bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SIRV: Spatial inference of RNA velocity at the single-cell resolution

Tamim Abdelaal[1,2,3], Boudewijn P.F. Lelieveldt[1,2], Marcel J.T. Reinders[2,3,4] and Ahmed Mahfouz[2,3,4,*]

[1]LKEB, Department of Radiology, Leiden University Medical Center, 2333ZC Leiden, The Netherlands
[2]Delft Bioinformatics Lab, Delft University of Technology, 2628 XE Delft, The Netherlands
[3]Leiden Computational Biology Center, Leiden University Medical Center, 2333ZC Leiden, The Netherlands
[4]Department of Human Genetics, Leiden University Medical Center, 2333ZC Leiden, The Netherlands

*Corresponding author: Ahmed Mahfouz (a.mahfouz@lumc.nl)

Abstract

Studying cellular differentiation using single-cell RNA sequencing (scRNA-seq) rapidly expands our understanding of cellular development processes. Recently, RNA velocity has created new possibilities in studying these cellular differentiation processes, as differentiation dynamics can be obtained from measured spliced and unspliced mRNA expression. However, to map these differentiation processes to developments within a tissue, the spatial context of the tissue should be taken into account, which is not possible with current approaches as they start from dissociated cells. We present SIRV (Spatially Inferred RNA Velocity), a method to infer spatial differentiation trajectories within the spatial context of a tissue at the single-cell resolution. SIRV integrates spatial transcriptomics data with reference scRNA-seq data, to enrich the spatially measured genes with spliced and unspliced expressions from the scRNA-seq data. Next, SIRV calculates RNA velocity vectors for every spatially measured cell and maps these vectors to the spatial coordinates within the tissue. We tested SIRV on the Developing Mouse Brain Atlas data and obtained biologically relevant spatial differentiation trajectories. Additionally, SIRV annotates spatial cells with cellular identities and the region of origin which are transferred from the annotated reference scRNA-seq data. Altogether, with SIRV, we introduce a new tool to enrich spatial transcriptomics data that can assist in understanding how tissues develop.

1. Introduction

Single-cell RNA-sequencing (scRNA-seq) enables the study of cellular differentiation dynamics at single-cell resolution[1]. Trajectory inference methods, such as Monocle[2], DPT[3] and PAGA[4], aim to define a developmental sequence of gene expression changes in a certain pool of cells. This developmental sequence should potentially model a trajectory of the cellular differentiation process. However, scRNA-seq only captures a static snapshot of the cellular states, which represents a major challenge for trajectory inference methods to correctly capture the dynamics of the cellular differentiation process. That challenge is addressed by RNA velocity[5,6], which estimates the dynamics of cellular differentiation using the expression

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

balance between the unspliced immature and spliced mature mRNA molecules captured by scRNA-seq protocols. Typically, for each cell, the RNA velocity is estimated for each gene individually, which together, define the new state of the cell. When calculated over all cells, a flow field can be calculated by some averaging of neighboring cells, which can be projected in a reduced visualization space.

To date, the majority of cellular differentiation studies only use dissociated cells from scRNA-seq. However, this neglects that cells are spatially organized, for example in tissues. Taking the spatial context into account can enhance our understanding of cellular differentiation processes[7]. Recent advances in spatial transcriptomics technologies have enabled the study of cellular heterogeneity of complex tissues while retaining spatial information, providing spatial gene expression patterns and organization of different cell types in the tissue[8,9]. Current protocols can be divided into two main categories: (1) sequencing-based technologies that detect and quantify the mRNA in situ, such as 10X Genomics Visium, Slide-seq[10,11] and ST[12]; or (2) imaging-based technologies using fluorescence in situ hybridization (FISH), such as smFISH[13,14], MERFISH[15] and seqFISH[16,17].

In principle, it is possible to apply RNA velocity analysis to spatial transcriptomics measured using sequencing-based protocols, as the spliced and unspliced expression ratios can be directly obtained from the sequencing data[11]. However, these protocols are still limited in spatial resolution measuring tissue spots consisting of a few cells. On the other hand, imaging-based protocols provide (sub)cellular resolution but lack the spliced and unspliced expressions which are necessary to study cellular differentiation using RNA velocity.

We and others have previously shown that high-resolution imaging-based spatial transcriptomics data can be enriched with predicted expression of spatially unmeasured genes when integrating with scRNA-seq data measured from the same tissue[18–22]. Building on the same concept, it should be possible to predict the spliced and unspliced expression of each spatially measured gene from a reference scRNA-seq dataset. With these imputed values RNA velocities can then be calculated for each cell in its spatial context, enriching cellular differentiation processes with true spatial coordinates.

Here, we present SIRV (Spatially Inferred RNA velocity), a method to derive cellular differentiation dynamics in a spatial context at the single-cell resolution. SIRV integrates spatial transcriptomics and scRNA-seq data in order to predict the spliced and unspliced expression of the spatially measured genes. Next it calculates RNA velocity vectors for each cell that are then projected onto the two-dimensional spatial coordinates, which are then used to derive flow fields by averaging dynamics of spatially neighboring cells. In addition, SIRV transfers various label annotations of the scRNA-seq to the spatial transcriptomics data, allowing us to richly annotate the spatial data with, for example, cellular identity or region of origin. We tested SIRV on a spatial dataset from the Developing Mouse Brain Atlas[23] and showed that SIRV produced biologically relevant spatial differentiation trajectories. Moreover, it accurately predicts the cellular identities and region of origin of the spatial cells. Our results show that SIRV is a valuable tool to gain deeper insights into cellular differentiation processes spatially.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2. Methods

2.1. SIRV algorithm

The SIRV algorithm requires two inputs, the spatial transcriptomics data represented by a gene expression matrix, and the scRNA-seq data having three expression matrices corresponding to the spliced (mature mRNA), unspliced (immature mRNA) and full mRNA expression. The scRNA-seq data may also contain relevant metadata like cellular identity annotations, tissue/region of origin, etc. Using the set of shared genes between the two datasets, SIRV enriches the spatial transcriptomics data with un/spliced expressions as predicted from the scRNA-seq. These un/spliced expressions are then used to calculate the RNA velocity of each gene for each cell. Additionally, SIRV transfers the metadata from the scRNA-seq data to the spatial transcriptomics data. The SIRV algorithm consists of four major parts: (i) integration of the spatial transcriptomics and scRNA-seq datasets, (ii) predictions of un/spliced expressions, (iii) label/metadata transfer (optional), and (iv) estimation of RNA velocities within the spatial context.

2.1.1. Integration

The spatial transcriptomics and scRNA-seq dataset are integrated by finding the common signal between the two datasets. Building on SpaGE[18], the integration step is performed using PRECISE to define a common latent space[24]. In brief, using the set of shared genes across the two datasets, we calculate a separate Principal Component Analysis (PCA) for each dataset, and then aligns these separate principal components, resulting in principal vectors (PVs). These PVs have a one-to-one correspondence between the two datasets, and the highly correlated PV-pairs represent the common signal. Finally, both the spatial transcriptomics and scRNA-seq datasets are projected onto the PVs of the reference dataset (scRNA-seq in this case), producing an integrated and aligned version of both datasets.

This integration step is performed using the full mRNA expression matrix from the scRNA-seq side, together with the expression matrix of the spatial transcriptomics data. Thus, the spliced and unspliced expressions are only used in the prediction (following) step.

2.1.2. Un/spliced expression prediction

After obtaining the aligned datasets, SIRV enriches the spatially measured genes with spliced and unspliced expression predicted from the scRNA-seq dataset. Such prediction is performed using a kNN regression[18]. For each spatial cell i, we calculate the k-nearest-neighbors from the (aligned) scRNA-seq dataset and assign a weight to each neighbor inversely proportional to its distance.

$$w_{ij} = 1 - \frac{dist\ (i,j)}{\sum_{j\,\in NN(i)} dist(i,j)} \tag{1}$$

$$w_{ij} = \frac{w_{ij}}{k-1}; \quad where \sum_{j\,\in NN(i)} w_{ij} = 1 \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with w_{ij} representing the weight between each spatial cell i and its j-th nearest neighbor, $dist(i,j)$ being the cosine distance between spatial cell i and scRNA-seq cell $j \in NN(i)$, and k equaling the number of nearest-neighbors used.

For every spatially measured gene g, the spliced (S_{ig}) and unspliced (U_{ig}) expression are predicted by:

$$S_{ig} = \sum_{j \in NN(i)} w_{ij} * SR_{jg} \text{ and } U_{ig} = \sum_{j \in NN(i)} w_{ij} * UR_{jg} \quad (3)$$

with SR_{jg} and UR_{jg} representing the spliced and unspliced expression of gene g from the scRNA-seq dataset, respectively.

2.1.3. Label (metadata) transfer

SIRV can annotate the spatial transcriptomics dataset with any relevant labels from the scRNA-seq dataset using the same kNN regression scheme as introduced earlier. Taking the cell identity annotation as an example: for each cell type c in the scRNA-seq annotation, we calculate the probability P_{ic} that the spatial cell i should be assigned to cell type c by aggregating the weights w_{ij} of the nearest neighbors annotated with c. The transferred cell type C_i for each spatial cell i is then selected based on the cell type with highest probability:

$$P_{ic} = \sum_{\substack{j \in NN(i) \\ j \in c}} w_{ij} \quad (4)$$

$$C_i = \underset{c}{argmax}\, P_{ic} \quad (5)$$

2.1.4. RNA velocity analysis

After enriching the spatial genes with spliced and unspliced expressions, we applied the RNA velocity[5] method (implemented in the scvelo[6] python package) to study the cellular development and differentiation. Following scvelo, first, we calculated the high-dimensional RNA velocity vectors for the spatially measured genes (set of genes originally measured in the spatial dataset), next we projected and visualized these vectors on the spatial coordinates of the cells in order to define directions of cellular differentiation in the spatial context.

2.2. Dataset description

We used both spatial transcriptomics and scRNA-seq datasets from the Developing Mouse Brain Atlas[23]. Datasets were downloaded from http://mousebrain.org/downloads.html. The spatial transcriptomics dataset profiled the expression of 119 genes in an E10.5 mouse embryo, measured using the HybISS protocol[25]. Out of 25 different spatial slices provided by the authors, we selected the '40 µm' slice as it contains a clear structure of the brain. Cell segmentation was not provided with the data, however, we used the voxel version of the data

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(provided by the authors) which summarizes the spatial gene expression in a 2D grid of 30,000 pixels.

The scRNA-seq data profiled a developing mouse brain tissue from E7 to E18. To match the HybISS data, we only used E10 and E11 having a total 47,639 cells expressing 31,053 genes. Additionally, the scRNA-seq data was annotated with several metadata labels, we focused on the labels indicating the region (Forebrain, Midbrain and Hindbrain) and cellular identity. The latter was provided at two different resolutions, the 'Class' annotation covering 19 major cell populations and the fine-grained 'Subclass' annotation with 104 subpopulations.

2.3. Data preprocessing

For the spatial HybISS dataset, the gene expression of each pixel (pseudo-cell) is the count of the spots detected for each gene in that pixel location. To separate tissue from background, we filtered out any pseudo-cell with total counts across all genes less than 4. Next, we manually segmented only the brain tissue, ending with a total of 4,628 spatial pseudo-cells (Supplementary Fig. S1). Next, the HybISS dataset was normalized by dividing the counts within each (pseudo-)cell by the total count within that cell, multiplied by a scaling factor equal to the median number of counts across cells, and log(x+1) transformed.

$$x_{ig} = \log\left(\left(\frac{x_{ig}}{\sum_g x_{ig}} * mc\right) + 1\right); \quad where\ \ mc = \underset{i}{\text{median}}(\sum_g x_{ig}) \tag{6}$$

with x_{ig} represents the expression of gene g in cell i, $\sum_g x_{ig}$ equals the total count in cell i, and mc is the median number of counts across cells.

For the scRNA-seq dataset, after selecting E10 and E11, cells having 'Class' annotation of 'Bad cells' or 'Undefined' were filtered out. Additionally, genes annotated as invalid ('Valid' = 0) were removed. Further, we filtered out genes expressed in less than 10 cells, ending up with an expression matrix of 40,733 cells and 16,907 genes. Finally, the dataset was normalized by dividing the counts within each cell by the total count within that cell, multiplied by a scaling factor of 10^6, and log(x+1) transformed.

$$x_{ig} = \log\left(\left(\frac{x_{ig}}{\sum_g x_{ig}} * 10^6\right) + 1\right) \tag{7}$$

with x_{ig} represents the expression of gene g in cell i, and $\sum_g x_{ig}$ equals the total count in cell i.

2.4. Implementation details

The HybISS spatial dataset and the scRNA-seq dataset shared 117 genes used for integration and prediction of spliced and unspliced expression. For the integration step we used 50 principal vectors explaining 62.9% and 62.4% of the variance in the HybISS and scRNA-seq datasets, respectively. For the prediction step, we used k = 50. For the RNA velocity analysis,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

we scale the data to zero mean and unit variance features, next we calculate the top 30 principal components (PCA) to build a neighborhood graph with 30 neighbors. This neighborhood graph is used to calculate a UMAP[26] embedding of the data and cluster the data using Leiden[27] graph-based clustering (resolution = 1). Next, we calculate the RNA velocity vectors using 30 principal components and 30 neighbors. Finally, we project the high-dimensional velocity vectors on the UMAP coordinates and the spatial (x,y) coordinates using the *velocity_embedding* and *velocity_embedding_stream* functions. We used the scanpy[28] python package (version 1.7.0) to perform data preprocessing, PCA, UMAP and Leiden clustering. While the scvelo[6] python (version 0.2.3) package was used to calculate RNA velocities and their projections in the UMAP and the spatial context.

3. Results

3.1. SIRV overview

We developed Spatially Inferred RNA Velocity (SIRV), a method to study cellular differentiation in the spatial context using RNA velocity (Fig. 1, Methods). Calculating the RNA velocity vectors requires measuring the proportion of mature (spliced) and immature (unspliced) mRNA expression. This information is usually missing in the spatial transcriptomics data especially for the imaging-based approaches. SIRV combines the spatial transcriptomics dataset with a reference scRNA-seq dataset having the spliced and unspliced expression for the whole transcriptome. SIRV first integrates both datasets to correct for technical differences using domain adaptation method PRECISE[24]. Next, SIRV predicts the spliced (mature) and unspliced (immature) expression for the spatially measured genes from the scRNA-seq data using kNN regression. The predicted spliced and unspliced expression can then be used to calculate the RNA velocity vectors[5,6], and project these vectors on the spatial coordinates of the tissue.

Additionally, SIRV transfers various label annotations from the scRNA-seq to the spatial transcriptomics data using the same kNN regression scheme. Considering cell identity labels, this label transfer feature offers an automated manner to annotate the spatial dataset. Since scRNA-seq captures the whole transcriptome, the transferred annotations represent more fine-grained cellular identities as compared the populations derived from the spatial dataset alone as this is based on a limited number of measured genes.

3.2. SIRV produces interesting spatial differentiation trajectories in the developing mouse brain

We applied SIRV on the HybISS spatial dataset from the developing mouse brain (E10.5). First, we clustered the HybISS dataset, using Leiden graph-based clustering, into 21 cell clusters visualized using UMAP (Fig. 2A). When mapping the cell clusters back to their spatial locations, we observed that the clustering agrees with the spatial organization, i.e. the majority of the cell clusters are indeed separated in the spatial domain and each cluster is mapped to a specific spatial location (Fig. 2B and Supplementary Fig. S2), except for clusters 0, 15, 19 and 20. Furthermore, cluster 1 is spatially divided into two groups of cells, one in the Midbrain and another in the Hindbrain.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Next, we integrated the HybISS spatial dataset with a scRNA-seq dataset of the E10 and E11 development stages of the embryonic mouse brain. Using SIRV, we integrated the two data sets and predicted the spliced and unspliced expression for the spatially measured genes. The two datasets shared 117 genes which were used for integration and prediction. With the predicted un/spliced transcripts for each of the 117 genes, we generated RNA velocity vectors for each cell. The RNA velocity analysis shows dynamic flow patterns matching the HybISS clusters (Fig. 2C) and reveals differentiation patterns for each of the clusters. For instance, cluster 9 differentiates into clusters 7, 8 and 13. Cluster 3 differentiates into clusters 5, 17 and part of 14, while the other part of cluster 14 is developed from cluster 13. Cluster 12 differentiates into cluster 16, while clusters 2 and 10 differentiate towards a common end point. Additionally, we observed that clusters 0, 15, 19 and 20 almost have zero magnitude velocity vectors. A possible explanation is that these clusters form the boundary of the brain (Supplementary Fig. S2) and are not further involved into cellular differentiation at this stage of development.

Next, we projected and visualized the velocity vectors onto the spatial coordinates (Fig. 2D), which reveals the developmental dynamics of the cells in their spatial context. For example, cluster 9 is roughly located between the Midbrain and the Hindbrain, and the velocity vectors branches in 3 directions towards clusters 7, 8 and 13. To obtain a more detailed view of the spatial RNA velocities, we visualized the velocity vectors on the cell level (Fig. 3A). Results show that the velocity vectors follow consistent spatial paths across the different cell clusters and different brain regions. If we consider cluster 9 again, the spatial velocity vectors on the cell level shows the same differentiation into 3 branches (Fig. 3B). Furthermore, we can clearly observe the differentiation of cluster 12 into 16, and cluster 13 into a part of cluster 14 (Fig. 3C). Moreover, cluster 2 forms a relatively long path of differentiation through the Hindbrain (Fig. 3D), and reaches a common end point together with cluster 10. Additionally, in some cases, the spatial organization of the cell clusters can provide more insights to the cellular differentiation. When considering cluster 3 in the Forebrain, the spatial velocity vectors suggest that cluster 3 differentiates towards cluster 5 and part of 14, however, it does indirectly differentiate to cluster 17 through cluster 5 (Fig. 3E). Moreover, from the spatial context, we observed that cluster 5 further differentiates to cluster 6. Together, these results show the potential of SIRV to define interesting spatial differentiation trajectories that cannot be obtained from the scRNA-seq data alone.

3.3. SIRV correctly transferred label annotation verified by spatial organization

In addition to the prediction of spliced and unspliced expression, we used SIRV to transfer annotation labels from the scRNA-seq dataset to the HybISS spatial dataset. First, we transferred the brain region annotation (Forebrain, Midbrain and Hindbrain). In the UMAP embedding of the HybISS spatial data, the three brain regions form three different groups with some overlap in the middle (Fig. 4A). When visualized in their spatial coordinates, we observe that the different brain regions are in agreement with the anatomical structure of the brain (Fig. 4B and Supplementary Fig. S1C-D).

Next, we annotated the HybISS dataset with cell identity labels from the scRNA-seq data, which contains two levels of annotation: fine-grained subpopulations ('Subclass' annotation)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

and high-level major populations ('Class' annotation). The transferred 'Subclass' annotation contained 49 subclasses which roughly formed distinct groups in the UMAP embedding of the spatial data (Fig. 4C). Spatially, the obtained Subclass labels also showed a well-separated and structured spatial organization (Fig. 4D and Supplementary Fig. S4), with many examples validating the correct estimation of these labels. For example, cells annotated with identity subclasses Forebrain, Midbrain and Hindbrain are spatially localized with the corresponding brain region. The Midbrain-hindbrain boundary cells, are correctly located exactly at the boundary between these two brain regions (as can be derived from the name). The Cortical hem cells indeed localize in the Forebrain as was previously shown by the authors of the Developing Mouse Brain data[23]. The diencephalon cells which are known to be part of the Forebrain[29] are also located there. Also, we observed a group of Mesenchyme cells covering the borders of the whole brain. This group of Mesenchyme cells provides an explanation for the overlap between different brain regions shown in the middle of the UMAP in Fig. 4A, as these Mesenchyme cells are present at the borders of all three brain regions.

When transferring the 'Class' annotation, we obtained 11 classes which are mainly composed of Radial glia cells and Neuroblasts (Supplementary Fig. S3A-B), which are both progenitor cells for the development of the nervous system and neurons, respectively[30,31]. The Mesenchyme cells, from the 'Class' annotation, form the borders of the different brain regions, again in agreement with the 'Subclass' annotation. Overall, SIRV shows reliable label transfer ability with correct spatial localization of the transferred labels, even at a subclass annotation level.

3.4. RNA velocities interpretation based on transferred cell labels

Having the transferred cell identity annotations, we aimed at adding biological insights to the spatial differentiation trajectories as predicted by SIRV. First, considering the transferred 'Subclass' annotations, we visualized the spatial RNA velocity vectors at the cell level colored by the different cell subclasses (Fig. 5A) and zoomed in on the same differentiation examples introduced earlier with the cell clusters (Fig. 5B-E). The Midbrain-hindbrain boundary cells show the three branched trajectories, shown earlier by cluster 9, and differentiating towards Midbrain cells, Dorsal midbrain cells and Dorsal hindbrain cells (Fig. 5B). Comparing the 21 cell clusters with their 'Subclass' annotation indeed shows that cluster 9 mostly maps to the Midbrain-hindbrain boundary group (Supplementary Fig. S5B). In the Midbrain, Midbrain basal plate cells differentiate into Midbrain floor plate cells (Fig. 5C). In the Hindbrain, Dorsal hindbrain cells differentiate towards Hindbrain cells (Fig. 5D). Considering the Forebrain region, cells annotated as Forebrain (mostly covering clusters 3, 5 and 11) differentiate into Dorsal diencephalon (cluster 17), Cortical hem (cluster 6) and Midbrain basal plate (cluster 13) (Fig. 5E and Supplementary Fig. S5A). While cluster 0, 15, 19 and 20 are mostly composed of Mesenchyme cells forming the borders of the brain (Supplementary Fig. S5A), no differentiation was associated with these clusters in agreement with Mesenchyme cells which also have almost zero magnitude RNA velocity vectors.

We repeated the same visualization with the 'Class' annotation (Supplementary Fig. S3C). As shown, almost all cells with clear RNA velocity vectors (cells undergoing differentiation) are either Radial glia or Neuroblasts, which is expected as these two classes are progenitor cells.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

When comparing the 21 cell clusters with their 'Class' annotation (Supplementary Fig. 5B), we observe that the majority of the cell clusters are Radial glia cells, and that cluster 1 and 4 contain more Neuroblasts.

4. Discussion

We demonstrated the potential of SIRV to define biologically relevant spatial differentiation trajectories. Using SIRV, it is possible to apply RNA velocity analysis on high-resolution imaging-based spatial data, as SIRV enriches the spatially measured genes with spliced and unspliced expression predicted from scRNA-seq data.

We tested SIRV on the Developing Mouse Brain data. SIRV identified spatial branching differentiation trajectories at the Midbrain-Hindbrain boundary and the Forebrain. SIRV found linear trajectories in the Hindbrain and the Midbrain. Additionally, SIRV accurately transferred the region label annotation from the scRNA-seq data to the spatial data, and correctly located the three brain regions spatially. Also, SIRV successfully transferred the cell identity annotation and assigned cell populations to their expected spatial location.

Since the HybISS spatial dataset was not originally annotated with cellular identities, the spatial differentiation trajectories obtained with SIRV could not be easily interpreted using the spatial dataset only. Transferring cell annotations with SIRV helped greatly to biologically interpret the predicted spatial differentiation trajectories.

Currently, we only evaluated SIRV using one dataset. In the near future, we aim to test SIRV on various datasets to evaluate the robustness of the produced spatial differentiation trajectories. The interpretation of the estimated spatial velocity vectors is, however, still challenging, i.e. do the differentiation directions only imply that cells differentiate from one state to another, or do they also hint that cells do migrate in the tissue during development?

Concluding, SIRV produces valuable spatial differentiation trajectories for high-resolution imaging-based spatial transcriptomics data and opens new possibilities to study cellular differentiation processes in their spatial context, which helps understanding the natural tissue development.

Data and code availability

The implementation of SIRV, processed datasets, as well as the code to reproduce the results, are available in the GitHub repository, at https://github.com/tabdelaal/SIRV. The code is released under MIT license.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* **14**, e8046 (2018).

2. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

3. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* **13**, 845–848 (2016).

4. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).

5. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

6. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

7. Pham, D. *et al.* stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv* 2020.05.31.125658 (2020) doi:10.1101/2020.05.31.125658.

8. Mantri, M. *et al.* Spatiotemporal single-cell RNA sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nat. Commun.* **12**, (2021).

9. Srivatsan, S. R. *et al.* Embryo-scale, single-cell spatial transcriptomics. *Science (80-.).* **373**, 111 LP – 117 (2021).

10. Rodriques, S. G. *et al.* Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science (80-.).* **363**, 1463–1467 (2019).

11. Stickels, R. R. *et al.* Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat. Biotechnol.* **39**, (2021).

12. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* vol. 353 78–82 (2016).

13. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, (2012).

14. Codeluppi, S. *et al.* Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).

15. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science (80-.).* **348**, (2015).

16. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, (2016).

17. Eng, C. H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).

18. Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, (2020).

19. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

20. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

21. Lopez, R. *et al.* A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv* (2019).

22. Chen, S., Zhang, B., Chen, X., Zhang, X. & Jiang, R. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *bioRxiv* 2021.04.16.440115 (2021) doi:10.1101/2021.04.16.440115.

23. La Manno, G. *et al.* Molecular architecture of the developing mouse brain. *bioRxiv* (2020) doi:10.1101/2020.07.02.184051.

24. Mourragui, S., Loog, M., Van De Wiel, M. A., Reinders, M. J. T. & Wessels, L. F. A. PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. in *Bioinformatics* vol. 35 i510–i519 (2019).

25. Gyllborg, D. *et al.* Hybridization-based in situ sequencing (HybISS) for spatially resolved transcriptomics in human and mouse brain tissue. *Nucleic Acids Res.* **48**, (2020).

26. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

27. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, (2019).

28. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, (2018).

29. Moini, J. & Piran, P. Chapter 8 - Diencephalon: Thalamus and hypothalamus. in (eds. Moini, J. & Piran, P. B. T.-F. and C. N.) 267–292 (Academic Press, 2020). doi:https://doi.org/10.1016/B978-0-12-817424-1.00008-2.

30. Beattie, R. & Hippenmeyer, S. Mechanisms of radial glia progenitor cell lineage progression. *FEBS Letters* vol. 591 (2017).

31. Kempermann, G., Song, H. & Gage, F. H. Neurogenesis in the adult hippocampus. *Cold Spring Harb. Perspect. Biol.* **7**, (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

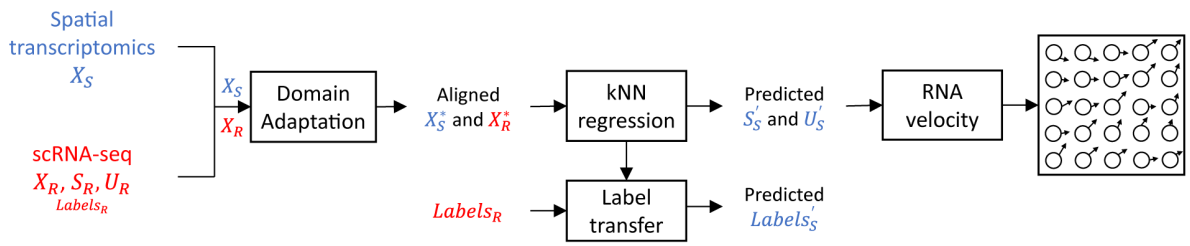


Fig. 1 SIRV pipeline. SIRV takes as input a spatial transcriptomics and a reference scRNA-seq datasets. The latter contains spliced S_R, unspliced U_R and full mRNA X_R expressions for each gene, and possibly label annotations $Labels_R$ e.g. cellular identity. SIRV integrates the two datasets X_S and X_R using domain adaptation producing aligned datasets X_S^* and X_R^*. Next, SIRV predicts the spatial spliced S_S' and unspliced U_S' expressions from the scRNA-seq data using kNN regression applied on X_S^* and X_R^*. SIRV transfers label annotations from scRNA-seq to spatial data ($Labels_S'$) using the same kNN regression. The predicted S_S' and U_S' expressions are used to calculate RNA velocity vectors, projected on the spatial coordinates of the tissue estimating spatial cellular differentiation trajectories.

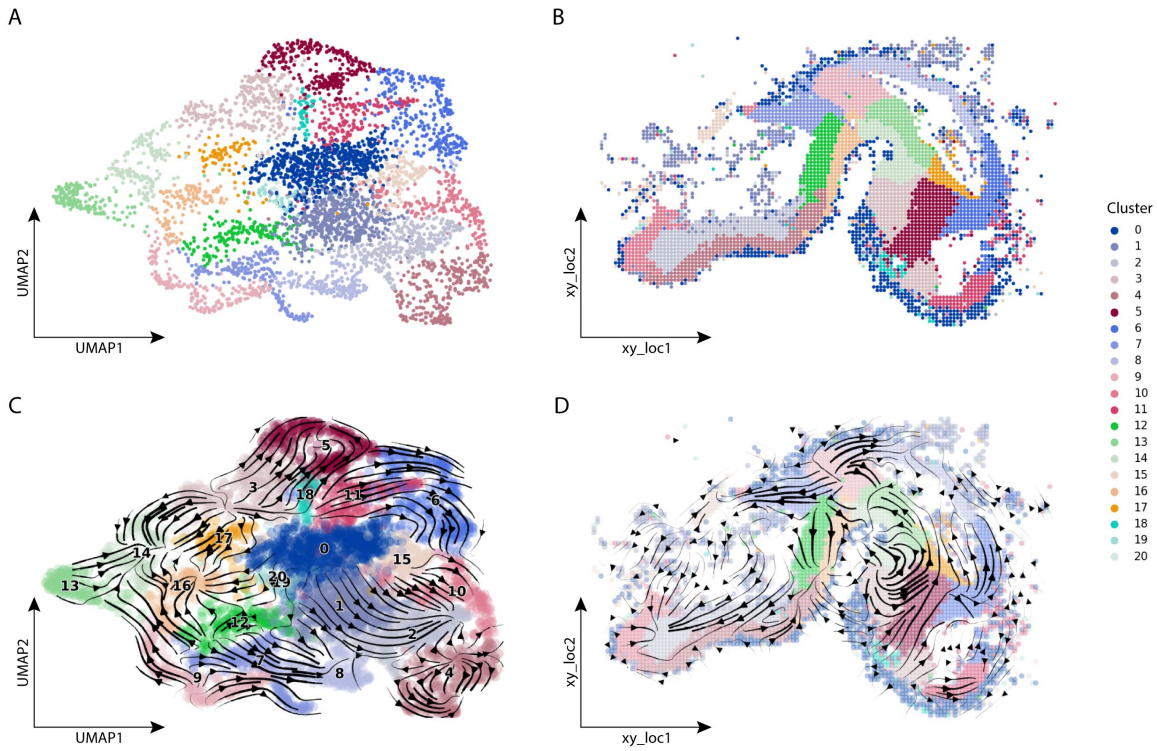


Fig. 2 (A) UMAP embedding of the HybISS spatial data colored according to 21 cell clusters obtained using Leiden clustering. **(B)** Spatial map of the HybISS spatial data showing spatial localization of the cell clusters. **(C-D)** Main flow of RNA velocities visualized by velocity streamlines, projected on **(C)** UMAP embedding and **(D)** spatial coordinated of the HybISS data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

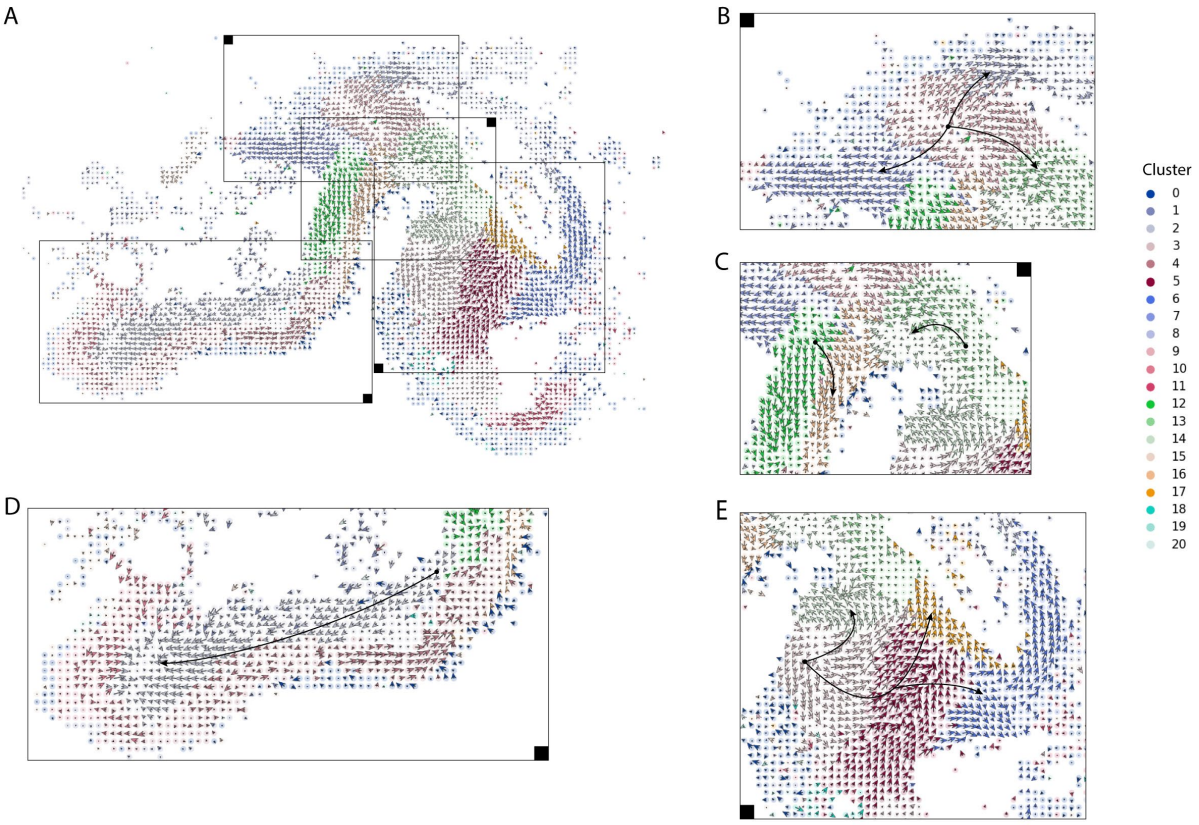


Fig. 3 (A) Cell-level RNA velocities projected on the spatial coordinates of the HybISS spatial data, colored according to the 21 cell clusters. **(B-E)** Zoom-in views on interesting spatial differentiation trajectories at **(B)** Midbrain-Hindbrain boundary, **(C)** part of Midbrain, **(D)** Hindbrain, and **(E)** Forebrain. Black arrows show branching or linear spatial differentiation directions between cell clusters.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

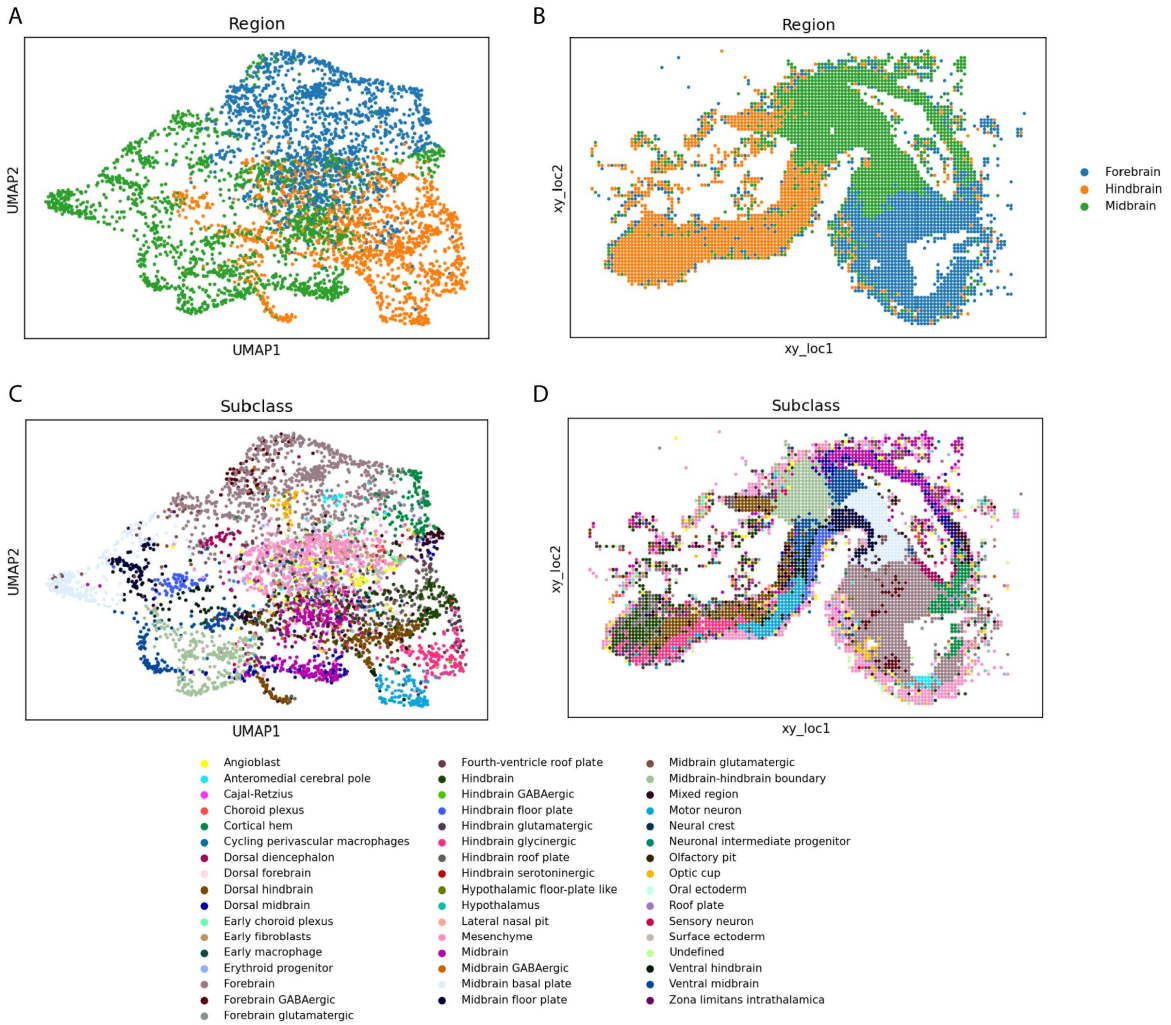


Fig. 4 (A,C) UMAP embedding of the HybISS spatial data colored according to **(A)** 'Region' and **(C)** 'Subclass' annotations transferred from the scRNA-seq data. **(B,D)** Spatial map showing the location of each **(B)** 'Region' and **(D)** 'Subclass' label in the tissue.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

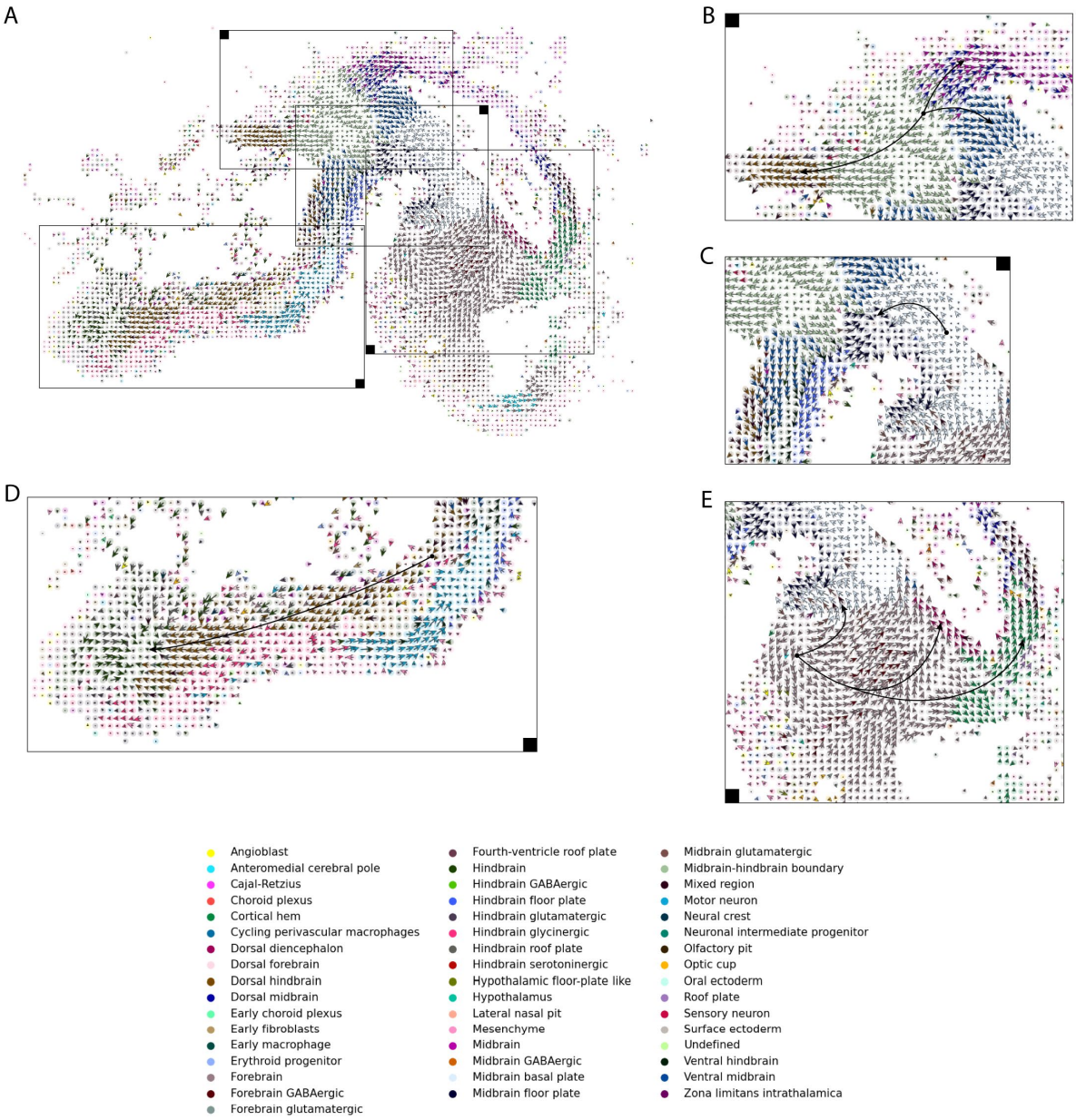
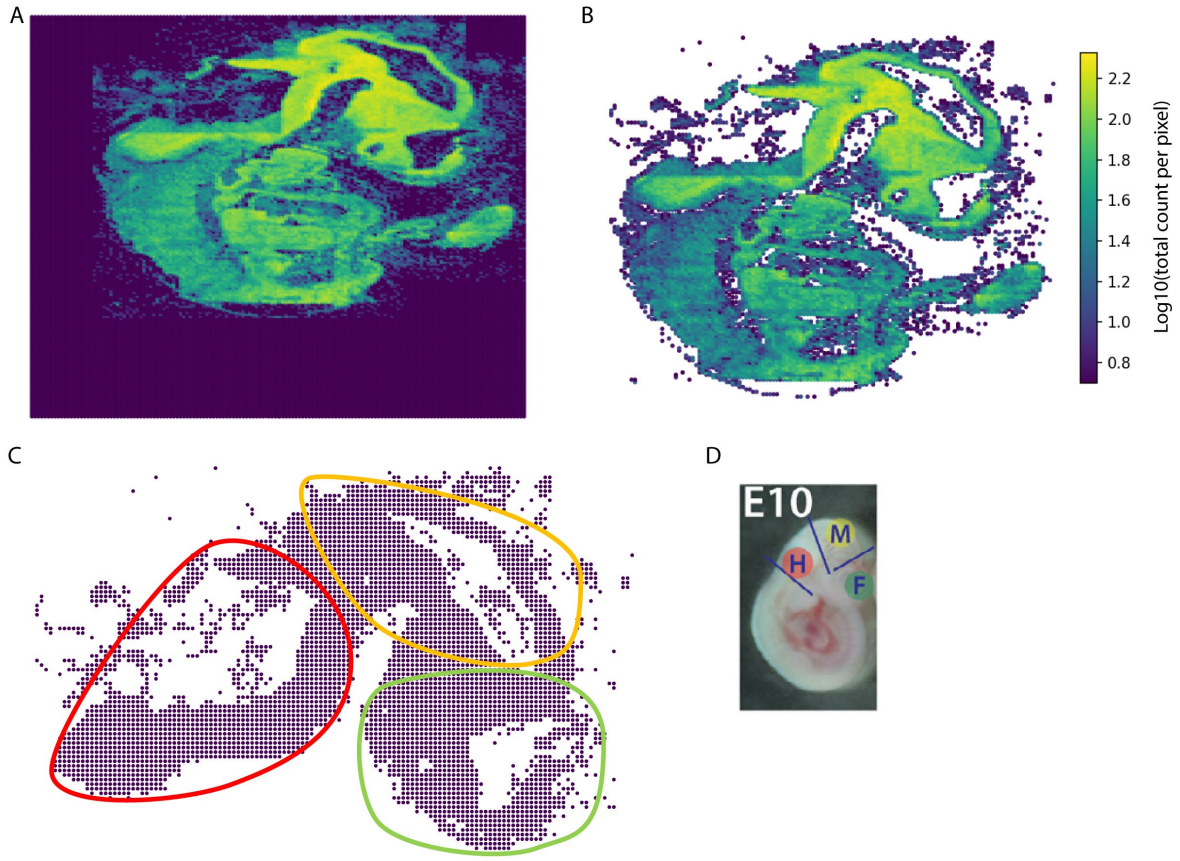


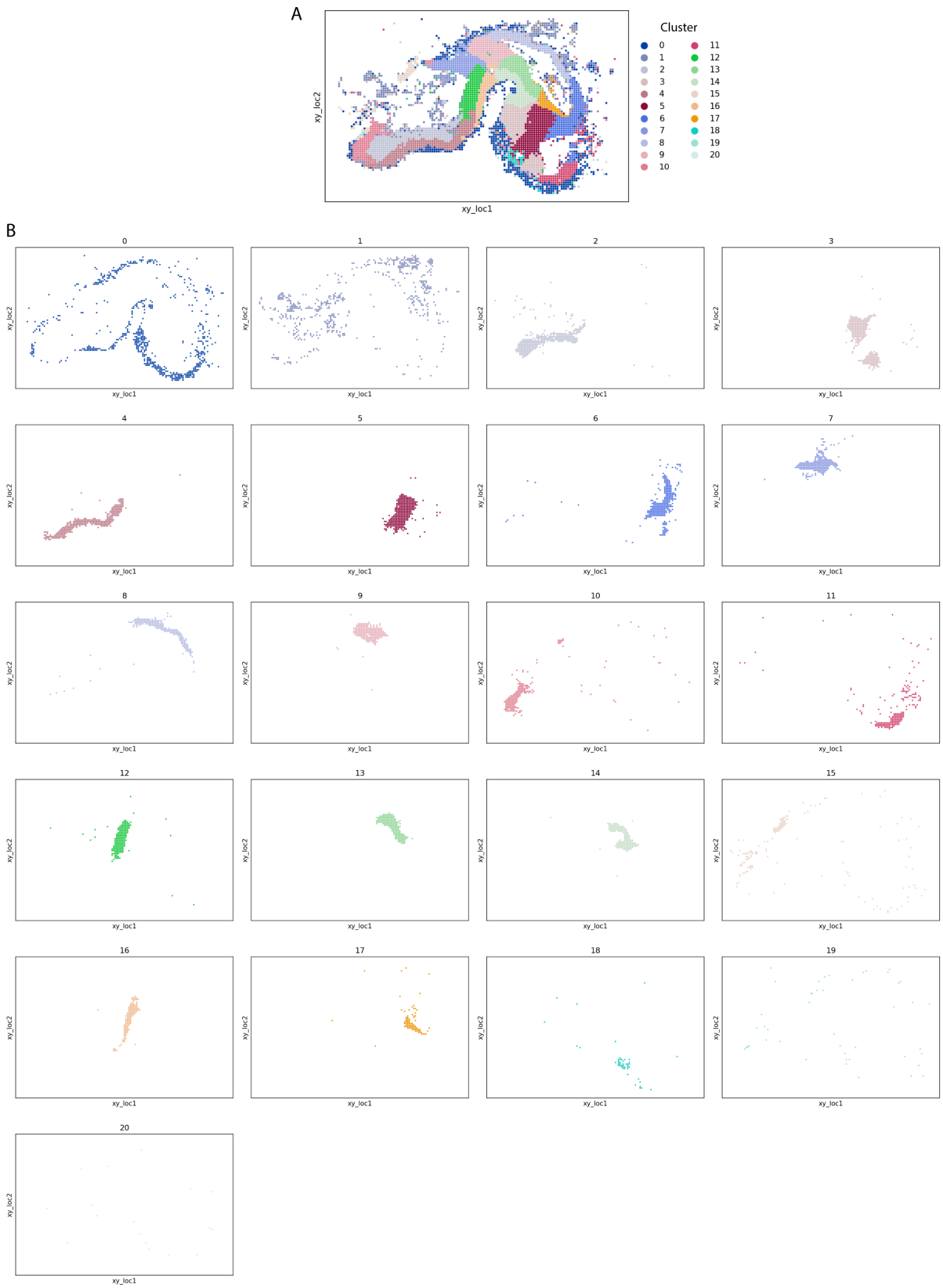
Fig. 5 (A) Cell-level RNA velocities projected on the spatial coordinates of the HybISS spatial data, colored according to the 'Subclass' annotation transferred from the scRNA-seq data. **(B-E)** Zoom-in views on the same spatial differentiation trajectories as in Fig. 3; **(B)** Midbrain-Hindbrain boundary, **(C)** part of Midbrain, **(D)** Hindbrain, and **(E)** Forebrain. Black arrows show branching or linear spatial differentiation directions between subclasses.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



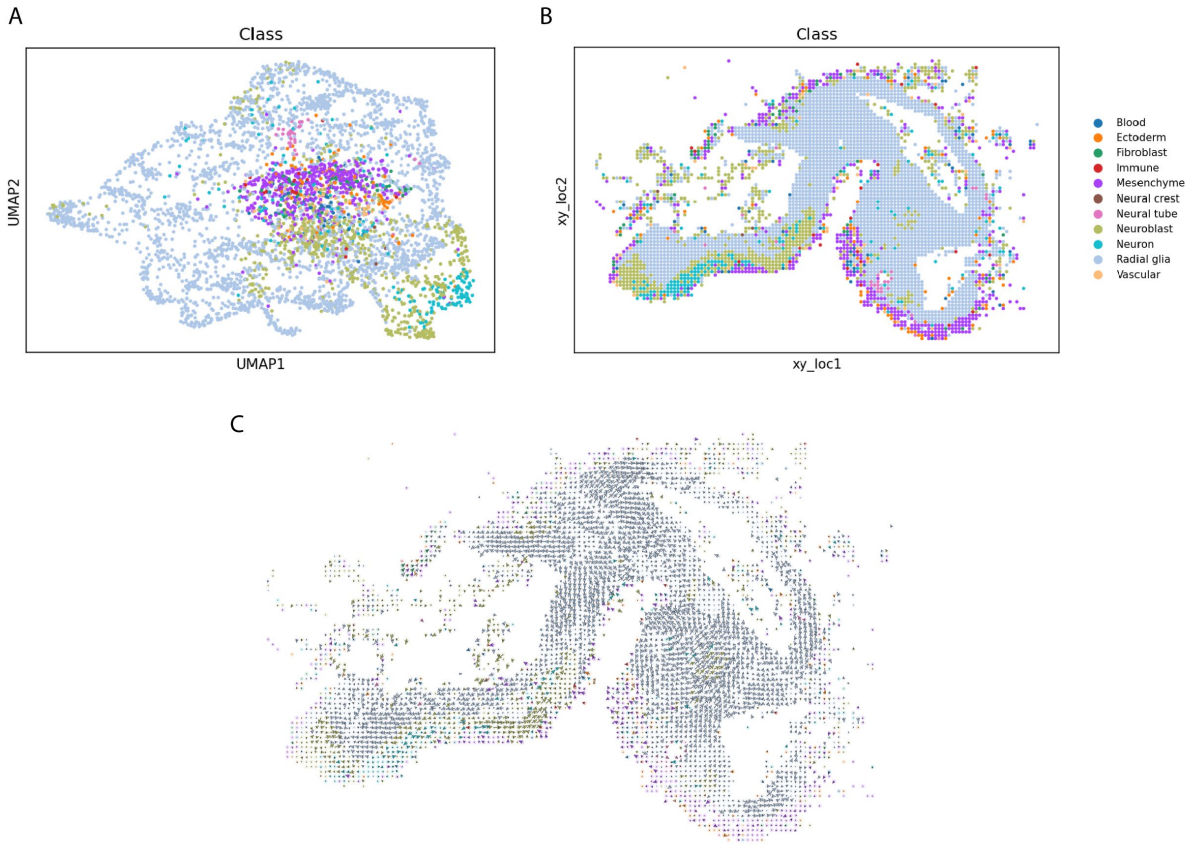
Supplementary Fig. S1 (A) Total count per pixel (pseudo-cell) in the HybISS spatial data separating mouse embryonic tissue from background. **(B)** Selecting only tissue pixels with a cutoff of total count per pixel ≥ 4. **(C)** Manual segmentation of only brain tissue (upper part in B), the three brain regions Hindbrain (red), Midbrain (yellow) and Forebrain (green) are highlighted according to D. **(D)** Figure adapted from G. La Manno et. al[23] illustrating the tissue dissection strategy and highlighting the location of the brain (and different regions in the brain) within the mouse embryo.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



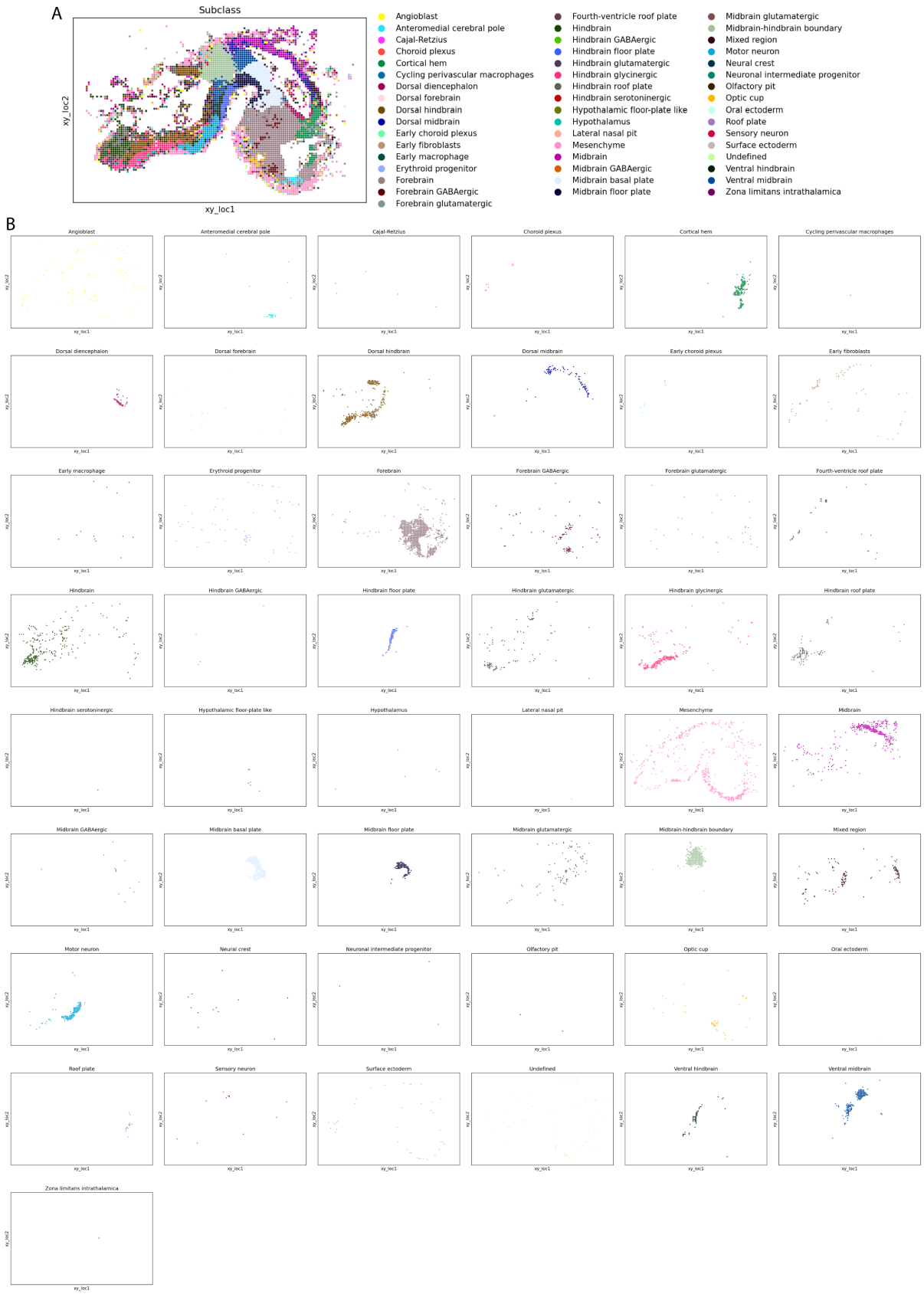
Supplementary Fig. S2 Spatial distribution of the HybISS 21 cell clusters (A) spatial map showing all clusters combined, **(B)** easier visualization of the spatial location of each individual cluster showing one cluster at a time.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



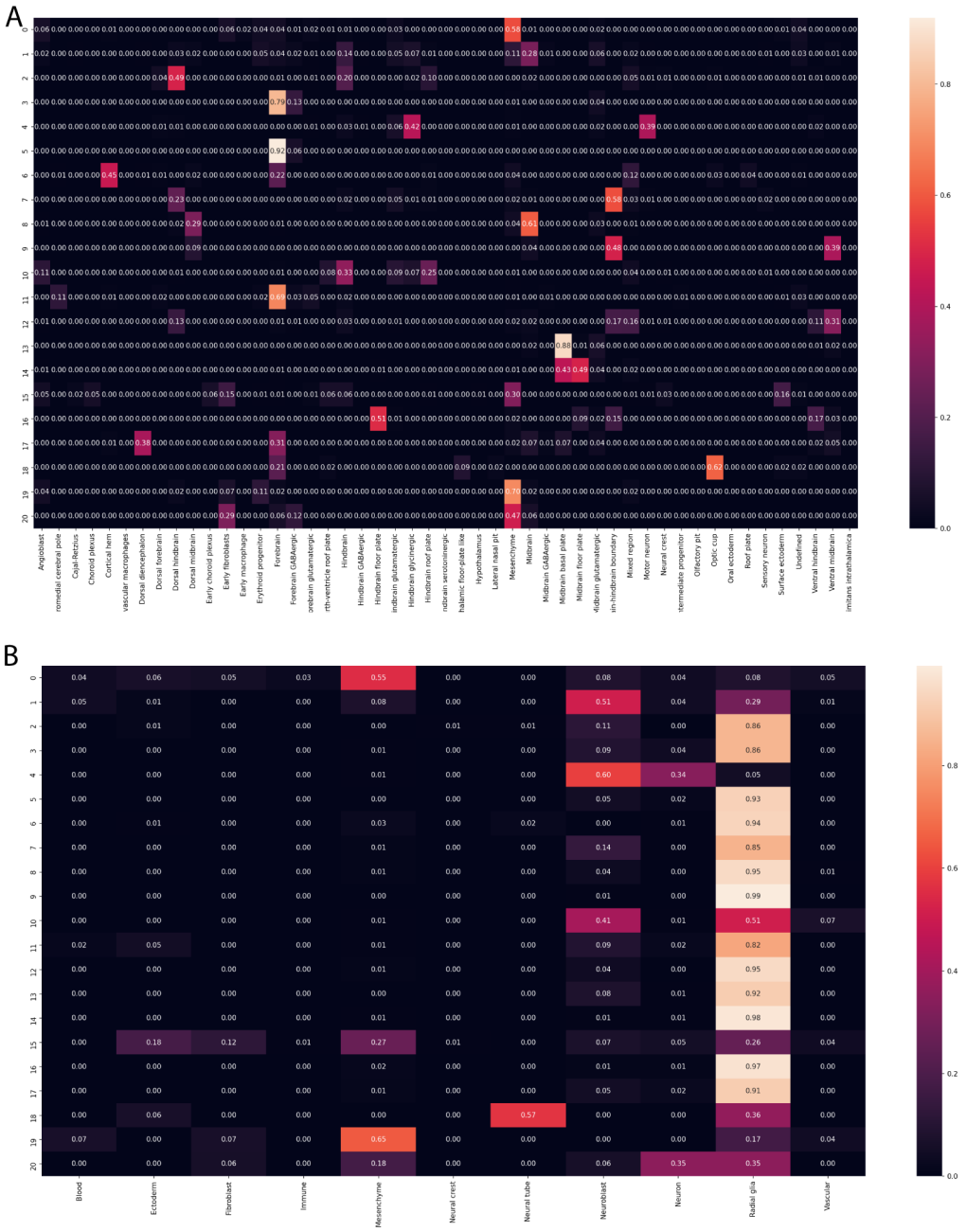
Supplementary Fig. S3 (A) UMAP embedding of the HybISS spatial data colored according to the 'Class' annotation transferred from the scRNA-seq data. **(B)** Spatial map showing the location of each 'Class' label in the tissue. **(C)** Cell-level RNA velocities projected on the spatial coordinates of the tissue, colored according to the 'Class' annotation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplementary Fig. S4 Spatial distribution of the 'Subclass' annotation transferred from the scRNA-seq data **(A)** spatial map showing all 49 subclasses combined, **(B)** easier visualization of the spatial location of each individual subclass showing one subclass at a time.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.26.453774; this version posted July 26, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Supplementary Fig. S5 Contingency matrices comparing the HybISS 21 cell clusters (rows) with their corresponding **(A)** 'Subclass' and **(B)** 'Class' annotations transferred from the scRNA-seq data (columns). Each row is normalized to sum up to 1.