CellPress

Breaking New Ground in the Landscape of Single-Cell Analysis

Kenji Kamimoto[1,2,3] and Samantha A. Morris[1,2,3,*]
[1]Department of Developmental Biology
[2]Department of Genetics
[3]Center of Regenerative Medicine
Washington University School of Medicine in St. Louis. 660 S. Euclid Avenue, Campus Box 8103, St. Louis, MO 63110, USA
*Correspondence: s.morris@wustl.edu
https://doi.org/10.1016/j.cels.2017.12.015

Here, we outline p-Creode, a new algorithm to construct multi-branching cell lineage trajectories from single-cell data. Application of this platform to diverse sources of single-cell data demonstrates its robustness and scalability, while the discovery of a new origin for rare gut tuft cells showcases the utility of p-Creode.

Single-cell technologies are entering the mainstream, having demonstrated their utility to elucidate a diverse range of complex biological phenomena. Single-cell analysis is fast becoming indispensable for the study of biology concerning stem cell and developmental processes, where heterogeneous cell populations undergo dynamic changes, differentiating toward many distinct identities. Amplifying this complexity, the passage of cells along these trajectories is not perfectly synchronous. The resulting diversity in cellular states and identities means that population-based analyses rarely provided the resolution that is necessary to dissect the true biological heterogeneity of a differentiating system despite efforts to enrich for cell populations of interest and experimental designs to enhance temporal resolution. Single-cell analyses provide context and resolution to this intricate cellular choreography, underpinning a new wave of discovery. In this issue of *Cell Systems*, Charles Herring, Ken Lau, and colleagues present a novel algorithm, p-Creode that facilitates the dissection of complex single-cell data collected from diverse sources (Herring et al., 2018). Via an unsupervised approach, p-Creode generates multi-branching trajectories whose robustness can be statistically assessed. The authors demonstrate the efficacy of p-Creode through the identification of a new developmental origin for tuft cells, a rare chemosensory cell that is situated in the gut.

Varied experimental approaches enabling single-cell analysis currently exist, ranging from single-cell RNA sequencing (scRNA-seq) to mass cytometry to multiplex immunofluorescence (MxIF) (de Vargas Roditi and Claassen 2015). These techniques enable the measurement of tens to thousands of variables from each individual cell, yielding what is termed "high-dimensional data." These advances have permitted simultaneous profiling of heterogeneous populations, where high-throughput single-cell technologies now enable transcriptomes from thousands of cells to be captured in parallel (Macosko et al., 2015; Klein et al., 2015). The collection of potentially thousands of measurements from multitudes of cells is accompanied by significant challenges for the analysis and interpretation of these data. For instance, how can this abundance of information be conceptualized? In this respect, much progress has been made to reduce the dimensionality of the data, clustering cells based on feature similarity to enable visualization and identification of cell subpopulations within heterogeneous tissues (Satija et al., 2015).

Surveying cells from developing and differentiating tissues represents a more complicated challenge, involving transitions between discrete fates, introducing potentially ambiguous cell identities that are difficult to define. Most high-dimensional single-cell capture techniques destroy cells, resulting in the capture of static "snapshot" data. As a consequence, invaluable spatial, temporal, and lineage information is lost. Elegant analytical approaches are emerging, though, focusing on recovery of this information to reconstruct differentiation hierarchies from snapshot data. Assuming that cell identity transitions gradually during differentiation, algorithms have been developed to position each profiled cell within a defined trajectory, termed a "pseudo-temporal" order. This approach has indeed enabled lineage progressions to be reconstructed from single-cell data, leveraging the fact that differentiation tends to be asynchronous, resulting in a diverse range of stages and transitions being captured within a single snapshot. One class of methods are based on minimum spanning tree (MST) algorithms, which effectively aim to "join the dots" between similar cells, mapping the longest path through the data to create a pseudo-temporal cell fate trajectory. These approaches can be unstable, though, having a tendency to generate different cell fate topologies from the same input data, in addition to producing misleading branches in the trajectory as a result of overfitting (Giecold et al., 2016).

Early analytical approaches were based on linear methods such as principal component analysis (PCA) and independent components analysis (ICA) (Trapnell et al., 2014). Although successful in some instances, linear methods are generally not well-suited to complex developmental datasets that encompass multiple branch-points toward distinct terminal cell identities. In this respect, non-linear data-embedding approaches have met with greater success for reconstructing multi-branching cell lineage trajectories. Even so, data structure, distribution, and size can impact the accuracy of these non-linear algorithms, producing variable results (Haghverdi et al., 2015). For example, differences in data dimensionality can confound analysis; MxIF and mass cytometry generate relatively

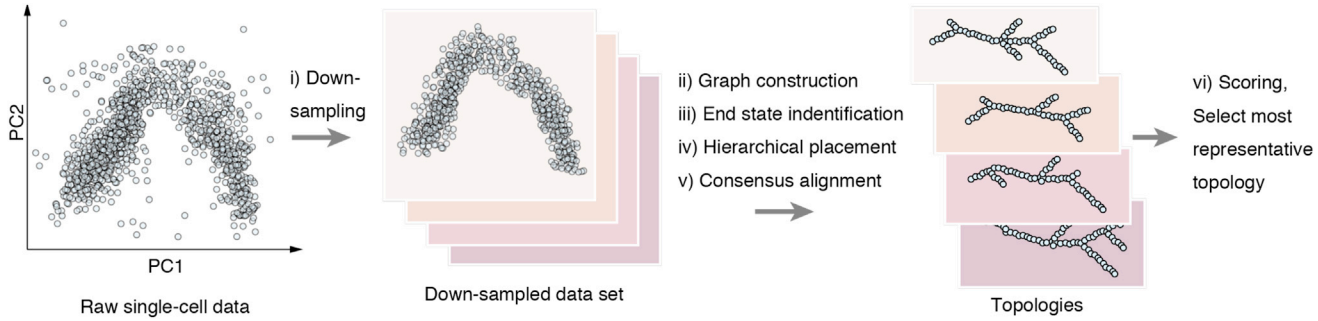

Figure 1. p-Creode Is Based on an Ensemble Method; the Final Result Is Generated by the Consolidation of Sub-outcomes of Many Iterations
(i) Single-cell data are downsampled to generate probabilistic sub-datasets. (ii) The down-sampled dataset is used to construct many preliminary trajectories, using a density-based k-nearest neighbor method that leverages not only the distance between individual cells but also information on the distribution of the population. (iii) Via this approach, transition and terminal differentiation states are identified, from which (iv) trajectories are constructed in an unsupervised manner. (v) p-Creode then deduces the consensus structure to produce a representative topology. (vi) As a result of random downsampling and iteration, this enables the statistical verification of the resulting pseudo-temporal cell trajectories.

low-dimensional data, while scRNA-seq produces high-dimensional data, accompanied by variations in data sparsity that can impact algorithm performance. In addition, depending on the degree of cell heterogeneity within a population, the trajectories of rare cell populations can easily be masked. Given these uncertainties, particularly considering the myriad cell types and platforms employed in profiling, methods to statistically validate algorithm performance are increasingly sought after. In this respect, Wishbone, an algorithm based on supervised random-walk over a cellular network, generates results that are statistically scored, although prior knowledge of the biological system is required for construction of multi-branch temporal order graphs (Setty et al., 2016). At present, many algorithms are unpredictable and highly sensitive to both biological and technological context. Thus, there is currently an unmet need for an unsupervised analytical approach that is versatile and that can be statistically validated to assess the robustness of the trajectories produced.

In an effort to address these current limitations, Herring et al. (2018) have developed a new algorithm, p-Creode. Named for the valleys carved into Waddington's landscape, p-(putative) Creode constructs multi-branched pseudo-time trajectories in an unsupervised manner from a wide variety of single-cell data streams. Similar to above approaches, p-Creode assumes that cell identity transitions gradually during differentiation and can, therefore, be pieced together from snapshot data. Broadly, the algorithm determines the geometric shape of dense single-cell data points in an effort to reveal underlying transition structures within the data, aiming to identify consensus routes representative of the differentiation process. Setting it apart from many current approaches, p-Creode statistically validates the trajectories it generates. Via a multistage process (Figure 1), the strategy of p-Creode is based on an ensemble method; the final result is generated by the consolidation of sub-outcomes of many iterations. First, single-cell data are downsampled to generate probabilistic sub-datasets. Subsequently, the downsampled dataset is used to construct many preliminary trajectories, using a density-based k-nearest neighbor method that leverages not only the distance between individual cells, but also information on the distribution of the population. Via this approach, transition and terminal differentiation states are identified, from which trajectories are constructed in an unsupervised manner. p-Creode then deduces the consensus structure to produce a representative topology. As a result of random downsampling and iteration, this enables the statistical verification of the resulting pseudo-temporal cell trajectories, in addition to reducing the effect of noise on single-cell data. Employing this new metric, the p-Creode score for estimating the differences in graph topology, the authors demonstrated the remarkable reproducibility of their approach relative to an existing pseudo-time analysis method, SPADE (Linderman et al., 2012). This introduces p-Creode as a new ensemble approach and statistical validation metric to ensure robustness and reproducibility in the unsupervised reconstruction of multi-branching trajectories.

Application of p-Creode to single-cell mass spectrometry analysis of hematopoiesis, a hierarchical and well-characterized differentiation process, validated the efficacy of the unsupervised algorithm, generating a multi-branching trajectory in agreement with previous biological knowledge. Following this validation, Herring et al. (2018) examined intestinal differentiation. They analyzed 39,000 small intestinal cells and 17,000 colonic epithelial cells via MxIF, using p-Creode to deduce multi-branching differentiation trajectories of the intestinal epithelium. Focusing on a rare chemosensory cell type, tuft cells, p-Creode was able to reveal key differences of tuft cell biological origins between the small and large intestine.

The next challenge for p-Creode was to test its performance with high-dimension scRNA-seq data. Using publicly available scRNA-seq data of alveolar epithelial cell differentiation obtained by the Fluidigm-C1 system and hematopoiesis data acquired by massively parallel single-cell RNA sequencing (MARS-seq), p-Creode accurately reconstructed trajectories from both datasets that are in agreement with previous reports. In fact, p-Creode identified not only major cell differentiation trajectories, but also sub-branches in hematopoietic pathways, demonstrating its potential to identify both global and local structures simultaneously. Moreover, application of p-Creode to these two datasets highlighted the versatility and robustness of the platform: while it was not developed for application to smaller numbers of single cells, it performed well with alveolar

development data consisting of a relatively small number of cells (< 500). This robust performance with smaller datasets may be due to the effective removal of noise as a result of downsampling and consensus alignment.

Advancing dramatically, single-cell technology is producing an ever-expanding array of multidimensional information and is fast becoming an indispensable source of information for delineating complex developmental processes. However, due to lack of analytical capability for large-scale multidimensional single-cell data, it is not trivial to fully utilize this information. p-Creode generates reproducible pseudo-time trajectory graphs after many rounds of sub-graph construction, consensus alignment, and statistical scoring. This new strategy endows p-Creode with robustness and reproducibility. Perhaps most exciting is the promise of this platform to provide mechanistic insights into differentiation process; the statistical power of p-Creode has the potential to define deterministic, probabilistic, or stochastic events during cell fate determination processes. Beyond development, the utility of p-Creode to assess cell fate trajectories will find wide application in the analysis of disease processes and cell fate engineering. Altogether, this launches p-Creode as a robust, flexible, and scalable platform that promises to form an essential component of the single-cell analytical toolkit.

REFERENCES

de Vargas Roditi, L., and Claassen, M. (2015). Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. Curr. Opin. Biotechnol. *34*, 9–15.

Giecold, G., Marco, E., Garcia, S.P., Trippa, L., and Yuan, G.C. (2016). Robust lineage reconstruction from high-dimensional single-cell data. Nucleic Acids Res. *44*, e122.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics *31*, 2989–2998.

Herring, C.A., Banerjee, A., McKinley, E.T., Simmons, A.J., Ping, J., Roland, J.T., Franklin, J.L., Liu, Q., Gerdes, M.J., Coffey, R.J., et al. (2018). Unsupervised trajectory analysis of single-cell RNA-seq and imaging data reveals alternative tuft cell origins in the gut. Cell Systems *6*, this issue, 37–51.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Linderman, M.D., Bjornson, Z., Simonds, E.F., Qiu, P., Bruggner, R.V., Sheode, K., Meng, T.H., Plevritis, S.K., and Nolan, G.P. (2012). CytoSPADE: high-performance analysis and visualization of high-dimensional cytometry data. Bioinformatics *28*, 2400–2401.

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

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

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

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

Broad Views of Non-alcoholic Fatty Liver Disease

Adil Mardinoglu,[1,2,*] Mathias Uhlen,[1] and Jan Borén[3]
[1]Science for Life Laboratory, KTH - Royal Institute of Technology, Stockholm, Sweden
[2]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[3]Department of Molecular and Clinical Medicine/Wallenberg Laboratory, University of Gothenburg, and Sahlgrenska University Hospital, Gothenburg, Sweden
*Correspondence: adilm@scilifelab.se
https://doi.org/10.1016/j.cels.2018.01.004

Multi-omics multi-tissue data are used to interpret genome-wide association study results from mice to identify key driver genes of non-alcoholic fatty liver disease. Non-alcoholic fatty liver disease (NAFLD) is the accumulation of fat (steatosis) in the liver due to causes other than excessive alcohol consumption. The disease may progress to more severe forms of liver diseases, including non-alcoholic steatohepatitis, cirrhosis, and hepatocellular carcinoma. In this issue of *Cell Systems*, Krishnan et al. (2018) reveal mechanisms underlying NAFLD by generating multi-omics data using liver and adipose tissues obtained from the Hybrid Mouse Diversity Panel, consisting of 113 mouse strains with various degrees of NAFLD. The study identified key driver genes of NAFLD that can be used in the development of efficient treatment strategies and illustrates the potential utility of systematic analysis of multi-layer biological networks.

In parallel with the increase in obesity, the prevalence of NAFLD has increased 5-fold over the past two to three decades, and NAFLD is now considered the most common cause of chronic liver disease in Western countries (Loomba and Sanyal, 2013). Despite the alarming prevalence of NAFLD, no single therapy has to date been approved for treating it, and