Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Systems Biology

Learning regulatory models for cell development from single cell transcriptomic data

Ann C. Babtie[1], Thalia E. Chan[1] and Michael P. H. Stumpf[1,2]

Abstract

Single cell transcriptomic data allow us to probe the transcriptional changes occurring during cell development in unprecedented detail. These complex datasets are driving the development of new computational and statistical tools that are revolutionizing our understanding of differentiation processes. Many clustering and dimensionality reduction methods exist to aid visualization and exploration of structure in these datasets. Increasingly, pseudotemporal ordering and network inference algorithms are emerging that aim to elucidate the regulatory mechanisms that drive and control changes in gene expression state. Combining multiple analytical approaches enables us to make best use of the complementary information they offer, and provides the detail needed to infer mathematical models describing the structure and dynamics of gene regulatory networks.

Addresses

[1] Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK
[2] MRC London Institute of Medical Sciences, Hammersmith Campus, Imperial College London, London W12 0NN, UK

Corresponding author: Babtie, Ann C (a.babtie@imperial.ac.uk)

Current Opinion in Systems Biology 2017, **5**:72–81

This review comes from a themed issue on **Development and differentiation (2017)**

Edited by **Philip Greulich** and **Ben MacArthur**

For a complete overview see the Issue and the Editorial

Available online 5 August 2017

http://dx.doi.org/10.1016/j.coisb.2017.07.013

Introduction

Advances in genomics technologies now enable high-throughput screening of many biomolecules within single cells. Perhaps the best developed techniques are those allowing measurement of gene expression levels in up to thousands of individual cells, enabling us to explore how cellular transcriptional states vary during e.g. developmental processes. These techniques are widely applied in stem cell and developmental biology research, where they are revolutionizing our understanding of the changes that occur as cells progress through development. Single cell data have, for example, enabled identification of rare cell types, provided insight into the subpopulation structure of developing cell populations, and challenged existing models of developmental hierarchies (see Refs. [1–3] for recent reviews).

Precisely controlled spatial and temporal patterns of gene expression accompany the differentiation of cells from multipotent progenitor states towards specialized cell lineages, as a multicellular organism develops from a single fertilized egg. Single cell transcriptomic data — cross-sectional data comprising 'snapshots' of mRNA expression levels that are generated using single cell RNA (scRNA) sequencing or quantitative PCR — provide unprecedented insights into individual cells and how they respond to environmental, developmental and physiological cues. However, analysing these data poses new computational and statistical challenges due to the technical and biological heterogeneity that characterise such data. Numerous methods — specifically tailored for single cel data — have been developed for pre-processing (including normalizing) and visualizing these high-dimensional data, characterising cell types and subpopulation structure, and detecting differentially expressed genes (reviewed in Refs. [4–8]). Here, we focus on recent computational methods for analysing scRNA data that address the challenge of learning temporal dynamics from static measurements, allowing us to examine potential functional interactions between genes, and move towards developing mathematical models describing the gene regulatory mechanisms controlling cell development and differentiation.

Gene regulatory networks and models of cell development

Complex gene regulatory networks (GRNs) comprising activating and repressing interactions between transcription factors and their targets control the transcriptional state of cells. In a dynamical systems framework, such networks are viewed as regulating the probability of cells occupying different gene expression states. Stable 'attractor' states are associated with discrete cell types observed experimentally, and the potential landscape determines probable transition routes between states [9–11]. The analogy of landscapes that dictate cellular developmental pathways has long been used as a conceptual framework for describing differentiation processes [12].

Single cell experiments effectively provide snapshots of these notional landscapes, enabling us to quantitatively

assess the distributions in gene expression space of cell populations undergoing differentiation. While the landscape analogy implies smooth, continuous transitions between stable states, single cell data allow much more detailed examination of the moments when cells commit to certain lineages and have led to proposals that we should refine our descriptions of these key bifurcation or cell fate decision events. Rather than smooth transitions, these may be discontinuous, stochastic transition events driven by the dynamic nature of the landscape (which changes in response to GRN activity and extracellular signals) and reflected in the observed increased transcriptional heterogeneity at these points [11,13–15]. There is great interest in analysing single cell data to understand the transcriptional changes that occur as cells differentiate and the genes and regulatory mechanisms controlling these processes [1,5,6,8,16].

Defining cell types and subpopulation structure

Single cell transcriptomic data are high-dimensional comprising information on up to thousands of genes and cells depending on the experimental protocol, so most analyses start by visualizing and exploring structure in these data using clustering and dimensionality reduction algorithms. The assumptions and inherent biases of different algorithms — e.g. the relative emphasis on preserving local versus global structure when reducing dimensions, or how to define similarities between cells or genes — affect our conclusions about structure and patterns in these data [6,7,17,18]; choices made during the preliminary steps will therefore influence any subsequent downstream analyses relying on these results (Figure 1).

When studying cell differentiation, detecting genes showing variable expression across different developmental stages is a first step towards identifying putative GRN components. Clustering genes by expression profile similarity (or bi-clustering by both gene and cell similarities) can identify gene modules showing coordinated expression changes associated with developmental progression (e.g. Refs. [19,20]). Several approaches to detect differential expression of genes between cell subsets are specifically tailored to deal with the complexities of single cell data, e.g. by accounting for the prevalence of 'dropouts' (where gene expression is undetected in a given cell due to low mRNA capture rates) [21–25].

These approaches are helpful for many downstream analyses — particularly by revealing any subpopulation structure — but may only provide cursory mechanistic insights. Here, we focus on efforts to gain more insight into the precise dynamics and regulation of gene expression changes.

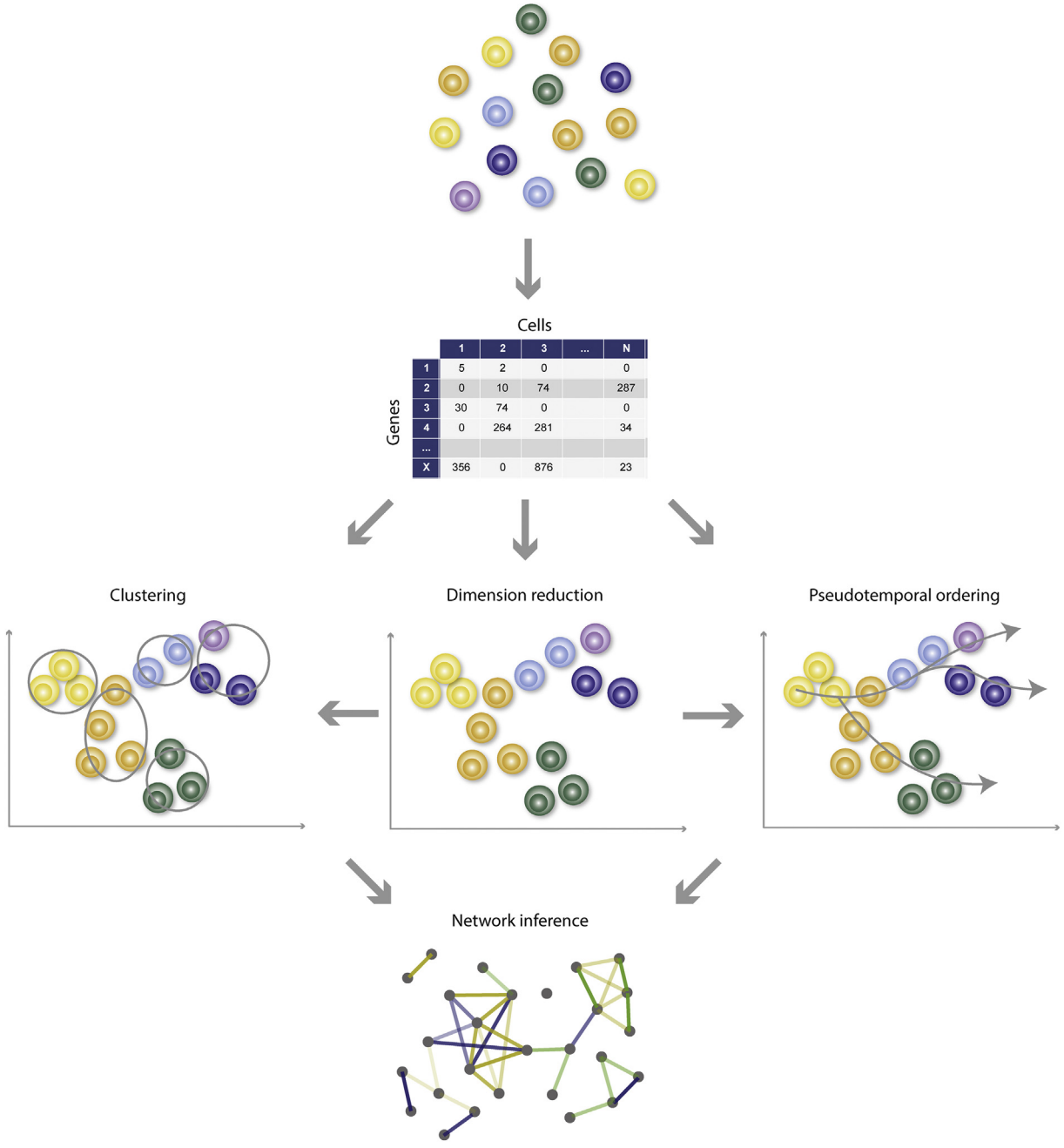
Inferring temporal progression through development

Longitudinal scRNA data cannot be collected straightforwardly (since cells are lysed for mRNA quantification) but, under certain assumptions, we can use samples of populations of cells undergoing differentiation to reconstruct probable developmental trajectories. The asynchronous behaviour of cells means that even when we trigger differentiation artificially, experimental sampling times will not necessarily reflect the extent of a cell's developmental progression. Instead, by assuming that transcriptional state reflects developmental stage, and that cells follow common trajectories, we can order cells according to similarities in expression state and infer their relative progression; individual cells are assigned a 'pseudotime' depending on their position in this inferred order. Many pseudotemporal ordering algorithms exist, with different capabilities in terms of the types of trajectories they can infer and requirements for prior knowledge (see Table 1) [26–39].

This re-ordering provides candidate temporal trajectories for each gene that, if correct, give a clearer view of developmental gene expression dynamics. This can show the relative timing of expression changes, reveal sets of genes with coordinated dynamics, identify gene expression signatures of specific developmental lineages, or indicate cellular processes that change systematically during development [26–28,30–32,40–42]. We can also study the transcriptional changes associated with bifurcation or cell fate decision events, enabling us to identify putative regulators of specific transitions from gene expression changes that accompany or immediately precede such events [26,28,30,40]. Determining the number and location of these bifurcation events remains a challenging problem [26,28–30,32,36,37]. Comparing gene expression dynamics may of course indicate the directionality of any regulatory interactions and, in a few cases, pseudotemporal trajectories have been used to infer dynamical models of GRNs [43,44].

Pseudotemporal ordering algorithms have generated much interest and provided insight into developmental processes. However, as with all inference and modelling, we should bear in mind the limitations and underlying assumptions: many algorithms rely on initial clustering and/or dimensionality reduction steps, and some require or can incorporate prior knowledge (e.g. number of lineages or experimental sampling times); these methods and choices will influence our results. Comparisons demonstrate that inferred trajectories can differ substantially between algorithms due to e.g. different susceptibilities to noise and data sparsity [26–28,30]. While such algorithm-dependent influences can be assessed through quantitative comparisons, we should always consider whether the

Figure 1



Common workflows for analysing single cell data. A wide array of computational methods are available for analysing single cell transcriptomic data, with the complementary aims of characterising cell subpopulations and their gene expression patterns, identifying genetic drivers of transition events and inferring (mechanistic) models of gene interactions. Following pre-processing steps such as eliminating poor quality data, normalizing, and correcting technical errors (not depicted), dimension reduction and clustering help characterise cell subtypes, and may provide the initial steps for pseudotemporal ordering. Using clustering or pseudotemporal ordering, it is then possible to identify genes that are differentially expressed in different states, or more ambitiously, infer a gene regulatory network (the arrows indicate common – but not all possible – analysis workflows).

assumptions are appropriate for the particular system under study. While initial analyses such as dimensionality reduction often appear to depict cells undergoing smooth, continuous changes in transcriptional state, the homogeneity, parsimony and irreversibility of developmental transitions are in fact assumptions that may or may not accurately reflect the true biological processes. In addition there may be a host of other factors — apart from development — that affect changing gene expression patterns [11,16,28].

Table 1

Overview of single cell pseudotemporal ordering algorithms. Summary of algorithms developed to infer lineage hierarchies and temporal ordering of cells from single cell transcriptomic data. A brief overview of each method and key features is provided, see the original references for more detail. N.B. This is not a comprehensive list of all algorithms, but aims to provide an overview of the most commonly used and recent algorithms, and illustrate the wide range of methods that have been applied to this type of problem. Abbreviations used in table: k-nearest neighbour graph (k-NNG), Gaussian process (GP), minimal spanning tree (MST), Latent Dirichlet Allocation (LDA).

Algorithm	Method summary	Key features	References
GPfates	First infers pseudotimes (and potentially reduces dimensions) using a GP latent variable model, using experimental capture times as prior information. Then identifies bifurcation using a nonparametric temporal mixture model.	Uses experimental cell capture times; provides uncertainty estimates; currently limited to a single bifurcation event.	Lönnberg 2017 [26]
Monocle 2	Selects genes showing differential expression between cell clusters. Uses reversed graph embedding to learn a mapping between high and low-dimensional space, and a spanning tree connecting cell clusters in low-dimensions.	Unsupervised method to select informative genes and branching structure; scalable for large datasets.	Qiu 2017 [27]
scTDA	Uses a topological data analysis (TDA) algorithm to construct low-dimensional network representation, where nodes represent cell clusters, and edges connect nodes with cells in common.	Scalable; infers multi-branching lineages; does not enforce common differentiation trajectories, allows more complex topological structures.	Rizvi 2017 [28]
Slingshot	Constructs MST between cell clusters to identify number and location of branch points. Infers pseudotemporal trajectories by fitting principal curves to each lineage (option for user to define lineage endpoints).	Infers multi-lineage structures; compatible with any upstream dimensionality reduction/clustering methods.	Street 2017 [29]
Mpath	Hierarchical clustering identifies 'landmark' cell clusters representing different states. Constructs a network between landmarks, with edges weighted by the number of transitioning cells.	Infers multi-branching lineages; relies on observing transitioning cells (i.e. assumes continuum).	Chen 2016 [30]
CellTree	Bayesian method to infer branching hierarchies and gene sets associated with developmental stages. Based on LDA – a model that assumes a mixture of unobserved 'topics' (gene sets) can explain the observed cell states.	Requires user-defined number of topics (provides heuristic guide); directly links gene expression patterns to cellular hierarchy.	duVerle 2016 [31]
Diffusion pseudotime	Introduces a distance metric describing transition probabilities between any cell pair by considering random walks of all lengths between cells in gene expression space. Automatically detects branching points.	Scalable to large datasets; identifies branch points and metastable cell states; requires user-defined number of branches.	Haghverdi 2016 [32]
TSCAN	Averages expression values in each cell for genes with similar expression profiles. Clusters cells in reduced dimensions by fitting mixture of multivariate normal distributions, and constructs MST linking clusters.	Uses clustering to improve robustness; (optionally) uses prior knowledge (e.g. no. of clusters or branches); allows multi-branching lineages.	Ji 2016 [33]
SCOUP	Initial temporal ordering based on MST in reduced dimensions. Refines ordering by optimising a mixture Ornstein-Uhlenbeck (OU) process model (models variables moving towards an attractor state with Brownian motion).	Currently applicable to linear or bifurcating trajectories; identifies putative regulatory interactions through correlation analysis.	Matsumoto 2016 [34]
deLorean	Uses GPs to model gene expression profiles and infer pseudotimes within a Bayesian inference framework, using experimental cell capture times as prior knowledge.	Uses experimental cell capture times; only infers linear trajectories; not scalable; provides uncertainty estimates.	Reid & Wernisch 2016 [35]
Wishbone	Reduces dimensions using diffusion maps and constructs a k-NNG between cells. Initial cell ordering based on shortest-paths. Refines trajectory and identifies branching structure using randomly selected 'waypoint' cells.	Limited to single bifurcation; relies on gene ontology annotations to select informative diffusion components; scalable to large datasets.	Setty 2016 [36]
SLICER	Selects genes varying systematically across cell population. Constructs k-NNG between cells in reduced dimensions. Infers pseudotimes and branching structure using geodesic distances and entropy respectively.	Unsupervised method to select informative genes and branching structure; allows multiple differentiation routes between two points.	Welch 2016 [37]
Waterfall	Reduces dimensions before clustering cells using a k-means algorithm. Constructs MST to link cluster centres, and assigns cell pseudotimes by projection onto trajectory.	Uses cell clustering to improve robustness; linear trajectories only.	Shin 2015 [38]
SCUBA	Fits a smooth curve in reduced dimensions using principal curve analysis. Divides cells into temporal clusters, before iteratively clustering cells at each time and mapping between different times to infer hierarchical structure.	Automatically infers trajectory endpoints and number of lineages; can detect multiple branching events.	Marco 2014 [39]

Developing gene regulatory network models

While approaches used for network inference from bulk transcriptomic data have been directly applied to single-cell data, newer dedicated methods have also been developed. Despite the challenges posed by technical noise, single cell data offer several potential advantages for inferring regulatory relationships — larger sample sizes; inherent biological heterogeneity provides the variability necessary to infer relationships without needing perturbation experiments; and the ability to visualise subpopulation structure avoids potential confounding effects from analysing mixed populations of cell types [3,8,9,16].

A common (and simple) approach is to calculate pairwise correlations between gene expression states, generating an undirected network indicating statistical relationships between genes that are interpreted as putative (co-) regulatory interactions. While this method has successfully identified regulatory relationships from single cell developmental data (e.g. Refs. [45–48]), it only detects linear (or monotonic) relationships and thus may overlook many biological interactions. Information theory provides alternative measures, e.g. mutual information, that can capture more complex non-linear statistical dependencies between variables and are widely applied for GRN inference from bulk data [49]. Calculating information measures typically requires estimating joint probability distributions from experimental data so these methods benefit hugely from the larger sample sizes afforded by new single cell technologies, particularly when using measures that quantify relationships between three or more variables [49,50]. Both pairwise and higher-order information theoretical measures have been successfully integrated into network inference algorithms to infer regulatory interactions from single cell data [51,52]. Without making further assumptions (e.g. temporal ordering) or integrating other data (e.g. transcription factor binding) these statistical models (whether correlation or information theoretic based) do not indicate the directionality of interactions.

Other methods aim to infer mathematical models of GRNs that represent the mechanistic nature and directionality of interactions, and allow system dynamics to be studied using Boolean or ordinary differential equation (ODE) models (see Table 2) [43,44,51–57]. Boolean models rely on discretized data which may increase their robustness to noise and provide benefits for computational efficiency, and, unlike ODE models, they make fewer assumptions about the nature of interactions and avoid the need to infer many parameters. However, Boolean discretization inherently results in some data loss and may be overly simplistic, and the method chosen to learn model structure may require certain dataset features — e.g. one of the earliest algorithms successfully applied to single cell data constructs a state-transition graph from binarised cell expression states and thus requires large numbers (e.g. thousands) of cells [56]. Several ODE-based models have been inferred from single cell data using differing assumptions about the nature of relationships, but in all cases relying on temporal gene expression data — either inferred pseudotemporal orderings or experimental sampling times — and the assumptions associated with these [43,44,53,57]. Temporal assumptions and assumptions of irreversibility in particular have strong implications on ODE-based analyses, because they directly inform the directionality of inferred relationships. So far, mechanistic models inferred from single-cell data tend to be limited to smaller GRNs (comprising tens of genes) than the statistical models outlined previously (which can easily scale to hundreds of genes), but they do provide the capability to simulate GRN dynamics and thus allow predictions of system behaviour under different scenarios.

All these statistical or mechanistic models of GRNs provide a set of putative functional interactions between genes (or modules of co-varying genes). While we of course aim to develop methods that provide the most reliable inference results, these networks should be viewed as hypotheses about the underlying regulatory mechanisms. These can guide further investigation and experiments, and allow us to test our current understanding but, like any models, they should be continually refined and improved as new information emerges. These models rely on some key assumptions: firstly, that mRNA expression levels are indicative of the corresponding protein levels (thus ignoring potential post-transcriptional influences), but also that differentiation is the dominant process driving the observed gene expression dynamics. Our choice of cells and genes to include in our analyses is critical to ensure this latter assumption is appropriate.

There are of course technical limits to what we can learn about a biological system by experimentally observing the system state. Some interactions will not be inferable, e.g. if they do not drive observable expression changes, or these changes are too transient to detect associations between the corresponding genes. It can be difficult to distinguish certain regulatory topologies, such as indirect versus direct regulation, depending on the inference method. Finally, while GRNs are sometimes defined as the complete collection of possible gene regulatory interactions within a given cell, we can of course only hope to infer the subset of interactions active under our specific experimental conditions. We expect to infer distinct networks using different cell subsets, depending on the variability present in the selected cell population, and thus should carefully choose which data to analyse [45,51,58].

Table 2

Overview of single cell network/model inference algorithms. Summary of the different types of statistical and modelling approaches that have recently been applied to single cell transcriptomic data in order to infer mathematical models of putative gene regulatory mechanisms. An overview of each class of method is given, with a few examples described in greater detail; some of the relative advantages and disadvantages of these approaches are noted. Abbreviations used in table: transcription factor (TF), ordinary differential equation (ODE).

Model	Method summary	Notes
Correlation/relevance networks		
Overview	Undirected edges connect pairs of genes exhibiting co-ordinated expression. Gene pairs are ranked by Pearson (or Spearman) correlation; positive or negative values indicate activation or repression respectively.	Simple to interpret and fast to calculate. Limited to detecting linear (or monotonic) relationships.
Information theory		
Overview	Undirected edges connect pairs of genes showing statistically dependent expression profiles. Dependence quantified using information theoretic measures, often mutual information or a more complex variant.	Relatively computationally efficient. Can detect non-linear dependencies thus avoids assumptions about the nature of interactions. Often requires discretising data, which may reduce noise, but may lose information. Perform best with large sample sizes.
PIDC [51]	Exploits large sample sizes to estimate a three-variable information measure; aims to distinguish direct interactions by decomposing mutual information between a pair of genes into contributions that are unique to that pair or shared with other genes.	Avoids several common assumptions regarding state space and temporal progression. An information-based approach that is applicable to networks of hundreds of genes.
MAGIC [52]	Imputes missing expression values using a diffusion process through cells, then estimates conditional densities for pairs of genes using a k-nearest neighbour approach. Calculates mutual information from conditional densities to score putative interactions.	Alleviates influence of dropouts and aims to recover information from sparsely populated regions of expression space. Assumes data inherently low-dimensional, and that signal overcomes noise in sparse regions.
ODE model inference		
Overview	ODEs represent the regulatory interactions controlling the expression of each gene; algorithms aim to infer parameters for these ODEs in a network where directed edges connect transcription factors to their targets.	Infer detailed mechanistic networks capturing direction and strength of regulation; provide information on system dynamics. May be computationally complex. Assume specific mathematical forms for interactions, and often rely on inferred temporal trajectories.
Jang et al., 2017 [53]	First identifies discrete cell states, transition lineages, and key gene modules. Creates a step-function based model of gene module interactions; estimates parameter probability distributions using linear programming with observed cell states determining constraints.	Coarse-resolution network connects modules of genes with similar expression profiles. Uses binarised data which may reduce noise or may lose information.
SCODE [44]	Infers linear ODE-model of regulatory interactions between TFs from pseudotemporal trajectories. Develops an efficient parameter estimation algorithm that relies on linear regression and a lower-dimensional transformation of the data.	Assumes linear relationships. Computationally efficient, compared to similar approaches.
Ocone et al., 2015 [43]	Modular approach combines dimension reduction, pseudotemporal ordering and network inference. Infers initial coarse network using random forest and correlation methods, then infers a Hill-function ODE model using Bayesian model selection and parameter inference.	Modularity allows substitution of different algorithms. Multiple steps introduce multiple sets of assumptions. Limited to small models.
Boolean model inference		
Overview	Binarised gene expression levels are governed by update functions that describe the combinatorial action of regulating genes in terms of Boolean logic rules. A state transition graph comprises the possible cell states arising from the governing Boolean network.	Infer detailed mechanistic networks capturing direction of interactions and combinatorial regulation; provide information on system dynamics. Binarising data may reduce sensitivity to noise or may lose information.
BTR [54]	Infers an asynchronous Boolean model, by iterative optimisation starting from an initial model (random or prior knowledge) using a swarming hill climbing strategy. Scores proposed models by comparing the experimental data and model state spaces.	Avoids making assumptions about temporal progression. Search strategy optimised for local searches, so performs best when informed by prior knowledge of regulatory interactions.
SingCellNet [55]	Infers an asynchronous probabilistic Boolean network. Uses genetic algorithms to optimise network topology and probabilities of Boolean update rules, based on consistency with known cell lineage hierarchies; update rules based on prior knowledge.	Limited to small networks. Assumes knowledge of cell lineage hierarchy and putative update rules.

(*continued on next page*)

Table 2 (*continued*)

Model	Method summary	Notes
SCNS [56]	Creates state transition graph from observed binarised transcriptional states of cells; initial/final cell states defined using prior knowledge. Infers Boolean update functions for each gene that are consistent with the observed transitions.	Assumes observed cell states represent state space of Boolean network. Requires large datasets (thousands of cells) to construct a connected state transition graph.
Linear regression		
SINCERITIES [57]	Infers directed network of TF-target gene interactions using linear regression. Assumes change in TF expression distribution causes proportional change in target gene distribution at subsequent time point. Distinguishes activation and repression using partial correlation.	Relies on experimental sampling times (i.e. cross-sectional time series data). Assumes linear relationships between changes at consecutive times.

Combining computational analyses

In general, we should use several of the approaches outlined above to gain insights into the regulatory mechanisms driving cell differentiation: they provide complementary information, and using them in combination can help redress some of the limitations and biases inherent to each method.

A preliminary descriptive analysis, using e.g. clustering, dimensionality reduction, and bioinformatics annotation of differentially expressed genes, can provide important information about any subpopulation structure within the data. This helps us to choose cell subsets to analyse that will be most informative about the biological process of interest — e.g. those where we believe that differentiation is the dominant driver of transcriptional variation. When detecting statistical relationships between genes for GRN inference, we might select cells undergoing a specific developmental transition as some statistical dependencies may be masked within more complex datasets comprising cells in multiple developmental lineages. Clustering or pseudotemporal ordering can help to identify subsets of non-responsive cells to exclude from subsequent analyses. Cells are likely to be simultaneously affected by multiple biological processes, so we may aim to account for any potential confounding factors, e.g. large-scale transcriptional changes associated with cell cycle stage may mask the variability linked to differentiation [59].

Although scRNA-sequencing provides information about thousands of genes, analysis is greatly aided by careful pre-processing and biologically guided selection of relevant genes. Basic filtering metrics allow us to remove genes expressed at very low levels (and therefore dominated by technical noise) or those showing little variation in expression. Clustering and pseudotemporal ordering can help us select genes associated with the biological process of interest, or identify gene modules demonstrating similar dynamics. Removing non-informative genes (or cells) aids all downstream analyses, but particularly those that seek to develop mechanistic — ODE and Boolean network — models.

Ideally, we should also consider the limitations and assumptions of each of the methods we include in our analysis, and aim to explore (at least to some extent) how our algorithm choices influence our conclusions. Many of the algorithms applied to single cell data do not allow us to quantify uncertainties in the outputs of earlier stages of analyses (e.g. clustering, dimensionality reduction, or inferred temporal orderings), but conclusions from any downstream analyses (e.g. inferred regulatory networks and models) are necessarily conditional on the accuracy of these initial results. In the absence of reliable methods to propagate uncertainties through the different stages of analysis, perhaps a pragmatic solution is to verify whether our conclusions are robust to some variation in the methods selected during earlier steps — e.g. we could explore how much inferred temporal orderings or network models vary when we subsample our data or use different dimensionality reduction methods.

Conclusions

The biological questions we seek to address using scRNA data are complex. We should carefully consider how to analyse such data — ideally prior to data collection to ensure suitable experimental design — and make optimal use of them by incorporating multiple analytical approaches. Particularly while these technologies and methods are relatively new, we should explore and compare different analytical frameworks, and continue to elaborate them. Flexible, open-source software implementations are essential to allow such comparisons and ensure algorithms are easy to adapt and integrate with other complementary methods.

To gain a more comprehensive picture of the regulatory mechanisms controlling differentiation, we need to incorporate other sources of information. We can design experiments to test and refine our putative hypotheses, and to verify that conclusions drawn from *in vitro* data

correspond to *in vivo* observations. We should develop ways of integrating other types of data into our analyses, such as incorporating information on transcription factor binding or chromatin accessibility when inferring GRN models. Genomics technologies are being adapted to measure multiple characteristics (e.g. chromatin accessibility and methylation) at single cell level with recent success at quantifying several features within the same cells [2,16]. These datasets will provide much richer information about the underlying biological processes and will demand dedicated computational and statistical methods to combine information from heterogenous data types.

We need to develop effective ways to integrate and compare data generated from independent experiments on similar biological systems, to ensure the conclusions we draw are robust and biologically meaningful. As experimental technologies advance we expect to see improved performance of many methods — e.g. pseudotemporal ordering and network inference methods should benefit from increasing sample sizes and become more robust to noise. Larger datasets will offer more comprehensive sampling of different cell states, providing better resolution of sparsely populated regions of gene expression space (e.g. during rapid state transitions). Finally, existing approaches that require large sample sizes for model inference [56,60] or data imputation [52] will become feasible to apply more widely.

Acknowledgements

ACB gratefully acknowledges support through a BBSRC Future Leaders Fellowship (Grant reference BB/N011597/1). TEC is funded through a BBSRC DTP PhD studentship.

References

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

* of special interest

1. Moignard V, Göttgens B: **Dissecting stem cell differentiation using single cell expression profiling**. *Curr Opin Cell Biol* 2016, **43**:78–86, http://dx.doi.org/10.1016/j.ceb.2016.08.005.

2. Wen L, Tang F: **Single-cell sequencing in stem cell biology**. *Genome Biol* 2016, **17**:1–12, http://dx.doi.org/10.1186/s13059-016-0941-0.

3. Kumar P, Tan Y, Cahan P: **Understanding development and stem cells using single cell-based analyses of gene expression**. *Development* 2017, **144**:17–32, http://dx.doi.org/10.1242/dev.133058.

4. Grün D, van Oudenaarden A: **Design and analysis of single-cell sequencing experiments**. *Cell* 2015, **163**:799–810, http://dx.doi.org/10.1016/j.cell.2015.10.039.

5. Stegle O, Teichmann SA, Marioni JC: **Computational and analytical challenges in single-cell transcriptomics**. *Nat Rev Genet* 2015, **16**:133–145, http://dx.doi.org/10.1038/nrg3833.

6. Woodhouse S, Moignard V, Göttgens B, Fisher J: **Processing, visualising and reconstructing network models from single-cell data**. *Immunol Cell Biol* 2016, **94**:256–265, http://dx.doi.org/10.1038/icb.2015.102.

7. Bacher R, Kendziorski C: **Design and computational analysis of single-cell RNA-sequencing experiments**. *Genome Biol* 2016, **17**:1–14, http://dx.doi.org/10.1186/s13059-016-0927-y.

8. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with single-cell genomics**. *Nat Biotechnol* 2016, **34**:1145–1160, http://dx.doi.org/10.1038/nbt.3711.

9. Trapnell C: **Defining cell types and states with single-cell genomics**. *Genome Res* 2015, **25**:1491–1498, http://dx.doi.org/10.1101/gr.190595.115.

10. Marr C, Zhou JX, Huang S: **Single-cell gene expression profiling and cell state dynamics: collecting data, correlating data points and connecting the dots**. *Curr Opin Biotechnol* 2016, **39**:207–214, http://dx.doi.org/10.1016/j.copbio.2016.04.015.

11. * Moris N, Pina C, Arias AM: **Transition states and cell fate decisions in epigenetic landscapes**. *Nat Rev Genet* 2016, **17**:693–703, http://dx.doi.org/10.1038/nrg.2016.98.
A thorough review of how scRNA data is shaped by the epigenetic landscape, and how it can help us elucidate mechanisms.

12. Waddington CH: **Canalization of development and the inheritance of acquired characters**. *Nature* 1942, **150**:563–565.

13. Rue P, Martinez Arias A: **Cell dynamics and gene expression control in tissue homeostasis and development**. *Mol Syst Biol* 2015, **11**, http://dx.doi.org/10.15252/msb.20145549. 792–792.

14. * Mojtahedi M, Skupin A, Zhou J, Castaño IG, Leong-Quong RYY, Chang H, *et al.*: **Cell fate decision as high-dimensional critical state transition**. *PLoS Biol* 2016, **14**, e2000640, http://dx.doi.org/10.1371/journal.pbio.2000640.
This study highlights how theoretical concepts can be used in the analysis of scRNA data on developmental processes.

15. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, *et al.*: **Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process**. *PLoS Biol* 2016, **14**, e1002585-35, http://dx.doi.org/10.1371/journal.pbio.1002585.

16. Tanay A, Regev A: **Scaling single-cell genomics from phenomenology to mechanism**. *Nature* 2017, **541**:331–338, http://dx.doi.org/10.1038/nature21350.

17. Ronan T, Qi Z, Naegle KM: **Avoiding common pitfalls when clustering biological data**. *Sci Signal* 2016, **9**, http://dx.doi.org/10.1126/scisignal.aad1932. re6–re6.

18. Haghverdi L, Buettner F, Theis FJ: **Diffusion maps for high-dimensional single-cell analysis of differentiation data**. *Bioinformatics* 2015, **31**:2989–2998, http://dx.doi.org/10.1093/bioinformatics/btv325.

19. Olsson A, Venkatasubramanian M, Chaudhri VK, Aronow BJ, Salomonis N, Singh H, *et al.*: **Single-cell analysis of mixed-lineage states leading to a binary cell fate choice**. *Nature* 2016, **537**:698–702, http://dx.doi.org/10.1038/nature19348.

20. Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, *et al.*: **Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq**. *Nature* 2014, **509**:371–375, http://dx.doi.org/10.1038/nature13173.

21. Kharchenko PV, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis**. *Nat Meth* 2014, **11**:740–742, http://dx.doi.org/10.1038/nmeth.2967.

22. Korthauer KD, Chu L-F, Newton MA, Li Y, Thomson J, Stewart R, *et al.*: **A statistical approach for identifying differential distributions in single-cell RNA-seq experiments**. *Genome Biol* 2016:1–15, http://dx.doi.org/10.1186/s13059-016-1077-y.

23. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, *et al.*: **Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis**. *Nat Meth* 2016, **13**:241–244, http://dx.doi.org/10.1038/nmeth.3734.

24. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, *et al.*: **MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data**. *Genome Biol* 2015:1–13, http://dx.doi.org/10.1186/s13059-015-0844-5.

25. Vallejos C: **Beyond comparisons of means: understanding changes in gene expression at the single-cell level**. *Genome Biol* 2016:1–14, http://dx.doi.org/10.1186/s13059-016-0930-3.

26. * Lönnberg T, Svensson V, James KR: **Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria**. *Science* 2017, **2**, http://dx.doi.org/10.1126/sciimmunol.aal2192. eaal2192.
This paper illustrates how a combination of experimental and computational approaches, including development of a new pseudotemporal ordering algorithm, can be used to characterise the transcriptional changes accompanying cell fate decisions and differentiation.

27. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner H, *et al.*: **Reversed graph embedding resolves complex single-cell developmental trajectories**. *bioRxiv* 2017, http://dx.doi.org/10.1101/110668.

28. * Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, *et al.*: **Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development**. *Nat Biotechnol* 2017, **35**:551–560, http://dx.doi.org/10.1038/nbt.3854.
This paper introduces a powerful topology-based computational method for unsupervised pseudotemporal ordering that makes fewer assumptions about the nature of developmental trajectories than many existing algorithms.

29. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, *et al.*: **Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics**. *bioRxiv* 2017:1–21, http://dx.doi.org/10.1101/128843.

30. 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, http://dx.doi.org/10.1038/ncomms11988.

31. duVerle DA, Yotsukura S, Nomura S, Aburatani H, Tsuda K: **CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data**. *BMC Bioinf* 2016, **17**:363, http://dx.doi.org/10.1186/s12859-016-1175-6.

32. Haghverdi L, Buttner M, Wolf FA, Buettner F, Theis FJ: **Diffusion pseudotime robustly reconstructs lineage branching**. *Nat Meth* 2016:1–6, http://dx.doi.org/10.1038/nmeth.3971.

33. Ji Z, Ji H: **TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis**. *Nucleic Acids Res* 2016, **44**, http://dx.doi.org/10.1093/nar/gkw430. e117–e117.

34. Matsumoto H, Kiryu H: **SCOUP: a probabilistic model based on the Ornstein-Uhlenbeck process to analyze single-cell expression data during differentiation**. *BMC Bioinf* 2016, **17**:232, http://dx.doi.org/10.1186/s12859-016-1109-3.

35. Reid JE, Wernisch L: **Pseudotime estimation: deconfounding single cell time series**. *Bioinformatics* 2016, **32**:2973–2980, http://dx.doi.org/10.1093/bioinformatics/btw372.

36. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, *et al.*: **Wishbone identifies bifurcating developmental trajectories from single-cell data**. *Nat Biotechnol* 2016, **34**:637–645, http://dx.doi.org/10.1038/nbt.3569.

37. Welch J, Hartemink A, Prins JF: **SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data**. *Genome Biol* 2016, **17**:106, http://dx.doi.org/10.1186/s13059-016-0975-3.

38. Shin J, Berg DA, Zhu Y, Shin JY, Song J, Bonaguidi MA, *et al.*: **Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis**. *Stem Cell* 2015, **17**:360–372, http://dx.doi.org/10.1016/j.stem.2015.07.013.

39. Marco E, Karp RL, Guo G, Robson P, Hart AH, Trippa L, *et al.*: **Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape**. *Proc Natl Acad Sci* 2014, **111**:E5643–E5650, http://dx.doi.org/10.1073/pnas.1408993111.

40. Cacchiarelli D, Qiu X, Srivatsan S, Ziller M, Overbey E, Grimsby J, *et al.*: **Aligning single-cell developmental and reprogramming trajectories identifies molecular determinants of reprogramming outcome**. *bioRxiv* 2017, http://dx.doi.org/10.1101/122531.

41. Qiu X, Hill A, Packer J, Lin D, Ma Y-A, Trapnell C: **Single-cell mRNA quantification and differential analysis with census**. *Nat Meth* 2017, **14**:309–315, http://dx.doi.org/10.1038/nmeth.4150.

42. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, *et al.*: **The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells**. *Nat Biotechnol* 2014, **32**:381–386, http://dx.doi.org/10.1038/nbt.2859.

43. Ocone A, Haghverdi L, Mueller NS, Theis FJ: **Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data**. *Bioinformatics* 2015, **31**:i89–96, http://dx.doi.org/10.1093/bioinformatics/btv257.

44. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, *et al.*: **SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation**. *Bioinformatics* 2017, **33**:2314–2321, http://dx.doi.org/10.1093/bioinformatics/btx194.

45. Moignard V, Macaulay IC, Swiers G, Buettner F, Schütte J, Calero-Nieto FJ, *et al.*: **Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis**. *Nat Cell Biol* 2013, **15**:363–372, http://dx.doi.org/10.1038/ncb2709.

46. Pina C, Teles J, Fugazza C, May G, Wang D, Guo Y, *et al.*: **Single-cell network analysis identifies DDIT3 as a nodal lineage regulator in hematopoiesis**. *Cell Rep* 2015, **11**:1503–1510, http://dx.doi.org/10.1016/j.celrep.2015.05.016.

47. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J: **Exploiting single-cell expression to characterize co-expression replicability**. *Genome Biol* 2016, **17**:1–19, http://dx.doi.org/10.1186/s13059-016-0964-6.

48. Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, *et al.*: **Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation**. *Cell Stem Cell* 2015, **17**:471–485, http://dx.doi.org/10.1016/j.stem.2015.09.011.

49. Villaverde A, Ross J, Banga J: **Reverse engineering cellular networks with information theoretic methods**. *Cells* 2013, **2**:306–329, http://dx.doi.org/10.3390/cells2020306.

50. Timme N, Alford W, Flecker B, Beggs JM: **Synergy, redundancy, and multivariate information measures: an experimentalist's perspective**. *J Comput Neurosci* 2013, **36**:119–140, http://dx.doi.org/10.1007/s10827-013-0458-4.

51. * Chan TE, Stumpf M, Babtie AC: **Gene regulatory network inference from single-cell data using multivariate information measures**. *Cell Syst* 2017 (in press).
This paper introduces a powerful information-theoretic based network inference algorithm targeted at scRNA data.

52. * van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, *et al.*: **MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data**. *bioRxiv* 2017, http://dx.doi.org/10.1101/111591.
This paper introduces a method for data imputation to overcome the issues of technical noise in scRNA data, and shows this enhances detection of cell clusters, developmental trajectories and gene regulatory interactions.

53. Jang S, Choubey S, Furchtgott L, Zou LN, Doyle A: **Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states**. *eLife* 2017, http://dx.doi.org/10.7554/eLife.20487.001.

54. Lim CY, Wang H, Woodhouse S, Piterman N, Wernisch L, Fisher J, *et al.*: **BTR: training asynchronous Boolean models using single-cell expression data**. *BMC Bioinf* 2016:1–18, http://dx.doi.org/10.1186/s12859-016-1235-y.

55. Chen H, Guo J, Mishra SK, Robson P, Niranjan M, Zheng J: **Single-cell transcriptional analysis to uncover regulatory circuits driving cell fate decisions in early mouse development**. *Bioinformatics* 2015, **31**:1060–1066, http://dx.doi.org/10.1093/bioinformatics/btu777.

56. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, *et al.*: **Decoding the regulatory network of early blood development from single-cell gene expression measurements**. *Nat Biotechnol* 2015, **33**:269–276, http://dx.doi.org/10.1038/nbt.3154.

57. Gao NP, Ud-Dean M, Gunawan R: **SINCERITIES: inferring gene regulatory networks from time-stamped single cell**

transcriptional expression profiles. *bioRxiv* 2016, http://dx.doi.org/10.1101/089110.

58. Stumpf PS, Smith RCG, Lenz M, Schuppert A, Müller F-J, Babtie A, *et al.*: **Stem cell differentiation is a stochastic process with memory**. *Cell Syst* 2017, http://dx.doi.org/10.1016/j.cels.2017.08.009 (in press).

59. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, *et al.*: **Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells**. *Nat Biotechnol* 2015, **33**: 155–160, http://dx.doi.org/10.1038/nbt.3102.

60. Fisher J, Köksal AS, Piterman N, Woodhouse S: **Synthesising executable gene regulatory networks from single-cell gene expression data**. In *Computer aided verification*. Edited by Kroening D, Păsăreanu C, *Lecture notes in computer science*, vol. 9206. Springer; 2015:544–560.