RNA Velocity in Single cells - An analysis of Publicly available datasets

Pallavi Surana, Ramana V Davuluri

Abstract

RNA velocity calculated from single-cell RNA sequence (scRNA-seq) data allows for the prediction of future states which give useful insights into transient stages during the developmental process in eukaryotic systems. RNA velocity thus holds promise to unveil dynamics from static RNA-seq data and is widely applicable across diverse organismal systems and technical platforms. By combining existing methods, we developed an informatics pipelin to perform comparative analysis of different scRNA-seq datasets to find overlapping genes between different human brain regions and mouse neural stem cells in different cell cycle states. We observe some genes upregulated in disease pathways. Here, we primarily used the velocity estimates to find important genes essential for cell cycle progression. Such analysis can help in understanding the misregulation of genes in diseased states like cancer.

Introduction

RNA velocity has enabled us to look at scRNA seq data with a different pair of lenses. A major component of transcriptional activity and cellular response to changing cell conditions is a shift of the transcriptome to new cell states. (Zeisel et al. 2011) The idea of estimating the velocity of single cells emerged from the challenge is that scRNA-seq does not provide the real time estimation of the cell states. Hence, change in states is measured in terms of velocity. In eukaryotes, the shift is due to epigenetic, co-transcriptional, and post-transcriptional processes which determine these transient states. The velocity is deduced by calculating the ratio of unspliced (pre-mRNA) to spliced mRNA (mature mRNA) content. Calculating this ratio for each gene will help in determining the state of that gene in terms of upregulation or downregulation which in terms of values with positive or negative values and thereby giving us an inference of possible future states of the cell.(Zeisel et al. 2011; Burgess 2018; La Manno et al. 2018; Bergen et al. 2021, 2020)

$$\frac{ds}{dt} = u - \gamma s \qquad 1$$

The velocity models as explained in (La Manno et al. 2018; Bergen et al. 2020), and represented in equation 1 captured the transcriptional dynamics. Here α captures transcription rates, β captures splicing rates and degradation rates by

γ. These are involved in production of unspliced (u) and spliced (s) mRNA products.

Models used

This work uses two types of velocity models: static (Zeisel et al. 2011; Burgess 2018; La Manno et al. 2018; Bergen et al. 2021, 2020) and dynamic (La Manno et al. 2018; Bergen et al. 2020).

The steady-state model (Zeisel et al. 2011; Burgess 2018; La Manno et al. 2018; Bergen et al. 2021, 2020) estimates velocities as the deviation of the observed ratio of unspliced to spliced mRNA from an inferred steady-state equilibrium. The steady state is a regression with the cells that have reached the steady state. The cells here usually belong to the extremes in their expression states. The 2 main assumptions made by the model are:

1. $\frac{ds}{dt} = \upsilon$, a constant. This ensures that current rates of pre-mature and mature mRNA continue in to the future states thereby, capturing the velocity of future states. This assumtion does not work well with downregul;ated genes as the $\upsilon < 0$.

2. That the data at least have some information about the steady-state expression levels in the samples sequenced.

An example is explained in (La Manno et al. 2018; Bergen et al. 2020) discuss where the steady-state model might fail. In heterogeneous populations of samples where the dataset has different subpopulations or when the system has no steady-state properties.

The dynamical model aims to explain the entire kinetics of the genes and does not rely on the assumptions made above. It uses the Expectation-Maximization algorithm. This algorithm aims to find local maximum likelihood where the model is believed to depend on transient states. The function then estimates the expectation of the log-likelihood estimates using the present state and then uses these features are then used to estimate the next state, iteratively. Genes with higher estimated values tend to have a higher significance. (Zeisel et al. 2011; Burgess 2018; La Manno et al. 2018; Bergen et al. 2021, 2020)

The analysis here aims to compare different datasets of the brain from humans and mice and to find the velocity of the genes which overlap with the cell cycle genes. This is important as in diseased states cells are able to enter the cell cycle without the presence of growth factors. These cells can evade growth suppressors and cell cycle checkpoints. This might lead to uncontrolled cell growth thereby leading to tumor formation. Therefore, studying the dynamics of cell cycle genes can help in understanding the profiles of genes that might be highly upregulated in disease states. (O'Connor et al. 2021)

Methods

Datasets analyzed

In this study, we used 2 public brain datasets and compared them with the human neural stem cells which describe the cell states of genes. The analysis lays a framework to use different scRNA-seq datasets from similar organisms to predict the possible role of genes involved in brain organ regulation. All the datasets were analyzed using the scvelo (La Manno et al. 2018; Bergen et al. 2020) and dynamo ("Mapping Transcriptomic Vector Fields of Single Cells" 2022) packages.

A mouse dentate gyrus (DG) dataset containing 5454 cells through different time points in development was used (Hochgerner et al. 2018). The DG is part of the hippocampus involved in learning, episodic memory formation, and spatial coding. This region has functions in memory and depression. It serves more as a pre-processing unit. It is one of a select few brain structures known to have significant rates of adult neurogenesis in many species of mammals, from rodents to primates. It is a site for new cell formation throughout life in the brain.

The human neural stem cells (HNSC) from (O'Connor et al. 2021) were used. These cells originate from the developing mammalian telencephalon which is the most developed part of the forebrain. Different cell cycle stages from this dataset were used to find markers of each cell stage.

The next dataset was a human and mouse neuronal splicing dataset from ("Mapping Transcriptomic Vector Fields of Single Cells" 2022) with about 55k cells. This dataset is important as timely expression and perhaps post-translational modification of neuron-specific splicing regulators play important roles in neuronal development. During neuronal differentiation, alternative splicing

modulates signaling activity, centriolar dynamics, and metabolic pathways.

Preprocessing of scRNA-seq datasets

In any scRNA seq pipeline, some basic steps of analysis are performed which were not in this project as the datasets were available as loom files which consisted of spliced and unspliced counts information with relevant metadata. There are many pipelines for this analysis and a common one is summarized in *Figure 1*. It starts by pre-processing .fastq files and mapping them to a human/mouse RNA velocity index. This is also pre-built for some sequencing technologies like 10X. This then generates an RNA count matrix of cells (rows) by genes (columns) in H5AD format which is a binary format of Anndata objects. Add metadata layers to the file format and merge other files together, if needed. The final anndata file can be saved in a loom format as well.

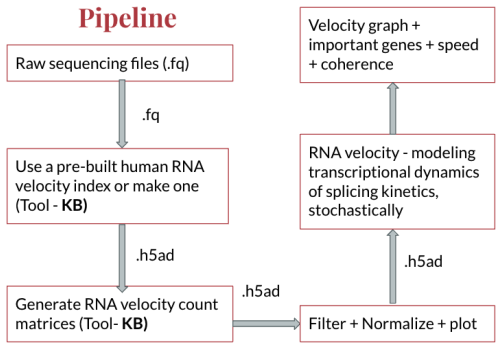


Figure 1: Pipeline to analyze and pre-process sc-RNA seq datasets to estimate RNA velocity.

Preprocessing of anndata from scvelo

The following steps of preprocessing were applied using the scvelo package:

1. Removing uninformative genes like in any sequencing analysis and selecting the ones by detection and high dispersions.
2. Normalization of the cells and applying log normalization.
3. Computing first and second-order moments. (La Manno et al. 2018; Bergen et al. 2020)

Velocity Analysis performed

The first step is to inspect the proportion of spliced and unspliced mRNA. This is followed by computing stochastic and dynamical models and make UMAP plots of the velocity embedding. This was followed by some plots to analyze the dynamics of top genes in cell types or top genes of the entire analysis. The speed and coherence give quantitative estimates of the speed of the cells.

Then the dynamical model was estimated for all the datasets and the kinetic rate parameters were recorded. This gives us numeric estimates of rates of transcription, splicing, and degradation of the cells along with switching time points, gene likelihood, and latent time points. This then helps to determine the cell's internal clock and determine its fate.

Overlapping genes inferred

The top genes from different analyses were merged and their overlaps were found. Further Enrichr (Chen et al. 2013) was used to estimate the pathways involved with the overlapping genes.

Results and Discussion

Depending on the protocol used (Drop-Seq, Smart-Seq), we typically have between 10%-25% of unspliced molecules containing intronic sequences. From *Figure 2A*, we find variations from the expected distribution. DG and HNSC show acceptable levels of mRNA contents. IN contrast, we see a very high incidence of unspliced RNA content in the NS dataset. Generally, having a large fraction of unspliced mRNAs is favorable even if unexpected. Would be nice to see if the data is

actually noisy later on as RNA velocity estimates usually do not do well in this case (due to timescale mismatch). ((La Manno et al. 2018; Bergen et al. 2020))

From *Figure 2B.* and as shown in ((La Manno et al. 2018; Bergen et al. 2020)) we observe that the dynamical model does better in finding the direction of cell states in the datasets. For instance, the G1 state cell direction is not clearly captured in the HNSC dataset by the stochastic model but works well with the dynamical model. *Figure 2C* represents the phase diagrams from a stochastic model of some overlapping genes further in the analysis. The black regression line corresponds to the estimated steady-state ratio. Velocity is determined by how much it deviates from this line. Positive velocity indicates that a gene is up-regulated, which occurs for cells that show a higher abundance of unspliced mRNA for that gene than expected in steady-state. Here, most genes for HNSC show upregulation through the cell cycle stages whereas, NS shows some level of downregulation in the *CDK1* gene and DG shows no clear trend for the selected genes.

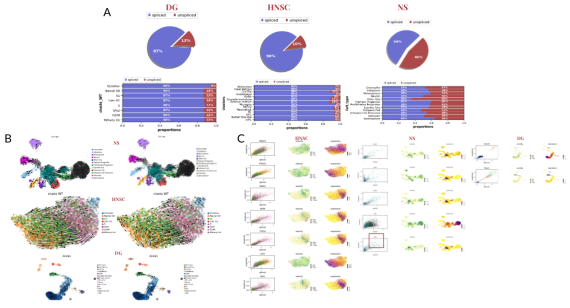


Figure 2: 2A: Spliced and unspliced mRNA content of the datasets. 2B: The left panel represents the stochastic models and the right panel represents the dynamic models of the datasets. 2C: Phase portraits of stochastic model overlapping genes.

Figure 3, provides insights into where cells differentiate at a slower/faster pace, and where the direction is un-/determined. For the HNSC dataset, on the clusters level, we find that differentiation is the fastest at the G1 stage and the next highest in the G2/M stage, keeping the pace during S/G2 and M/ Early G1 phases. For the NS dataset, we see that the Iridiophores show low velocity and pick up the pace at Schwann cells. Schwann cells help speed up the conduction of nerve impulses in the peripheral nervous system.

In the DG dataset, the highest velocity in astrocytes is observed. They play a vital role in information processing by altering neurotransmission speeds and changing the thickness of myelin. This explains the high velocity as compared to nIPC, which are progenitor cells - help in further differentiation further to create specialized cell types.

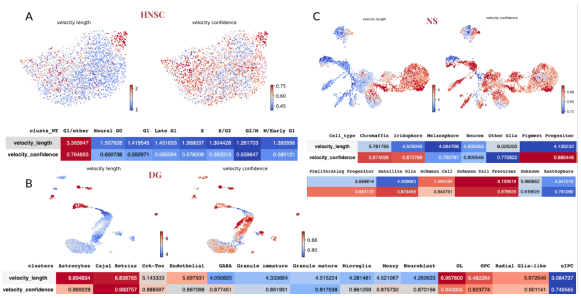
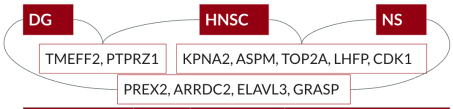


Figure 3: Velocity length and confidence estimates speed and coherence for all the datasets from the dynamical model



Overlapping genes

Cell cycles involved	Genes	Dataset	Cell types involved (top 10 only)
Early G1 dominant, S, G2	TMEFF2	Dentate gyrus	Oligodendrocytes
G1 dominant, Neural G0	PTPRZ1	Dentate gyrus	intermediate progenitor cells for neurons
G2/M transition dominant	KPNA2	Neuronal splicing data	neuron
G1, Neural G0 dominant	ASPM	Neuronal splicing data	none which are significant
S/G2 dominant	TOP2A	Neuronal splicing data	chromaffin
M/early G1 dominant	LHFP	Neuronal splicing data	Proliferating progenitor cells
M/early G1 dominant	CDK1	Neuronal splicing data	Satellite glia

Figure 4: Overlapping genes and the cell types they belong to summarized with the specific cell cycles they are dominant in.

Driver genes display pronounced dynamic behavior and are systematically detected via their characterization by high likelihoods in the dynamic model. Moreover, partial gene likelihoods can be computed for each cluster of cells to enable cluster-specific identification of potential drivers. For instance, the gene *KPNA2* is a marker of the G2/M phase. This gene has a role in nucleocytoplasmic transport. The G2-phase checkpoint, also known as the G2/M-phase checkpoint, has the function of preventing cells with damaged DNA, lasting from the G1 and S phases or generated in G2, from undergoing mitosis. *KPNA2* is overexpressed in various cancers, which is associated with poor prognosis. In addition, it has been shown to promote tumor formation and progression by participating in cell differentiation, proliferation, apoptosis, immune response, and viral infection. The gene *TOP2A* predominant in the S/G2 pathway and *CDK1* in early G1 show up as a retinoblastoma gene in cancer in the pathway analysis. *CDK1* is essential for G1/S and G2/M phase transitions of the eukaryotic cell cycle and is also involved in the lung, pancreatic, breast, and bladder cancers. Higher expression of *TOP2A* was higher in proliferative subtypes of breast cancers such as triple-negative and HER2-enriched diseases than in luminal type. These analyses show us that understanding the transcriptional dynamics at the single-cell level can help infer biology about disease states in future studies. (*Figure 4*)

Therefore, RNA velocity is a powerful tool to apply to scRNA seq datasets to infer the dynamics of the cell states. It is advanced as compared to the prior trajectory analysis which finds paths between aggregates of cells based on the most variable genes. It does not provide direction estimates like velocity analysis. Whereas, velocity analysis used here describes the direction, speed, and rates of transcription which make it clear to observe the transition between states of the cells in real-time at various time points. (Saelens et al. 2019)

Code and Presentation

https://github.com/PallaviSurana1/RNA-velocity-analysis

References

Bergen, Volker, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. 2020. "Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling." *Nature Biotechnology* 38 (12): 1408–14.

Bergen, Volker, Ruslan A. Soldatov, Peter V. Kharchenko, and Fabian J. Theis. 2021. "RNA Velocity—current Challenges and Future Perspectives." *Molecular Systems Biology*. https://doi.org/10.15252/msb.202110282.

Burgess, Darren J. 2018. "Full Speed Ahead for Single-Cell Analysis." *Nature Reviews. Genetics* 19 (11): 668–69.

Chen, Edward Y., Christopher M. Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R. Clark, and Avi Ma'ayan. 2013. "Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool." *BMC Bioinformatics* 14 (1): 1–14.

Hochgerner, Hannah, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. 2018. "Conserved Properties of Dentate Gyrus Neurogenesis across Postnatal Development Revealed by Single-Cell RNA Sequencing." *Nature Neuroscience* 21 (2): 290–99.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, et al. 2018. "RNA Velocity of Single Cells." *Nature* 560 (7719): 494–98.

"Mapping Transcriptomic Vector Fields of Single Cells." 2022. *Cell* 185 (4): 690–711.e45.

O'Connor, S. A., H. M. Feldman, S. Arora, P. Hoellerbauer, C. M. Toledo, P. Corrin, L. Carter, et al. 2021. "Neural G0: A Quiescent-like State Found in Neuroepithelial-Derived Cells and Glioma." *Molecular Systems Biology* 17 (6). https://doi.org/10.15252/msb.20209522.

Saelens, Wouter, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. 2019. "A Comparison of Single-Cell Trajectory Inference Methods." *Nature Biotechnology* 37 (5): 547–54.

Zeisel, Amit, Wolfgang J. Köstler, Natali Molotski, Jonathan M. Tsai, Rita Krauthgamer, Jasmine Jacob-Hirsch, Gideon Rechavi, et al. 2011. "Coupled Pre-mRNA and mRNA Dynamics Unveil Operational Strategies Underlying Transcriptional Responses to Stimuli." *Molecular Systems Biology* 7: 529.