The study of the post-mitotic retinal cells using single-cell transcriptional data from paper «Single-cell transcriptional logic of cell-fate specification and guidance in early-born retinal neurons»by Lo Giudice et al. [1]

BIO-463 Genomics and bioinformatics, EPFL

Garance Haefliger

May 2022

1 Background

Single-cell RNA-seq measures the mRNA concentration of hundreds of genes simultaneously to get the gene expression level of individual cells. It has become an effective and robust tool, since it provides a high-resolution and high-throughput transcriptomic analysis of single cells. Moreover, it is a relatively cheap method and it is used for several purposes: unravelling complex cell populations, reconstructing developmental trajectories, and modelling transcriptional dynamics [2]. This method distinguishes the molecular pattern of all cell types within a complex population ; unlike bulk RNA seq which only provides an average of the cells within a population [3]. In this project, we aimed to reproduce some figures from the reasearch article «Single-cell transcriptional logic of cell-fate specification and guidance in early-born retinal neurons»written by *Lo Giudice et al* [1].

In this paper, the authors wanted to show the diversity of the early developing mouse retina by performing single-cell RNA-seq analysis of 6067 retinal cells. To understand the logic of neuronal network assembly, we need to know how the different neuron types arise in the mammalian central nervous system. To begin with a simple dataset, they chose the retina because it contains only six neuron classes that are well-identified. The six classes are the following : retinal ganglion cells (RGC), the horizontal cells (HC), the amacrines cells (AC), the cones, the bipolar cells (BC) and the rod photoreceptors. It is known that the early-born RGC, HC, AC and cones appear from embryonic days 10 to 17, while the early-born BCs, rods and still AC, appear from embryonic day 14 to postnatal day 5. However, how all these neuron classes emerge from the retinal progenitor cells remains

unclear. Additionally, the progenitor cells performed the cell cycle, they divide into other progenitor cells and occasionally into nondividing neuroblasts that can differentiate into neurons like RGC [4].

In this paper, we can identify three main purposes: (1) track the origin of early-born retinal cell fates, (2) deduce the spatial relationships across retinal neurons, and, finally, (3) point out RGC groups with transcriptional signatures linked to different projections in the developing brain [1].

In this project, we were specifically concentrated on a subset of the data ; on the differentiation trajectories of the retinal progenitor cells. Therefore, we reproduced figure 3G and a part of the figure 3H of the original paper [1]. Then, we performed additional analyses that were not in the paper to explain some of the divergences obtained between the original figures and the reproduced ones.

2 Data description

We first got a **.csv** file containing the single-cell data from two merged replicated experiments that were generated with the 10x Genomics protocol. Then, we got two **.loom** files (one for each replicated experiment) that contained the RNA velocity information.

First, we processed the 10x **.csv** file following the procedure stated in the "Methods" section in the paper. We created a *Seurat* object in Rstudio and then, we normalized and scaled the data, selected the top 2000 highly-variable features. Afterwards, we performed a PCA and run harmony to regress out the technical variation between both replicates. Then, we computed the 14 clusters and we finally performed dimensionality reduction, UMAP, for visualisation. Our plot closely reproduces figure 2B of the original paper [1], however, we encountered some differences.

After that, we extracted a subset of the data by excluding the progenitor cells (cell cycle cells) and did the same pre-processing as on the whole dataset. Finally, we registered all the information in four different files to switch from R to python as the *Velocyto* package did not work on Rstudio. Then, we loaded them and **.loom** files on python. We had some pre-processing to do on **.loom** files : we renamed the id of the cells to get the same as the 10x Genomics data. Then, we concatenated these two files and we merged the RNA velocity data with the 10x Genomics thanks to *Anndata* package. We normalized and scaled the resulting object. Finally, we got only one *Anndata* object at the end of the preprocessing containing all useful information to reproduce figures with the help of *Scanpy* and *Scvelo* packages [5].

3 Data analysis 1

By reproducing the UMAP of the post-mitotic cells with the RNA velocity field, we saw some differences between the original figure 3G (UMAP) and 3A (velocity field) that we decided to investigate further [1]. First, we have the same number of clusters, but in the original figures, none of the clusters is separated from the others like the yellow one

in our figure. Moreover, the RGC and AC/HC branches are inverted ; the RGC branch should be right and the AC/HC branch should be left. This difference is not important because UMAP is agnostic to rotation and reflexion. Concerning the velocity vectors, the left branch (RGC) seems to be the starting point of a new branching according to the arrows in the dark blue cluster and it is not the case in the original figure. Then, the orange cluster shows opposite velocity directions ; it should normally go down towards the green cluster as the progenitor cells were located just above it. We tried to explain these differences in the next paragraphs.

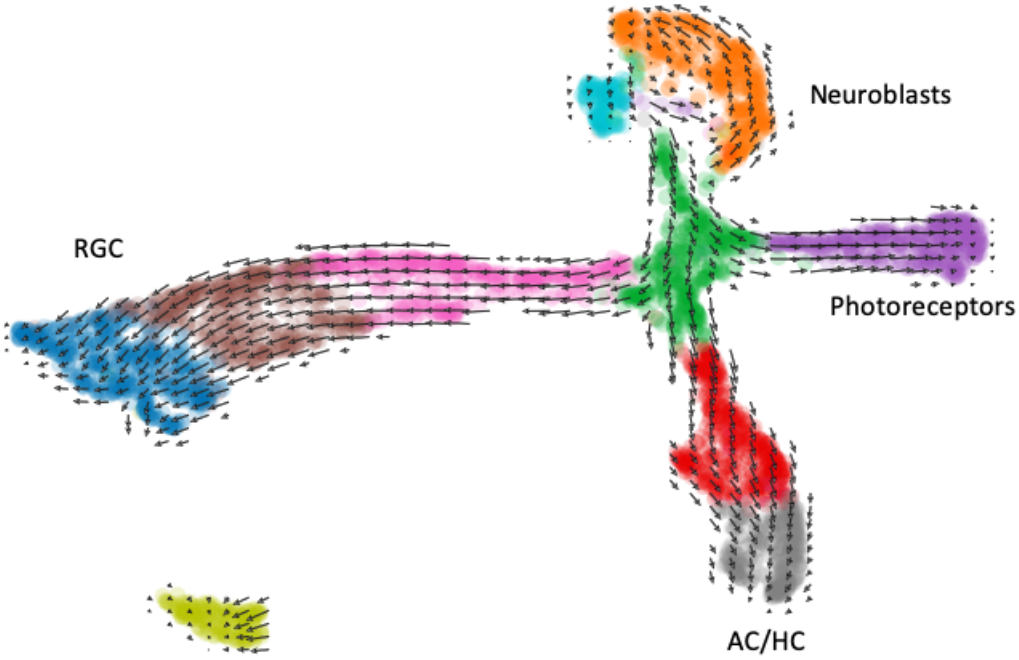

Figure 1: The UMAP of the post-mitotic cells separated in different clusters and with the RNA velocity field on it, based on the figure 3G of the original paper [1]

We decided to plot the top 4 marker genes distribution of the yellow cluster on the UMAP (second row of figure 2) and we saw that its marker genes distribution is often correlated with the light blue and dark blue clusters. We can also see that with the top 4 marker genes distribution of dark blue cluster (third row of figure 2). As in the original figure, there were 4 clusters in the RGC branch, the yellow cluster might belongs to this cell family. Moreover, the light blue cluster is called unknown/RGC-like (U/RGC) because it contains many RGC gene markers.

Then, we searched the top 30 marker genes of the yellow cluster and we uploaded them on *geneontology* to look into cell annotations. We found only ribosomal and translational functions that are basic functions of the cell. These annotations might be an indication of low-quality cells. To test this hypothesis, we did a quality-control analysis of the clusters

by plotting the QC metrics (violin plots). The number of unique genes detected in each cell and the total number of molecules detected within a cell of the yellow cluster are lower than in the other clusters ; indicating that indeed this cells might be low quality compared to the rest [6]. So, the authors probably removed this cluster from their analysis. Another argument in favor of this hypothesis is that three of the top 4 marker genes of the yellow cluster (second row of figure 2) are mitochondrial genes. Usually cells with high mitochondrial expression refers to stress or dying cells and are filtered.

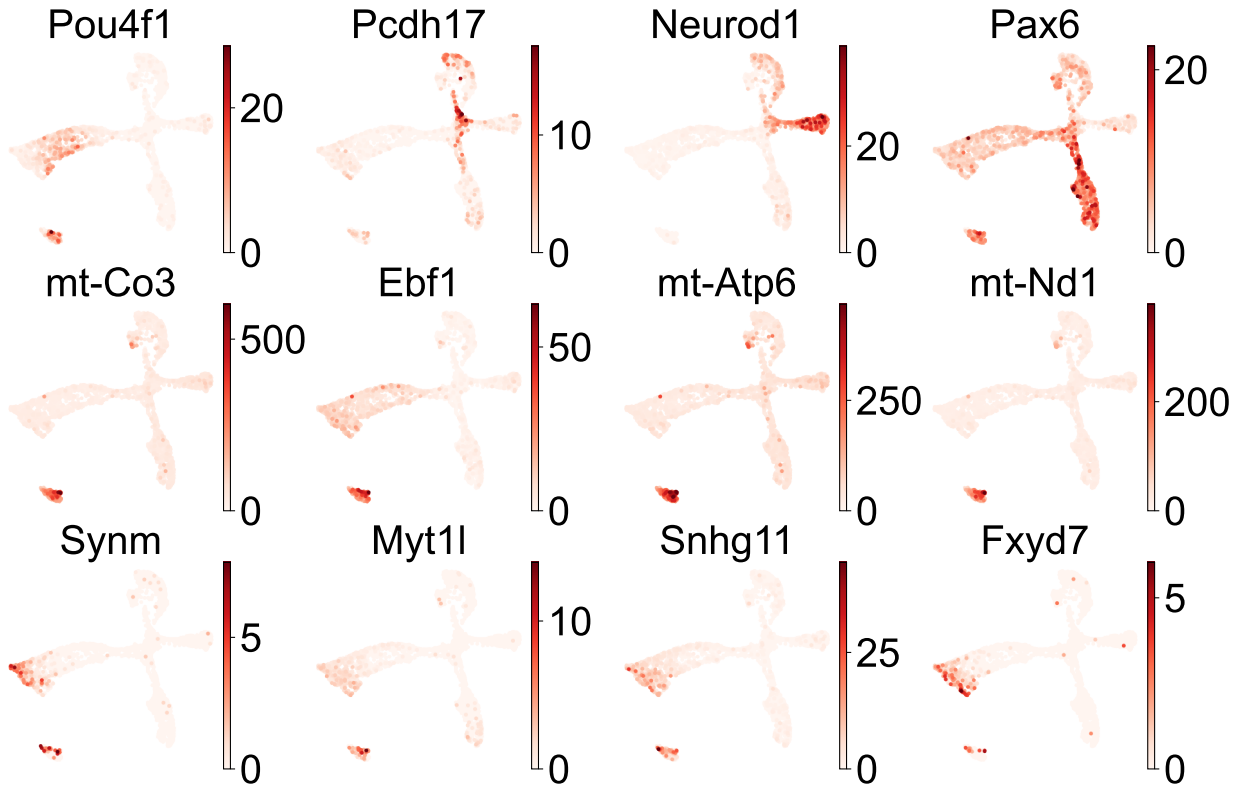

Figure 2: The first row is the pattern of the driver genes of RGC (Pou4f1, Pcdh17, Pax6), PR (Neurod1), AC/HC (Pax6, Pcdh17), based on the figure 3H of the original paper [1]. The second row is the pattern of the top 4 marker genes of the yellow cluster. The third row is the 4 marker genes of the blue cluster. The red bar represents the mRNA count.

To explain the differences in arrows' direction within the dark blue cluster, we plotted the top 4 marker genes distribution of this cluster on the UMAP (third row of figure 2). We can see that the expression of the first gene *synm* is concentrated only in the upper part of the dark blue cluster, while the others are well distributed between the upper and the lower part. We looked at more marker genes and most of them were well distributed between the two parts except a few genes like *igf1*. So, this could partly explain the difference in the velocity field. We looked at the functions of *synm* and *igf1* genes to have an idea of the biological differences within the two sub-branches. *Synm* codes for a

structural support protein and *igf1* for a protein implicated in growth and development [7]. Maybe the upper sub-branch contains more robust cells or it could simply be two different subtypes of RGC.

The UMAP of the post-mitotic cells shows the three branches : photoreceptors, RGC, AC/HC originate from the neuroblasts branch which comes from the progenitor cells (cell cycle). The progenitor cells are not shown in this UMAP. The cell fate explained in the background section seems to be correct except for the fact that the orange cluster has some cell cycle activity which is normally not the case for neuroblasts that are non-diving cells. Maybe some progenitor cells belong to the orange cluster and they can recycle themselves or it could just be a technical artefact.

Afterwards, we tried to replicate some plots of figure 3H (first row of figure 2). If we do not take into account the differences for the UMAP form, the four feature plots replicated very well the ones in the original paper.

4 Data analysis 2

In this section, we performed further analyses to look more in-depth into the replication problems that we underwent in the previous section.

We looked at the previous velocity disparities by identifying genes that explain the vector field. We plotted the phase portraits that show the unspliced (pre-mature) vs spliced (mature) proportion of mRNA in cells coloured by cluster for a given gene. First, we have to understand how the velocities are computed : the velocities are the residuals from the steady-state ratio of the unspliced vs spliced mRNA counts (black dot line in figure 3). A positive velocity means higher unspliced mRNA than the steady-state and it is an indication of up-regulation of the gene. A negative velocity means a down-regulation of the gene (higher spliced mRNA) [8]. We performed this analysis on the top genes that explain the vector field of dark blue and orange clusters (see top 2 of the blue cluster and top 2 of the orange cluster in figure 3). We saw more unspliced than spliced mRNA in the orange and dark blue clusters ; it agrees with their relatively big arrow size. By looking at the distribution of these top 10 genes of the dark blue cluster, we noticed only one gene (*meg3*) that was deferentially distributed between the two sub-branches (more expression in the lower part). *Meg3* is a tumour suppressor gene [7] ; so, it is not very useful to distinguish subcluster function.

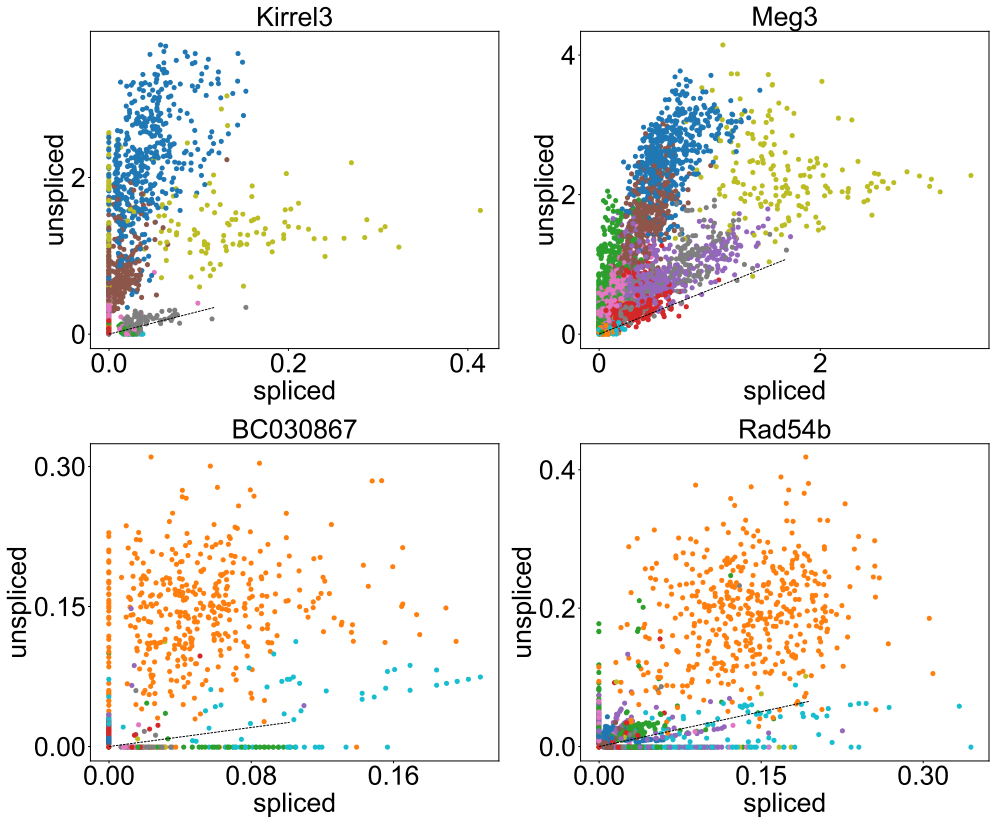

Figure 3: Phase portraits of the top two genes that explain the vector field of dark blue cluster (first row) and the orange cluster (second row). To get more information, please read the above paragraph.

To do a latent time analysis, we had first to run the dynamical model to learn the full transcriptional dynamics of splicing kinetics. The latent time analysis approximates the real time lived by the cells while they are differentiating ; it shows the cell's internal clock [9]. Then, in figure 4, we saw the scheme that we described in the background section. RGC, AC/HC and photoreceptors come from neuroblasts. Photoreceptors take more time to appear ; their branch contains more yellow than the others, which confirms what the biological background said. The cycling pattern of the orange cluster seems to not occur long because we observed only red and purple in this zone. However, it is hard to be sure about this because we do not have temporal data. The lower part of the dark blue cluster seems to appear before the upper part ; maybe the upper part of this cluster would be about to diverge from it. We noticed also the strange behaviour of the U/RGC cluster (light blue) as latent analysis shows that it contains the first synthesized cells.

In the neuroblasts heatmap 5, we can see the non-linear path between the orange and green cluster because the two colours remain all along with the plot while in the other branches heatmap (6 and 7) we have an almost linear path between clusters (colour separation in the cluster bar and a smooth yellow line).

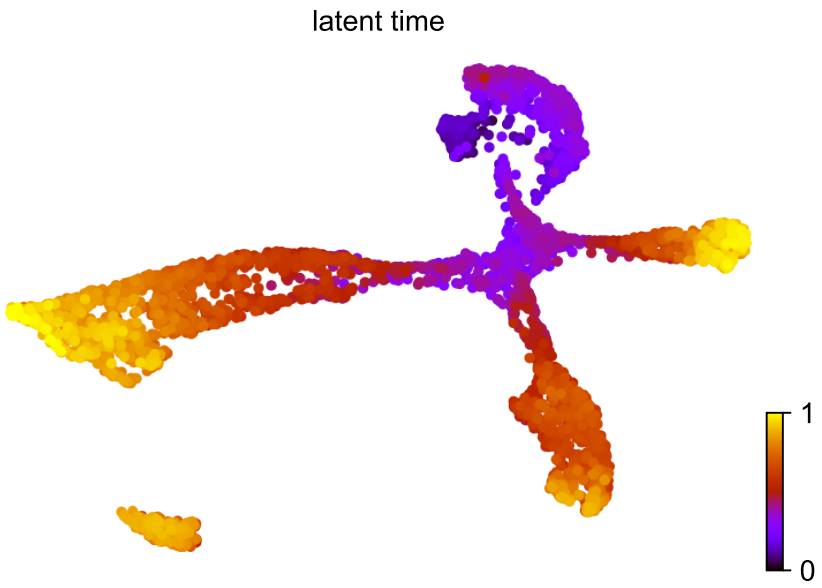

Figure 4: Latent time analysis. The latent time of the underlying cellular mechanisms is recovered by the dynamical model. Based on the transcriptional dynamics, it is possible to approximate the real time experienced by cells as they differentiate [9].

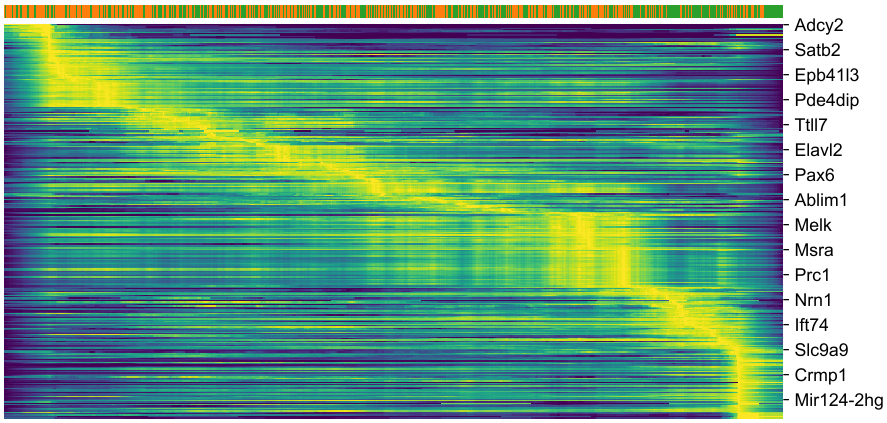

Figure 5: Neuroblast branch heatmap of its high likelihoods genes in the dynamic model.

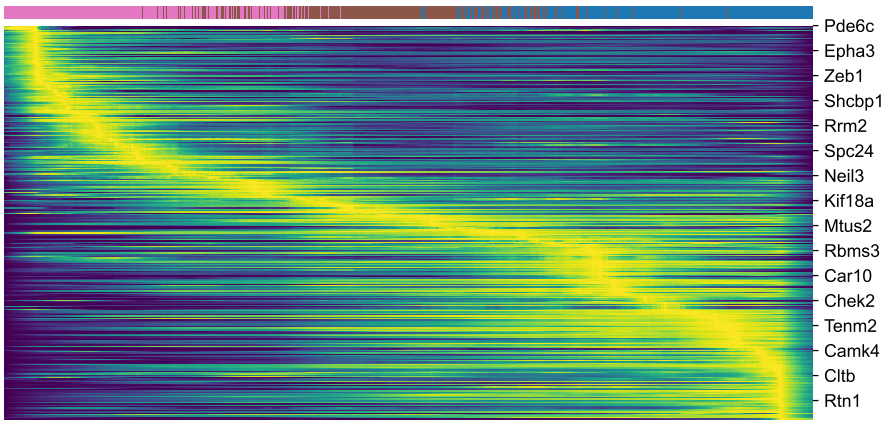

Figure 6: RGC branch heatmap of its high likelihoods genes in the dynamic model.

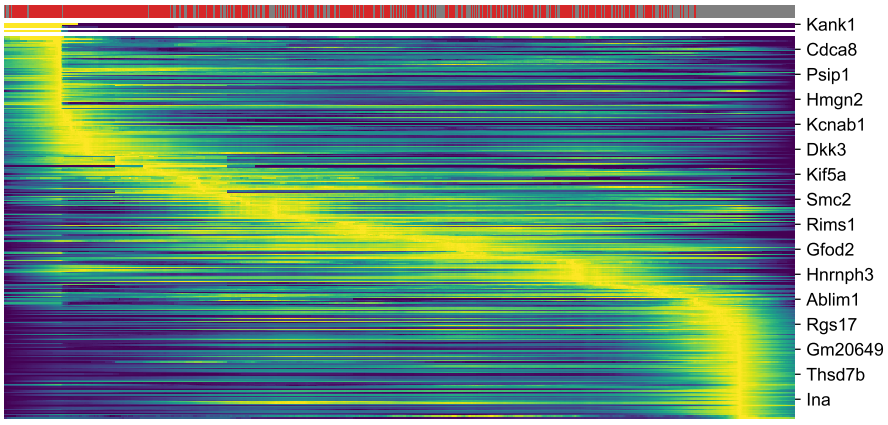

Figure 7: AC/HC branch heatmap of its high likelihoods genes in the dynamic model.

Finally, we decided to check if some clusters have cell cycle activity by using a list of genes that are specific to the S phase and another specific to the G2/M phase. We plotted the results in the figure 8. Indeed, some cells of the orange, green and light blue clusters contain G2/M specific genes. As none of them is composed of S specific genes, it could be that some of the progenitors have just not finished their cell cycle yet and already corresponds to neuroblast markers ; like an in-between progenitor and neuroblast cell.

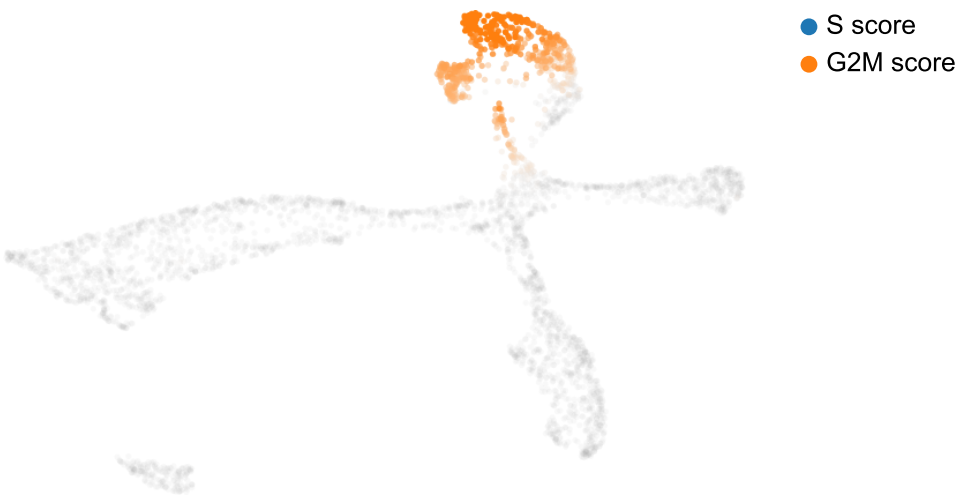

Figure 8: Maps of the location of gene markers of cell cycle. In our graph, some cells contain cell cycle genes specific to G2/M phase.

5 Conclusion

In this project, we reproduced the figure 3G (1) and 3H (2) from the original paper [1]. Then, we tried to explain some of the differences we encountered between our reproduced figures and the original ones. The most likely explanations are the following :

- The cells that composed the yellow cluster were removed from the analysis in the original paper because they had a strong indication of low-quality cells.
- The dark blue cluster probably contains two different subtypes of RGC.
- In the neuroblast branch of the UMAP, there is still cell cycle activity (G2/M phase), especially for the orange cluster.

References

[1] Q. Lo Giudice, M. Leleu, G. La Manno, and P. J. Fabre, "Single-cell transcriptional logic of cell-fate specification and axon guidance in early-born retinal neurons," *Development*, vol. 146, no. 17, 09 2019, dev178103. [Online]. Available: https://doi.org/10.1242/dev.178103

[2] S. Liu and C. Trapnell, "Single-cell transcriptome sequencing: Recent advances and remaining challenges," *F1000Research*, vol. 5, 02 2016.

[3] S. S. Potter, "Single-cell RNA sequencing for the study of development, physiology and disease," pp. 479–492, aug 2018. [Online]. Available: /pmc/articles/PMC6070143//pmc/articles/PMC6070143/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6070143/

[4] D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A.-S. LaMantia, J. O. McNamara, and S. M. Williams, *Neuroscience. 2nd edition. The Initial Formation of the Nervous System: Gastrulation and Neurulation.* Sinauer Associates, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK10993/

[5] "Sam Morabito | RNA velocity analysis with scVelo." [Online]. Available: https://smorabit.github.io/tutorials/8{_}velocyto/

[6] "Seurat - Guided Clustering Tutorial • Seurat." [Online]. Available: https://satijalab.org/seurat/articles/pbmc3k{_}tutorial.html

[7] "GeneCards - Human Genes | Gene Database | Gene Search." [Online]. Available: https://www.genecards.org/

[8] "RNA Velocity Basics — scVelo 0.2.5.dev6+g1805ab4.d20210829 documentation." [Online]. Available: https://scvelo.readthedocs.io/VelocityBasics/

[9] "Dynamical Modeling — scVelo 0.2.5.dev6+g1805ab4.d20210829 documentation." [Online]. Available: https://scvelo.readthedocs.io/DynamicalModeling/

[10] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan, J. Fan, L. E. Borm, Z. Liu, D. van Bruggen, J. Guo, X. He, R. Barker, E. Sundström, G. Castelo-Branco, P. Cramer, I. Adameyko, S. Linnarsson, and P. V. Kharchenko, "RNA velocity of single cells," *Nature*, vol. 560, no. 7719, pp. 494–498, aug 2018. [Online]. Available: https://doi.org/10.1038/s41586-018-0414-6