OXFORD UNIVERSITY PRESS | Bioinformatics

Consensus Representation Estimation of Lineage Expression (CREoLE) Algorithm for scRNA-seq

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-0518
Category:	Original Paper
Date Submitted by the Author:	05-Apr-2016
Complete List of Authors:	Schau, Geoffrey; Oregon Health and Science University, Department of Medical Informatics and Clinical Epidemiology McWeeney, Shannon; Oregon Health and Science University, Department of Medical Informatics and Clinical Epidemiology Adey, Andrew; Oregon Health and Science University, Department of Molecular and Medical Genetics
Keywords:	HitSeq, Algorithms, Bioinformatics, Signal processing

"MainDocument" — 2016/4/5 — 21:46 — page 1 — #1

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Sequence Analysis

Consensus Representation Estimation of Lineage Expression (CREoLE) Algorithm for scRNA-seq

Geoffrey F. Schau [1,*], Shannon McWeeney [1,2], Andrew Adey [3,4,*]

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR, 97239, USA and
[2]Knight Cancer Institute, Oregon Health and Science University, Portland, OR, 97239, USA and
[3]Department of Molecular and Medical Genetics, Oregon Health and Science University, Portland, OR, 97239, USA and
[4]Knight Cardiovascular Institute, Oregon Health and Science University, Portland, OR, 97239, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Single-cell RNA-sequencing (scRNA-seq) is a promising technology widely used to recapitulate gene expression trends through developmental progression of heterogeneous biological tissue. Although several methods have sought to estimate pseudo-temporal gene expression trends, a number of technical limitations presented by scRNA-seq remain, including high expression variability and drop-out measurements, which complicate accurate trend estimation. Consensus Representation Estimation of Lineage Expression (CREoLE) is an efficient and robust algorithm for automatically detecting an underlying branching lineage structure and estimating smooth developmental trends of gene expression by consensus without the need for model fitting or data filtering.
Results: Applied to synthetic data, CREoLE recapitulates underlying gene expression for each gene and across each lineage with an average Pearson correlation coefficients of 0.983 ± 0.009. The impact of simulated technical noise, drop-out measurements, and cell count reduction are evaluated. Our analysis suggests that Creole is robust to additive noise and smaller initial cell populations (as low as 25% initial population). CREoLE correlates with synthetic expression trends with a mean Pearson's correlation coefficient above 0.9 in all cases. Applied to real data, CREoLE accurately identifies lineage structure and computes high-resolution consensus trends that align closely with published findings.
Availability: CREoLE is free and open-source software available from https://github.com/schaugf/creole
Contact: schau@ohsu.edu, adey@ohsu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single-cell RNA-sequencing (scRNA-seq) is widely used to estimate time-series gene expression trends through developmental progression of heterogeneous biological tissue at the unit biological level (Macaulay *et al*., 2014; Moignard *et al*., 2015). Though a number of methods have been designed to estimate expression trends from single-cell data (Trapnell *et al*., 2014; Giecold16 *et al*., 2016; Bendall *et al*., 2014; Shin *et al*., 2015), several technical challenges remain (Gawad *et al*., 2016; Fan *et al*., 2015; Ning *et al*., 2014). Among these are increased variability and persistent drop-out measurements, which complicate accurate trend estimation. To address these and other computational challenges, we have developed CREoLE, a Consensus Representation Estimation of Lineage Expression algorithm. CREoLE leverages scRNA-seq measurements and provides a comprehensive set of utilities to automate several key computational steps in estimating developmental expression trends, including identification of the biological lineage structure, lineage origin, and lineage pathways.

CREoLE compensates for inherent technical noise by calculating consensus trends over bootstrapped iterations of expression estimations, rendering smooth, high-resolution estimators of transcriptional processes. The manner by which this is done requires few subjective inputs (e.g.

"MainDocument" — 2016/4/5 — 21:46 — page 2 — #2

developmental way-point selection as in Bendall *et al*., 2014, exclusion criteria based on cellular annotation as in Trapnell *et al*., 2014, or fitted mathematical model parameters as in Shin *et al*., 2015), possibly making CREoLE less susceptible to user bias than other methods while potentially offering a preferred alternative in the absence of expert annotation. CREoLE is designed to be robust to noise by calculating consensus estimations over multiple iterations, smoothing out confounding technical noise presented by non-iterative estimators. CREoLE is implemented in R and is open-source, free to use, and hosted on GitHub at https://github.com/schaugf/creole.

2 Approach

CREoLE is designed to leverage iterative estimations of developmental gene expression to overcome inherent technical noise and variability in single-cell RNA-seq measurements through a two-stage process. Each stage is executed by two complementary functions, `creole_map` and `creole_lines`. A third visualization function, `creole_plots`, generates illustrative figures to convey CREoLE results. A overview flow diagram of the CREoLE algorithm is shown below in Figure 1.

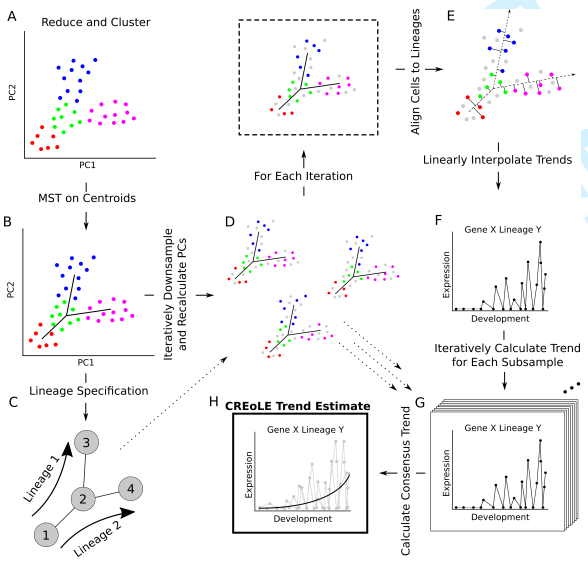

Fig. 1. CREoLE estimates developmental lineage hierarchies and calculates smooth consensus expression trends. (A) Raw scRNA-seq data is reduced by a the user's choice of supported reduction algorithm of either Principal Component Analysis (PCA), Independent Component Analysis (ICA), or Diffusion Mapping (DM). The reduced space is clustered by k-means clustering. (B) A Minimum Spanning Tree (MST) is calculated through cluster centroids. (C) The MST informs lineage specification. A putative origin for the biological system is calculated and all extensible lineages are identified. (D) The raw single-cell data is iteratively subsampled to produce singular estimates of gene expression for each sample. Because a reduced dimensional space is recalculated for each sample, each cell retains its initial cluster identity, which permits carrying over of lineage specification. (E) Cells are aligned to the joining vectors that define lineage pathways from origin to each terminus. For each sample, expression trends are estimated over each gene and through each lineage. (F) Expression trends are linearly interpolated through point measurements along the developmental time-line. (G) Expression trend estimations for each gene and through each lineage are generated for each sampling iteration. (H) Consensus estimations are calculated as the mean expression estimation value for each point in developmental time across all sampling iterations, smoothing the expression trend signal from inherent scRNA-seq technical noise.

2.1 CREoLE Mapping

The first component of the CREoLE algorithm, `creole_map`, utilizes a supported method of data dimensionality reduction, including Principal Component Analysis (PCA) (by default), Independent Component Analysis (ICA), or Diffusion Mapping (DM), which has been shown to be an effective method of non-linear dimensionality reduction for single-cell data (Coifman *et al*., 2005; Haghverdi *et al*., 2015). CREoLE automatically estimates the supportive lineage structure within the data by associating cluster centroids in low-dimensional space by a Minimum Spanning Tree (MST). CREoLE calculates a putative origin cluster node to define the beginning cluster of the developmental lineage tree as the cluster with maximal connectivity. Lineage termini nodes are defined as any remaining non-origin nodes in the MST with exactly one edge.

2.2 CREoLE Lineage

The second component of the CREoLE algorithm, `creole_lines`, calculates fine-resolution estimations of gene expression through developmental progression. The algorithm iteratively sub-samples the raw scRNA-seq data set and recalculates cell position in reduced dimensional space while retaining the cells' initial cluster identity and preserving the lineage specification calculated by `creole_map`. A Minimum Spanning Tree (MST) links these clusters together and joining cluster geometric centroids by low-dimensional vectors, thereby defining lineage progression in low-dimensional space. For each iteration, the sub-sampled cells are projected onto the appropriate vectors that connect the sets of clusters that define distinct lineages. Cell position, once projected onto the centroid-joining vector, is normalized from zero to one to define lineage-specific developmental progression and to estimate gene expression trends. By iteratively sub-sampling the raw data and generating distinct expression trend estimations, CREoLE generates a series of gene expression trend estimates through parallel and divergent developmental lineages. The average expression through each lineage and for each gene defines expression consensus, smoothing out individual estimates to address inherent technical noise in the data.

2.3 CREoLE Plots

The `creole_plots` function generates a series of illustrative figures for the user to visualize results from CREoLE. Examples of some, but not all, are shown in the following analysis of a synthetic data set. The MST and reduced dimension spaces are both generated by `creole_map`. Following expression trend estimation by `creole_lines`, `creole_plots` generates boxplots to illustrate expression distribution for a given gene between each cluster, line plots of expression amplitude for specific genes throughout all developmental lineage, and line plots of all genes through each specific lineage.

3 Methods

The CREoLE algorithm is applied to a synthetic data set as well as publicly available scRNA-seq data sets. Synthetic data analysis is desired to verify that CREoLE results align with known expectation of a simplified biological system.

3.1 Synthetic Data Design

A synthetic scRNA-seq data set was generated to reflect a branching biological lineage system, such as a developing embryo or cancer tumor, where a root niave state differentiates through development time. In this example, 700 cells are artificially generated for down-stream analysis. The inherent branching lineage topology, shown in Figure 2, represents

"MainDocument" — 2016/4/5 — 21:46 — page 3 — #3

a double-bifurcation event that results in four distinct cellular lineages. Each bifurcation event is distinguished by activation of a unique gene whose expression trend is defined as a ramp function from zero to unit expression within the context of distinct lineages. For example, referring to Figure 2, a distinct gene activates exclusively in cluster 2 and stays active in both clusters 4 and 5. Similarly, a separate gene activates exclusively in cluster 4, distinguishing that lineage from that which terminates at cluster 5. The origin cluster is also defined by an exclusive gene, whose activation is evident in all down-stream cells. In this manner, each of the four lineages is defined by a pattern of gene activation beginning at the origin and terminating at each of the four terminal clusters. The code used to generate the synthetic data set is available at https://github.com/schaugf/creole/R/makeSyntheticData.R

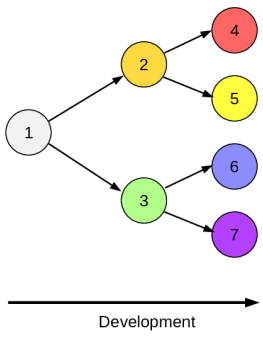

Fig. 2. The synthetic scRNA-seq data set is designed to exhibit a double bifurcation event. For each conceptual cluster, a gene is associated exclusively to that cluster and all down-stream lineage clusters, ensuring a one-way propagation of information through the developmental time-line.

3.2 CREoLE Analysis of Synthetic Data

The CREoLE algorithm is applied to the synthetic data set. The algorithm is run through the RStudio environment Version 0.99.491 on a Dell Latitude E7250 running 64-bit Ubuntu 15.10.

3.2.1 Estimated Lineage Structure for Synthetic Data

The synthetic single-cell dataset is first analyzed by `creole_map`, which maps raw data into a low-dimensional space, clusters the cells, and estimates an underlying lineage structure within the data. The number of principal components is automatically calculated to account for a user-defined proportion of variance in the data; in this example, 90% variance is accounted for by five components. Following reduction, an MST is calculated between each of the centroids. Cluster number is automatically chosen by Silhouette Analysis (Rousseeuw *et al*., 1987) with parameters that may be set by the user. CREoLE uses the established lineage system to inform iterative estimations of expression by preserving initial cluster identity and storing lineages as series of clusters. CREoLE predicts a putative origin by identifying the lineage node of greatest connectively and can be updated by the user. A series of two-dimensional projections of the data following reduction by PCA and the calculated MST are shown in Figure 3

3.2.2 Synthetic Expression Trend Estimation

The CREoLE algorithm establishes a twice-bifurcation lineage system as expected. Choosing cluster nine as the origin establishes four developmental lineages, terminating at clusters seven, five, three, and eight. For each of the four lineages, consensus expression trend estimations

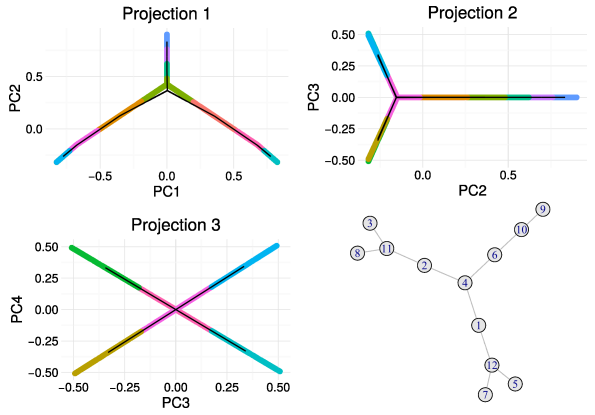

Fig. 3. Low-dimensional projections following Principal Component Analysis on raw synthetic data and calculated double-bifurcating MST. In this example, cluster 4 is identified as the putative origin cluster according to maximal node connectivity. However, knowledge of the underlying biology informs the user to select cluster 9 as the origin.

are calculated through development time as shown in Figure 4. In this example, each of the four lineages are accurately defined by activation sequences of distinct genes. For example, we observe that because we know Gene 1 is associated with the origin cluster, it should be expressed in each of the four lineages. Similarly, each downstream pair of clusters illustrates activation of the gene associated exclusively with those lineages precisely as defined in the synthetic data set. In this example, estimations are calculated over one hundred iterations by sampling 70% of the initial data set.

Pearson's correlation coefficients are calculated for each gene through each lineage relative to synthetic expression expectation as used previously to evaluate single-cell expression trend accuracy (Bendall *et al*., 2014). The output from CREoLE correlates well to expression trends inherent in the generated synthetic data set. The average correlation coefficient across each gene and through each lineage of the synthetic data set is 0.983 ± 0.009, implying both a strong and consistent estimation of expression trend throughout the entire synthetic system.

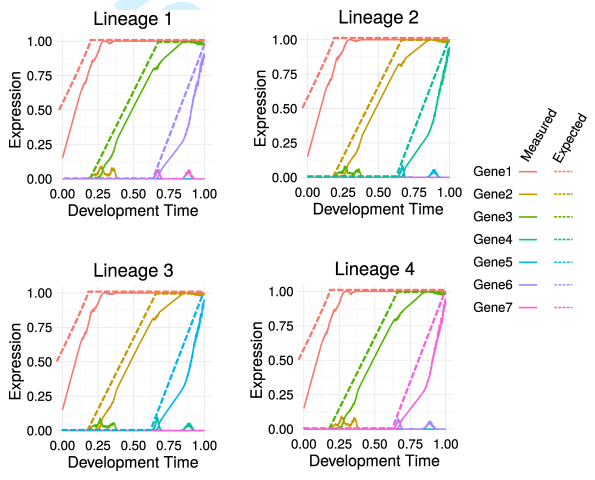

Fig. 4. Composite trends of gene expression through development time define each lineage. In all cases, the double bifurcation event is conveyed as a branching deviation of distinct cell lines.

"MainDocument" — 2016/4/5 — 21:46 — page 4 — #4

3.2.3 Impact of Variable Iterations

The quality of estimated expression trend is dependent on the number of sub-sampling iterations computed by CREoLE. As shown in Figure 5, increasing the number of computation iterations produces a smoothing effect on the signal by reducing noise through consensus estimation.

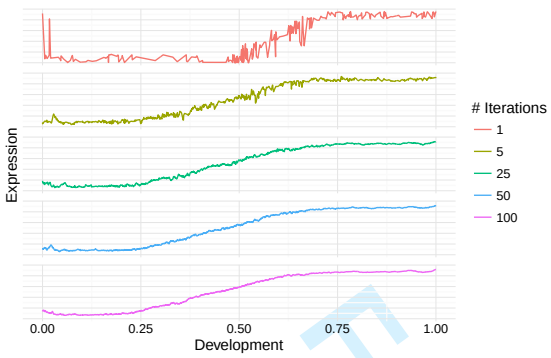

Fig. 5. Increasing the number of iterations performed by CREoLE exerts a smoothing effect on the estimated expression signal.

3.2.4 Noise and Drop-out Measurements

Similar evaluation was performed to evaluate how well consensus estimations correlate with known expectation following addition of additive noise and drop-out measurements. Technical noise is simulated by adding uniformly distributed values from zero to 40% unit gene expression. Similarly, drop-out measurements are simulated by randomly setting expression levels for portions of cells to zero. CREoLE appears to accurately recapitulate expected expression patters despite significant additional additive technical noise which reduced mean Pearson correlation coefficient from 0.983 ± 0.009 to 0.964 ± 0.018. The distributions of Pearson's correlation coefficients between each gene and through each lineage with the addition of additive noise are shown in the top half of Figure 6.

3.2.5 Cell Count Reduction

As the quantity of cells sequenced is associated with pragmatic experimental considerations, we sought to evaluate how well CREoLE estimations correlate with known expectations at reduced cell counts. The initial data set is randomly sub-sampled without replacement to fractions of the initial cell count. CREoLE appears to accurately recapitulate expected expression patterns with high correlation despite significant reductions in total cell count from 0.983 ± 0.009 with no reduction to 0.959 ± 0.032 with a 75% reduction in total cell count. The distributions of Pearson's correlation coefficients between each gene and through each lineage with fractionally reduced cell counts are shown in the bottom half of Figure 6.

3.3 CREoLE Recapitulates Published Findings

CREoLE is applied to the publicly available scRNA-seq data set presented by Shin *et al.*, 2015 which, following filtering by annotation, contains expression values for 23207 genes over 132 cells. The method of ordering single cells proposed by the authors is recreated on the left side of Figure 7 along with the CREoLE estimation of gene expression on the right side. Pearson's' correlation coefficients between the distinct points on the left hand side and the accordingly sub-sampled CREoLE estimate are provided. Although certain correlation coefficients are not strong, trend visualization illustrates how the general shape of the expression curve is evident despite significant technical noise in certain samples. The smooth

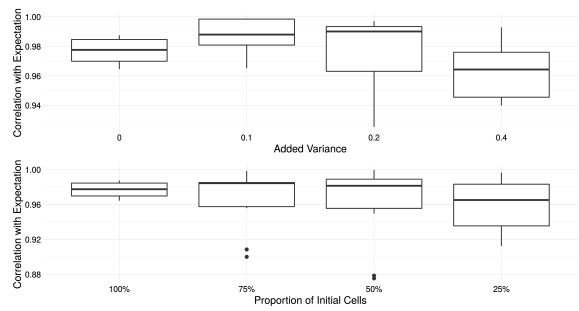

Fig. 6. Distribution of Pearson's correlation coefficients across all genes and through each lineage. In the top row of boxplots, additive noise and drop-out measurements are injected at increasing rates. In the bottom row of boxplots, the total cell count used for analysis is reduced by a fraction of the total number of initial cells. In both cases, CREoLE appears robust to common technical perturbations presented by scRNA-seq.

and high-resolution trends generated by CREoLE may prove valuable in computational characterizations of transcription dynamics of complex biological systems.

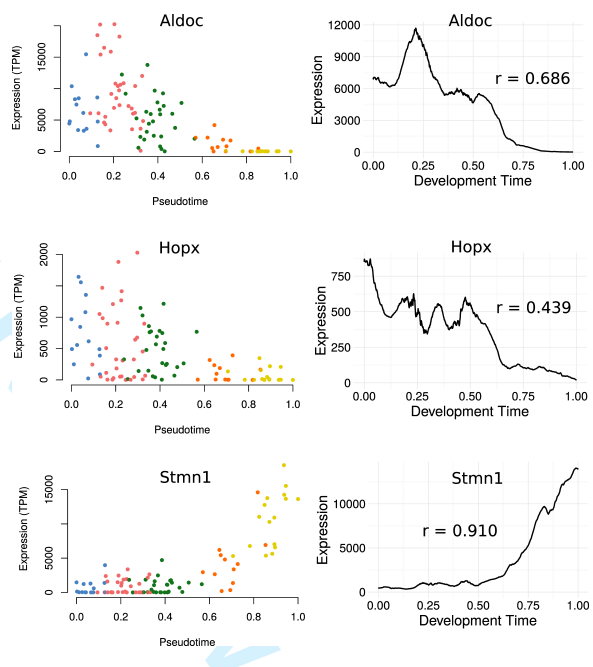

Fig. 7. Expression trend results presented in Shin *et al.*, 2015 and by CREoLE on the left and right, respectively.

4 Discussion

We have presented CREoLE, an algorithm designed to generate smooth consensus trends of gene expression through a complex branching lineage system. CREoLE minimizes subjective user input and addresses a number of inherent technical limitation of scRNA-seq, including additive noise, drop-out measurements, and cell count reduction. CREoLE is fast and efficient, requiring mere minutes of desktop computational time to generate consensus estimations. CREoLE has shown to be capable of both estimating inherent lineage structure within synthetic scRNA-seq data and rendering smooth lineage-dependent expression trends even with additive technical perturbation. Figures generated with `creole_lines` communicate that the CREoLE algorithm can accurately reconstitute a

"MainDocument" — 2016/4/5 — 21:46 — page 5 — #5

synthetic lineage system and calculated smooth, high-resolution consensus trends that align closely with synthetic expression trends.

While single-cell RNA-seq provides incredibly rich data sets, many computational challenges remain unmet. Estimations of gene expression made by CREoLE may power down-stream analyses to address additional biological questions. Several areas of active investigation desire computational methodologies to identify key genes that regulate lineage specification, differentiation, and progression. The smooth signals generated by CREoLE may be well-suited for network or system identification studies designed to address these fundamental questions regarding transcriptional biology.

Software

CREoLE is implemented in R and runs on standard desktop operating systems. The open-source software is available for free at https://github.com/schaugf/creole.

Acknowledgements

We thank Guanming Wu, Eisa Mahyari, Nathan Lazar, Julian Egger, Kristóf Törkenczy, Ryan Mulqueen, Sarah Vitak, and Taylor Mighell for their helpful technical discussions.

Funding

This work was supported by the National Library of Medicine of the National Institutes of Health under [Award Number T15LM007088]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Bendall, S. C., Davis, K. L., Amir, E. A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Pe'Er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3), 714-725.

Coifman, R. R., Lafon, S., Lee, a B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21), 7426-31.

Fan, J.-B. J., Salathia, N., Liu, R., Kaeser, G., Yung, Y., Herman, J. L., Kharchenko, P. V. (2015). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *bioRxiv*, (May 2015), 026948.

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*.

Giecold, G., Marco, E., Trippa, L., & Yuan, G.-C. (2016). Robust Lineage Reconstruction from High-Dimensional Single-Cell Data. *bioRxiv*, 036533.

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

Macaulay, I. C., & Voet, T. (2014). Single cell genomics: advances and future perspectives. *PLoS Genetics*, 10(1), e1004126.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, S., Piterman, N., Kouskoff, V., Theis, F. J., Fisher, J., Göttgens (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33, 269-276.

Ning, L., Liu, G., Li, G., Hou, Y., Tong, Y., & He, J. (2014). Current challenges in the bioinformatics of single cell genomics. *Front Oncol*, 4, 7.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C), 53-65.

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Song, H. (2015). Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell*, 17(3), 360-372.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4), 381-6.

Wang, Q., Zhu, X., Feng, Y., Xue, Z., & Fan, G. (2013). Single-cell genomics: An overview. *Frontiers in Biology*, 8(6), 569-576.