Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Revealing routes of cellular differentiation by single-cell RNA-seq

Dominic Grün

Abstract

Differentiation of multipotent stem cells is controlled by the intricate regulatory interactions of thousands of genes. It remains one of the major challenges to understand how nature has designed such robust and reproducible regulatory mechanisms. Knowing the detailed structure of the underlying lineage trees is the basis for investigating the molecular control of this process. The recent availability of large-scale sensitive single-cell RNA-seq protocols has enabled the generation of snapshot data covering the entire spectrum of cell states in a system of interest. Consequently, a large number of computational methods for the reconstruction of cellular differentiation trajectories have been developed. Here, I will provide a detailed overview of the concepts and ideas behind some of these algorithms and discuss the particular aspects addressed by each method.

Addresses

Max Planck Institute of Immunobiology and Epigenetics, 79108 Freiburg, Germany

Corresponding author: Grün, Dominic (gruen@ie-freiburg.mpg.de)

Current Opinion in Systems Biology 2018, **11**:9–17

This review comes from a themed issue on **Development and differentiation**

Edited by **Stanislav Shvartsman** and **Robert Zinzen**

For a complete overview see the Issue and the Editorial

Available online 21 July 2018

https://doi.org/10.1016/j.coisb.2018.07.006

2452-3100/© 2018 Elsevier Ltd. All rights reserved.

Keywords

Single-cell RNA-sequencing, Differentiation trajectory, Lineage tree, Stem cell differentiation.

Introduction

The emergence and maintenance of a complex multicellular organism requires a multitude of cellular differentiation decisions. These have to be executed with spatial and temporal precision in order to ensure that the appropriate number of cells of each type is produced in a tissue at any developmental stage. Given the stochasticity of the molecular processes underlying the interactions of thousands of genes in each cell, it is remarkable that stem cell differentiation is extremely robust and reproducible, always giving rise to the same complex organismal architecture even under highly variable external conditions [1–4]. It is one of the major goals of contemporary molecular biology to decipher the cellular processes underlying stem cell differentiation [5–7]. Although scientists have explored stem cell differentiation in great detail for decades, e.g. the early embryonic development of mammals [8] or the development and the homeostasis of the immune systems [6], fundamental aspects of cell fate decisions in these and other systems remain to be elucidated.

The recent establishment of a number of advanced single-cell RNA-sequencing protocols [9–18] for the large-scale analysis of single-cell transcriptomes [19–23] has begun to reveal unprecedented insight into the heterogeneity of organs and tissues and spawned the endeavor to characterize every cell type in the human body [24]. Seminal studies have resolved cell types in a variety of tissues, including lung [25], brain [26], skin [27], intestine [28] and bone marrow [29]. One of the main goals is to understand the process by which mature cell types are generated from multipotent cells during development or tissue homeostasis in the adult organism, requiring the inference of cellular differentiation trajectories. It comes as a major disadvantage of single-cell RNA-seq that the tissue has to be dissociated at a particular point in time and cells are lysed during the process and cells are lysed during the process, prohibiting the direct inference of ancestral relations between cells at distinct stages of differentiation. Although cutting-edge multiplexed lineage-tracing techniques utilizing CRISPR/Cas9 [30–33], recombination-based approaches [34,35] or lentiviral barcoding [36] allow the indirect inference of this information, these methods are challenging to establish, depend on the availability of cell type-specific inducible markers, and are limited in resolution.

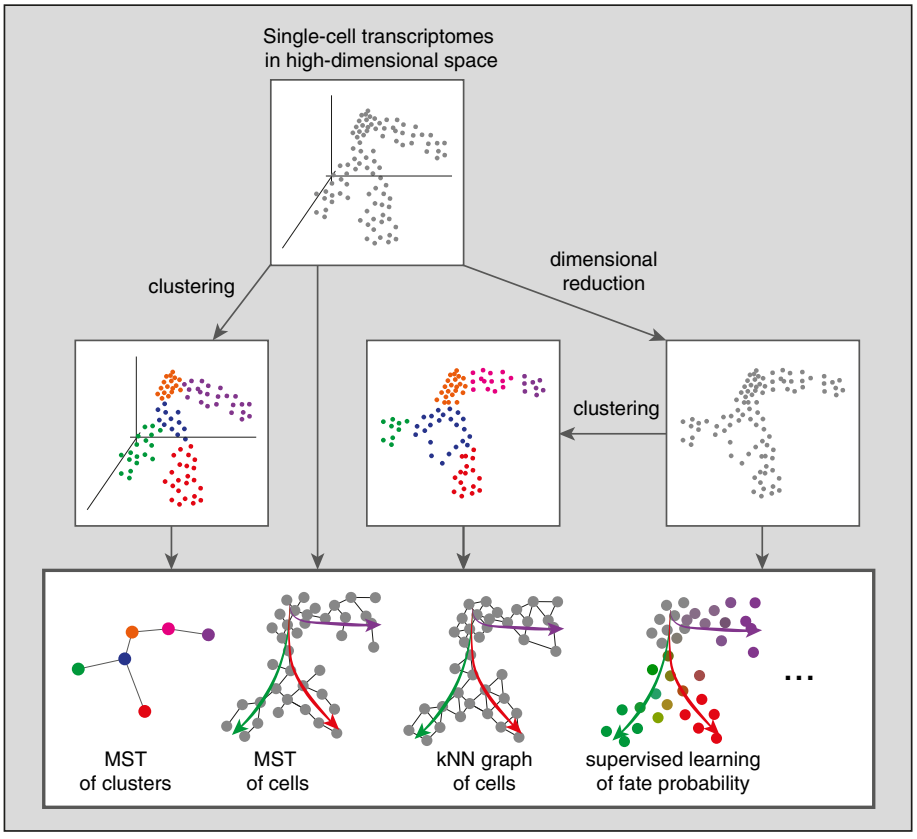
As a common alternative approach, pseudo-temporal ordering permits the inference of differentiation trajectories from single-cell RNA-seq snapshot data of a given tissue. Here, cells are ordered by transcriptome similarity on a continuous trajectory or on a branched structure representing a lineage tree. Such methods assume continuity of transcriptome changes during differentiation and rely on the presence of all intermediate stages connecting the stem cell to the mature cell types in the sample. Since transcript numbers change continuously upon receiving an activation or repression signal, the assumption of continuous changes in

transcriptome space is a reasonable one. However, the presence of all intermediates is much more critical and depends on the differentiation dynamics. Short-lived differentiation stages might be rare and thus be absent from the sample due to cell drop-outs. Related to this, single-cell RNA-seq suffers from substantial technical variability of transcript levels [37,38], making it very challenging to infer a pseudo-temporal ordering. Given sufficient coverage of all intermediate stages, a number of computational methods have been developed which permit the inference of intricate lineage trees from complex snapshot samples comprising single-cell transcriptomes of distinct progenitor stages across multiple different lineages.

Inference of differentiation trajectories by tree-based methods

Over the years a number of studies have applied fundamentally different strategies for the derivation of lineage trees from single-cell RNA-seq data. Common approaches of these methods are exemplified in Figure 1. As one of the first algorithms, Monocle predicts differentiation trajectories on a tree with a specified number of branches of developmentally related cells from connected paths of a minimal spanning tree (MST) computed after reducing the dimensionality of the data by independent component analysis [39]. MSTs are graphs that connect all vertices without any cycles, minimizing the total edge weight. This strategy had already been applied to flow cytometry data in the past, and a number of dedicated algorithms for single-cell RNA-seq trajectory inference have utilized this approach. For multi-branched lineage trees MSTs lack the robustness to reliably resolve branches of transcriptomically similar cell types. Since cell type prediction by dedicated clustering algorithms appears to be a more reliable and less complex task compared to the de novo inference of branched lineage trees, utilizing clustering information helps to reduce complexity and enhance the robustness of lineage tree inference.

Figure 1



Computational approaches to the inference of differentiation trajectories from single-cell RNA-seq data. The starting point of each method is a high-dimensional gene expression matrix indicating the transcript level of each gene in every cell. The data points described by this matrix populate a manifold in high-dimensional space. Cell fate transitions during differentiation follow trajectories within this manifold. To predict these trajectories published algorithms utilize related strategies. In many cases, the manifold is projected into a low dimensional space by one of many different dimensionality reduction algorithms. An initial clustering step is incorporated in many algorithms to reduce complexity and guide trajectory inference. To derive trajectories from the raw or dimensionally reduced data, with or without cluster guidance, published algorithms construct minimal spanning trees or explore connectivity in k-nearest neighbor networks. More recent methods also apply (semi-) supervised strategies to learn trajectories starting from mature end states.

Consequently, a number of algorithms derive MSTs on groups of cells obtained by a prior clustering step. For instance, Waterfall performs k-means clustering on the first two components of a PCA and builds an MST on these clusters [40]. A similar approach of cluster-guided MST-inference is employed by the TSCAN method [41]. Monocle 2 introduced another strategy of cluster-guided lineage tree inference. Starting from an initial dimensional reduction, it constructs an MST on cluster centroids derived by k-means and updates cell positions by shifting towards the nearest vertices. This procedure is iterated until it converges to a stable configuration [42]. Slingshot [43] is the most recent algorithm within this group of methods: it starts by building an MST based on an arbitrary input clustering and refines trajectories by fitting principal curves onto the MST structure. This allows to freely choose methods for prior dimensionality reduction and clustering. Due to the secondary principal curve inference Slingshot is very robust to the choice of the clustering method and the cluster number. Moreover, it permits the integration of prior knowledge by the definition of differentiation endpoints. The StemID method identifies differentiation trajectories as sequences of links between cluster medoids, which are more populated then expected by chance [44]. Instead of using MSTs, StemID assigns individual cells to links by maximizing the projection coordinate of a vector connecting a cell to the medoid of its cluster onto the links to all other clusters. It could correctly predict lineage tree topologies of intestinal epithelial cells and of the hematopoietic system. Similarly, the Mpath algorithm predicts multi-branched lineage trees by first identifying landmark cell states by clustering followed by the identification of highly populated transitions between landmarks in order to build a neighborhood network [45].

Cluster-guidance arguably increases robustness, but at the same time limits the resolution of branching points. In general, clustering methods per se are not ideal for resolving branched structures. A notable exception is the dedicated K-branches method, which applies local fits of K half-lines with a common center, utilizing a strategy akin to the K-means algorithm, and identifies tip regions, intermediate regions and branching regions by model selection [46]. K-branches showed excellent performance in resolving branching regions, e.g. within myeloid progenitor single-cell RNA-seq data [29].

Decoding of differentiation trajectories by graph-based methods

Another class of algorithms leverages the power of k-nearest neighbor (kNN) graphs. The Wishbone algorithm measures developmental distance between cells by shortest paths within a kNN graph [47]. It initially starts ordering cells by the distance to an early input cell, and refines ordering by averaging distances from randomly selected cells that act as waypoints with weights reflecting the proximity of a cell to a waypoint. Inconsistencies in the lengths of paths to a cell crossing different waypoints enable the identification of bifurcations. An important aspect of Wishbone is the avoidance of short circuits by applying an initial dimensional reduction using diffusion maps, which determine the distance between two cells by considering all possible paths connecting these cells. Wishbone recovered the branching of myeloid and erythroid progenitors in a single-cell RNA-seq data set [29]. The SLICER algorithm [48] was designed to improve the prediction of nonlinear differentiation trajectories by constructing a kNN graph after performing nonlinear dimensionality reduction using locally linear embedding of an initial kNN graph constructed in the original space based on an inferred set of genes with systematic variation. To identify the number and location of branches, a metric called geodesic entropy was defined, based on evaluating shared vertices across the collection of shortest paths from a starting cell. The most recent algorithm belonging to this group, p-Creode [49], computes a kNN graph after density normalization by downsampling and defines a graph attribute termed closeness centrality to reveal end states. These are defined by k-means clustering on cells with low closeness centrality values, and connected via path nodes in a hierarchical manner to reconstruct the topology of the tree. The final topology is obtained after iterative repositioning of the path nodes to ensure that paths are aligned with dense regions of the data. p-Creode revealed the complex multi-lineage tree of colonic epithelial cells and enabled the identification of a novel pathway of tuft cell development.

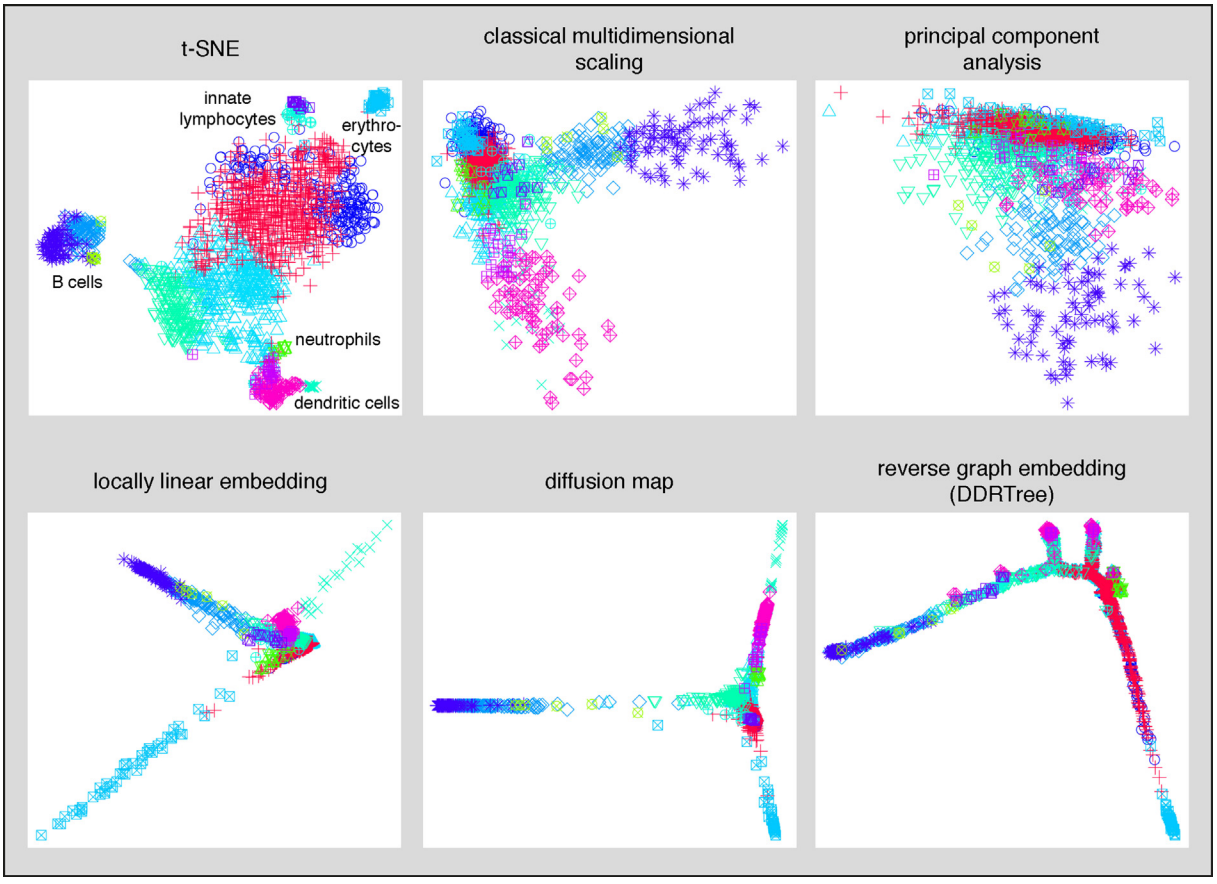
Dimensional reduction reveals underlying manifold of lineage trees

Diffusion maps have become very popular as a tool for dimensionality reduction and visualization of single-cell RNA-seq data [47,50,51] due to their favorable robustness and scalability, permitting the analysis of tens of thousands of cells. To leverage the power of this method for trajectory reconstruction, a metric called diffusion pseudo-time (DPT) was developed [51]. DPT convolves Gaussians centered at nearby cells to construct a weighted nearest neighbor graph. The probability of transitioning between any two cells is then measured by the ensemble of random walks of any lengths connecting these cells on the graph. The method identifies branching by a change from anti-correlation to positive correlation between the paths connecting two distinct cells to a third cell. The favorable robustness of DPT in comparison to alternative methods was demonstrated on single cell qPCR data of early blood development [52], and large-scale single cell RNA-seq data of myeloid progenitors [29] and embryonic stem cells [11]. The recent MAGIC algorithm [53] exploits the weighted

graph structure of diffusion maps in order to impute the expression profile of each cell from the weighted average of its neighbors, leading to smoothening of the data and enhancement of cluster and trajectory structure, in particular for single-cell RNA-seq data with high transcript drop-out rate, e.g. generated by droplet microfluidics.

Dimensional reduction algorithms, such as t-SNE, multidimensional scaling, diffusion maps, or principal component analysis emphasize specific aspects of the data and the resulting trajectory inference will strongly depend on the choice of the dimensional reduction method (Figure 2). Moreover, these methods have limited ability to preserve complex topologies of differentiation trajectories in the original high-dimensional space. The scTDA method [54] was specifically developed to ameliorate this short-coming by applying clustering of cells within the original space for cells extracted from bins in a dimensional reduction representation and connecting nodes of clusters with shared cells, yielding a low-dimensional network representation. Based on gene expression patterns on this network, i.e., common distance from a root cell and similar expression in connected sets of cells, scTDA identifies transient states. In particular, scTDA can reveal periodic topological structures such as trajectories arising from the cell cycle, which was demonstrated for in vitro differentiation of embryonic stem cells. The cell cycle is one example of a source of variability considered as a confounding factor for trajectory inference. An alternative to infer circular trajectories is the removal of gene expression variability associated with the cell cycle or other unwanted hidden factors. For this purpose, single-

Figure 2



Dimensionality reduction affects data structure. Dimensionality reduction of high-dimensional single-cell RNA-seq data is a common pre-processing step for cell type identification and differentiation trajectory inference. However, different algorithms emphasize distinct aspects of the data. The figure depicts single-cell transcriptome data of adult mouse bone marrow cells from Herman et al. [64]. Cell types are annotated based on specific marker genes. The central cloud in the t-SNE map comprises multipotent progenitors, which populate a much denser region in the original space compared to more mature populations. While the t-SNE map resolves the structure of this population well and reveals how rare populations such as innate lymphocytes and neutrophils (under-represented in this dataset) emanate from these progenitors, it separates more advanced stages from this cloud (B-cells and erythrocytes). Classical multidimensional scaling conserves distances in the original space well, connects B cells and erythrocytes to the progenitor cloud, but cannot resolve the structure of the progenitor population. The first two principal components are only sufficient to resolve B cell differentiation. Locally linear embedding and diffusion maps reflect continuous trajectories of B cell, dendritic cell, and erythrocyte differentiation, but fail to resolve rare populations. Reverse graph embedding by DDRTree [68] shows a similar structure, but splits the dendritic cell trajectory into two sub-branches.

cell latent variable models have been utilized [55]. In this approach, the covariance structure of a set of genes representing a hidden factor such as the cell cycle is used to decompose gene expression variability into a technical component, a component for each hidden factor, and a residual biological component. The latter can be used in downstream analyses, effectively eliminating the variability associated with technical noise and other hidden factors.

Identifying stem cells from lineage trees and decoding directionality of differentiation trajectories

Once a lineage tree has been predicted from the computational analysis of single-cell RNA-seq data, a non-trivial challenge is the identification of the root of the tree corresponding to a multipotent or stem cell. While heuristic screening of the tree structure can reveal the stem cell if prior information on stem cell marker genes is available, unstudied systems require an unsupervised de novo approach. A number of methods have been developed to address this challenge. As the first method of this kind, StemID [44] introduced a score proportional to the number of links of a cluster of cells to other clusters reflecting the level of multipotency. This number is multiplied by the transcriptome entropy of the cluster. High entropy reflects a state of more unspecific gene expression, while low entropy indicates that the transcriptome is dominated by a few highly expressed genes, a situation often observed in specialized mature cell types. For instance, erythrocytes are specialized towards the production of hemoglobin and pancreatic beta cells dedicate most expressed transcripts to the synthesis of insulin. Following a similar idea, SLICE [56] computes the functional entropy of a cell based on the expression of genes associated with given functional annotations such as GO terms, and reconstructs a network of stable states represented by local minima of the entropy. After locally grouping cells by network-based community detection or clustering, differentiation trajectories connecting these groups are inferred by building an MST only permitting transitions leading from high to low entropy states and thus yielding a lineage tree with a candidate multipotent cell type at its root. Yet another concept of entropy as a proxy of differentiation potency has been employed in the SCENT method [57]. Here, the transcriptome of a cell is integrated with a high-confidence protein−protein interaction network to define a cell-specific signaling process. The core idea is that proteins are more likely to interact if the corresponding transcripts are present at larger numbers. The entropy on the network derived from these expression-based interaction probabilities, termed signaling entropy, is expected to be high for multipotent cells, simultaneously activating diverse pathways, and low for specialized mature cell types. After identifying potency states by a Bayesian mixture model applied to the entropy values and deriving cell states by bi-clustering in potency-coexpression state, a lineage trajectory network is predicted from partial correlation of cell states. The result is a directed lineage tree with a candidate stem cell at its root.

Another elegant strategy to infer directionality of differentiation trajectories and identify root and end states relies on the estimation of transcriptome velocity [58]. RNA velocity estimates the rate and direction of expression change for each gene from relative read counts of spliced versus unspliced transcripts, in essence modeling the lifecycle of a transcript, to enable the extrapolation of the future state of a cell. Since the timescales of an RNA lifecycle are often comparable to the timescales of differentiation processes, the vector field predicted as RNA velocity reflects the movement of cells along differentiation trajectories connecting distinct cell states and allows the identification of root and terminal states as source and sinks of the velocity field, respectively. The algorithm was applied to describe fate decisions of major neural lineages in the hippocampus.

Towards a probabilistic understanding of cell fate emergence

Most available computational methods for the inference of lineage trees from single-cell RNA-seq data are deterministic in their assignment of each cell to an individual branch. This view is agnostic to the probabilistic nature of cell fate decision, assuming that a given progenitor state could give rise to a number of fates with different probabilities in a stochastic manner. Potentially, gene expression variability of master regulators could be an underlying mechanism, requiring that a random fluctuation of transcript levels crosses a given threshold in order to drive differentiation towards a particular fate [2,59−61]. A probabilistic modeling of cell differentiation in general leads to a better understanding of the commitment process by revealing the stages at which a progenitor looses potency for alternative fates. A beautiful example of this approach was implemented in the GPfates algorithm [62] for modeling the bifurcation into T_H1 and T_{FH} sub-types of T helper cells during blood-stage *Plasmodium* infection in mice. After dimensional reduction and pseudotime inference within the Gaussian process framework, GPfates models cell states along a trajectory branching into multiple fates by a Gaussian Process Latent Variable Model, i.e. an overlapping mixture of Gaussian processes each corresponding to a distinct fate. For the studied system, this model demonstrated the gradual bifurcation into two fates.

The STEMNET algorithm [63] represents another approach to the probabilistic analysis of lineage priming.

This supervised method relies on prior knowledge of terminal cell states, which can be, for example, unambiguously identified based on specific marker gene expression. STEMNET predicts the fate probability of naïve multipotent cells from these mature states by a robust elastic-net regularized general linear model. Applied to human hematopoietic cells, this algorithm predicts direct emergence of unilineage-restricted cells from low-primed hematopoietic stem and progenitor cells. Following a related strategy, the semi-supervised FateID algorithm [64] also starts from defined end states to learn the fate bias, i.e. the probability to differentiate into each lineage, by random forests-based classification. However, in contrast to STEMNET, this algorithm maintains a dynamic training set by iteratively moving "backward" in differentiation time from the mature states into the naïve multipotent compartment. This strategy accounts for the activity of distinct gene modules and regulatory pathways at sub-sequent stages of differentiation and ensures that naïve cells are not classified based on genes only expressed at terminal stages. The application of FateID to mouse hematopoietic progenitors showed that the multipotent progenitor population segregates into domains with a predominant bias towards a particular lineage. Existing transition zones between these domains represent oligopotent progenitor states.

Using time course data to infer complex developmental trees

With the availability of microfluidic-based high-throughput single-cell RNA-seq it became feasible to acquire dense time course data covering subsequent stages of embryonic development in vertebrates. Two studies analyzed differentiation trajectories in developing frog [65] and zebrafish [66] embryos, respectively, by performing large-scale single-cell RNA-seq at subsequent timepoints. The authors applied graph-based analysis strategies utilizing the actual developmental timepoint information. After identifying local neighborhoods within graphs for each timepoint separately, ancestral relations between the states represented by these timepoint-specific neighborhoods are inferred in the frog study: for each cell residing in a given state at a particular timepoint, nearest neighbors are identified within the dataset of the subsequent timepoint and state transitions are predicted based on consensus, i.e. based on the most frequently connected states. In the zebrafish study, k-nearest neighbor graphs are constructed for each timepoint and subsequently connected based on nearest neighbor links across timepoints, giving rise to a single graph connecting all timepoints, amenable to formal graph-based methods.

Another similar study on zebrafish embryonic developmental trajectories utilizing single-cell RNA-seq data [67] introduces the URD method based on an extension of the diffusion map framework. URD applies simulated diffusion on a k-nearest neighbor graph connecting cells across all timepoints, followed by the identification of root and tip states as starting and end points, respectively, based on actual developmental time. Cells are then ordered by developmental progress based on simulated diffusion starting from the root. Finally, branched lineage trajectories are inferred by simulated biased random walks with decreasing pseudotime starting from the tip. Cells visited on backward trajectories starting from distinct tips enable the identification of branching points.

All three strategies allow successful reconstruction of the complex developmental lineage trees of frog and zebrafish, respectively, and demonstrate how actual developmental time information can be incorporated into lineage inference algorithms.

Conclusions

The challenging task of deriving differentiation trajectories from single-cell RNA-seq snapshot data covering a multi-branched lineage tree has been addressed by a large number of methods. Although the major goal of lineage tree inference remains the same, most methods focus on a specific aspect, such as high-resolution analysis of branching regions [46,47], topological structure of the data capturing highly non-linear and circular trajectories [54], the prediction of stem cells [44,56,57], velocity and directionality of differentiation [58], or the probabilistic quantification of multi-lineage bias in individual cells [62–64]. The field will continue to establish novel experimental methods to integrate other relevant aspects into the analysis of cellular differentiation at single-cell resolution, such as ancestral information or spatial context. This progress will require the development of sophisticated algorithms for the multimodal analysis of large-scale single-cell data. Hence, exciting challenges lie ahead of us promising a fundamentally deeper understanding of the fascinating and complex process of stem cell differentiation as a reward.

Conflict of interest statement

Nothing declared.

Acknowledgements

I thank Nina Cabezas-Wallscheid, Roman Sankowski, and Josip Herman for critical reading of the manuscript. The work was financially supported by the Max Planck Society.

References

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Eldar A, Elowitz MB: **Functional roles for noise in genetic circuits**. *Nature* 2010, **467**:167–173.

2. Raj A, van Oudenaarden A: **Nature, nurture, or chance: stochastic gene expression and its consequences**. *Cell* 2008, **135**:216–226.

3. Munsky B, Neuert G, van Oudenaarden A: **Using gene expression noise to understand gene regulation**. *Science* 2012, **336**:183–187.

4. Lucchetta EM, Lee JH, Fu LA, Patel NH, Ismagilov RF: **Dynamics of Drosophila embryonic patterning network perturbed in space and time using microfluidics**. *Nature* 2005, **434**:1134–1138.

5. Jaenisch R, Young R: **Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming**. *Cell* 2008, **132**:567–582.

6. Orkin SH, Zon LI: **Hematopoiesis: an evolving paradigm for stem cell biology**. *Cell* 2008, **132**:631–644.

7. van der Flier LG, Clevers H: **Stem cells, self-renewal, and differentiation in the intestinal epithelium**. *Annu Rev Physiol* 2009, **71**:241–260.

8. Yamanaka Y, Ralston A, Stephenson RO, Rossant J: **Cell and molecular regulation of the mouse blastocyst**. *Dev Dynam* 2006, **235**:2301–2314.

9. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R: **Smart-seq2 for sensitive full-length transcriptome profiling in single cells**. *Nat Methods* 2013, **10**:1096–1098.

10. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, *et al.*: **CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq**. *Genome Biol* 2016, **17**:77.

11. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, *et al.*: **Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets**. *Cell* 2015, **161**:1202–1214.

12. Klein CA, Schmidt-Kittler O, Schardt JA, Pantel K, Speicher MR, Riethmuller G: **Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells**. *Proc Natl Acad Sci* 1999, **96**:4494–4499.

13. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, Ueda HR: **Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity**. *Genome Biol* 2013, **14**:R31.

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

15. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, Fortune S, Love JC, Shalek AK: **Seq-well: portable, low-cost RNA sequencing of single cells at high throughput**. *Nat Methods* 2017, **14**:395–398.

16. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, Qiu X, Lee C, Furlan SN, Steemers FJ, *et al.*: **Comprehensive single-cell transcriptional profiling of a multicellular organism**. *Science* 2017, **357**:661–667.

17. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, Mildner A, Cohen N, Jung S, Tanay A, *et al.*: **Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types**. *Science (80-)* 2014, **343**:776–779.

18. • Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, *et al.*: **Mapping the mouse cell atlas by microwell-seq**. *Cell* 2018, **172**. 1091–1107.e17.

Microwell-Seq allows RNA-seq of tens of thousands of individual cells at low cost and minimal equipment and was utilized to build a cell type atlas of several mouse organs establishing a valuable resource for future studies.

19. Grün D, van Oudenaarden A: **Design and analysis of single-cell sequencing experiments**. *Cell* 2015, **163**:799–810.

20. Saliba A-E, Westermann AJ, Gorski SA, Vogel J: **Single-cell RNA-seq: advances and future challenges**. *Nucleic Acids Res* 2014, **42**:8845–8860.

21. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in single-cell transcriptomics**. *Nat Rev Genet* 2015, **16**:133–145.

22. Picelli S: **Single-cell RNA-sequencing: the future of genome biology is now**. *RNA Biol* 2017, **14**:637–650.

23. Papalexi E, Satija R: **Single-cell RNA sequencing to explore immune cell heterogeneity**. *Nat Rev Immunol* 2017, **18**:35–45.

24. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, Carninci P, Clatworthy M, *et al.*: **Science forum: the human cell atlas**. *Elife* 2017, **6**.

25. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, Desai TJ, Krasnow MA, Quake SR: **Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq**. *Nature* 2014, **509**:371–375.

26. Zeisel A, Manchado ABM, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, Marques S, Munguba H, He L, Betsholtz C, *et al.*: **Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq**. *Science (80-)* 2015, **347**:1138–1142.

27. Joost S, Zeisel A, Jacob T, Sun X, La Manno G, Lönnerberg P, Linnarsson S, Kasper M: **Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity**. *Cell Syst* 2016, **3**. 221–237.e9.

28. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A: **Single-cell messenger RNA sequencing reveals rare intestinal cell types**. *Nature* 2015, **525**:251–255.

29. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, *et al.*: **Transcriptional heterogeneity and lineage commitment in myeloid progenitors**. *Cell* 2015, **163**:1663–1677.

30. • Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, Ninov N, Junker JP: **Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars**. *Nat Biotechnol* 2018, **36**:469–473.

Simultaneous CRISPR-Cas9-base lineage tracing and single-cell RNA-seq recapitulates cell lineage decisions during zebrafish development.

31. • Alemany A, Florescu M, Baron CS, Peterson-Maduro J, van Oudenaarden A: **Whole-organism clone tracing using single-cell sequencing**. *Nature* 2018, **556**:108–112.

The ScarTrace method enables simultaneous lineage tracing utilizing imperfect repair of CRISPR-Cas9 mediated breaks and allows deciphering cell fate decisions during zebrafish embryonic development and fin regeneration.

32. McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J: **Whole-organism lineage tracing by combinatorial and cumulative genome editing**. *Science* 2016, **353**. aaf7907.

33. • Raj B, Wagner DE, McKenna A, Pandey S, Klein AM, Shendure J, Gagnon JA, Schier AF: **Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain**. *Nat Biotechnol* 2018, **36**:442–450.

Combining lineage tracing capabilities of inducible CRISPR-Cas9 mutagenesis at multiple time points with single-cell RNA-seq reveals cell types and lineage histories in juvenile zebrafish brain.

34. •• Pei W, Feyerabend TB, Rössler J, Wang X, Postrach D, Busch K, Rode I, Klapproth K, Dietlein N, Quedenau C, *et al.*: **Polylox barcoding reveals haematopoietic stem cell fates realized in vivo**. *Nature* 2017, **548**:456–460.

The unperturbed inference of ancestral relations of cell type progeny emerging from the same stem cell population is the gold standard for lineage tree inference. The polylox system utilizes a recombination-based barcoding approach to label individual clones emerging from the murine hematopoietic stem cell compartment with high diversity. The analysis supports the existence of an early split into myeloid and lymphoid branches of the hematopoietic tree.

35. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, Jankovic M, Sun J, Calogero RA, Klein AM, Camargo FD: **Clonal analysis of lineage fate in native haematopoiesis**. *Nature* 2018, https://doi.org/10.1038/nature25168.

36. Perié L, Duffy KR, Kok L, de Boer RJ, Schumacher TN: **The branching point in erythro-myeloid differentiation**. *Cell* 2015, **163**:1655–1662.

37. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, *et al.*: **Accounting for technical noise in single-cell RNA-seq experiments**. *Nat Methods* 2013, **10**:1093–1095.

38. Grün D, Kester L, van Oudenaarden A: **Validation of noise models for single-cell transcriptomics**. *Nat Methods* 2014, **11**: 637–640.

39. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL: **The dynamics and regulators of cell fate decisions are revealed by pseudo-temporal ordering of single cells**. *Nat Biotechnol* 2014, **32**: 381–386.

40. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, Enikolopov G, Nauen DW, Christian KM, Ming G, *et al.*: **Single-cell RNA-seq with Waterfall reveals molecular cascades underlying adult neurogenesis**. *Cell Stem Cell* 2015, **17**:360–372.

41. Ji Z, Ji H: **TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis**. *Nucleic Acids Res* 2016, **44**. e117–e117.

42. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C: **Reversed graph embedding resolves complex single-cell trajectories**. *Nat Methods* 2017, **14**:979–982.

43. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S: **Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics**. *BMC Genom* 2018, **19**:477.

44. Grün D, Muraro MJ, Boisset J-C, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, *et al.*: **De novo prediction of stem cell identity using single-cell transcriptome data**. *Cell Stem Cell* 2016, **19**: 266–277.

45. Chen J, Schlitzer A, Chakarov S, Ginhoux F, Poidinger M: **Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development**. *Nat Commun* 2016, **7**:11988.

46. Chlis NK, Wolf FA, Theis FJ: **Model-based branching point detection in single-cell data by K-branches clustering**. *Bioinformatics* 2017, **33**:3211–3219.

47. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D: **Wishbone identifies bifurcating developmental trajectories from single-cell data**. *Nat Biotechnol* 2016, **34**:637–645.

48. Welch JD, Hartemink AJ, Prins JF: **SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data**. *Genome Biol* 2016, **17**:106.

49. Herring CA, Banerjee A, McKinley ET, Simmons AJ, Ping J, Roland JT, Franklin JL, Liu Q, Gerdes MJ, Coffey RJ, *et al.*: **Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut**. *Cell Syst* 2018, **6**. 37–51.e9.

50. Haghverdi L, Buettner F, Theis FJ: **Diffusion maps for high-dimensional single-cell analysis of differentiation data**. *Bioinformatics* 2015, **31**:2989–2998.

51. • Haghverdi L, Büttner M, Wolf FA, Buettner F, Theis FJ: **Diffusion pseudotime robustly reconstructs lineage branching**. *Nat Methods* 2016, **13**:845–848.

Diffusion pseudotime is introduced for pseudo-temporal ordering of cells based on diffusion-like random walks. This method identifies branching events, meta-stable states, and differentiation endpoints. It is highly robust and scales up to tens of thousands of cells.

52. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, Buettner F, Macaulay IC, Jawaid W, Diamanti E, *et al.*: **Decoding the regulatory network of early blood development from single-cell gene expression measurements**. *Nat Biotechnol* 2015, **33**:269–276.

53. • Van Dijk D, Sharma R, Nainys J, Wolf G, Krishnaswamy S, Pe'er Correspondence D, Gene GA: **Recovering gene interactions from single-cell data using data diffusion in brief population analysis archetypal analysis gene interactions**. *Cell* 2018, **174**:1–14.

MAGIC (Markov affinity-based graph imputation of cells) shares information across similar cells via data diffusion to fill in missing transcript for elucidating gene–gene relationships without perturbations.

54. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R: **Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development**. *Nat Biotechnol* 2017, **35**:551–560.

55. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O: **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells**. *Nat Biotechnol* 2015, **33**:155–160.

56. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y: **SLICE: determining cell differentiation and lineage based on single cell entropy**. *Nucleic Acids Res* 2016, **45**. gkw1278.

57. Teschendorff AE, Enver T: **Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome**. *Nat Commun* 2017, **8**:15599.

58. •• Manno G La, Soldatov R, Hochgerner H, Zeisel A, Petukhov V, Kastriti M, Lonnerberg P, Furlan A, Fan J, Liu Z, *et al.*: **RNA velocity in single cells**. *bioRxiv* 2017, https://doi.org/10.1101/206052.

RNA velocity is derived from the ratio of spliced versus unspliced transcripts quantified by single-cell RNA-seq to determine the stage of a transcript within its lifecycle, which is informative on the cell state. This information is used to infer directionality on differentiation trajectories and permits the prediction of starting and end points on lineage trees.

59. Raj A, Rifkin SA, Andersen E, van Oudenaarden A: **Variability in gene expression underlies incomplete penetrance**. *Nature* 2010, **463**:913–918.

60. Chubb JR, Trcek T, Shenoy SM, Singer RH: **Transcriptional pulsing of a developmental gene**. *Curr Biol* 2006, **16**: 1018–1025.

61. Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F: **Mammalian genes are transcribed with widely different bursting kinetics**. *Science* 2011, **332**:472–474.

62. • Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, Soon MSF, Fogg LG, Nair AS, Liligeto U, *et al.*: **Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria**. *Sci Immunol* 2017, **2**. eaal2192.

Latent variable modeling of single-cell RNA-seq data by overlapping mixtures of Gaussian processes provides a high-resolution picture of the cell fate bifurcation into TH1 and TFH sub-types upon Plasmodium infection, which is causing malaria.

63. • Velten L, Haas SF, Raffel S, Blaszkiewicz S, Islam S, Hennig BP, Hirche C, Lutz C, Buss EC, Nowak D, *et al.*: **Human haematopoietic stem cell lineage commitment is a continuous process**. *Nat Cell Biol* 2017, **19**:271–281.

Using a regression analysis of single-cell RNA-seq data, lineage priming of multipotent progenitors is learned from more mature stages of differentiation. The analysis reveals a low level of lineage priming of human hematopoietic stem and progenitor cells.

64. • Herman JS, Sagar, Grün D: **FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data**. *Nat Methods* 2018, **15**:379–386.

FateID is a semi-supervised method for the inference of cell fate bias in multipotent progenitors from single-cell RNA-seq data. Heterogeneity within the murine lymphoid progenitor compartment is revealed and segregation into domains of predominant fate bias towards a particular lineage is observed for the multipotent progenitor compartment.

65. •• Briggs JA, Weinreb C, Wagner DE, Megason S, Peshkin L, Kirschner MW, Klein AM: **The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution**. *Science* 2018, **360**. eaar5780.

High-resolution single-cell RNA-seq time course analysis of Xenopus development, revealing early cell fate priming and lineage-defining regulators. Together with [66] this study provides an evolutionary analysis of vertebrate development at single-cell resolution.

66. •• Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM: **Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo**. *Science* 2018, **360**: 981–987.
Time course analysis of the first day of zebrafish development by single-cell RNA-seq in combination with a transposon-based barcoding approach. A graph-based approach is utilized for successful lineage tree reconstruction.

67. •• Farrell JA, Wang Y, Riesenfeld SJ, Shekhar K, Regev A, Schier AF: **Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis**. *Science* 2018, **360**. eaar3131.
Large-scale single-cell RNA-seq time course analysis of early zebrafish development, introcuding the URD method for lineage reconstruction. This study highlight canalization and plasticity of early fate specification.

68. Mao Q, Wang L, Goodison S, Sun Y: **Dimensionality reduction via graph structure learning**. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining – KDD '15*. ACM Press; 2015:765–774.