Network Analysis of Single-cell Transcriptome

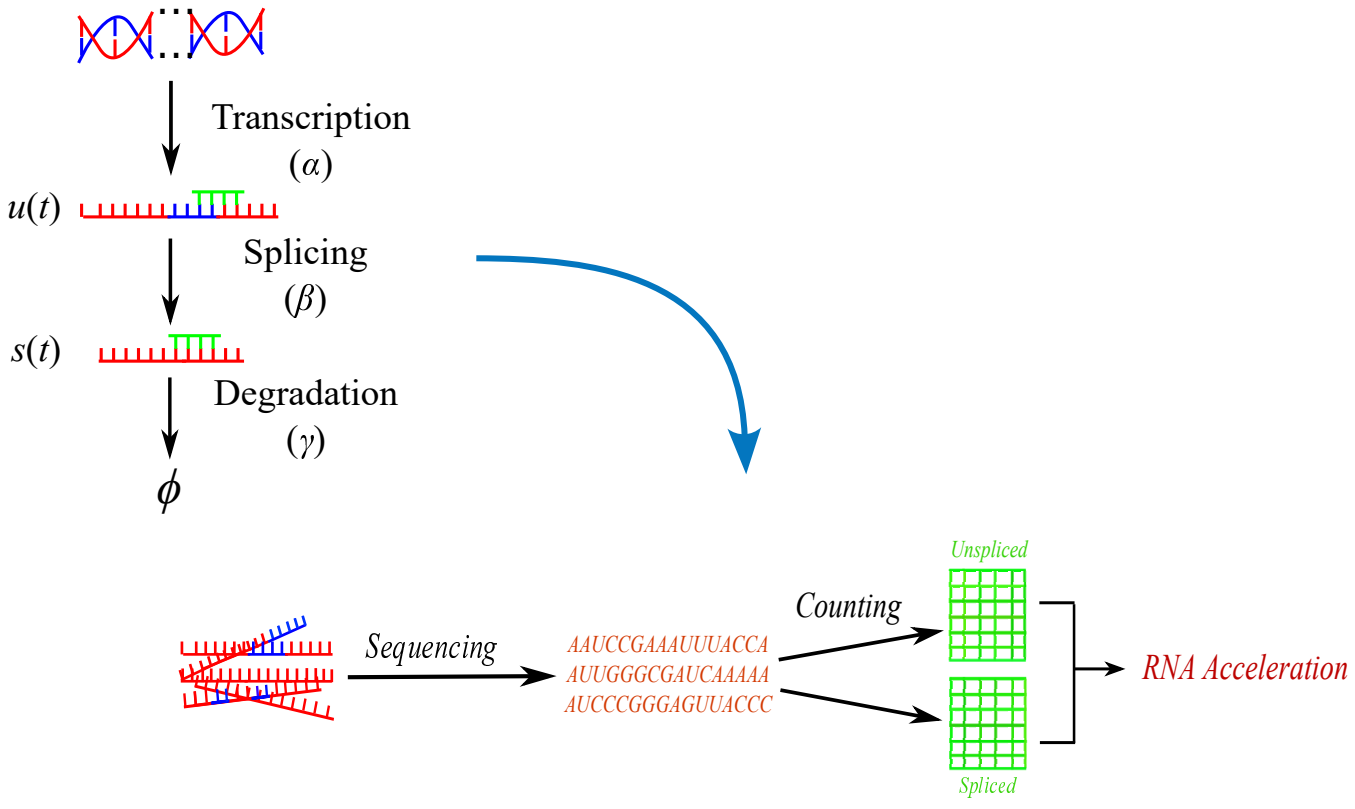
Pradip Das
M.Tech (CS), Roll No: CS2115

Under the supervision of
Dr. Malay Bhattacharyya

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

June 20, 2023

Overview of RNA-acceleration



Cellular Biology to computation

Background

- The use of single-cell RNA sequencing (scRNA-seq) has significantly enhanced our understanding of biological systems by representing them as complex interconnected networks.
- Single-cell mRNA sequencing allow us unbiased, high throughput, and high-resolution transcriptomic analysis of individual cells.
- The analysis of RNA velocity is a commonly utilized approach to deduce temporal dynamics in single-cell gene expression data.
- RNA molecules in a cell have both unspliced and spliced forms.

Background

- Comparing the relative abundance of unspliced and spliced RNA forms we get RNA velocity.
- RNA velocity estimates the speed and trajectory of changes in gene expression, allowing us to infer the future state of individual cells.
- The initial meaning of RNA velocity is the derivative of gene expression state with respect to time.
- In the previous work, the major assumption is that the connection between the levels of spliced and unspliced pre-mRNA expression can be utilized to deduce whether a particular gene is undergoing up-regulation or down-regulation or is in a steady state of expression.

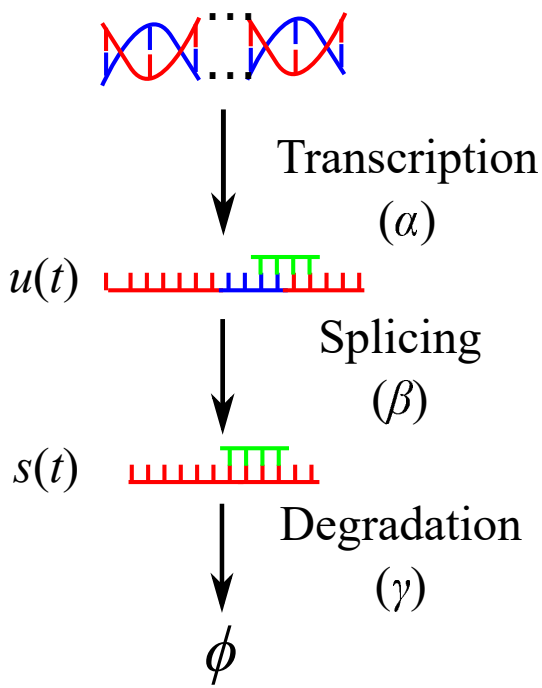
The Problem

- The objective of this problem is to capture the rate at which specific cells undergo differentiation into other cell types.
- The process begins with DNA being transcribed into unspliced RNA or pre-mRNA (u) at transcription rate α.
- The unspliced mRNA molecule undergoes further processing and produce spliced mRNA (s) at a splicing rate β.
- After that, the mature mRNA molecule may undergo degradation at a rate γ.
- The degradation of mRNA is an essential part of gene regulation, as it helps control gene expression levels and maintain cellular homeostasis.

The Problem

- The rate and mechanisms of mRNA degradation can vary depending on cellular conditions, external signals, and specific mRNA sequences.
- RNA sequences are generated from both spliced and unspliced RNA.
- By counting the occurrences of genes corresponding to spliced and unspliced RNA, we obtain the gene counts for each.
- We compute the RNA acceleration using these two dataframes.

Modeling RNA-acceleration



$$\frac{du}{dt} = \alpha - \beta u$$

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

$$\frac{d^2 s}{dt^2} = \alpha\,\beta - \beta(\beta + \gamma)u + \gamma^2 s$$

Modeling the dynamics of transcription captures the process of both induction and repression of unspliced pre-mRNAs, their subsequent transformation into mature or spliced mRNAs, and their eventual degradation.

Proposed Method

This whole process (transcription, splicing and degradation) can be represented by two differential equations:

$$\frac{du(t)}{dt} = \alpha(t) - \beta\ u(t) \tag{1}$$

$$\frac{ds(t)}{dt} = \beta\ u(t) - \gamma\ s(t) \tag{2}$$

corresponding to each gene and independent of all other genes.

Generally equation 2 represents the RNA velocity.

Proposed Method

Differentiating equation (2) and re-arranging of its term we get

$$\frac{d^2 s(t)}{dt^2} = \alpha(t)\beta - \beta(\beta + \gamma)\ u(t) + \gamma^2\ s(t) \tag{3}$$

Define equation (3) as RNA acceleration.

The challenges are to estimate the parameters α, β and γ.

Pre-processing

- The count matrices were normalized for size by dividing them by the median of the total molecule counts across cells.
- The best n (generally $n = 2{,}000$) highly variable genes are chosen out of those that pass a least threshold of 20 expressed counts commonly for spliced and unspliced mRNA.
- Calculate a nearest neighbor graph in PCA space on logarithmized spliced counts based on Euclidean distances.

Parameter Estimation

- In the velocity framework, the author claims that α is unknown and cannot be easily estimated.
- In steady-state, where $\frac{du}{dt} = 0$, we can determine α and we assume a common, constant rate of splicing:

$$\alpha = u$$
$$\beta = 1$$

- For estimating γ we consider three approaches: Steady-state, Stochastic and Revised Stochastic.

Estimation of γ: Steady-state Model

Steady-state scenario for equation (2) is $\frac{ds}{dt} = 0$. Form this we get $\gamma = \frac{\beta u(t)}{s(t)}$. Setting $\beta = 1$, $\gamma = \frac{u(t)}{s(t)}$. To estimate γ we use equation $\gamma = \frac{u(t)+1}{s(t)+1}$.

Estimation of γ: Stochastic Model

Regarding transcription, splicing, and degradation as probabilistic occurrences, we can analyze the probabilities of all potential reactions associated with these events occurring within a short time interval $(t,\ t + dt]$ are given as follows:

$$\begin{aligned} \mathbb{P}(u_{t+dt} = u_t + 1,\ s_{t+dt} = s_t) &= \alpha\ dt \\ \mathbb{P}(u_{t+dt} = u_t - 1,\ s_{t+dt} = s_t + 1) &= \beta u_t\ dt \\ \mathbb{P}(u_{t+dt} = u_t,\ s_{t+dt} = s_t - 1) &= \gamma s_t\ dt \end{aligned} \tag{4}$$

where we represented $u_t = u(t)$ and $s_t = s(t)$.

Estimation of γ: Stochastic Model

From equation 4 the time derivative for the uncentered moment $< u_t^l s_t^k >= \mathbb{E}[u_t^l s_t^k]$is derived as

$$\begin{aligned}\frac{d < u_t^l s_t^k >}{dt} = &< \alpha((u_t + 1)^l s_t^k - u_t^l s_t^k) > + \\ &< \beta u_t((u_t - 1)^l (s_t + 1)^k - u_t^l s_t^k) > + \\ &< \gamma s_t(u_t^l (s_t - 1)^k - u_t^l s_t^k) >\end{aligned}$$

Estimation of γ: Stochastic Model

Thus, the first and second order dynamics are given by

$$\begin{aligned}
\frac{d < u_t >}{dt} &= \alpha - \beta < u_t >, \\
\frac{d < s_t >}{dt} &= \beta < u_t > -\gamma < s_t >, \\
\frac{d < u_t^2 >}{dt} &= \alpha + 2\alpha < u_t > +\beta < u_t > -2\beta < u_t^2 >, \\
\frac{d < u_t s_t >}{dt} &= \alpha < s_t > +\beta < u_t^2 > +\beta < u_t s_t > -\gamma < u_t s_t >, \\
\frac{d < s_t^2 >}{dt} &= \beta < u_t > +2\beta < u_t s_t > +\gamma < s_t > -2\gamma < s_t^2 >,
\end{aligned} \tag{5}$$

Estimation of γ: Stochastic Model

The moments for each cell are calculated based on a predefined number of closest neighboring cells to the specific cell. These expansions can be readily applied to the steady-state model. By utilizing first and second-order moments, the steady-state ratio can be derived from the following system.

$$\begin{pmatrix} < u_t > \\ < u_t > +2 < u_t s_t > \end{pmatrix} = \gamma \begin{pmatrix} < s_t > \\ 2 < s_t^2 > - < s_t > \end{pmatrix} + \vec{\epsilon}$$

where $E[\vec{\epsilon}|\vec{s}] = 0$, $Cov[\vec{\epsilon}|\vec{s}] = \Omega$ and $\vec{u} = \begin{pmatrix} < u_t > \\ < u_t > +2 < u_t s_t > \end{pmatrix}$, $\vec{s} = \begin{pmatrix} < s_t > \\ 2 < s_t^2 > - < s_t > \end{pmatrix}$

Estimation of γ: Stochastic Model

The steady-state ratio can be explicitly determined through the use of generalized least squares, resulting in the following expression.

$$\gamma = (\vec{s}^T \Omega^{-1} \vec{s})^{-1} (\vec{s}^T \Omega^{-1} \vec{u}) \tag{6}$$

Estimation of γ: Revised Stochastic Model

We obtain γ by same method used in stochastic model but here we take a new spliced/unspliced count matrices defined by $u_{new}(t) = u(t) + 1$ and $s_{new}(t) = s(t) + 1$.

Visualization of the RNA acceleration

- Create low-dimensional embeddings for the cells.
- Chose a set of principal components (PCs) and applied Uniform Manifold Approximation and Projection (UMAP) using the top 30 PCs for all the datasets.
- Construct Partition-based Graph Abstraction (PAGA) graph to captures the global structure and connectivity of cells corresponding to RNA acceleration.

Gene set enrichment analysis (GSEA)

- Created volcano plot corresponding to the cells that have low generation/exchange rate.
- GSEA is performed based on two different metrics, namely (i) False Discovery Rate (FDR) and (ii) Fold enrichment.
  - FDR is determined using the nominal p-value obtained from the hypergeometric test and a threshold value of 0.05 was considered for the current analysis.
  - Fold enrichment is defined as the percentage of genes in a given list that belong to a specific pathway, divided by the corresponding percentage in the background.

Gene set enrichment analysis (GSEA)

To better understand the significance of the genes upregulated and downregulated in a particular cells, we studied the Gene Ontology (GO) results from GSEA.

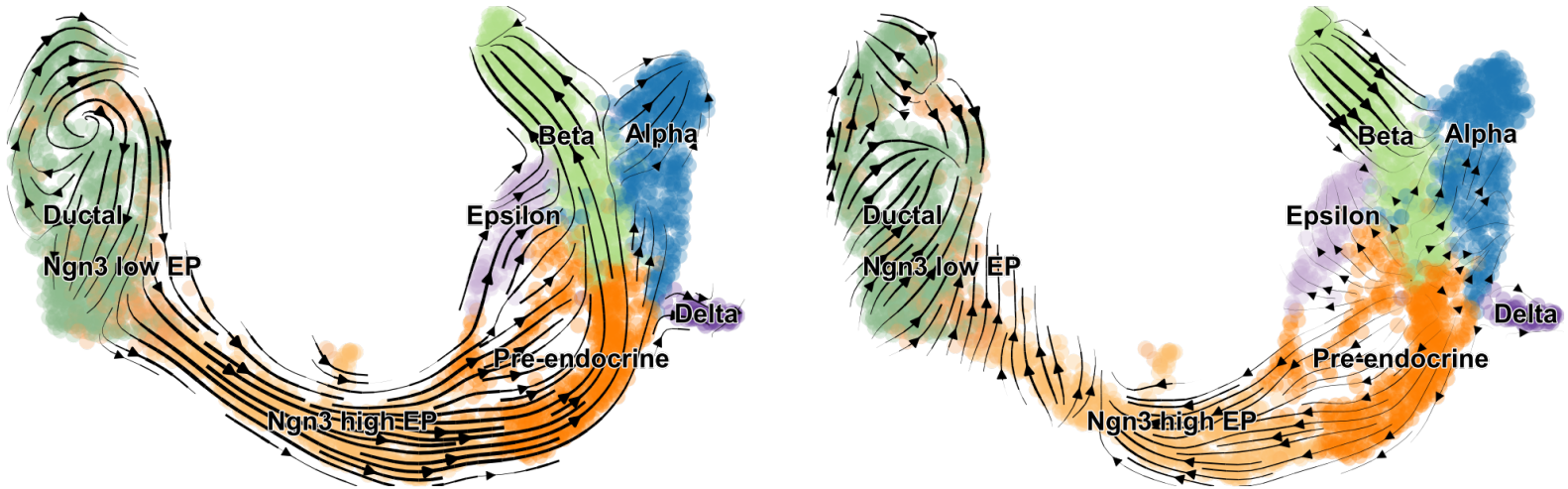
- GO Biological Processes
- GO Cellular Component
- GO Molecular Function
- KEGG Pathways

Dataset Details

The essential metadata, physiological information and parameters employed for the three datasets are detailed below. Notably, all the three datasets are available in the scVelo library.

- **Pancreatic Endocrinogenesis**
- **Dentate Gyrus Neurogenesis**
- **Mouse Gastrulation Subset to Erythroid Lineage**

Analysis on the Pancreatic Endocrinogenesis dataset



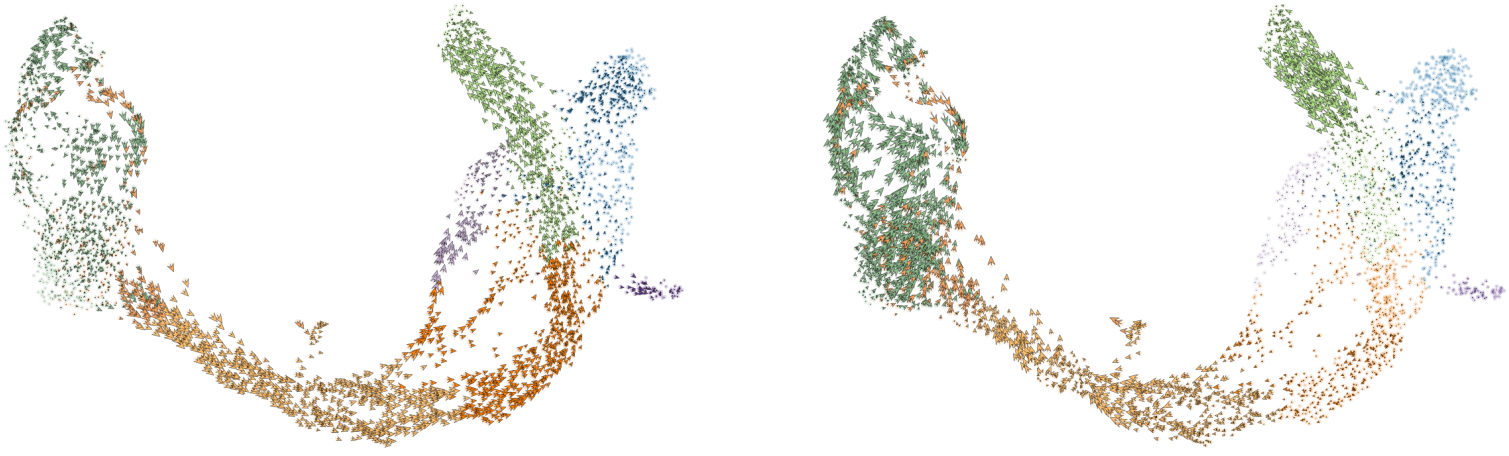
(a) (b)

Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** (a) Velocity obtained from the stochastic model for pancreatic endocrinogenesis are visualized as lines of motion in a UMAP-based embedding. (b) Accelerations obtained from the stochastic model for pancreatic endocrinogenesis are visualized as lines of motion in a UMAP-based embedding.

Analysis on the Pancreatic Endocrinogenesis dataset

- On carefully comparing the velocity output with the acceleration output, we are able to highlight that the change of velocity at which β-cells are generated are different than the others.
- The rate at which β-cells degrade is in contrast to the rate at which α, δ and ϵ-cells degrades (and so the rate at which pre-endocrine degrades).
- The degradation of β-cells is much more controlled than the others.
- It is interestingly known from the existing literature that β-cells produce and secrete insulin in a firmly regulated fashion, to maintain the circulation of glucose concentrations in the physiological range.

Analysis on the Pancreatic Endocrinogenesis dataset



(a) (b)

Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** (a) The foremost fine-grained determination of the velocity vector field we get at single-cell level. (b) The foremost fine-grained determination of the acceleration vector field we get at single-cell level.

Analysis on the Pancreatic Endocrinogenesis dataset

- Our model highlights the quantities of α-, β-, and δ-cells that are generated.
- It is known that β-, α- and δ-cells make up approximately 70%, 15-20% and 5-10%, respectively of the cells in islets for the species mouse.

Analysis on the Pancreatic Endocrinogenesis dataset

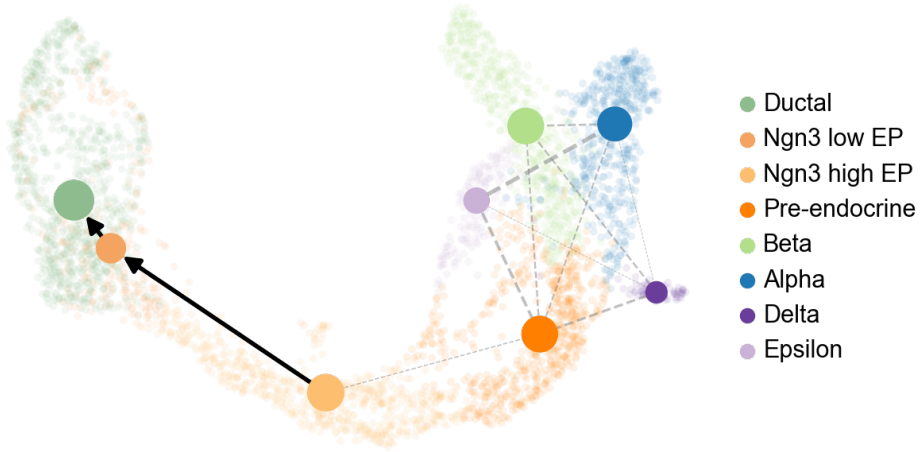


Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** PAGA graph corresponding acceleration model.

Analysis on the Pancreatic Endocrinogenesis dataset

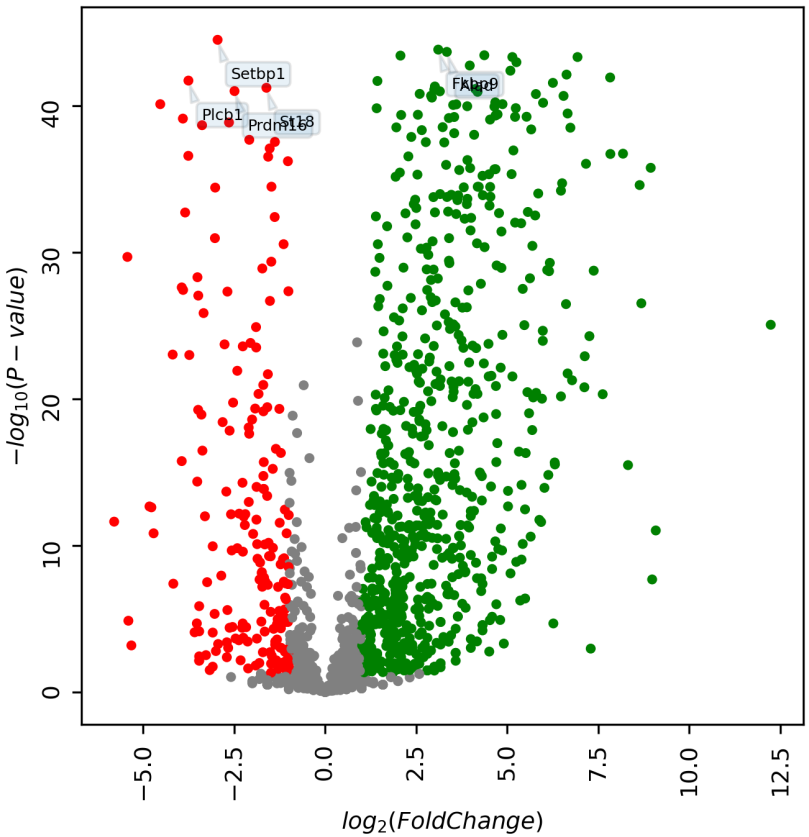


Figure: Volcano plot showing differential expression for each gene in spliced β-cells compared with unspliced β-cells. Up-regulated genes are highlighted in green and down-regulated genes are highlighted in red.

Analysis on the Pancreatic Endocrinogenesis dataset

Top three enrichment of β-cell up-regulated genes in GO biological process

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
0.001311025	4	12	30.9	Pyrimidine deoxyribonucleotide catabolic proc.
0.000395734	5	17	27.3	Response to interleukin-7
0.000395734	5	17	27.3	Cellular response to interleukin-7

Top three enrichment of β-cell down-regulated genes in GO biological process

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
0.000479963	5	20	23.18	Vocalization behavior
0.000688333	8	84	8.83	Protein localization to synapse
0.000688333	11	178	5.73	Signal release from synapse

Analysis on the Pancreatic Endocrinogenesis dataset

Top three enrichment of β-cell up-regulated genes in GO Cellular Component

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathway**
9.00E-06	5	13	35.66	Oligosaccharyltransferase complex
0.000250906	7	64	10.14	Cytosolic large ribosomal subunit
5.14E-07	12	112	9.93	Cytosolic ribosome

Top three enrichment of β-cell down-regulated genes in GO Cellular Component

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathway**
0.003014333	4	25	14.84	Inhibitory synapse
0.002668594	6	70	7.95	Excitatory synapse
8.89E-06	19	393	4.48	Postsynaptic density

Analysis on the Pancreatic Endocrinogenesis dataset

Top three enrichment of β-cell up-regulated genes in GO Molecular Function

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
0.01338297	2	3	61.82	Protein-disulfide reductase (glutathione) activity
0.018982353	2	4	46.36	Inorganic diphosphatase activity
0.005539047	3	8	34.77	Deoxyribonucleotide binding

Top three enrichment of β-cell down-regulated genes in GO Molecular Function

Enrichment FDR	nGenes	Pathway Genes	Fold Enrichment	Pathway
0.012155208	2	3	61.82	Calmodulin-dependent cyclic-nucleotide phosphodiesterase activity
0.012155208	2	3	61.82	Calcium- and calmodulin-regulated 3 ,5 -cyclic-GMP phosphodiesterase activity
0.012155208	3	14	19.87	Gamma-catenin binding

Analysis on the Pancreatic Endocrinogenesis dataset

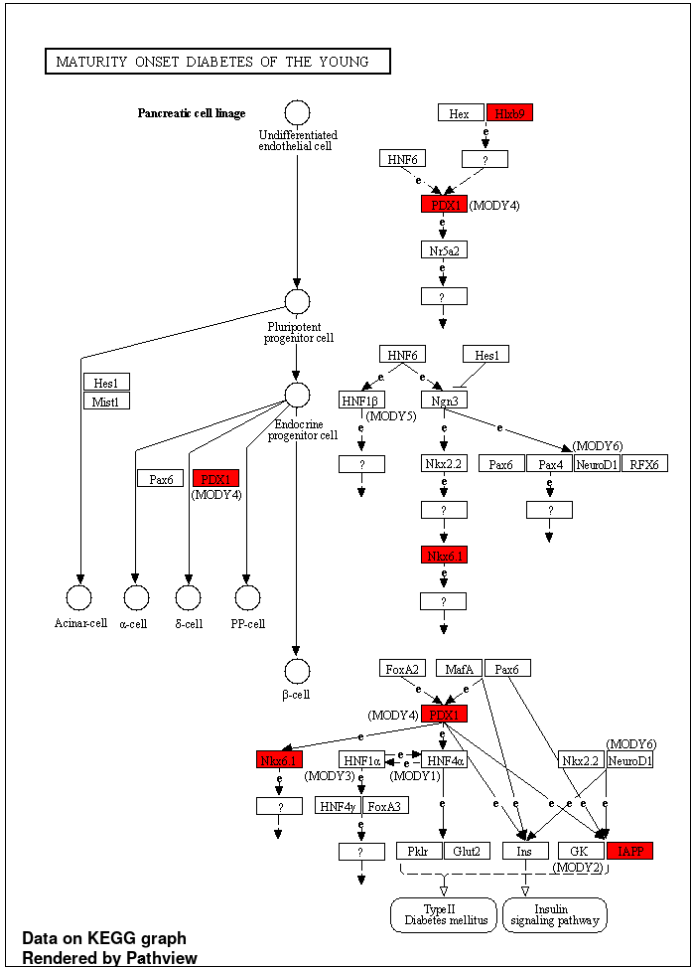


Figure: The *Maturity onset diabetes of the young* KEGG pathway enrichment of genes upregulated in the β-cells of pancreas.

Analysis on the Pancreatic Endocrinogenesis dataset

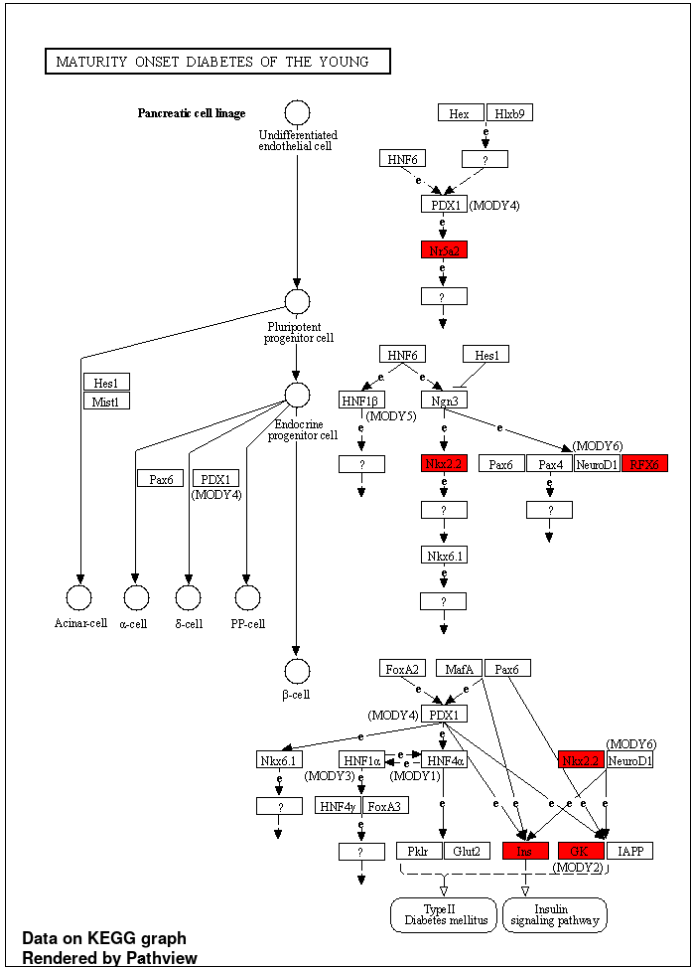


Figure: The *Maturity onset diabetes of the young* KEGG pathway enrichment of genes downregulated in the β-cells of pancreas.

Analysis on the Pancreatic Endocrinogenesis dataset

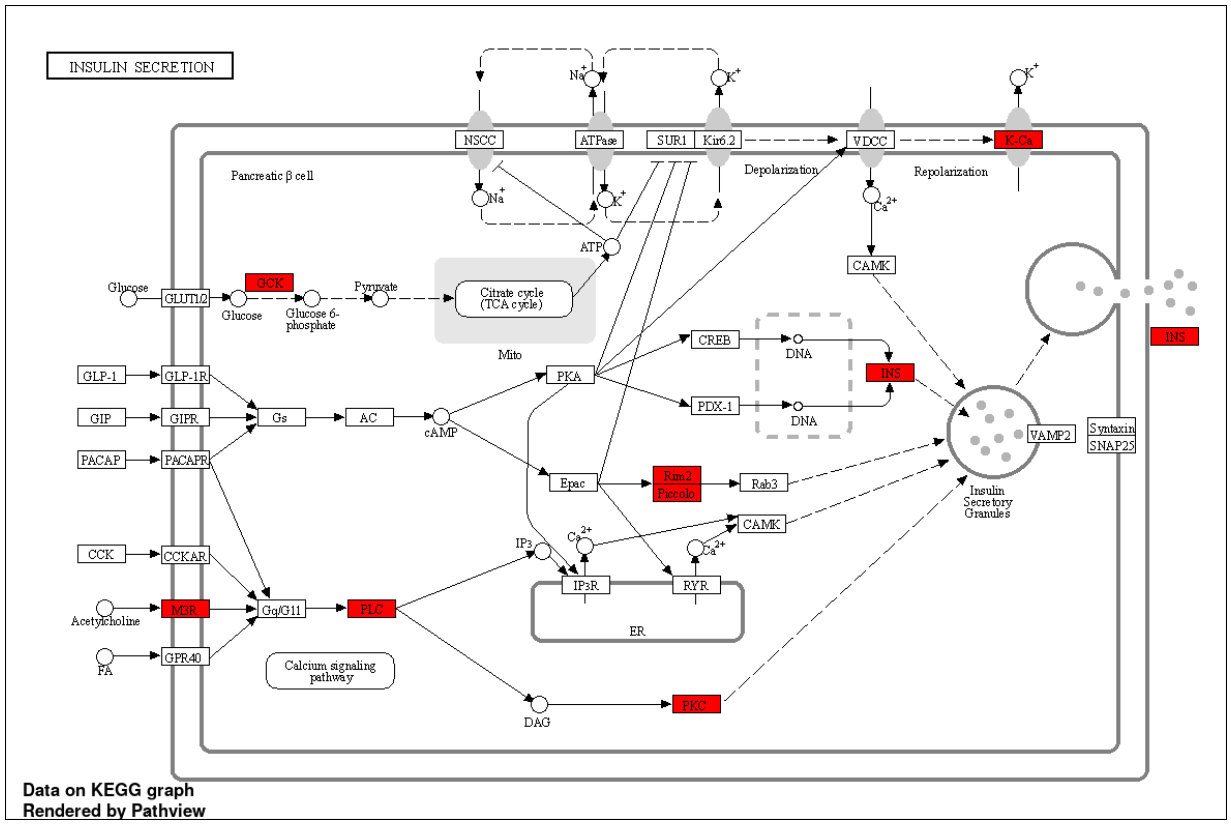


Figure: The *Insulin secretion* KEGG pathway enrichment of genes downregulated in the β-cells of pancreas.

Analysis on the Dentate Gyrus Neurogenesis dataset

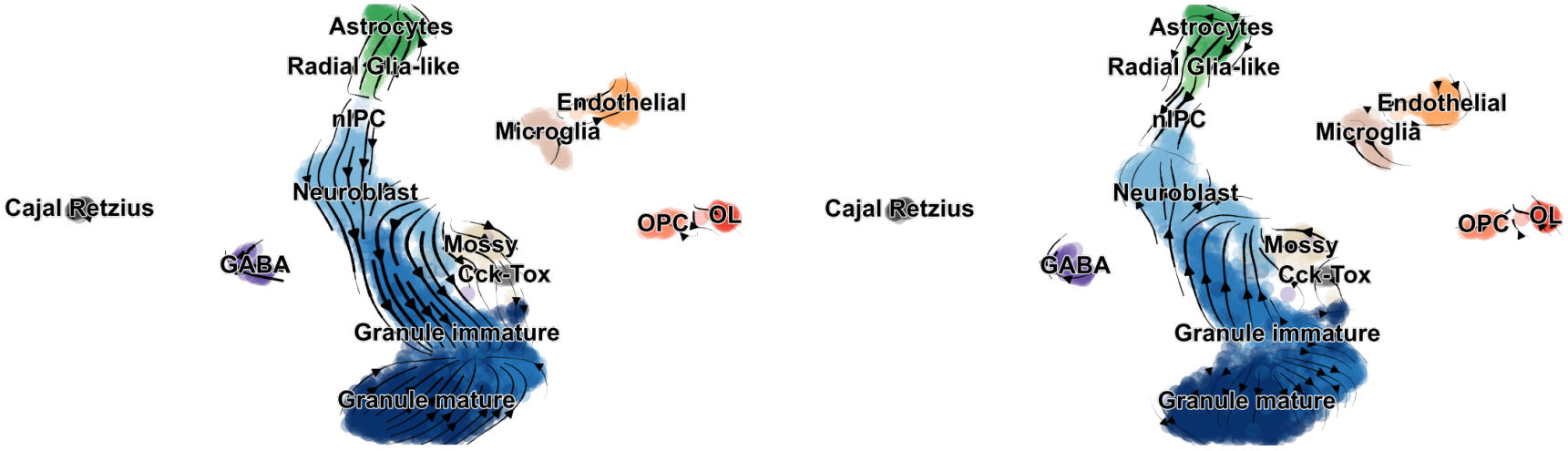


Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** (a) Velocity obtained from the stochastic model for dentate gyrus neurogenesis are visualized as lines of motion in a UMAP-based embedding. (b) Accelerations obtained from the stochastic model for dentate gyrus neurogenesis are visualized as lines of motion in a UMAP-based embedding.

Analysis on the Dentate Gyrus Neurogenesis dataset

(a) (b)

Figure: **RNA acceleration recapitulates dynamics of Dentate Gyrus cell differentiation.** (a) The foremost fine-grained determination of the velocity vector field we get at single-cell level. (b) The foremost fine-grained determination of the acceleration vector field we get at single-cell level.

Analysis on the Dentate Gyrus Neurogenesis dataset

- On carefully comparing the velocity output with the acceleration output, we are able to highlight that the change of velocity at which oligodendrocytes (OLs)-cells are generated are different.
- It is known that after the initial population of oligodendrocytes is established during childhood.
- Studies have shown a significantly low rate of generation/ exchange of oligodendrocytes in the white matter of the brain.
- Our model correctly identifies the fact that generation/ exchange rate of oligodendrocytes is low, in RNA-acceleration, the direction observed in oligodendrocyte cells is opposite to the direction observed in RNA-velocity in oligodendrocyte cells.

Analysis on the Dentate Gyrus Neurogenesis dataset

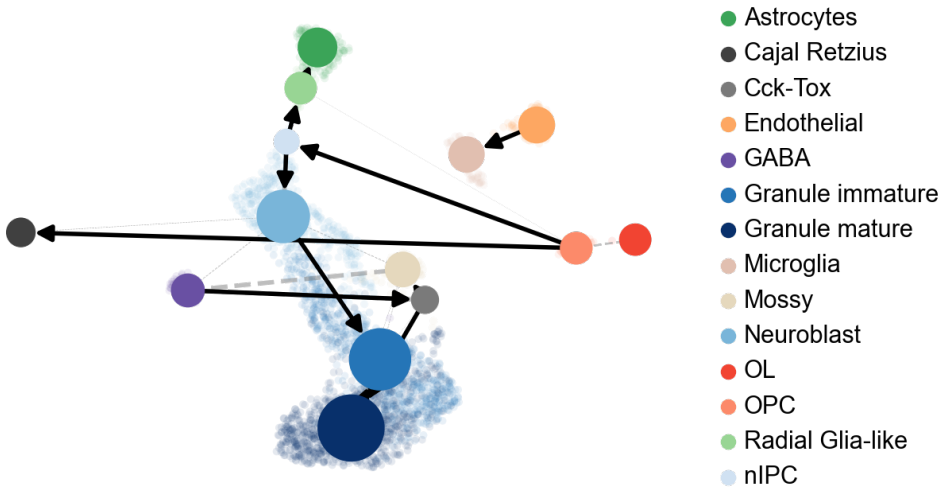


Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** PAGA graph corresponding acceleration model.

Analysis on the Dentate Gyrus Neurogenesis dataset

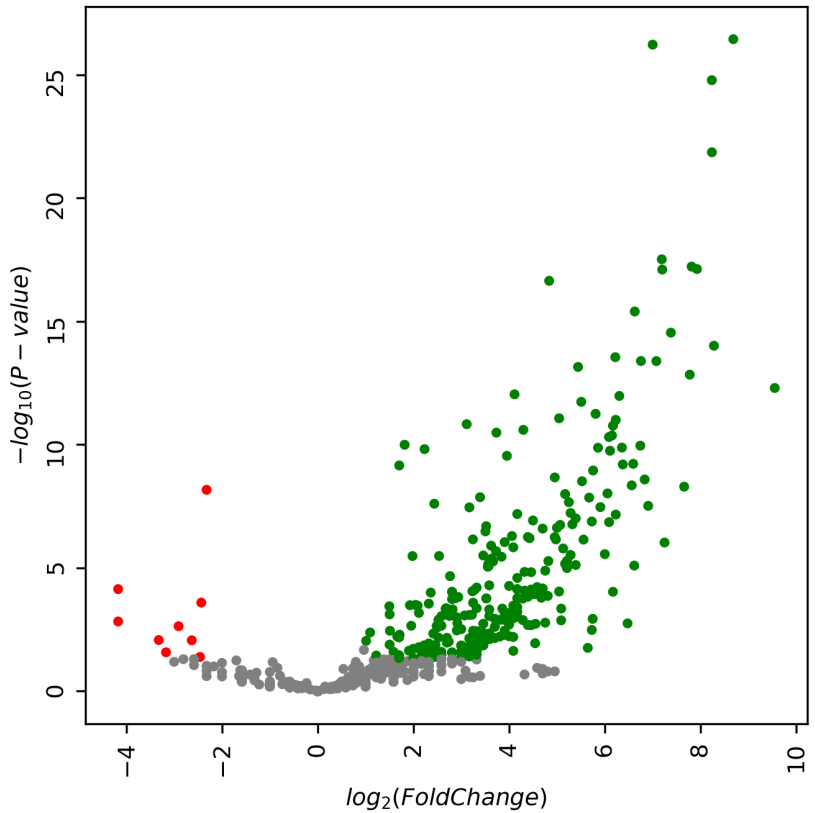


Figure: Volcano plot showing differential expression for each gene in spliced oligodendrocyte cell cells compared with unspliced oligodendrocyte cells. Up-regulated genes are highlighted in green and down-regulated genes are highlighted in red.

Analysis on the Dentate Gyrus Neurogenesis dataset

Top three enrichment of oligodendrocyte-cell up-regulated genes in GO biological process

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathway**
4.67E-13	17	101	15.48	Cytoplasmic translation
7.45E-08	11	69	14.66	Ribosomal small subunit biogenesis
2.48E-06	13	151	7.92	Neg. reg. of protein-containing complex assembly

Analysis on the Dentate Gyrus Neurogenesis dataset

Top three enrichment of oligodendrocyte-cell up-regulated genes in GO Cellular Component

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathway**
5.17E-07	6	17	32.45	Proton-transporting two-sector ATPase complex, catalytic domain
1.13E-12	11	32	31.61	Polysomal ribosome
2.08E-15	14	45	28.61	Cytosolic small ribosomal subunit

Analysis on the Dentate Gyrus Neurogenesis dataset

Top three enrichment of oligodendrocyte cell up-regulated genes in GO Molecular Function

Enrichment FDR	**nGenes**	**Pathway Genes**	**Fold Enrichment**	**Pathway**
0.00037938	3	4	68.96	Rho GDP-dissociation inhibitor binding
2.70E-30	34	172	18.18	Structural constituent of ribosome
0.000782339	5	30	15.32	G protein activity

Analysis on the Dentate Gyrus Neurogenesis dataset

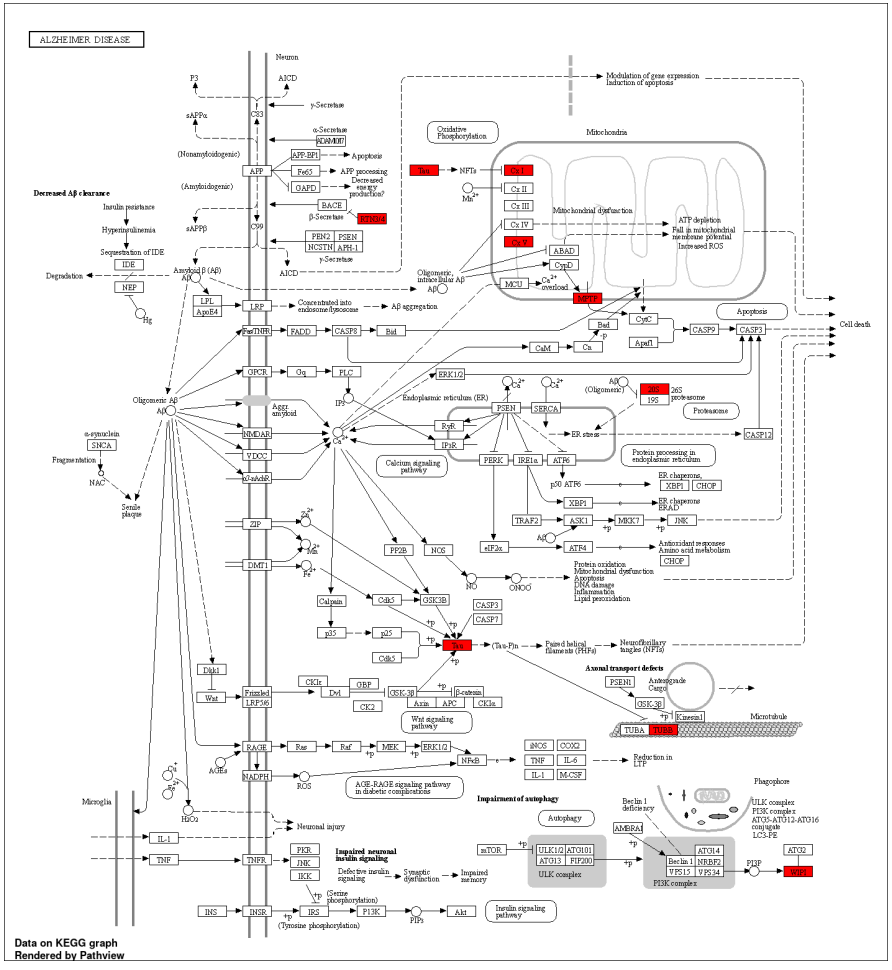


Figure: The *Alzheimer's disease* KEGG pathway enrichment of genes upregulated in the oligodendrocytes cell.

Analysis on the Dentate Gyrus Neurogenesis dataset

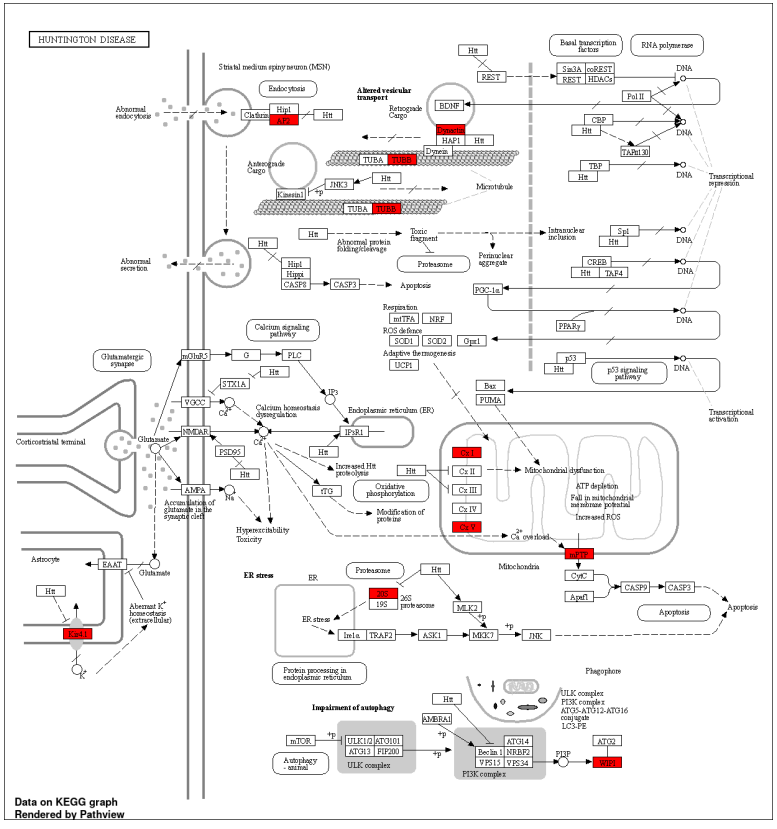


Figure: The *Huntington's disease* KEGG pathway enrichment of genes upregulated in the oligodendrocytes cell.

Analysis on the Dentate Gyrus Neurogenesis dataset

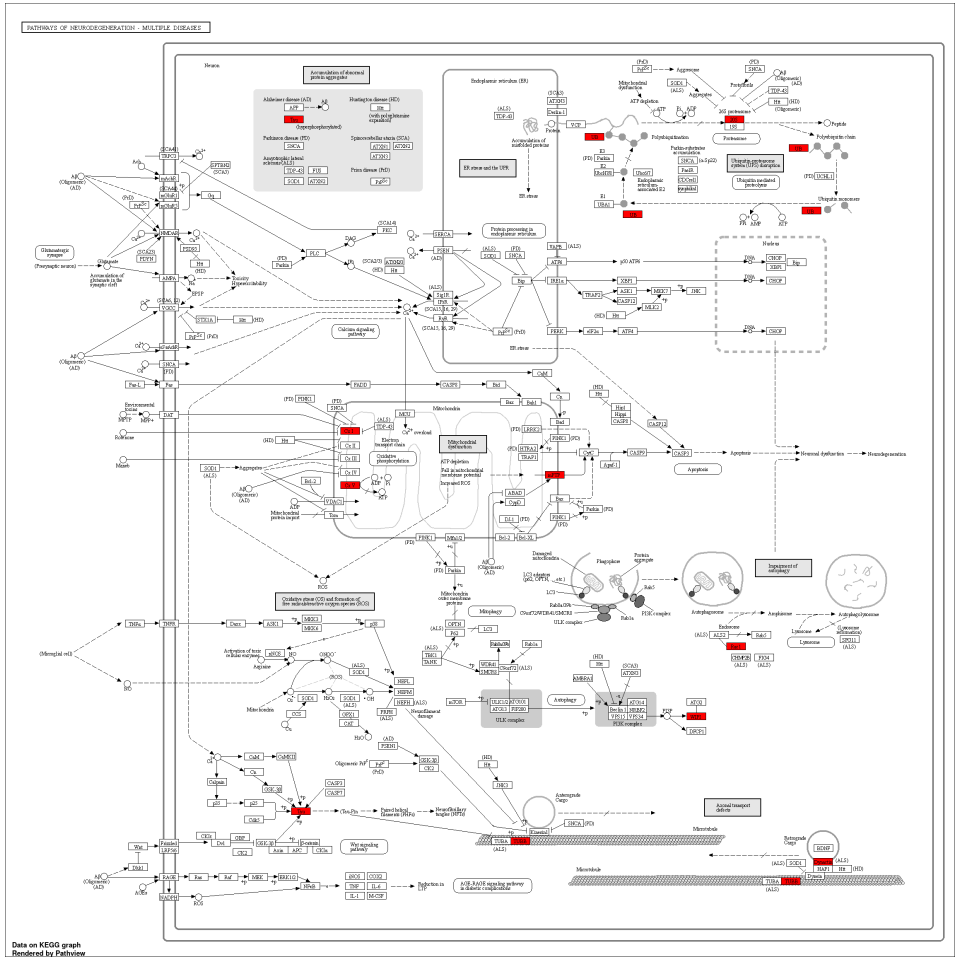


Figure: The *Pathways of neurodegeneration* KEGG pathway enrichment of genes upregulated in the oligodendrocytes cell.

Analysis on the Dentate Gyrus Neurogenesis dataset

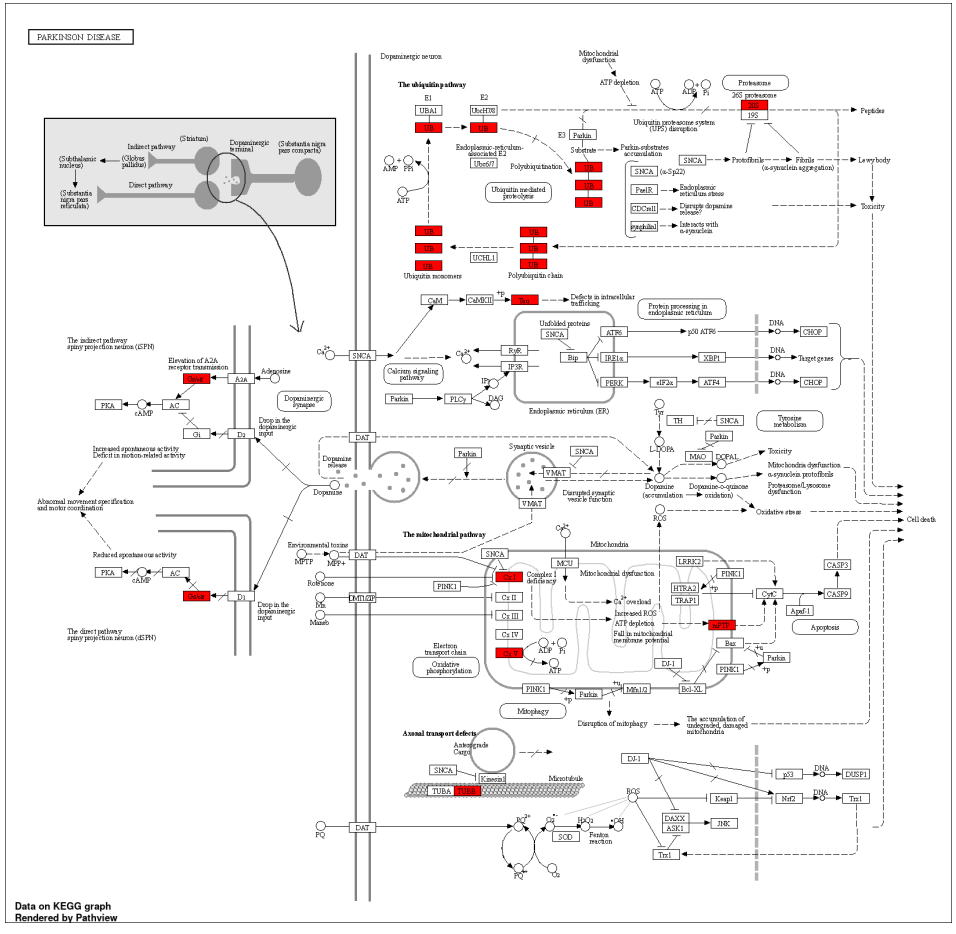


Figure: The *Parkinson's disease* KEGG pathway enrichment of genes upregulated in the oligodendrocytes cell.

Analysis on the Dentate Gyrus Neurogenesis dataset

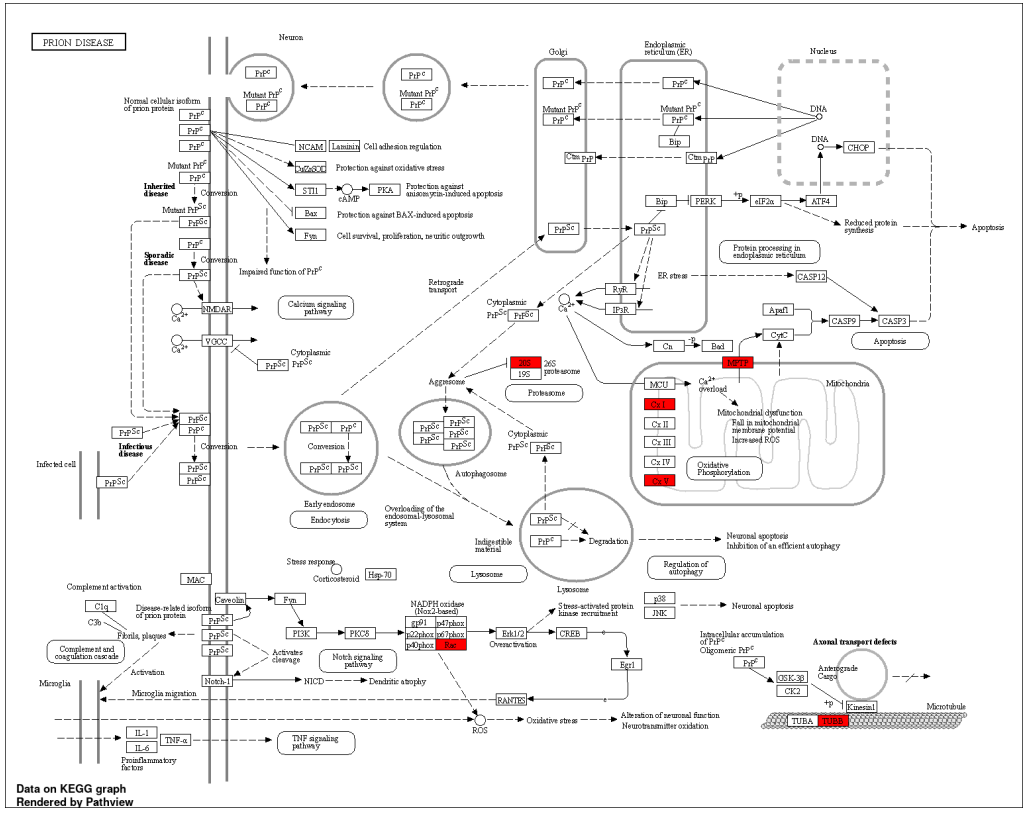
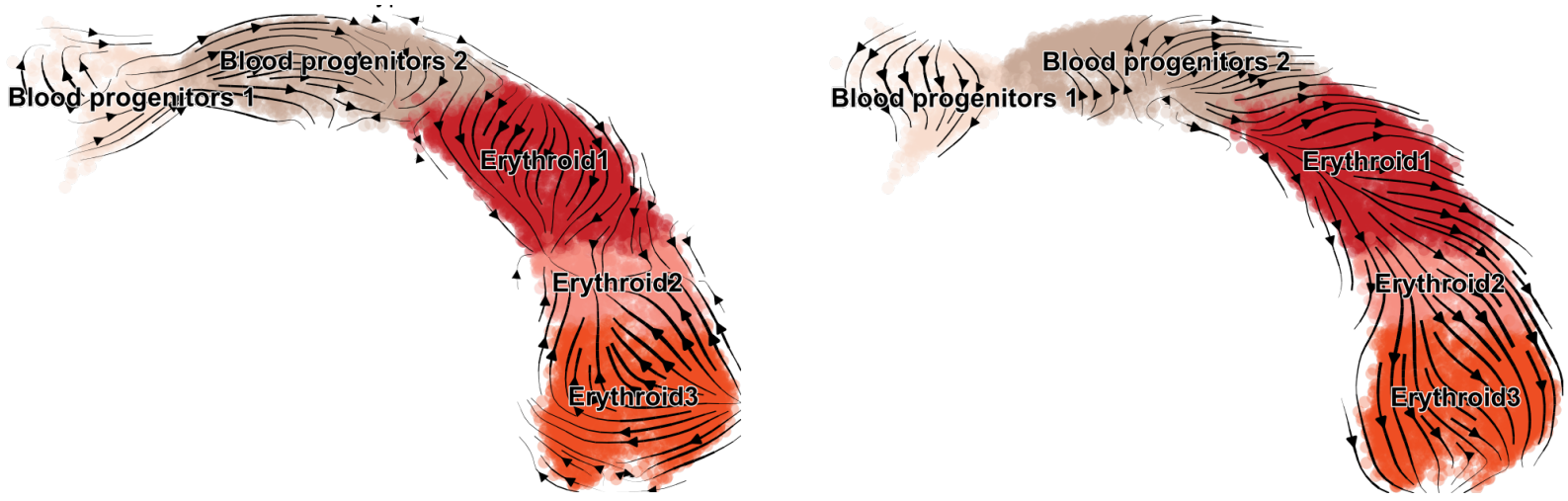


Figure: The *Prion disease* KEGG pathway enrichment of genes upregulated in the oligodendrocytes cell.

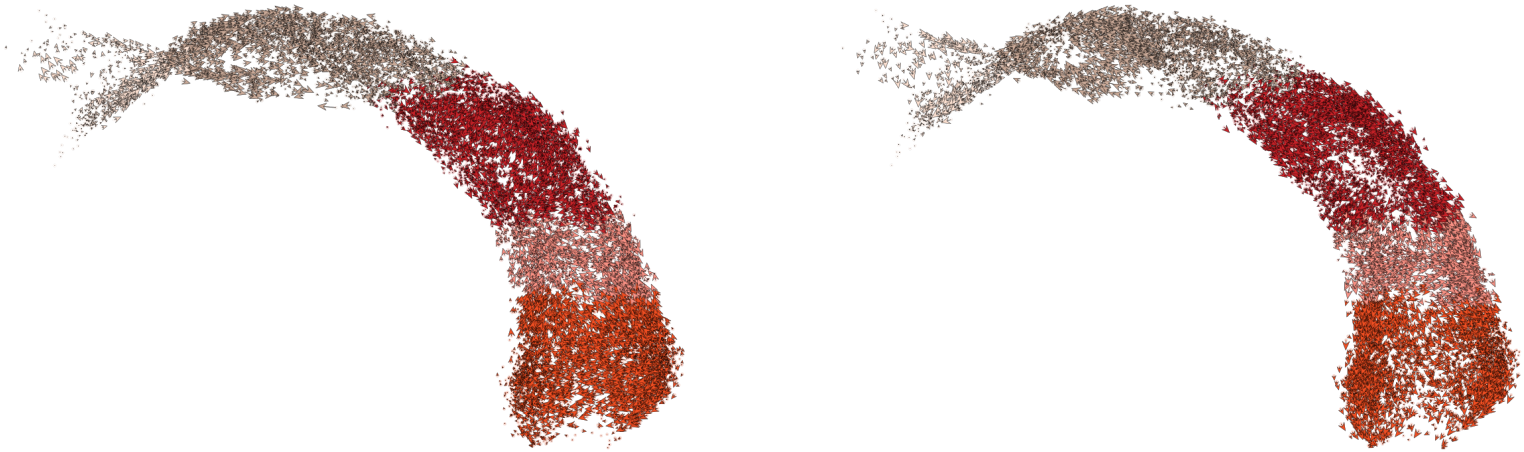
Analysis on the Mouse Gastrulation Subset to Erythroid Lineage dataset



(a) (b)

Figure: **RNA acceleration recapitulates dynamics of endocrine pancreas cell differentiation.** (a) Velocity obtained from the stochastic model for Erythroid Lineage are visualized as lines of motion in a UMAP-based embedding. (b) Accelerations obtained from the stochastic model for Erythroid Lineage are visualized as lines of motion in a UMAP-based embedding.

Analysis on the Mouse Gastrulation Subset to Erythroid Lineage dataset



(a) (b)

Figure: **RNA acceleration recapitulates dynamics of Erythroid cell differentiation.** (a) The foremost fine-grained determination of the velocity vector field we get at single-cell level. (b) The foremost fine-grained determination of the acceleration vector field we get at single-cell level.

Analysis on the Mouse Gastrulation Subset to Erythroid Lineage dataset

- On carefully comparing the velocity output with the acceleration output, we are able to highlight that the change of velocity remains the same for all the intermediate cells getting generated.
- The acceleration takes place in the entire pathway of generating erythroid cells from the progenitors in blood cells.
- In the time of embryonic development blood progenitor 2 cells differentiate into erythroid 1 cells, erythroid 1 cells differentiate into erythroid 2 cells and erythroid 2 cells differentiate into erythroid 3 cells.
- However, the RNA velocity model cannot correctly capture the cell generation from erythroid 2 to erythroid 3.
- Our model captures the same generation rate from erythroid 1 to erythroid 2 and erythroid 2 to erythroid 3.

Conclusion

- For the first time, we proposed a RNA acceleration to examine the DNA expression and our model predicts the rate of cell generation.
- We used manifold learning algorithm (UMAP) that at the same time fit a manifold and the kinetics on that manifold, on the premise of RNA acceleration.
- We create volcano plot corresponding the cells which have low generation/exchange rate and extract the up-regulated and down-regulated genes for performing Gene set enrichment analysis.
- We identified characteristic gene signatures by conducting tests for differential expression between a subgroup and all other cells.

Future Work

- Downside of our model is assuming the existence of steady states or a constant (generally $\beta = 1$) splicing rate across genes.
- α is a time-dependent parameter, so we assume $\alpha = u$, where u denotes the count of unspliced RNAs.
- Though scVelo tries to predict the parameters α and β in their Dynamical Model, however many NaN (Not a Number) values appear in their estimation.
- Hence, further attention is required to develop methods for accurately estimating the parameters α and β.

References

1. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A. and Fan, J., 2018. RNA velocity of single cells. Nature, 560(7719), pp.494-498.
2. Bergen, V., Lange, M., Peidli, S., Wolf, F.A. and Theis, F.J., 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. Nature biotechnology, 38(12), pp.1408-1414.
3. Byrnes, L.E., Wong, D.M., Subramaniam, M., Meyer, N.P., Gilchrist, C.L., Knox, S.M., Tward, A.D., Ye, C.J. and Sneddon, J.B., 2018. Lineage dynamics of murine pancreatic development at single-cell resolution. Nature communications, 9(1), p.3922.
4. Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C., Ho, D.L.L. and Reik, W., 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. Nature, 566(7745), pp.490-495.

THANK YOU