells (Scavuzzo et al., 2018). Therefore, these sampling times depend largely on the developmental speed and accessibility of the respective tissues.

Other factors that are key to the success of any trajectory inference method are the input features, which must be informative for the developmental process we wish to model. This means it is important how we filter the measured features or information relating to a cell. Usually, a subset of genes or components of a lower-dimensional representation of the entire gene space ('low-dimensional embedding'; see Glossary, Box 4) are included for the modelling. However, although such feature selection can reduce noise or remove biological variation irrelevant for the developmental process, it can also introduce a strong bias, as cells will be ordered based only on the selected factors. Similarly, methods that include any prior clustering of cells assume that the input clustering is correct. For example, we cannot finely resolve a branching point and decompose a population into cells with distinct fates if the clustering is too broad. Therefore, if we put too much prior knowledge into a model, we will not unbiasedly generate novel information but will simply recapitulate known information, reducing pseudotime to a visualisation tool.

Besides considering these more technical elements, the general assumptions underlying pseudotime inference need to be carefully checked. Most importantly, we assume that (transcriptomic) similarity between two cells implies a developmental relation. This is certainly reasonable, as many examples including the

---

**Box 4. Glossary: the jargon of single cell modellers**
**Dynamic time warping (DTW).** DTW was originally developed for speech recognition (Vintsyuk, 1968). Here, DTW measures the similarity of two time series datasets, which may vary in speed, such as audio tracks. In genomics, DTW allows two pseudotime trajectories to be mapped to one coordinate.
**K-nearest neighbour graph.** A neighbourhood graph consists of a set of vertices and connecting edges among the vertices; each vertex connects to its $k$ closest neighbours where $k$ is a constant.
**Low-dimensional embedding.** High-dimensional data can be projected on a low-dimensional space via dimensionality reduction methods such as PCA, t-SNE, diffusion map or UMAP. Whereas the high-dimensional data reflects the entire (biological) variation, the low-dimensional embedding only preserves important features of the data and removes noise, i.e. the relations between cell populations are recovered.
**Manifold.** A manifold is the mathematical term for a (topological) space, which locally looks like Euclidean space. For example, a city map locally approximates Earth's surface, which altogether is a two-dimensional manifold different from a linear plane.
**Random walk.** A random walk is a stochastic process that is used to describe various phenomena such as diffusion. For example, during diffusion, a particle moves a fixed step size in a random direction at discrete time points. At each time point, a new direction is chosen. If the choice of directions is uniform, the particle moves around its starting point forever.
**Spline-basis.** A non-linear function can be decomposed into a weighted sum of pre-defined non-linear ('wave'-like) functions. This set of pre-defined functions is the spline-basis space and allows the fitting of a non-linear gene expression trajectory in the context of a linear model as a weighted sum of its dimensions.

DEVELOPMENT

222222222

2222222

2222222222222222222222222222222222222222222222222222222

222222 22 2222222222

---

aforementioned have shown, but is only testable by experimental validation. This assumption also raises several other points that we must consider. First, inferred trajectories depend on global similarities in the transcriptomic profile. Highly expressed genes may dominate over lowly expressed factors in these profiles. Thus, the similarity derived from all genes potentially does not take into account the expression of a specific gene relevant for cell fate commitment. Second, the developmental process can be masked by another more dominant continuous expression pattern. This can distort pseudotime estimation and it is important to account for, and potentially correct for, such confounding effects. A typical example of a confounding process is cell cycle progression (Singh et al., 2014; Buettner et al., 2015; Blasi et al., 2017). Indeed, the inferred pseudotime may differ if cell cycle effects are removed or not. Other confounding effects include, for example, spatial patterning, cellular stress or cell size. However, this also implies that the concept of pseudotime can be directly adopted to non-temporal continua of cell states. For example, cells can be ordered along an axis in pseudospace that is characterised by gradual changes in gene signatures (Pijuan-Sala et al., 2019; Scialdone et al., 2016). The ordering may also be inaccurate if gene expression oscillations are the main driver of fate decisions (Weinreb et al., 2018b). Oscillations in cell state cannot be distinguished from non-directed and non-periodic fluctuations. Still, for processes determined not only by oscillations we should in principle be able to find oscillating genes via feature selection. For instance, if only the genes that show gradual expression changes are used for the pseudotemporal ordering, oscillatory patterns of other genes not included at first could be described. Such an approach has been applied to find wave-like expression along the anteroposterior pseudospace axis during murine somitogenesis (Ibarra-Soria et al., 2018) and could be transferred to temporal trajectories. Finally, it should be noted that the simple ordering of cells from snapshot data cannot capture cell dynamics, i.e. in which direction and at which pace a cell and its progeny transitions and traverses a trajectory (Haghverdi et al., 2016; Weinreb et al., 2018b). A common metaphor for a trajectory is a route in the (high-dimensional) transcriptomic landscape that connects the spots reflecting measured cell states. It is up to us to define the starting point of the route and the direction of transitioning cells. Also, in the case of stochastic fluctuations, only state-space normalised time differences, also known as universal time, and not actual temporal dynamics, can be expected to be learnt (Haghverdi et al., 2016). The pseudotime coordinate, a point along such a route, should therefore be interpreted as a measure of cellular differentiation state, or maturity, rather than one of real time.

When a developmental process is not linear but diverges into multiple lineages, the cell fate decisions are often viewed as branching points in the transcriptomic landscape. Trajectories are then modelled as trees or, in more general terms, as graphs. A branching point in the graph – a single spot or cell state in the transcriptomic landscape – is a zero-dimensional node, which implies that lineage commitment happens instantaneously at that point. Graph-based trajectory inference methods, such as PAGA (Wolf et al., 2019), use this assumption to reduce the complexity of the developmental process. However, they are therefore limited to describe fate decisions as an immediate event, and gradual lineage commitment cannot be modelled. This assumption certainly requires careful checking in the future as it was challenged by two recent studies (Setty et al., 2019; Schiebinger et al., 2019). On a side note, it should also be mentioned here that many pseudotime methods assume a fixed topology of the trajectory or are restricted in the type of topology they can model (Saelens et al., 2019). This can be important, as recently the classical definition of cell lineage, a tree, was questioned (Wagner et al., 2018). Cells can follow more complex topologies, including loops or alternative paths during development, and only few methods can model such trajectories.

The basic idea of pseudotime – that a transition likely occurs from a cell state to its most related state – also relies on the assumption that we observe a continuum of cellular states. It presumes that from each cell there could, in principle, be a transition to all other measured cells, i.e. the cells co-exist. This is not necessarily the case if cells are sampled over multiple time points or from multiple locations, and if the distribution of cells changes over time or space. A simple example would be if the same cell types emerge at different time points during development, e.g. the primary and secondary transition of endocrine cell birth in the pancreas, or in separated spatial domains, such as in the dorsal and ventral pancreatic bud (Johansson et al., 2007; Bramswig and Kaestner, 2011; Bastidas-Ponce et al., 2017). Thus, this concern is highly relevant in a scenario in which timing or location of cell state transitions matter and are thereby constrained (Schiebinger et al., 2019; Fischer et al., 2019). It is certainly also an issue for the inference of complex trajectories and lineage trees from an embryonic time series. Classical pseudotime methods do not perform well here; this is apparent, for example, in the large disagreement between the predicted developmental trajectories during zebrafish embryogenesis (Wagner et al., 2018; Farrell et al., 2018). By including additional covariates, such as experimental time or spatial information, this limitation could be overcome (Schiebinger et al., 2019); any inferred trajectory must go forward in experimental time or, similarly, cannot jump across tissue locations.

The restriction to transcriptional cell states is another limitation that is important to consider. In pseudotime inference, cellular trajectories are often defined solely by the current state of a cell, as defined by its gene expression profile. However, gene expression alone might not completely reflect a cell's state. For example, it is clear that morphogenetic movements and the location of a cell are important for cell fate decisions in development, but this type of spatial information is usually not available in scRNA-seq. Likewise, cell memories (e.g. chromatin state, metabolic state or post-translational modifications) are not considered. Also, processes such as asymmetric cell division can give rise to transcriptionally closely related cells with distinct fates (Knoblich, 2008). Indeed, it was shown that bifurcations in the transcriptomic landscape can lag behind real fate decisions because the relevant aspects of a cell's state are not measured by scRNA-seq alone (Weinreb et al., 2018a preprint). Thus, combining pseudotime with different pieces of information, such as lineage tracing, chromatin state, protein expression and phosphorylation, real time or spatial arrangement (see Box 1 and, also in this issue, McKenna and Gagnon, 2019; Ludwig and Bintu, 2019, Mayr et al. 2019) will eventually provide a more complete picture across multiple molecular levels and is crucial for the identification and validation of novel cell transitions and lineage relationships (Colomé-Tatché and Theis, 2018; Weinreb et al., 2018b).

## Inferring directionality and dynamics by superimposing additional data

To move towards more predictive trajectory models, in particular with regards to inferring dynamic information from snapshot data, we need to constrain the space of possible dynamics that could give rise to the same (undirected) trajectory (Weinreb et al., 2018b).

Recently, several experimental and computational approaches have been devised to mitigate the identifiability issues of the underlying models by superimposing additional information. These include cell labelling through lineage-tracing assays, incorporation of experimental time, distinction of mRNA from precursor mRNA, and integration of population size measurements (Fig. 2).

Lineage tracing is important not only to determine the direction of a process but also to validate the existence of predicted cell transitions. In lineage-tracing assays, the offspring of a cell are tracked via heritable marks. In general, one can distinguish between marks that are initially introduced and label cells over a certain time span, and marks that are successively introduced over multiple cell divisions or states and accumulate over time (Fig. 2A). Cell labels that are induced once, e.g. using reporter transgenes, can provide information about clonal expansion and, in combination with single cell phenotyping, cell fates. Using an elegant way to genetically barcode single cells and by measuring both their 'early' and 'late' daughter cells Weinreb et al. were even able to map cell fates from a

continuous set of starting cell states (Weinreb et al., 2018a preprint). By contrast, the repeated barcoding of single cells at multiple time points enables researchers to track back fate decisions at different steps of development allowing reconstruction of multilevel clonal trees. In these clonal trees, a node represents a cell and an edge a mother-daughter cell relationship. If clonal information and gene expression profiles are measured simultaneously, we can quantify the clonal relationships between differentiated cell states or types that are defined by their transcriptomic profile. The number of shared barcodes provides additional information for lineage tree inference: cell states that share few barcodes have diverged earlier during development than states that have many overlapping barcodes. Vice versa, we can also map transcriptional fates, i.e. the cell state of the cells that show a certain genetic barcode, onto the clonal tree, and thereby reconstruct cell fate decision hierarchies. Moreover, it should, in principle, be possible to condense such a clonal tree to resemble a lineage tree, if the clonal hierarchy aligns with the average differentiation path cells follow over time (Weinreb
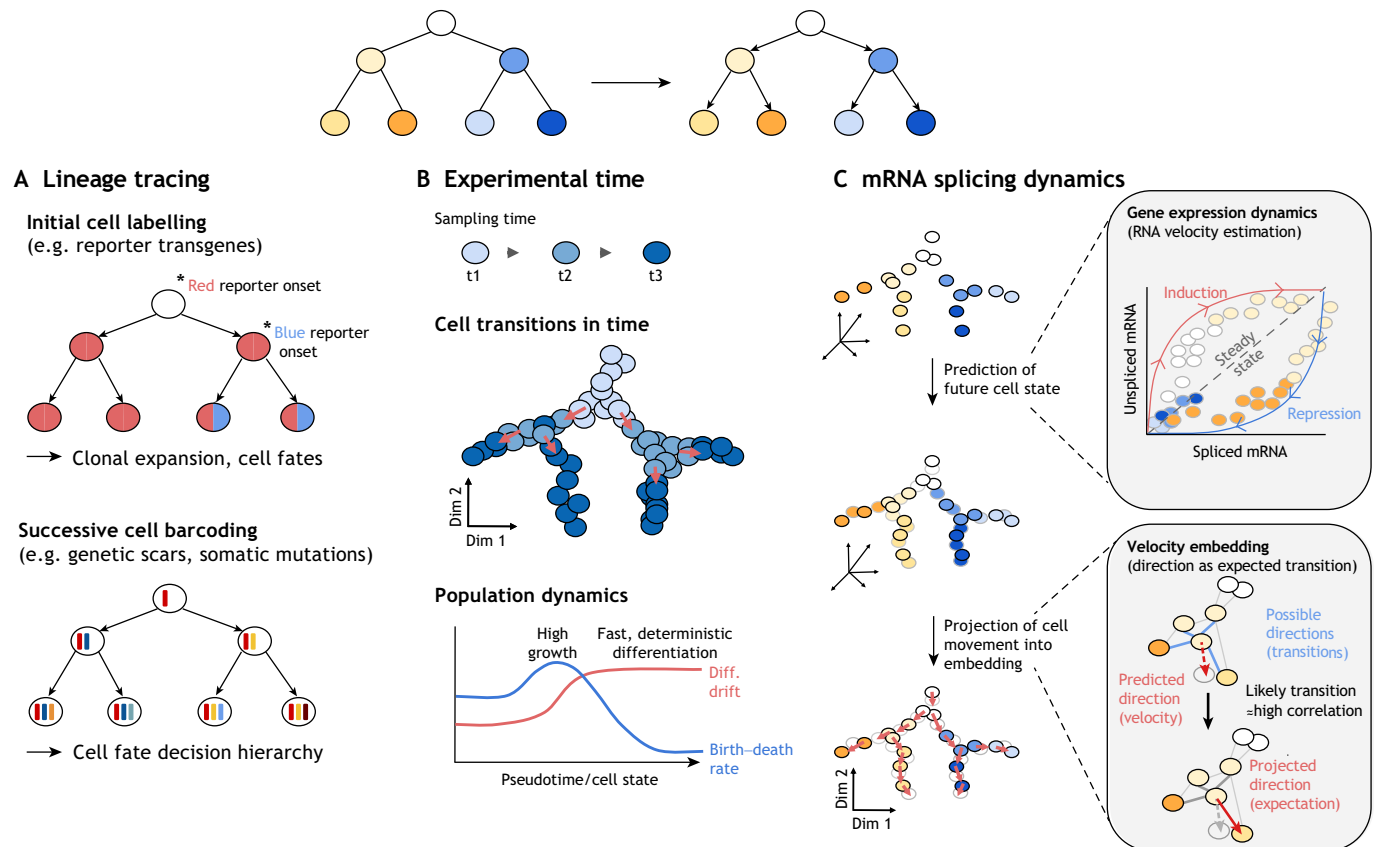


**Fig. 2. Superimposing additional data layers for predictive trajectory models.** Several approaches can be taken to infer directionality and dynamics in trajectories and graphs. (A) In lineage-tracing assays, cells are tracked experimentally via heritable marks to uncover genetic relationships. In combination with single cell genomics, fine-grained clonal trees can be reconstructed and experimentally verified. Cells are either initially labelled at a specific time point to study their clonal expansion and fates, or genetic marks are successively introduced at multiple time points, allowing computational reconstruction of the origin and fate of any cell. (B) The general direction of a developing population is implicit in time-course data and can be directly deduced from superimposing the sampling time on the trajectory. When experimental time is integrated into the trajectory model, more complex dynamic information can be extracted, such as population dynamics. Diff., differentiation. (C) mRNA splicing dynamics add an additional temporal dimension and are directly estimated from single cell transcriptomics data. The timescale of cellular development is comparable to the kinetics of the mRNA life cycle, from the transcription of precursor mRNA to the production of mRNA via splicing to mRNA degradation (La Manno et al., 2018). The ratio of unspliced to spliced mRNA can be leveraged to predict the rate and direction of change in gene expression (i.e. the RNA velocity). A balance indicates homeostasis (steady state) whereas imbalance indicates induction or repression in gene expression. Thus, RNA velocity is an estimate of how mRNA levels evolve over time and can be used to predict the future cell state. The RNA velocity vector of each individual cell can then be projected into a low-dimensional embedding to predict the direction and speed of potential cell movement. For the embedding, the probabilities for each possible cell transition in the neighbourhood graph are computed. A high probability corresponds to a high correlation with the actual velocity vector and indicates a likely transition. The projection is then obtained by the expected mean direction given these transition probabilities. The arrows in an RNA velocity plot, therefore, show the directional flow of cells in the low-dimensional embedding.

et al., 2018a preprint). Note that this requires continuous cell division along the trajectory, such that single clones can expand. This also implies that fate decisions faster than one cell cycle cannot be inferred (Weinreb et al., 2018a preprint). Accordingly, it is likely that some information is lost in the trajectories inferred from clonal hierarchies alone. Recent methods have indeed showcased the potential to unite the phenotypic information obtained from single cell genomics and randomly induced genetic marks (Spanjaard et al., 2018; Frieda et al., 2017; Raj et al., 2018; Alemany et al., 2018; Yao et al., 2017). However, so far they are still technically limited, e.g. because of low labelling efficiency and small clone sizes, and are restricted to a specific experimental system and one time point. Clonal groups also span multiple tissues of origin/germ layers or are spatially organised instead of completely reflecting lineage relationships (Wagner et al., 2018; Spanjaard et al., 2018). Moreover, the clonal tree inference problem scales badly with the size of the system, highlighting that reconstruction of genetic relationships requires advanced computational methods. For a detailed discussion of lineage tracing approaches, we refer the reader to Kester and van Oudenaarden (2018), and to McKenna and Gagnon (2019) in this issue.

Time series experiments offer alternative approaches for learning the dynamics of cell differentiation in non-steady state systems such as developing tissues (Fig. 2B). The simplest way is to overlay time upon the trajectory to derive the direction of cell transitions, or use it as a constraint in the visualisation (Buettner and Theis, 2012). More elegantly, temporal information can be integrated into the trajectory inference. For example, Schiebinger et al. recently leveraged single cell time-course data to reconstruct both an approximation to Waddington's landscape and probabilistic single cell trajectories along a developmental time course (Schiebinger et al., 2019). In contrast to trajectory inference methods based only on transcriptomic profiles, their 'Waddington's-OT' tool computes the distribution of cell states in adjacent time points and matches them. To that end, they built upon the mathematical concept of optimal transport (OT), which efficiently computes a distance between distributions. This concept can be leveraged to compute state-resolved cell flux between adjacent time points and allows a cell to be mapped to its most probable progenies in the consecutive time points. Thus, it exactly addresses the issue of directionality across space and time. Schiebinger et al. used Waddington's-OT to characterise a multitude of developmental programmes during the reprogramming of mouse embryonic fibroblasts into induced pluripotent stem cells. Moreover, it allowed them to predict cell fate decisions of downstream progenitors at consecutive time points. This framework is limited by effects (other than cell flux) that change the distribution of cells between time points, for example proliferation and death (Fischer et al., 2019). Moreover, time is treated as a categorical variable, so uniform sampling of the time course needs to be ensured to adequately capture the range of fluxes. Likewise, fine granularity of the time course is needed to preclude large cell state changes between consecutive time points. Furthermore, OT may produce erroneous directionality when data are asynchronous, e.g. when a fully differentiated cell is already sampled at the first time point and thus is forced to find its descendant at a later stage of differentiation.

In an alternative approach, La Manno et al. explored inference of directional trajectories by examining mRNA metabolism (La Manno et al., 2018) (Fig. 2C). Individual cells contain both newly transcribed, unspliced precursor mRNA and spliced mRNA; the former can be detected in many scRNA-seq assays by the presence

of introns. Intuitively, if a gene shows high amounts of precursor mRNAs, this implies a recent induction of transcription of the particular gene. Conversely, low levels or absence of precursor mRNAs but high amounts of mature mRNAs imply the absence of transcription of that particular gene. Leveraging this intuition in a dynamic model, the authors showed that the change in expression, called RNA velocity (the change in cell state over time), can be approximated from the relationship of mRNA production and its degradation (Fig. 2C). Consequently, RNA velocity gives a local approximation of the behaviour of the system in time, which allows prediction of the future state of a cell. Note that this approximation of the time dependence of the cell state is completely absent in single cell RNA-seq data alone. RNA velocity was successfully applied to inform the directionality of differentiation in regenerating planaria and, in combination with partition-based graph abstraction (Wolf et al., 2019), to reconstruct the lineage tree for the whole animal with all identified cell types rooted in a single stem cell group (Plass et al., 2018). However, the RNA velocity model uses several simplifications. In order to quantify the relationship of mRNA production and degradation, the same constant rate of splicing and degradation is considered in all cells. Furthermore, the estimation of these rates is based on the assumption that the cells come from steady-state populations, which may be realistic only for genes expressed in populations of terminally differentiated cells. Finally, velocities are computed on a gene-wise level. Thus, although the velocity vector can be projected onto a lower-dimensional embedding to show the direction and speed of movement of individual cells, whether the vector field in the embedding truthfully retains the velocities of the high-dimensional space usually depends on sampling, i.e. whether predicted future states are represented adequately by actual cells in the data (see Fig. 2C). Taken together, RNA velocity can greatly aid trajectory analysis, although gene-specific results and low-dimensional representations have to be interpreted with care.

In addition to directionality inference, a recent approach utilised by Fischer et al. leveraged the concept of population dynamics to infer how cells transition between states in a time-resolved manner (Fischer et al., 2019). In this example, the authors modelled the distribution shifts along a pseudotemporal trajectory and the total number of cells in a system with a set of partial differential equations. Including the total number of cells allowed the authors to infer state-resolved birth-death rates in addition to the population dynamics, and deconvolve cell flux from proliferation and death events. The combined modelling approach revealed how proliferative bursts, fast developmental phases and population shrinkage through selection pressure drive T-cell maturation in the mouse thymus (Fig. 2B).

### The identification of genes, pathways and networks driving cell fate decisions

Trajectory or graph descriptions of lineages yield gene expression profiles along a developmental coordinate. These gene expression profiles often correlate with fate decisions and differentiation states (Fig. 3A). It is therefore possible to identify novel marker genes for particular stages of development based on gene expression profiles (Farrell et al., 2018; Plass et al., 2018). Moreover, differential expression can be used to describe the expression changes along a lineage (Qiu et al., 2017a). These changes in gene expression can be summarised to describe changes in gene modules, pathways or annotated gene sets through gene set enrichment tests (Cordero and Stuart, 2017; Fan et al., 2016; Farrell et al., 2018).
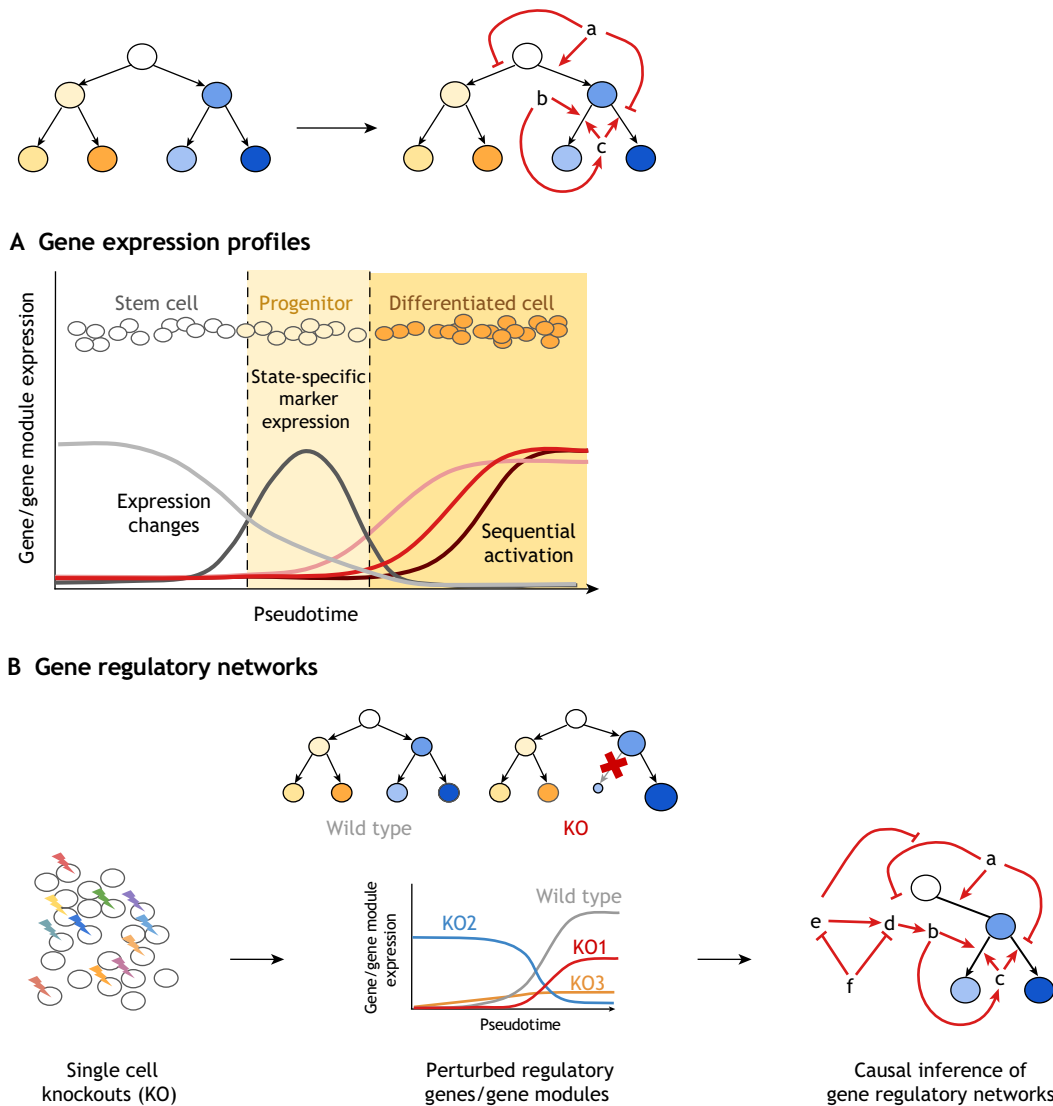
**Fig. 3. Identifying factors and regulatory networks that control cell fate decisions.** Gene expression profiles can be analysed to obtain insights into the regulatory factors (e.g. a, b and c) and pathways (red arrows) underlying cell fate decisions. (A) Potential driving factors of differentiation and fate decisions can be derived from gene expression profiles based on their correlation with pseudotime. (B) Single cell perturbation screens pave the way for causal inference of gene regulation and mechanistic insights into fate decisions.

Gene expression profiles are relatively easy to fit to trajectories. One can divide a trajectory into intervals and fit piecewise constant models via generalised linear models or similar frameworks. Alternatively, one can fit the expression of each gene as a continuous function of the trajectory. Such a continuous fit can be achieved with generalised linear models and a spline-basis (see Glossary, Box 4) transformed continuous cell state. Here, the continuous cell state could, for example, be pseudotime along a trajectory. Sequential gene activation along trajectories can be described based on fits of expression as a function of the trajectory; this was recently employed to characterise the expression of maturation markers along pseudotime in a trajectory model of T-cell maturation (Fischer et al., 2019). Note that one can achieve a higher resolution of sequential gene activation based on continuous cell states compared with discretised approximations of trajectories (Fischer et al., 2019).

Although high-resolution profiling of gene expression in developmental space can provide key insights into a given system, it should be noted that gene expression profiles only yield descriptive information and cannot be used for causal inference of gene regulatory networks. Several methods have been proposed for single cell regulatory network inference (Aibar et al., 2017; Matsumoto et al., 2017; Chan et al., 2017), of which some are even based on pseudotemporally ordered cell states (Hamey et al., 2017). Still, these are merely correlative and based on co-expression networks, and have been shown to perform relatively poorly in predicting accurate network structures (Chen and Mar, 2018). In principle, one could use perturbation screens (Dixit et al., 2016; Datlinger et al., 2017; Adamson et al., 2016; Jaitin et al., 2016) to gain causal insights into gene regulatory networks along trajectories (Tanay and Regev, 2017), although this has not yet been successfully shown. However, from these early CRISPR-Cas9-based methods only the CROP-seq technology (Datlinger et al., 2017) performs well in large-scale experiments, whereas the others suffer from very low sensitivity due to technical problems (Xie et al., 2018), which demand for experimental improvements (Feldman et al., 2018 preprint; Hill et al., 2018; Adamson et al., 2018 preprint). We anticipate that single cell knockouts will yield

DEVELOPMENT

the first proof-of-concept studies for gene regulatory network inference in the near future, as experimental and computational methods mature (Fig. 3B).

### Alterations in lineage allocation and cell differentiation: insights into disease

Diseases often arise due to, or result in, novel or altered cell type states as well as changes in tissue composition and differentiation defects. Although multiple methods have been developed for identifying differential expression between discrete states, approaches for comparing continuous structures such as pseudotime or lineage trees to unravel altered lineage allocation are still largely unexplored. Here, we discuss examples in which embeddings, such as pseudotime, and subtype distributions within lineages trees have been compared between different conditions.

Using the lineage tree derived from healthy zebrafish embryos as a control, Wagner et al. derived a highly resolved map to examine the effect of a *chordin* mutation on the patterning of the zebrafish dorsal-ventral axis (Wagner et al., 2018). In particular, they projected mutant samples onto the wild-type map to compute an over-/under-representation score for each cell type (Fig. 4A). As

expected, ventral tissue cell types were expanded at the expense of dorsal tissue cell types, but no novel cell types emerged. Such changes in patterning are only detectable via unbiased sampling of the complete tissue or embryo. Furthermore, projecting a new dataset into the embedding of control data could be problematic because the transcriptomic profile of a potential new cell type could be orthogonal to the control data; in this case, projection of the new cell type could result in a lower-dimensional embedding to a neighbourhood of cells, which are transcriptionally different, and the new cell type could remain undiscovered. Thus, recomputing a new embedding from a joint dataset may yield better results than projecting new datasets to a pre-existing reference. Note that, in this context, the analysis only used cell-type clustering of the lineage tree for comparison and did not involve a differential pseudotime trajectory (see Box 4) analysis of the two conditions.

In general, it is difficult to map pseudotime trajectories to a universal time unit (Haghverdi et al., 2016) to enable comparisons between conditions, especially if the cell density in the data does not fully reflect the dynamics of the developmental process. An intuitive approach to match time courses involves using the dynamic time warping (DTW) algorithm (see Glossary, Box 4; Fig. 4B). The
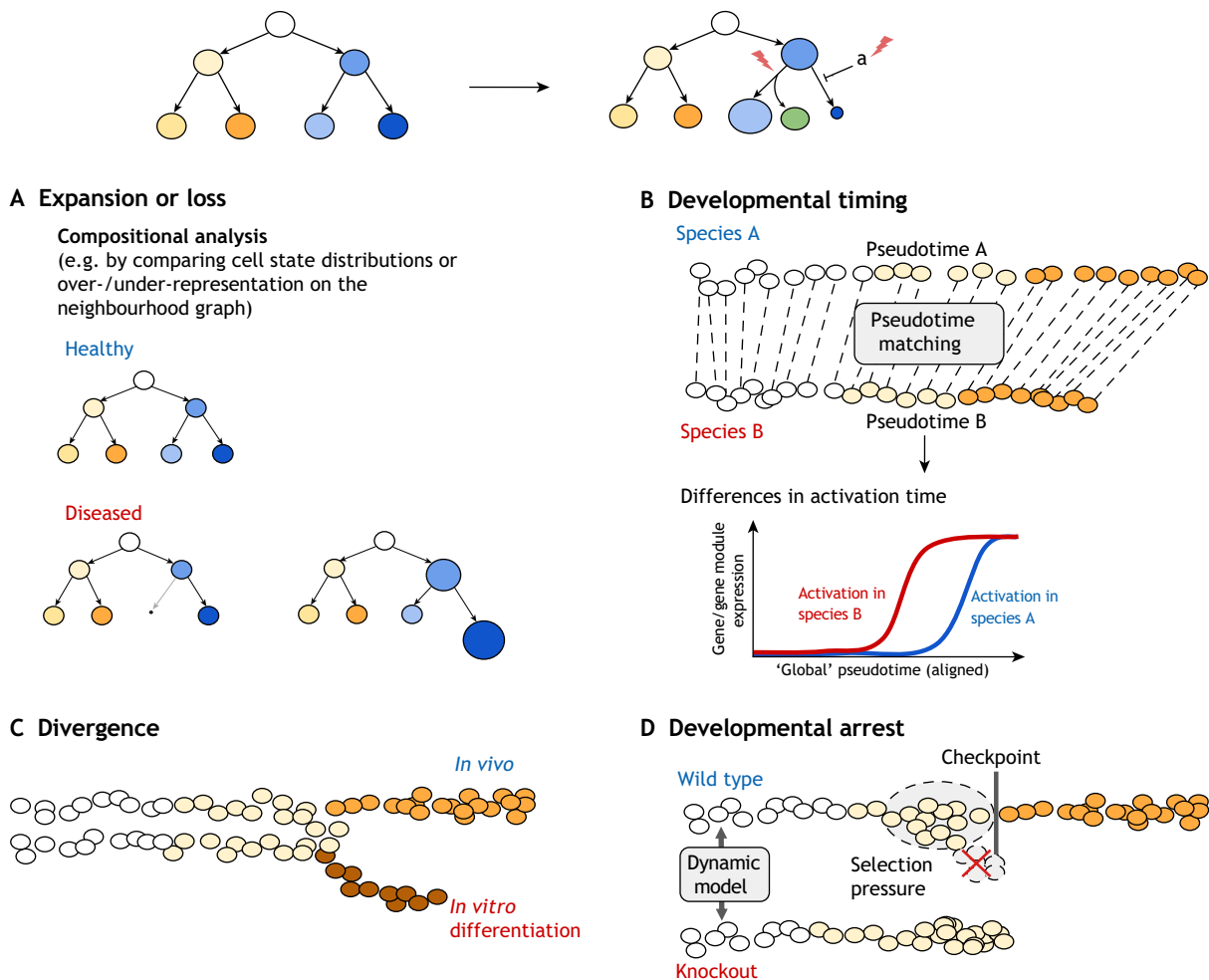


**Fig. 4. Using pseudotime to model disease.** (A) Compositional data analyses can reveal loss or expansion of cell states in disease, either by looking at cell state distributions or over-/under-representation in the single cell neighbourhood graph. (B) For comparison, two trajectories can be matched and a globally aligned pseudotime computed. This allows the identification of timing differences in gene or module activation between, for example, different species or conditions. (C) Using trajectory alignment, divergent processes in diseases or between *in vivo* and *in vitro* models of differentiation can be identified. (D) Dynamic modelling along pseudotime coordinates, together with the integration of knockout data, can enable the description of developmental arrest, exemplified here as a point in development with high selective pressure at which most cells die and only few pass the checkpoint.

tool cellAlign uses DTW to match two pseudotime trajectories and compute a global alignment (Alpert et al., 2018). Such an alignment can reveal timing differences in embryonic development across species. Moreover, cellAlign allows researchers to compare biologically similar processes, such as the response to different stimuli, and unrelated processes such as differentiation of myoblasts and neurogenesis (Fig. 4C). In a similar approach, DTW has been used to match a differentiation and reprogramming process, respectively, to identify a 'failure to exit' branching point in the reprogramming process (Cacchiarelli et al., 2018). However, the identification of such branching points depends on the size of the input gene set, and DTW algorithms are thus yet to be tested for robustness. Notably, the authors of this study also provided a detailed discussion of other caveats of DTW analysis. Most importantly, DTW may remove too much of the biologically relevant information, similar to batch effect correction (Büttner et al., 2019).

In addition to identifying the emergence of new populations, the study of disease models has been used to characterise developmental arrest (Fig. 4D). For example, using a dynamical model along the pseudotime coordinate, Fischer et al. showed that T cells arrest at the beta-selection point during development, based on data from *Rag1/2* knockout mice (Fischer et al., 2019). This example illustrates how pseudotime analysis not only facilitates the observation of changes in cell type composition upon perturbation, but also can serve as a landscape of cell states onto which developmental events can be mapped.

Finally, in order to compare differentiation and development in different conditions, all data should ideally be processed together with a single embedding to reduce systematic biases such as batch effects. Indeed, an extensive evaluation of batch effect correction methods and their applicability for reliable data integration has shown that current batch effect correction methods do not remove experimental bias or tend to overcorrect (Büttner et al., 2019). Thus, the increasing amount of single cell data, and global efforts such as the Human Cell Atlas (Regev et al., 2017), will hopefully encourage the development of robust, scalable and conservative data integration methods.

## Conclusions and perspectives

Single cell omics has been revolutionising the way dynamic developmental processes and cellular lineage hierarchies are being studied. Using trajectory models, the process of cell specialisation can be elucidated. By incorporating a series of interesting concepts from machine learning and computational biology, these models now allow us to precisely track cells and gene expression during differentiation, and reconstruct cell lineage maps of tissues and even whole, complex animals. Still, the largely descriptive and phenomenological information contained in trajectory models offers only a glimpse of the complex processes occurring during development.

From correlation alone, specifically transcriptional similarity of cells or gene expression correlation, we cannot infer causal links but merely can generate hypotheses. A sequential gene expression pattern does not necessarily imply a cell state transition, nor does correlated (coordinated) gene expression a regulatory mechanism. The most direct approaches for the validation of trajectories and causal inference of gene regulation are perturbation experiments and lineage tracing, which now can be combined with single cell transcriptomics. Advances in functional perturbation screening in combination with computational trajectory inference will no doubt play an important role in the future to reconstruct gene regulatory networks controlling cell fates.

The sample size of single cell sequencing data has steeply increased in the past years (Angerer et al., 2017). With increasing sample size, we gain statistical power to fit more complex models, but we also face classical 'big data' challenges, such as the scalability of algorithms and memory-efficient implementation, especially when we aim to integrate data from various sources. Therefore, data integration is moving to the forefront of the single cell field (Stuart and Satija, 2019; Butler et al., 2018; Haghverdi et al., 2018). This will play an important role for developmental research as, by integrating multiple data sets – potentially at some point generating a cell atlas and a road map of development (Regev et al., 2017) – more robust estimates of cell states can be made. Recent work on mouse development have highlighted the potential to obtain such a global view of developmental processes from single cell sequencing data (Cao et al., 2019; Pijuan-Sala et al., 2019; Nowotschin et al., 2018). Moreover, with neural network models for estimating a latent state slowly coming to fruition in single cell omics (Eraslan et al., 2019; Lopez et al., 2018), we can potentially combine data sets using dimension-reduced latent spaces, which represent biologically relevant features. Moreover, understanding dynamic transitions using neural networks is on the horizon. The first examples of applying such deep-learning approaches to integrate distinct datasets and simulate differentiation (Ghahramani et al., 2018preprint) and perturbation effects (Lotfollahi et al., 2018preprint) have shown exciting outcomes and pave the way for future research.

Moving forward, another grand challenge remains the study of the additional layers of regulation that are involved in generating a functional tissue or organ. Recent advances in single cell technologies and computational methods have illustrated the potential to integrate spatial information (Salmén et al., 2018; Satija et al., 2015; Chen et al., 2015), cell-cell interactions (Vento-Tormo et al., 2018 preprint; Schapiro et al., 2017) or population size dynamics (Fischer et al., 2019). Together with our knowledge on cell fates, these pieces of information will give us a more complete perspective of the essential processes that create a whole body from a single cell during development.

## References
**Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nuñez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y. et al.** (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867-82.e21. doi:10.1016/j.cell.2016.11.048

**Adamson, B., Norman, T. M., Jost, M. and Weissman, J. S.** (2018). Approaches to maximize sgRNA-barcode coupling in Perturb-seq screens. *bioRxiv*doi:10.1101/298349

**Ahmed, S., Rattray, M. and Boukouvalas, A.** (2019). GrandPrix: scaling up the Bayesian GPLVM for single-cell data. *Bioinformatics* **35**, 47-54. doi:10.1093/bioinformatics/bty533

**Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J. et al.** (2017).

DEVELOPMENT

SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083-1086. doi:10.1038/nmeth.4463

**Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. and van Oudenaarden, A.** (2018). Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112. doi:10.1038/nature25969

**Alpert, A., Moore, L. S., Dubovik, T. and Shen-Orr, S. S.** (2018). Alignment of single-cell trajectories to compare cellular expression dynamics. *Nat. Methods* **15**, 267-270. doi:10.1038/nmeth.4628

**Angerer, P., Simon, L., Tritschler, S., Wolf, F. A., Ischer, D. and Theis, F. J.** (2017). Single cells make big data: new challenges and opportunities in transcriptomics. *Curr. Opin. Syst. Biol.* **4** Suppl. C, 85-91. doi:10.1016/j.coisb.2017.07.004

**Baran, Y., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A. and Tanay, A.** (2018). MetaCell: analysis of single cell RNA-seq data using K-NN graph partitions. *bioRxiv* doi:10.1101/437665

**Bastidas-Ponce, A., Scheibner, K., Lickert, H. and Bakhti, M.** (2017). Cellular and molecular mechanisms coordinating pancreas development. *Development* **144**, 2873-2888. doi:10.1242/dev.140756

**Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F. and Newell, E. W.** (2018). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38-44. doi:10.1038/nbt.4314

**Blasi, T., Buettner, F., Strasser, M. K., Marr, C. and Theis, F. J.** (2017). cgCorrect: a method to correct for confounding cell–cell variation due to cell growth in single-cell transcriptomics. *Phys. Biol.* **14**, 036001. doi:10.1088/1478-3975/aa609a

**Bramswig, N. C. and Kaestner, K. H.** (2011). Transcriptional regulation of α-cell differentiation. *Diabetes Obes. Metab.* **13**, 13-20. doi:10.1111/j.1463-1326.2011.01440.x

**Buettner, F. and Theis, F. J.** (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* **28**, i626-i632. doi:10.1093/bioinformatics/bts385

**Buettner, F., Natarajan, K. N., Paolo Casale, F., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. and Stegle, O.** (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155-160. doi:10.1038/nbt.3102

**Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R.** (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411-420. doi:10.1038/nbt.4096

**Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. and Theis, F. J.** (2019). A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43-49. doi:10.1038/s41592-018-0254-1

**Byrnes, L. E., Wong, D. M., Subramaniam, M., Meyer, N. P., Gilchrist, C. L., Knox, S. M., Tward, A. D., Ye, C. J. and Sneddon, J. B.** (2018). Lineage dynamics of murine pancreatic development at single-cell resolution. *Nat. Commun.* **9**, 3922. doi:10.1038/s41467-018-06176-3

**Cacchiarelli, D., Qiu, X., Srivatsan, S., Manfredi, A., Ziller, M., Overbey, E., Grimaldi, A., Grimsby, J., Pokharel, P., Livak, K. J. et al.** (2018). Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of myogenic reprogramming outcome. *Cell Syst.* **7**, 258-68.e3. doi:10.1016/j.cels.2018.07.006

**Campbell, K. and Yau, C.** (2015). Bayesian Gaussian process latent variable models for pseudotime inference in single-cell RNA-seq data. *bioRxiv* doi:10.1101/026872

**Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B. N., De Preter, K. and Saeys, Y.** (2016). SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. doi:10.1101/079509

**Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L. et al.** (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385. doi:10.1126/science.aau0730

**Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J. et al.** (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502. doi:10.1038/s41586-019-0969-x

**Chan, T. E., Stumpf, M. P. H. and Babtie, A. C.** (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst.* **5**, 251-67.e3. doi:10.1016/j.cels.2017.08.014

**Chen, S. and Mar, J. C.** (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics* **19**, 232. doi:10.1186/s12859-018-2217-z

**Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. and Zhuang, X.** (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090. doi:10.1126/science.aaa6090

**Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C. et al.** (2018). scNMT-Seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781. doi:10.1038/s41467-018-03149-4

**Colomé-Tatché, M. and Theis, F. J.** (2018). Statistical single cell multi-omics integration. *Curr. Opin. Syst. Biol.* **7**, 54-59. doi:10.1016/j.coisb.2018.01.003

**Cordero, P. and Stuart, J. M.** (2017). Tracing co-regulatory network dynamics in noisy, single-cell transcriptome trajectories. *Pacific Symp. Biocomput.* **22**, 576-587. doi:10.1142/9789813207813_0053

**Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D. and Bock, C.** (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297-301. doi:10.1038/nmeth.4177

**Ding, J., Condon, A. and Shah, S. P.** (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* **9**, 2002. doi:10.1038/s41467-018-04368-5

**Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R. et al.** (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-66.e17. doi:10.1016/j.cell.2016.11.038

**Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. and Theis, F. J.** (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390. doi:10.1038/s41467-018-07931-2

**Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J. et al.** (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13**, 241-244. doi:10.1038/nmeth.3734

**Farrell, J. A., Wang, Y., Riesenfeld, S. J., Shekhar, K., Regev, A. and Schier, A. F.** (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131. doi:10.1126/science.aar3131

**Feldman, D., Singh, A., Garrity, A. J. and Blainey, P. C.** (2018). Lentiviral co-packaging mitigates the effects of intermolecular recombination and multiple integrations in pooled genetic screens. *bioRxiv* doi:10.1101/262121

**Fischer, D. S., Fiedler, A. K., Kernfeld, E. M., Genga, R. M. J., Bastidas-Ponce, A., Bakhti, M., Lickert, H., Hasenauer, J., Maehr, R. and Theis, F. J.** (2019). Inferring population dynamics from single-cell RNA-sequencing time series data. *Nat. Biotechnol.* **37**, 461-468. doi:10.1038/s41587-019-0088-0

**Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B. and Cai, L.** (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111. doi:10.1038/nature20777

**Ghahramani, A., Watt, F. M. and Luscombe, N. M.** (2018). Generative adversarial networks simulate gene expression and predict perturbations in single cells. *bioRxiv* doi:10.1101/262501

**Griffiths, J. A., Scialdone, A. and Marioni, J. C.** (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14**, e8046. doi:10.15252/msb.20178046

**Haghverdi, L., Büttner, M., Wolf, F. A., Büttner, F. and Theis, F. J.** (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845-848. doi:10.1038/nmeth.3971

**Haghverdi, L., Lun, A. T. L., Morgan, M. D. and Marioni, J. C.** (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421-427. doi:10.1038/nbt.4091

**Hamey, F. K., Nestorowa, S., Kinston, S. J., Kent, D. G., Wilson, N. K. and Göttgens, B.** (2017). Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc. Natl. Acad. Sci. USA* **114**, 5822-5829. doi:10.1073/pnas.1610609114

**Hill, A. J., Mcfaline-Figueroa, J. L., Starita, L. M., Gasperini, M. J., Matreyek, K. A., Packer, J., Jackson, D., Shendure, J. and Trapnell, C.** (2018). On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271-274. doi:10.1038/nmeth.4604

**Ibarra-Soria, X., Jawaid, W., Pijuan-Sala, B., Ladopoulos, V., Scialdone, A., Jörg, D. J., Tyser, R. C. V., Calero-Nieto, F. J., Mulas, C., Nichols, J. et al.** (2018). Defining murine organogenesis at single-cell resolution reveals a role for the leukotriene pathway in regulating blood progenitor formation. *Nat. Cell Biol.* **20**, 127-134. doi:10.1038/s41556-017-0013-z

**Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A. and Amit, I.** (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883-96.e15. doi:10.1016/j.cell.2016.11.039

**Johansson, K. A., Dursun, U., Jordan, N., Gu, G., Beermann, F., Gradwohl, G. and Grapin-Botton, A.** (2007). Temporal control of Neurogenin3 activity in pancreas progenitors reveals competence windows for the generation of different endocrine cell types. *Dev. Cell* **12**, 457-465. doi:10.1016/j.devcel.2007.02.010

**Kester, L. and van Oudenaarden, A.** (2018). Single-cell transcriptomics meets lineage tracing. *Cell Stem Cell* **23**, 166-179. doi:10.1016/j.stem.2018.04.014

**Knoblich, J. A.** (2008). Mechanisms of asymmetric stem cell division. *Cell* **132**, 583-597. doi:10.1016/j.cell.2008.02.007

**La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A. et al.** (2018). RNA velocity of single cells. *Nature* **560**, 494-498. doi:10.1038/s41586-018-0414-6

Lein, E., Borm, L. E. and Linnarsson, S. (2017). The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science* **358**, 64-69. doi:10.1126/science.aan6827

Lönnberg, T., Svensson, V., James, K. R., Fernandez-Ruiz, D., Sebina, I., Montandon, R., Soon, M. S. F., Fogg, L. G., Nair, A. S., Liligeto, U. N. et al. (2017). Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves Th1/Tfh fate bifurcation in malaria. *Sci. Immunol.* **2**, eaal2192. doi:10.1126/sciimmunol.aal2192

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053-1058. doi:10.1038/s41592-018-0229-2

Lotfollahi, M., Wolf, F. A. and Theis, F. J. (2018). Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. *bioRxiv* doi:10.1101/478503

Ludwig, C. H. and Bintu, L. (2019). Mapping chromatin modifications at the single cell level. *Development* **146**, dev170217. doi:10.1242/dev.170217

Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M. S. H., Ko, S. B. H., Gouda, N., Hayashi, T. and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* **33**, 2314-2321. doi:10.1093/bioinformatics/btx194

Mayr, U., Serra, D. and Liberali, P. (2019). Exploring single cells in space and time during tissue development, homeostasis and regeneration. *Development* **146**, dev176727. doi:10.1242/dev.176727

McGinnis, W. and Krumlauf, R. (1992). Homeobox genes and axial patterning. *Cell* **68**, 283-302. doi:10.1016/0092-8674(92)90471-N

McKenna, A. and Gagnon, J. A. (2019). Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730. doi:10.1242/dev.169730

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J. et al. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* **15**, 363-372. doi:10.1038/ncb2709

Moor, A. E. and Itzkovitz, S. (2017). Spatial transcriptomics: paving the way for tissue-level systems biology. *Curr. Opin. Biotechnol.* **46**, 126-133. doi:10.1016/j.copbio.2017.02.004

Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C. et al. (2018). Charting the emergent organotypic landscape of the mammalian gut endoderm at single-cell resolution. doi:10.1101/471078

Pijuan-Sala, B., Guibentif, C. and Göttgens, B. (2018). Single-cell transcriptional profiling: a window into embryonic cell-type specification. *Nat. Rev. Mol. Cell Biol.* **19**, 399-412. doi:10.1038/s41580-018-0002-5

Pijuan-Sala, B., Griffiths, J. A., Guibentif, C., Hiscock, T. W., Jawaid, W., Calero-Nieto, F. J., Mulas, C., Ibarra-Soria, X., Tyser, R. C. V., Ho, D. L. L. et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495. doi:10.1038/s41586-019-0933-9

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F. J., Kocks, C. and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723. doi:10.1126/science.aaq1723

Qiu, W.-L., Zhang, Y.-W., Feng, Y., Li, L.-C., Yang, L. and Xu, C.-R. (2017a). Deciphering pancreatic islet β cell and α cell maturation pathways and characteristic features at the single-cell level. *Cell Metab.* **25**, 1194-1205.e4. doi:10.1016/j.cmet.2017.04.003

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A. and Trapnell, C. (2017b). Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979-982. doi:10.1038/nmeth.4402

Raj, B., Wagner, D. E., Mckenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A. and Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442-450. doi:10.1038/nbt.4103

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. et al. (2017). The human cell atlas. *eLife* **6**, e27041. doi:10.7554/eLife.27041

Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547-554. doi:10.1038/s41587-019-0071-9

Salmén, F., Ståhl, P. L., Mollbrink, A., Navarro, J. F., Vickovic, S., Frisén, J. and Lundeberg, J. (2018). Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections. *Nat. Protoc.* **13**, 2501-2534. doi:10.1038/s41596-018-0045-2

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495-502. doi:10.1038/nbt.3192

Scavuzzo, M. A., Hill, M. C., Chmielowiec, J., Yang, D., Teaw, J., Sheng, K., Kong, Y., Bettini, M., Zong, C., Martin, J. F. et al. (2018). Endocrine lineage biases arise in temporally distinct endocrine progenitors during pancreatic morphogenesis. *Nat. Commun.* **9**, 3356. doi:10.1038/s41467-018-05740-1

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R. T., Schulz, D., Giesen, C., Catena, R., Varga, Z. and Bodenmiller, B. (2017). histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873-876. doi:10.1038/nmeth.4391

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P. et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 1517. doi:10.1016/j.cell.2019.02.026

Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C. and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature* **535**, 289-293. doi:10.1038/nature18633

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N. and Pe'er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637-645. doi:10.1038/nbt.3569

Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L. and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451-460. doi:10.1038/s41587-019-0068-4

Singh, A. M., Chappell, J., Trost, R., Lin, L., Wang, T., Tang, J., Matlock, B. K., Weller, K. P., Wu, H., Zhao, S. et al. (2014). Cell-cycle control of developmentally regulated transcription factors accounts for heterogeneity in human pluripotent cells. *Stem Cell Rep.* **2**, 398. doi:10.1016/j.stemcr.2014.02.009

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N. and Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469-473. doi:10.1038/nbt.4124

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. and Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865-868. doi:10.1038/nmeth.4380

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477. doi:10.1186/s12864-018-4772-0

Stuart, T. and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257-272. doi:10.1038/s41576-019-0093-7

Tanay, A. and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331-338. doi:10.1038/nature21350

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S. and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381-386. doi:10.1038/nbt.2859

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., Park, J., Stephenson, E., Polański, K., Payne, R. P. et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347-353. doi:10.1038/s41586-018-0698-6

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics* **4**, 52-57. doi:10.1007/BF01074755

Waddington, C. H. (1957). *The Strategy of the Genes, a Discussion of Some Aspects of Theoretical Biology.* London: Allen & Unwin.

Wagner, D. E., Weinreb, C., Collins, Z. M., Briggs, J. A., Megason, S. G. and Klein, A. M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987. doi:10.1126/science.aar4362

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. and Klein, A. M. (2018a). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *bioRxiv* doi:10.1101/467886

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. and Klein, A. M. (2018b). Fundamental limits on dynamic inference from single-cell snapshots. *Proc. Natl. Acad. Sci. USA* **115**, E2467-E2476. doi:10.1073/pnas.1714723115

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L. and Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59. doi:10.1186/s13059-019-1663-x

Xie, S., Cooley, A., Armendariz, D., Zhou, P. and Hon, G. C. (2018). Frequent sgRNA-barcode recombination in single-cell perturbation assays. *PLoS ONE* **13**, e0198635. doi:10.1371/journal.pone.0198635

Yao, Z., Mich, J. K., Ku, S., Menon, V., Krostag, A.-R., Martinez, R. A., Furchtgott, L., Mulholland, H., Bort, S., Fuqua, M. A. et al. (2017). A single-cell roadmap of lineage bifurcation in human ESC models of embryonic brain development. *Cell Stem Cell* **20**, 120-134. doi:10.1016/j.stem.2016.09.011

DEVELOPMENT