Contents lists available at ScienceDirect

Molecular Aspects of Medicine

journal homepage: www.elsevier.com/locate/mam

Constructing cell lineages from single-cell transcriptomes

Jinmiao Chen*, Laurent Rénia, Florent Ginhoux

*Singapore Immunology Network (SIgN), A*STAR, 8A Biomedical Grove, Immunos Building, Level 4, Singapore 138648, Singapore*

ARTICLE INFO

Article history:
Received 1 June 2017
Received in revised form
23 October 2017
Accepted 25 October 2017
Available online xxx

Keywords:
RNA-Sequencing
Differentiation
Single-cell analysis
Algorithm
Lineage mapping
Progenitor

ABSTRACT

Advances in single-cell RNA-sequencing have helped reveal the previously underappreciated level of cellular heterogeneity present during cellular differentiation. A static snapshot of single-cell transcriptomes provides a good representation of the various stages of differentiation as differentiation is rarely synchronized between cells. Data from numerous single-cell analyses has suggested that cellular differentiation and development can be conceptualized as continuous processes. Consequently, computational algorithms have been developed to infer lineage relationships between cell types and construct developmental trajectories along which cells are re-ordered such that similarity between successive cell pairs is maximized. Here, we compare and contrast the existing computational methods, and illustrate how they may be applied to build mouse myeloid progenitor lineages from massively parallel RNA single-cell sequencing data.

1. Introduction

Cellular differentiation and development is a temporal dynamic process whereby early progenitor cells develop into terminally differentiated cells via multiple transitional stages. In order to fully understand the kinetics of cellular differentiation, the identity of the cell types present at different stages need to be determined. Cells can be extracted at different stages and time points of development for analysis, but cells are rarely perfectly synchronized during differentiation. As such, cells sampled at the same time point are often found in various differentiative states (Trapnell et al., 2014a). Some developmental stages, however, can be identified by the expression of unique cellular markers, such as the various human hematopoietic stem and progenitor cells including *hematopoietic stem cells* ($CD45RA^{-}CD90^{+}CD49f^{+}$), multipotent progenitors ($CD45RA^{-}CD90^{-}CD49f^{-}$), multi-lymphoid progenitors ($CD45RA^{+}CD10^{+}CD7^{-}$), common myeloid progenitors *(*$CD45RA^{-}CD135^{+}CD10^{-}CD7^{-}$*)* and etc" to "*hematopoietic stem cells* (Lin^{-}CD34+CD38−CD45RA−CD90+), multi-potent progenitors (Lin^{-}CD34+CD38−CD45RA−CD90−), multi-lymphoid progenitors (Lin^{-}CD34+CD38−CD45RA+CD10+CD7+), common myeloid progenitors *(* Lin^{-}CD34+CD38+CD45RA−CD10−CD123int*)* and etc where Lin^{-} means lineage-negative (CD3^{-}CD19^{-}CD56^{-}CD14^{-}CD16^{-}CD66b^{-}CD1c^{-}CD303^{-}CD141^{-}) and etc (https://www.ncbi.nlm.nih.gov/pubmed/28650480), and thus these marker-specific cells can be isolated from the entire population. Bulk samples isolated based on known marker expression are also often a mixture of heterogeneous cells (Schlitzer et al., 2015; See et al., 2017). Moreover, a number of studies performed at the single-cell resolution have even suggested that every cell is unique and different from each other (See et al., 2017; Junker and van Oudenaarden, 2014). These findings pose new challenges on the research of cellular differentiation and the underlying regulatory events. Conventional bulk assays that use a time-point or marker-based approach are inadequate at capturing the entire developmental process as they underestimate cell-to-cell variability. Unfortunately, most of the current models for cellular differentiation have been based on bulk assay-generated data and thus need to be revised and complemented by new approaches (Doulatov et al., 2012; Paul et al., 2015).

The advent of single-cell omic-based techniques has provided new approaches to address the challenges in accurately delineating cellular differentiation pathways. Single-cell genomics (Navin et al., 2011; Wang et al., 2014; Behjati et al., 2014; Lodato et al., 2015; Woodworth et al., 2017), epigenomics (Naumova et al., 2013), transcriptomics (Trapnell et al., 2014a; Haghverdi et al., 2016), and proteomics (Setty et al., 2016) have all recently been applied to help reconstruct cell lineages. For instance, single-cell genome sequencing has detected somatic mutations that may be used as naturally occurring lineage markers for lineage tracking (Navin

* Corresponding author.
E-mail address: chen_jinmiao@immunol.a-star.edu.sg (J. Chen).

https://doi.org/10.1016/j.mam.2017.10.004
0098-2997/© 2017 Published by Elsevier Ltd.

et al., 2011; Wang et al., 2014; Behjati et al., 2014; Lodato et al., 2015; Woodworth et al., 2017). Somatic mutations, including retrotransposons, copy-number variants, single-nucleotide variants and microsatellites mark the progeny of the dividing parent cells. Cells bearing these lineage marks can, therefore, be identified and used to reconstruct cell genealogy (Navin et al., 2011; Wang et al., 2014; Behjati et al., 2014; Lodato et al., 2015; Woodworth et al., 2017). Lineage information can also be inferred from mRNA expression analysis by single-cell RNA-sequencing (RNA-seq). Single-cell RNA-seq has enabled an unbiased characterization of heterogeneous cellular states during differentiation (Schlitzer et al., 2015; See et al., 2017). More importantly, computational analysis of single-cell transcriptomic data has suggested that the cellular heterogeneity is not a disordered or chaotic process, as it may seem (Trapnell et al., 2014a; Haghverdi et al., 2016). Cells in heterogeneous transcriptional states can be positioned in temporal order along seemly continuous trajectories, and as such, cellular differentiation can be conceptualized as a continuous process rather than a series of discrete steps (Bendall et al., 2014; Buettner and Theis, 2012). As a result of developmental asynchrony, single cells sampled at one time point will be found in various developmental stages. If a sufficient number of single cells are sampled, one static snapshot has the potential to capture all cellular states along the entire developmental continuum. This continuous concept has opened new opportunities for the development of computational methods to construct developmental trajectories from single-cell transcriptomic data.

2. Computational lineage construction

Innovative computational methods have been recently developed to infer developmental trajectories from single-cell transcriptomic data (Trapnell et al., 2014a; Haghverdi et al., 2016; Setty et al., 2016; Bendall et al., 2014; Buettner and Theis, 2012; Shin et al., 2015; Marco et al., 2014; Xiaojie Qiu et al., 2017; Haghverdi et al., 2015a; Angerer et al., 2016; Welch et al., 2016; Moignard et al., 2015; Ji and Ji, 2016; Giecold et al., 2016; Chen et al., 2016; Gong et al., 2017; Velten et al., 2017; Matsumoto and Kiryu, 2016; Lonnberg et al., 2017; Campbell and Yau, 2017; Furchtgott et al., 2017). The majority of these methods have been built on the assumption that during cellular differentiation, transcriptional changes are gradual and continuous. Based on this assumption, the algorithms calculate distances or dissimilarities between cells and re-order single cells along a continuum such that similar cells are positioned next to each other in a successive fashion. These computational methods involve several steps of analysis including gene selection, dimension reduction, cluster analysis, pseudotime inference, and branch detection which are briefly introduced as follows.

Single-cell transcriptomic data is intrinsically noisy owing to the low amount of starting material (mRNA), hence it is crucial to perform proper gene selection prior to cell lineage construction. Several existing methods utilize supervised gene selection, such as differential expression analysis from bulk or single-cell transcriptomic data. For instance, initial application of Monocle was based on differentially expressed genes (DEG) detected from bulk RNA-seq data (Trapnell et al., 2014b). Further applications using various gene sets have shown that the performance of Monocle is dependent on the genes used (Chen et al., 2016). NBOR (Schlitzer et al., 2015) and Mpath (Chen et al., 2016) algorithms have also utilized gene signatures derived from bulk transcriptomic data of two classical dendritic cell (DC) lineages to infer lineage-commitment of single DC precursor cells in mouse bone marrow. However, supervised gene selection based on DEG analysis relies on prior knowledge as well as data with known cell types or time points. Unsupervised gene selection is therefore adopted as an alternative approach. Most existing unsupervised gene selection methods select genes that are highly variable and meanwhile expressed at an adequate level to provide meaningful signals. For example, Seurat (Macosko et al., 2015a) selects genes for which average expression and dispersion are above user-specified thresholds; Monocle (Trapnell et al., 2014b) fits a curve between average expression and dispersion of genes, and selects genes that are above the curve; SINGuLAR toolset makes use of principal component analysis (PCA) to select genes with the highest PCA loadings. BackSPIN (Zeisel et al., 2015) provides an unsupervised gene selection mechanism based on a bi-clustering algorithm which simultaneously optimizes clustering of genes and cells. However, highly variable genes are representative of cellular heterogeneity but not necessarily associated with the underlying cellular trajectories. Methods including Seurat (Macosko et al., 2015a), Monocle (Trapnell et al., 2014b), SINGuLAR and BackSPIN (Zeisel et al., 2015) select informative genes mainly for identifying distinct clusters of cells, but not the temporal relationship between cells. Intuitively, if a gene is involved in progression along a cellular trajectory, the expression of this gene should change gradually across neighboring cells along the trajectory; otherwise fluctuate in a manner independent of the trajectory. Based on this intuition, SLICER (Welch et al., 2016) method identifies genes likely involved in sequential progression as those that exhibit more gradual variation across k nearest neighbor cells than at global scale. Although these several gene selection approaches have been utilized for cell lineage construction and pseudotime inference, the impact of gene selection and the significance of various gene selection mechanisms remain to be evaluated.

Single-cell RNA-seq measures expression of several thousand genes per cell depending on cell type and sequencing depth. The measured several thousand genes define a high-dimensional space wherein each dimension is represented by the expression of one gene. Individual cells distribute in this space and cell's position is represented by its gene-expression vector. Assuming cellular differentiation is a continuous process and successive cells along the differentiation process have similar gene expression profiles, the distribution of cells in the high-dimensional space is expected to display a continuous structure or pattern. However, human vision is good at recognizing patterns in low dimensions (<=3), but cannot perceive high-dimensional (>3) relationship. Data projection via dimension reduction enables us to visualize high-dimensional data for better understanding the underlying structure. Therefore a number of cell lineage algorithms first project cells from the original high-dimensional space onto a low-dimensional space using dimension reduction approaches. The dimension reduction approaches are able to preserve the high-dimensional proximity relationship between cells in the low dimensions, such that similar cells in the original dimensions are positioned next to each other in the reduced dimensions. Various dimension reduction approaches have been utilized for single-cell data, including principal component analysis (PCA), multi-dimensional scaling (MDS), independent component analysis (ICA), diffusion map, reversed graph embedding (RGE), locally linear embedding (LLE), Gaussian Process Latent Variable Model (GPLVM) and etc. The most frequently used dimension reduction approach; principal component analysis (PCA) projects individual cells into a latent space spanned by principal components which represent the directions of the largest variance. The projection is linear as the coordinates of cells in the low-dimensional latent space are a weighted sum of the coordinates in the original high-dimensional space. PCA applied to single-cell transcriptomic data of stimulated dendritic cells (DCs) showed that DCs spread along a continuum of expression variation in each principal component (Shalek et al., 2014). The continuum