Transcriptional heterogeneity, cell identities, and trajectories

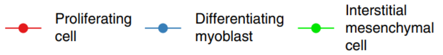
Kamil Krakowski

There are a few processes that change eukaryotic cell transcription profiles. Cells can **proliferate, differentiate and reprogramme**. Each of the cells can have other rates of transcriptional changes, because those processes are controlled by different stimuli and cell to cell interactions. Moreover, cells even from the clonal population, might undergo a different sequence of intermediate stages during differentiation and finally converge on the same state. So, the cells of **the same type can proceed the same changes in different time**. On the other hand, a cell can change its differentiation trajectory creating a sublineage (branch). That's why analysis of transcriptomic experiments is so difficult. Furthermore those assays (especially bulk expression measurements) can be affected by mixture effects like **Simpson's paradox** (a phenomenon in statistics, in which averages of groups show a different trend from a one that describes each group) (Trapnell et al., 2014).

There is also a technical noise, resulting from sequencing, wet-lab protocol and its equipment. It can be divided into shot noise, mRNA loss (dropout), capture efficiency, sequencing efficiency. This noise is especially present in lowly expressed genes. To reduce this one, external RNAs called spike-ins are introduced into cell lysis buffer. The concentration of spike-ins should be equal in all cells in the experiment, thus stochastic, technical noise can be easily modeled. In allele-specific expression patterns studies, technical noise can be described as about 80% of total noise (Kim et al., 2015).

SPADE (**S**panning-tree **P**rogression **A**nalysis of **D**ensity-normalized **E**vents) - machine learning algorithm that is used to reconstruct differentiation lineages and intermediate states. Requires knowledge of marker genes. Applied to flow and mass cytometry data it can help with the identification of cell types or with the analysis of heterogeneity (Qiu et al., 2011).

Monocle - unsupervised algorithm that uses single-cell RNA-Seq measurements collected at multiple time points. It increases the temporal resolution of transcriptome dynamics. Using the learned process of differentiation, Monocle orders an unsynchronised population of the cells into specific sublineages in pseudotime (quantitative measure of a biological process). It pinpoints genes that are differentially expressed and clusters them according to their kinetic trend in order to identify significant events occurred during biological processes. It does not require a priori knowledge of known transcription markers. Thus, it can be used to discover markers and regulators of uncharacterized transition processes .



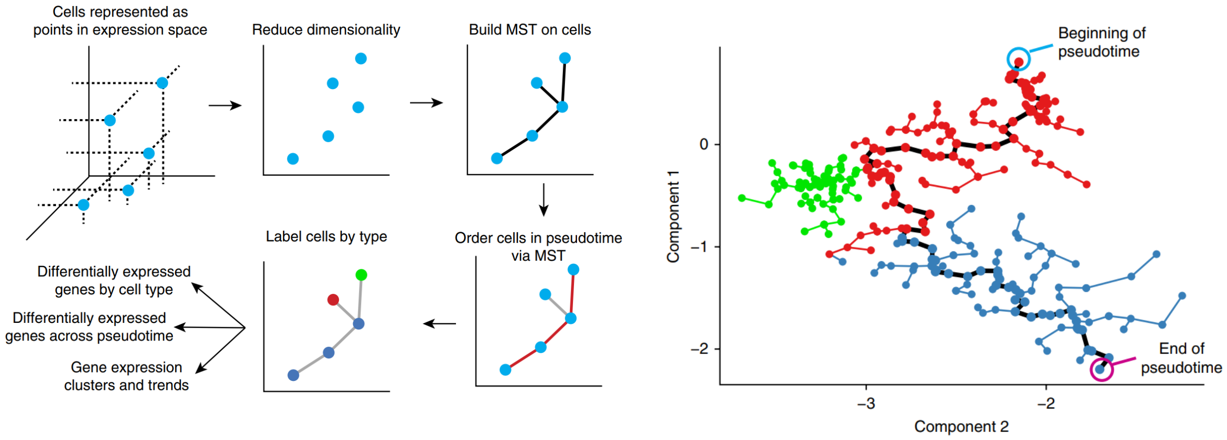

Cells represented as points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime via MST
Label cells by type
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends
Beginning of pseudotime
End of pseudotime
Component 1
Component 2
0
−1
−2
−3
−2

Figure 1. representing an overview of the Monocle workflow (Trapnell et al., 2014)

Monocle algorithm main steps (Figure 1):

1. choose manually genes according to experiment needs (up to 48 genes)
2. represent each cell as a vector of R^d , where d is a number of genes
3. reduce dimensionality through ICA (independent component analysis) to find one dimensional function of differentiation in R^d
4. find a polygonal reconstruction of differentiation (as a continuous, smooth function):
    a. construct a weighted complete graph (nodes -> cells, edges -> weighted distances between cells)
    b. calculate MST (minimum spanning tree) and find longest path
    c. construct rooted, ordered PQ tree that represents a family of good orderings of the cells
    d. search orderings that complies with the constraints and minimizes the total distance of polygonal reconstruction in the embedding geometry
5. identify genes that are differentially expressed using generalized additive models (GAMs)
6. perform K-medioid clustering on the predicted for each gene pseudotime (GAM)

Monocle was tested on human myoblasts and has contributed to identification of eight previously unknown transcription inhibitors in those cells. Some of those discovered transcription factors can repress differentiation by competition with promyogenic activators. These results showed how proliferation and differentiation are controlled during development and tissue regeneration (Trapnell et al., 2014).

Moreover Monocle is still an ongoing project (Qiu et al., 2017b). The new version of this software takes advantage of **reversed graph embeddings** to robustly define the cells trajectories. Another difference is another clustering method, that is inspired by Seurat strategy (Satija et al., 2015). **Monocle 2** also utilizes **Census** to support **mRNA counts**, which can be more accurate and easier to manipulate in comparison with the expression described by conventional measures like transcript per million - TPM. Those improvements allow the use of new regression models, such as **BEAM** (branch expression analysis modeling) (Qiu et al., 2017a).

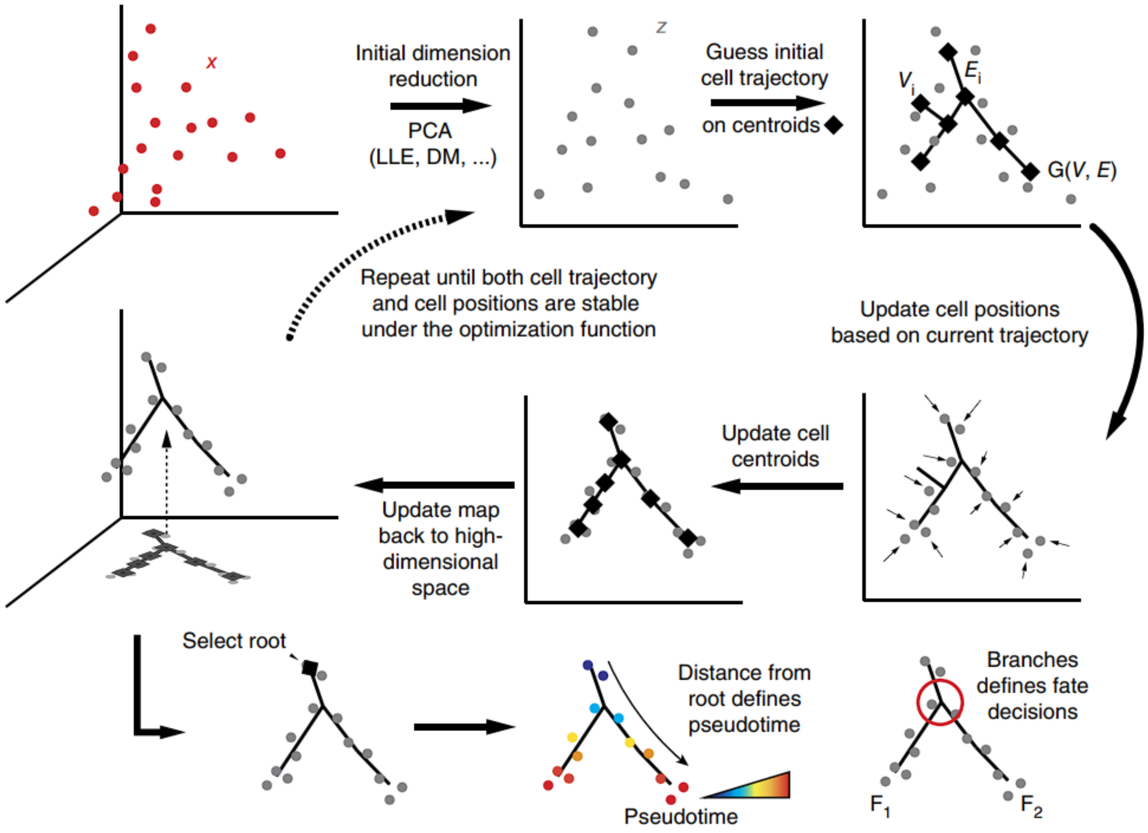

x
Initial dimension reduction
PCA
(LLE, DM, ...)
z
Guess initial cell trajectory
on centroids ◆
V_i
E_i
$G(V, E)$
Repeat until both cell trajectory and cell positions are stable under the optimization function
Update cell positions based on current trajectory
Update cell centroids
Update map back to high-dimensional space
Select root
Distance from root defines pseudotime
Pseudotime
Branches defines fate decisions
F_1
F_2

Figure 2. representing an overview of the Monocle 2 workflow (Qiu et al., 2017b)

Census - converter of TPM or FPKM (Fragments Per Kilobase Million), derived from single-cell RNA-seq **without the *spike-in* control**, to RPC (mRNAs per cell).

BEAM - generalized linear modeling strategy, that aims at finding genes with different expression between two branches

There is also other software available for single-cell gene expression analysis. One of them is **SCANPY**, which extends Monocle functionality. It is written in python language and offers a range of methods for differential expression testing, pseudotime and trajectory inference, clustering, visualization and gene regulatory networks simulations. The authors claim that it can handle more than one million datasets generated from single cell experiments (Wolf et al., 2018).

Bibliography:

Kim, J.K., Kolodziejczyk, A.A., Ilicic, T., Teichmann, S.A., Marioni, J.C., 2015. Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. Nat. Commun. 6, 8687. https://doi.org/10.1038/ncomms9687

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., Plevritis, S.K., 2011. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. 29, 886–891. https://doi.org/10.1038/nbt.1991

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., Trapnell, C., 2017a. Single-cell mRNA quantification and differential analysis with Census. Nat. Methods 14, 309–315. https://doi.org/10.1038/nmeth.4150

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C., 2017b. Reversed graph embedding resolves complex single-cell trajectories. Nat. Methods 14, 979–982. https://doi.org/10.1038/nmeth.4402

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., Regev, A., 2015. Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. 33, 495–502. https://doi.org/10.1038/nbt.3192

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., 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. https://doi.org/10.1038/nbt.2859

Wolf, F.A., Angerer, P., Theis, F.J., 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15. https://doi.org/10.1186/s13059-017-1382-0