Analysis of human hematopoietic cells using Scanpy and Dynamo

Seyone Chithrananda

University of California, Berkeley

Bioengineering 190/290

Professor Liana Lareau

December 16, 2022

Analysis of human hematopoietic cells using Scanpy and Dynamo

Introduction

The emergence of single-cell RNA sequencing, RNA velocity, and metabolic labeling has resulted in several techniques aiming to reveal cellular cell states and transitions at an unprecedented resolution. In this work, we examine raw hematopoiesis data from the Cell paper, *'Mapping transcriptomic vector fields of single cells'* [1]. We aim to first analyze the dataset using clustering tools such as the Leiden algorithm in Scanpy, and find marker genes by analyzing the most differentially expressed genes in specific clusters with regards to their respective cell type. Next, we use dynamo's ability to compute RNA velocity for sc-RNA seq data to better understand cell fate transition in the dataset, and compute in-silico perturbation tests to investigate cell fate outcomes after perturbing key regulators like GAT1.

Human hematopoiesis cells; an overview

In the tutorial, we use Dynamo to perform absolute total RNA velocity analysis with the metabolic labeling datasets that were reported in the dynamo Cell paper. The paper uses a metabolically labeled human hematopoiesis scRNA-seq dataset to showcase their technique's ability to overcome fundamental limitations of conventional splicing-based RNA velocity analyses to enable accurate velocity estimations. They use single-cell metabolically labeled new RNA tagged sequencing (scNT-seq), as a method for massively parallel analysis of newly transcribed and pre-existing mRNAs from the same cell. Using scNT-seq, a droplet microfluidics-based method, allows for high-throughput chemical conversion on barcoded beads, efficiently marking newly transcribed mRNAs with T-to-C substitutions. They generate a time-resolved scRNA-seq dataset of primary human HSPC (hematopoietic stem and progenitor cells), profiling these cells occurring multi-lineage differentiation in-vitro across day 4 and 7. Using this approach, for each gene they are

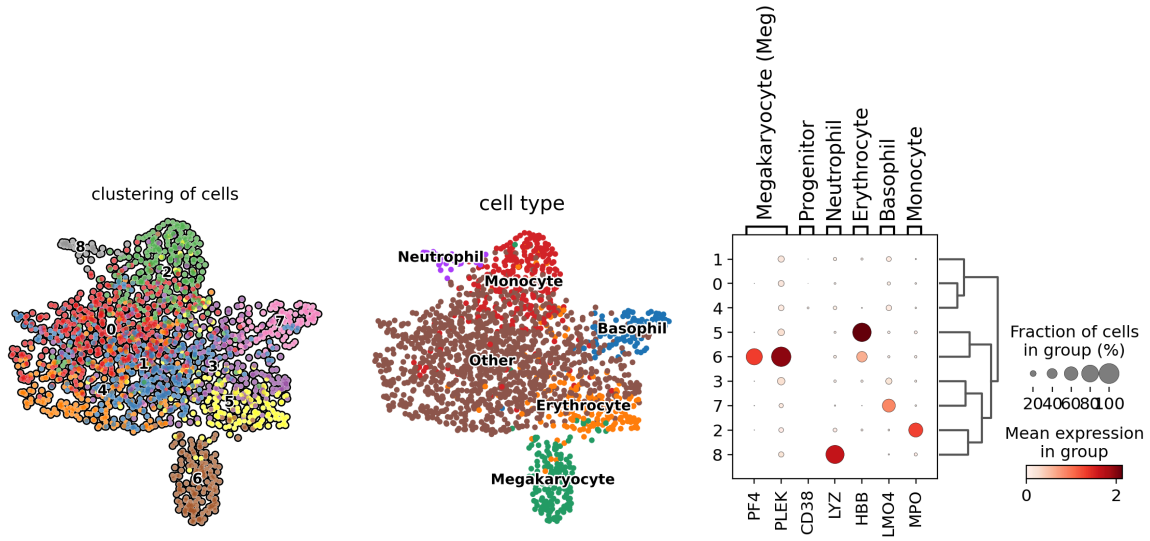
able to quantify unspliced, spliced, new and total RNA levels for each gene. This enables the RNA velocity analyses in dynamo's modeling framework.

We load the raw hematopoiesis dataset, where '*new*' is defined as the labeled RNA, and '*total*' corresponds to the total RNA for velocity estimation. Time denotes the RNA metabolic labeling duration (measuring RNA degradation and splicing rates). First, we use a well-established list of the highly variable genes and known marker genes based on previous reports [2] to filter the genes we use. Since the paper labels cells from the same harvest time (day 7, 10, etc) for a single time pulse, the experiment_type is set to **'one-shot'.**

Differential Expression Analysis & Gene marker annotation using Scanpy

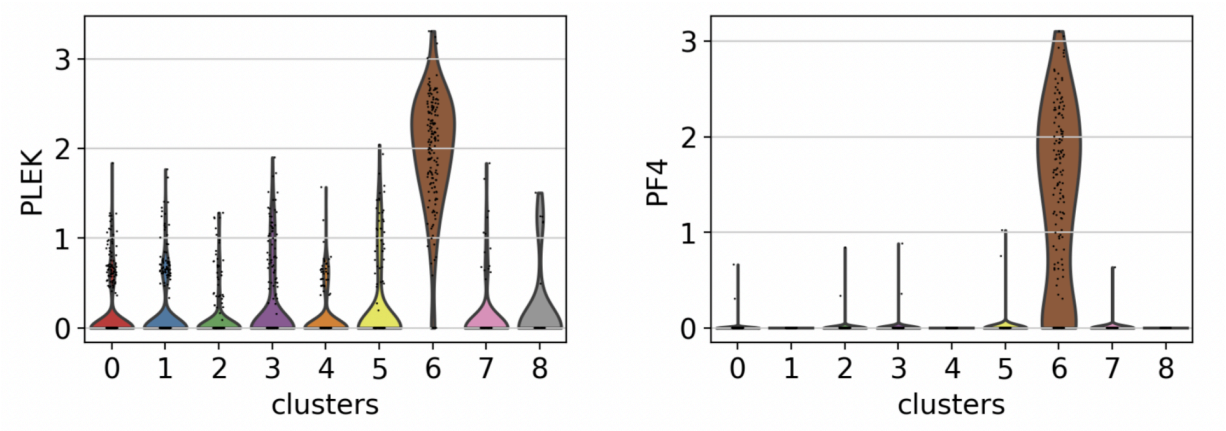
Next, we use Scanpy to look at the clusters generated by the Leiden clustering algorithm, and see if distinct clusters capture different transcriptomic profiles based on key gene markers. The Leidein method is an improved version of the Louvain hierarchical clustering algorithm that recursively aims to cluster cells into subgroups. The leiden algorithm aims to improve on shortcomings of the Louvain method, namely that the algorithm can yield arbitrarily badly connected communities/subgroups that may even be internally disconnected. Using this approach, we observe some clear clustering, but the method doesn't perform as well as we wish despite tuning the resolution. Instead, we look to see if distinct clusters appear to capture different transcriptomic profiles based on key gene markers, and thus different cell lineages.

Namely, we look at PF4 and PLEK as gene markers for Megakaryocytes cells, a type of hematopoietic cell responsible for the production of blood platelets. We study several other gene markers, for types of cells ranging from progenitor cells, neutrophils, erythrocytes and basophils.



Left: UMAP representation of cells, where the color labels represent each cluster from the Leiden algorithm. **Center**: Same UMAP visualization, but each color label represents the cell-type predominantly represented by the respective cluster. **Right:** Dotplot by cluster and target gene, where each circle's color represents the mean expression of a gene in the given cluster, and the size of the circle indicates the fraction of cells in the cluster.

Interestingly using the simple example above,, we see that each cluster (besides 0 & 1) uniquely captures a higher mean expression for a specific marker. As an example, cluster 5 captures cells with a higher mean expression of gene HBB, a marker for Erythrocyte cell types. In one case, we see that cluster 6 captures cells with a higher expression of two markers for the Meg lineage. Another way to explore the presence of gene markers in specific clusters in the dataset is by looking at violin plots. We plot genes PLEK and PF4 once again, and observe significantly higher expression of the gene in cluster 6. All in all, this example demonstrates the ability for simple clustering methods to capture distinct transcriptomic profiles across different groups, enabling us to find marker genes by looking at the most differentially expressed genes in distinct cell-types.



Above: Plotting the variation of mean expression of PLEK and PF4 (markers for the Meg cell lineage) across clusters, we observe significantly higher expression in cluster 6.

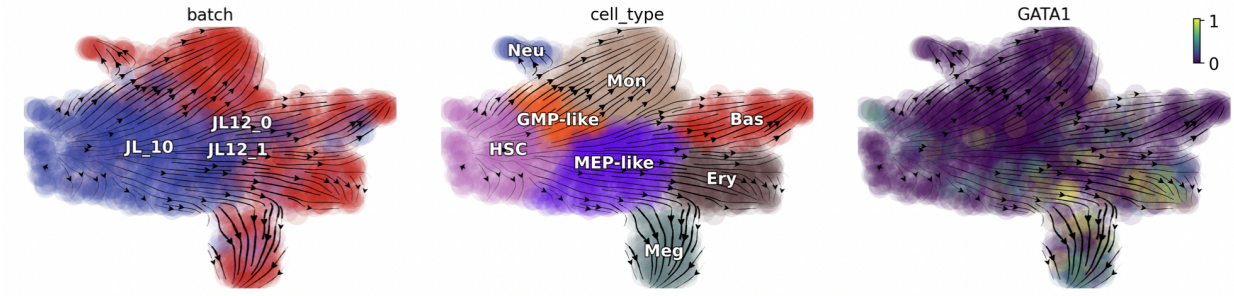
Examining RNA velocity as a proxy for cellular state progression using Dynamo

Dynamo infers absolute RNA velocity and reconstructs continuous vector fields to predict cell fates, utilizing differential geometry to understand the underlying regulatory interactions in a cell. In their work, they introduce differential geometry analyses to single-cell geometry, describing the idea of a vector field that can take in input coordinates ***x*** in a transcriptomic space, and output a vector **v** in the same space, representing a **differentiable velocity vector field**. Using the Jacobian (a d-by-d matrix encoding a function **f**'s derivatives), we can infer how the velocity of a specific **gene x_i is impacted by changes in gene x_j.** This enables the model's dynamo constructs to identify activating/inhibitory regulatory activities.

We use dynamo's Model 2 (not Monacle) to estimate RNA-velocity for time-resolved scRNA-seq data, the main difference being that dynamo's models do not rely on RNA splicing rate but rather labeling in estimating velocity unlike previous approaches. We group cells for gene expression cells using labeling time (see *dyn.tl.moments* [3]), splitting the dataset based on the two time points in the hsc dataset. Given both the labeling and splicing information, we can define total RNA velocity using the following:

$$\dot{r} = n/(1 - e^{-rt}) \cdot r - \gamma s$$

Where ***r, n, t, γ, s*** are total RNA, new RNA, labeling time, splicing rate, and spliced RNA, respectively. Given total RNA velocities, we project them to two-dimensional UMAP space and visualize the total RNA stream line available in dynamo, reproducing the plots in the original Cell paper, but for GATA1.



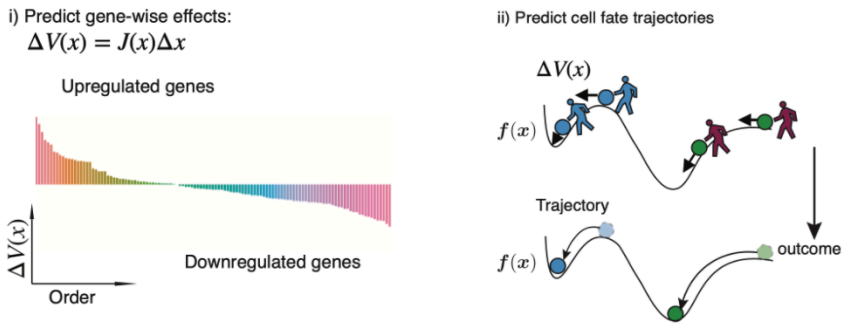
Above: Total RNA stream line plot for entire dataset, followed by GATA1 TF regulator on UMAP space.

One of the exciting facets of the RNA velocity approach outlined by Dynamo is the ability to get an idea of how cell fate transition works on a large scale for any scRNA-seq experiment, and even get an idea of how mean gene expression differs across these different cell types (shown above for GATA1 regulator). This, for example, can inform key knockdown experiments to see if a cell-type changes accordingly with the RNA stream line plots.

Investigating the effects of in-silico perturbations on key regulators using Dynamo

Finally, we'll be diving deeper into the ideas introduced by using RNA velocity to study cell fate transitions, by introducing in-silico perturbations and viewing the corresponding velocity vectors. Leveraging the previously described analytical Jacobian of the vector field that dynamo reconstructs, we can make in-silico genetic perturbations and try to predict cell-fate outcomes after the perturbation by integrating the displacement of velocities across cells.

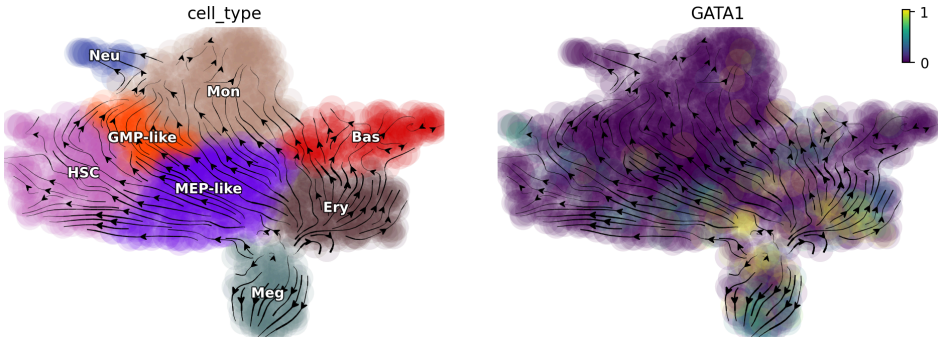
Analysis of human hematopoietic cells using Scanpy and Dynamo



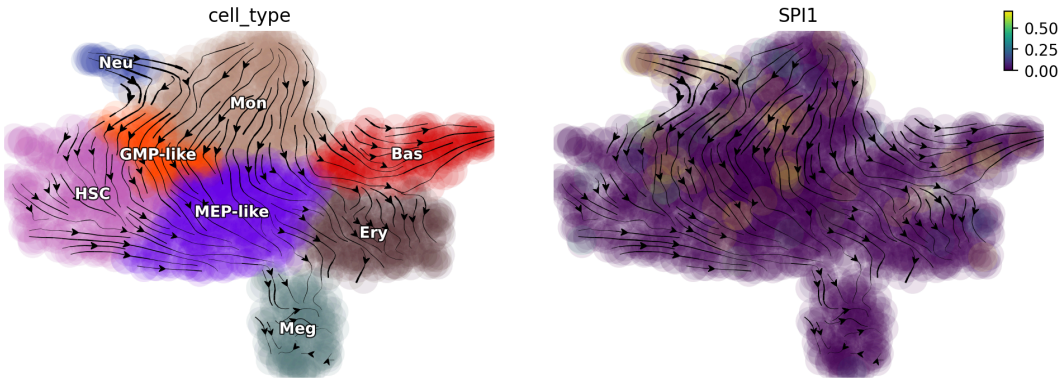
To simulate genetic perturbation effects, dynamo uses a perturbation vector to introduce perturbations to the system, and then lets the perturbations further propagate across the gene regulatory network represented by our Jacobian. The genetic perturbation effects/changes in the velocity of a gene i with respect to small perturbations in the expression of all genes in the network is represented as the following differential:

$$\mathrm{d}f_i = \frac{\partial f_i}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f_i}{\partial x_2}\mathrm{d}x_2 + \ldots + \frac{\partial f_i}{\partial x_n}\mathrm{d}x_n.$$

We attempt to compute in-silico perturbations on GATA1, a master regulator of the GMP cell lineage during hematopoiesis. We perturb GATA1 in-silico by under-expressing it. To view the velocity vectors computed by dynamo in-silico predictions, we project them to a lower-dimensional basis using UMAP. The arrows indicate cell type progression, and we can see that the suppression of GATA1 causes cell type progression from MEP-related lineages to GMP-like:



Repeating the same for SP1, we observe a reverse diversion from GMP to MEP, and suppressing both regulators creates a dual effect - essentially trapping cell states in the middle with regards to driving GMP/MEP lineages.



Above: Suppressing GATA1 results in a cell-type progression from MEP-like lineages to GMP.

Conclusion

In review, the field of computational single-cell genomics remains at an exciting point in time. The availability of time-resolved, metabolically labeled scRNA-seq datasets has enabled the emergence of several techniques aiming to reveal cellular cell states and transitions at an unprecedented resolution. We use dynamo to compute splicing rate-agnostic RNA velocities for the raw hematopoiesis data from their Cell paper, revealing key cell-type progressions in regulators like GATA1 and SP1 through in-silico perturbations. We also highlight the ability of clustering methods in Scanpy to do gene marker annotation. A future direction with this dataset and approach would be to **construct potential landscapes of single-cells datasets**, to better understand the transition rate between different cell types, and the optimal paths of cell fate transition.

References

[1] Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., ... & Weissman, J. S. (2022). Mapping transcriptomic vector fields of single cells. *Cell*, *185*(4), 690-711.

[2] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, *367*(6479), eaaw3381.

[3] Dynamo Documentation